BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

User-Centric Security

HABILITATION THESIS
(Collection of articles)

Mgr. Kamil Malinka Ph.D.

Brno 2025

Abstract

This thesis focuses on usable security, which links computer science, human-computer interaction, and information technology security. The research focuses on understanding user behavior and its impact on security measures. It shows that even the best security mechanism can fail if it is not understandable and usable by the user.

The first part of the thesis focuses more generally on usable security and presents the author's results in this area. These include user perceptions of security policies, educating future IT professionals, and the impact of new technologies on user security. In this case, the authors focus on deepfakes and people's ability to detect them.

The second part focuses on user authentication. It provides context for its various uses and then examines voice and facial biometrics and their resilience to deepfake attacks in detail. It gives examples of different types of attacks utilizing deepfakes and their impact. Next, we discuss the possible protection mechanisms which can range from deepfake detection methods to legislative measures.

Keywords

Security, usable security, human perception, security policy, university education, bug bounty, cybersecurity specialization, secure coding, experience report, authentication, artificial intelligence, deepfake, speech synthesis, biometric authentication, spoofing attacks, speaker recognition, face recognition, deepfake detection, robustness, evaluation framework.

Acknowledgment

First, I would like to thank Vashek Matyáš and Anton Firc, who have played a significant role in my research activities. Vashek showed me how to do science systematically and with quality. He has greatly inspired me while building my research team by allowing me to look over his shoulder and follow his best practices. Anton, as my first PhD student, helped me define new research directions, which we explored together. He was also a great supporter of my megalomaniacal tendencies. We motivated each other to a high level of commitment to work, which enabled us to achieve many interesting results, often going beyond scientific activity.

Secondly, I would like to thank all my co-authors, especially Martin Perešíni, Milan Šalko, Ondra Hujňák, and Pavel Loutocký, and also to my supervisor Petr Hanáček and colleagues and friends at FIT BUT and FI MU, for many stimulating discussions and interesting research ideas.

Third, I would like to thank all Ph.D./MSc/BSc students I have had the chance to work with.

Last but not least, I would like to thank my beloved Katka for supporting me throughout my career, my kids Eliška and Kuba for just being awesome, and my parents for inspiring me to work on things I enjoy.

Over time, I have been supported by a number of projects, in particular, by EU HORIZON EUROPE projects, the internal grant agency of Masaryk University, the Brno University of Technology internal project, the Technology Agency of the Czech Republic, and the Security Research Programme of the Czech Republic.

Contents

Part I

COMMENTED RESEARCH

Chapter 1

Introduction

IT security is a complex field, which, in practice, is often simplified to mere technical measures. However, well-managed security encompasses many areas, from process definition, risk analysis, setting the context of the environment, technical measures, the security documentation, to end users. Unfortunately, we see that the consideration of the human factor is often neglected despite its crucial importance. Thus, my research focuses primarily on areas of security that are closely linked to the end users of technologies.

A field that focuses, among other things, on exploring the usability of security technologies and understanding how users handle these technologies is usable security [1]. Usable security is a subfield of computer science, human-computer interaction, and cybersecurity. One of the fundamental narratives of usable security is that a secure technical element should ideally not allow a user to make a wrong security decision.

Research in user behavior is essential, as understanding real user behavior is the key to correctly designing protective measures to have the intended effect in real-world environments. It is crucial to avoid the false sense of security that can be created, for example, by using a high-quality and complicated security measure that the user cannot operate in reality and so bypasses it. However, usable security is not exclusively aimed at end-users; it encompasses a wide range of IT expertise, including IT professionals.

I have spent a significant part of my career outside of academia working on developing and operating IT technologies and leading various IT teams with varying expertise. On one hand, this has caused a visible publication gap, on the other hand, I have gained valuable practical experience. Often, I had the opportunity to observe first-hand the impact on security when the human component is not adequately addressed. As an example, I will share the findings from one of the projects I participated in.

The aim of the project was to update the university security policy. During its update, one of the goals was to ensure that users understood it better. This activity was also combined with research, where we measured the impact of user education in this area alongside the policy development.

It turned out that we could not properly evaluate the approach used and its impact because ordinary users, in the end, did not access and read the policy at all after it was published and promoted. The work put into making the policy more understandable to users has gone to waste, as the element of informing users that a new guideline exists has failed. The fact that the directive was binding on users played no role.

This example only confirms that a more comprehensive approach is needed to ensure user security. In addition to deploying appropriate technical tools, this includes understanding how the user works with these tools and ensuring ongoing user education. The next crucial step is transferring the identified weaknesses and vulnerabilities to the system designers, which enables them to improve their solutions and better target the training content and various supporting methodologies for working with the tools.

My research has dealt with, among other things, the areas mentioned above, and this thesis presents my research contributions. This thesis is designed as a collection of works accompanied by an explanatory commentary. The aim is not to present new scientific results but to explain the context of the author's scientific activity in detail, highlight the relationship of the individual results, and relate them to the existing literature.

In my Ph.D. thesis, I addressed selected issues of behavioral patterns in computer security, where I was primarily engaged in developing a novel concept of biometric authentication based on visually evoked potentials [A1, A2]. Secondly, I investigated the effects of email user behavior on the effectiveness of anonymization systems [A3, A4, A5].

This habilitation thesis loosely builds on my previous work with other related topics - user behavior when using selected security tools, future IT professionals' education, and deepfakes' impact on voice and face biometrics.

Some of my works [A6, A7, A8, A9] also cover areas of legislation topics, as the law is also part of the broader context in which IT security needs to be addressed. I have chosen not to include these works in this thesis because they are mostly focused on Czech law and thus have a limited impact.

The work presented in this thesis is based on research conducted with co-authors from our laboratory and research collaborations with co-authors from several different faculties of Masaryk University. I acknowledge the use of DeepL as a support tool for writing in English and the use of Grammarly for grammar correction.

The rest of the thesis is organized as follows.

Chapter 2 introduces usable security, focuses on user behavior, and investigates how users handle the selected security mechanisms. Specifically, we cover three areas: how ordinary users perceive security policy as the primary security management tool, how to educate future IT professionals in secure coding effectively, and finally, we focus on the properties of user authentication security mechanisms.

In the last mentioned area, we explore the phenomenon of deepfakes, which introduces many additional security risks. To better understand how serious the problem is, we conducted a series of experiments to test people's ability to recognize deepfakes. The results revealed major weaknesses that need to be addressed.

In the next part in Chapter 3, we focus on user-related security measures. We take a closer look at the resilience of authentication mechanisms. To thoroughly map this issue, we have researched possible attacks based on deepfakes, evaluated the resilience of current authentication against these attacks, evaluated existing detection mechanisms, and proposed new approaches.

Chapter 4 concludes the thesis and outlines our future research directions.

At the end of each chapter is the list of the author's publications contributing to the topic, with several representative papers attached in Part II. In Part II, we also describe the author's contributions to the papers contained in this thesis.

Chapter 2

Usable Security

Usable security focuses on integrating security in an end-user-friendly way [2]. Usable security research has been ongoing for more than twenty years. The community has already understood that users are part of the solution how to increase security and that the problem is not them but the lack of usability of security mechanisms [3]. Also, psychological effects such as fatigue from security decisions need to be taken into account [4].

Major areas and challenges in usable security include authentication, encryption, social engineering, security dialogs and warnings, and privacy [4, 5]. However, to show that the whole field is much more diverse, we will give a few more examples to illustrate the diversity of the topic: Green et al. [6] show the importance of usable API design, Fischer et al. [7] pointed to the issue of stack overflow code reuse, which without deeper understanding results in security issues, and Chiasson et al. [8] focused on the process part, namely guidelines for security management interfaces. To add to the variety, we can also mention our research focused on antivirus software users [A10]. We have proven that a simple text change can provide a clearer presentation of the security benefits to the user and lead to greater use of more advanced security solutions.

Despite the fact that usable security research is growing and yielding many results, it typically focuses on end users and lacks research results focused on IT professionals [5]. In our research, we therefore also focused on user groups other than ordinary users.

2.1 Security Policy

In this section, we take a closer look at security policies and their application as their usability is considered crucial to influence users to behave securely [9]. Many observations and experiments [10, 11, 12] show that while policies have been with us for years, there is still a big gap between their mere existence and their actual use. This could be the reason why most security professionals still consider users as the top data breach risk and deem that users are negligent or just simply break the security policy[1]. On the contrary, overloading users with security requirements can lead to a negative effect [13].

[1] https://www.darkreading.com/cyber-risk/despite-rise-of-third-party-concerns-end-users-still-the-biggest-security-risk

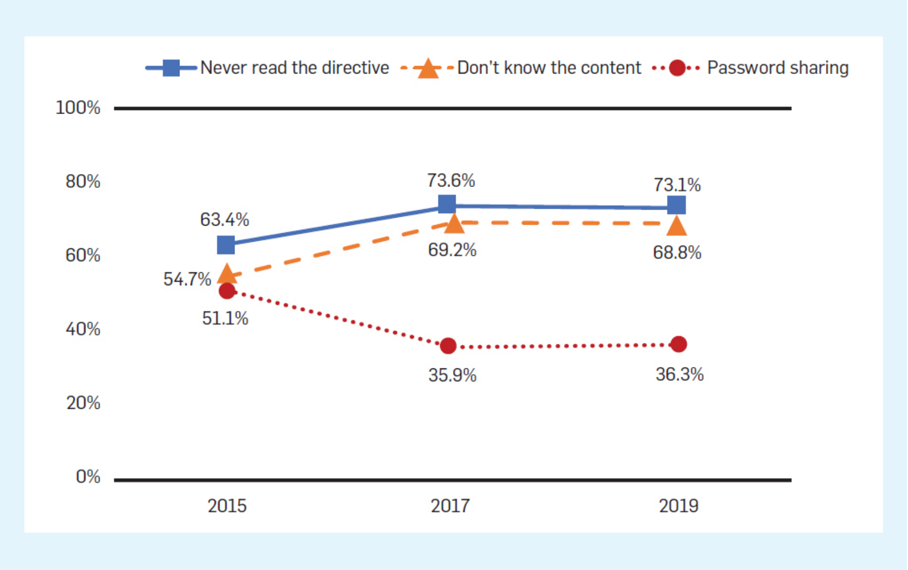


Figure 2.1: The graph reports the share of users who never read the directive and the share of users reporting not knowing the content. These trends contrast with the decreasing share of users who shared their university passwords. Figure taken from [A11].

Authors contribution

In our research [A11], we were interested in how users at the university work with the security policy. We had the great opportunity to participate in modernizing the security directive at Masaryk University, so we decided that besides implementing modern trends into the directive, we would also design surveys to measure the impact of these changes on the user-reported security behavior.

We designed a long-term study to find out if users are complying with the security policy. Three times in five years we surveyed a sample of students (from 600 in the first year to 1300 in the last year of the survey) to find out whether they read the directive and knew its content. We were also interested in their security behavior in different situations such as sharing a password or if they regularly apply security updates. I was responsible for setting up the data collection environment and implementing the survey.

We measured three phases, awareness before the publication of the new directive, awareness after the publication of the new directive and the notification through standard channels, and awareness after the implementation of the additional information campaign.

The additional campaign focused on several security issues, such as password sharing or the scope of malware, and was launched two months after the policy was issued through standard institutional channels; we also highlighted the existence of the new security directive. We ran the campaign in the university magazine as well as on several university Facebook groups.

The final results were inconclusive. Despite all the campaigning, users were reading the directive less and less. The number of people who have never seen it at all has increased by 10% over the years – to an alarming 73%.

In contrast, the reported user security behavior was surprisingly quite reasonable – and often improving. Over time, students began to behave more safely on their own. For example, password sharing between students has decreased from nearly 50% to just 36% (see Figure 2.1). We can say that in areas of endpoint protection or handling passwords users behave reasonably. Whether this was due to exposure to external sources of relevant information or a more naturally increased adoption of technologies, remains to be investigated in future work.

Our results showed that education has completely missed the mark in this area of security. The additional objectives were then irrelevant - it is impossible to evaluate efforts to improve the readability of the directive if nearly nobody reads it. This only demonstrates the need to adequately inform users when deploying security measures.

2.2 Education of future IT professionals

In this section, we will look at another area where security education has been very effective for change and has had additional positive impacts. Education is one of the few security mechanisms that affects people directly. Continuous security awareness is a fundamental pillar of security. However, it needs to reflect the current situation and threats to be effective. Education must also consider the target audience, where educating the general population against common cybercrime will be designed differently than, for example, educating future IT professionals.

The topic of education and its effectiveness is very relevant to me because, in addition to research activities, I am also responsible for teaching the cybersecurity specialization courses at our faculty. I believe that to provide quality education and increase its effectiveness, it is necessary to bring new concepts of teaching, to integrate teaching and practice more closely, and, where appropriate, to investigate the impact of new technologies on teaching and learning [A12].

Education in computer science degrees puts a strong emphasis on practicality and due to a broad curriculum it is often difficult to cover more than fundamentals of each subject [14, 15, 16]. When it comes to IT students who are not security-focused, they usually obtain only a brief introduction to cybersecurity. The problem remains even if the student's focus is purely on cybersecurity, where we must also consider another factor - the curriculum time allocation structure, which is often limited. Thus, more general topics such as encryption, authentication, or IT security management usually have priority over narrow areas such as ethical hacking [A13].

However, even this narrow area can bring many benefits - e.g. it allows the student to better understand how an attacker thinks, and what tools are available to him, which in practice makes it easier to create secure products.

Given that the educational system, especially in regions such as the Czech Republic, is insufficient in addressing the ethical aspects of hacking [14], we have further focused on this area since it is considered an important piece of cybersecurity professional skillset by a substantial portion of the community [17, 18, 19]. At the same time, we wanted to emphasize the involvement of practical tasks, which provide students with valuable experience, allow them to test their technical knowledge, and further develop non-technical skills [20].

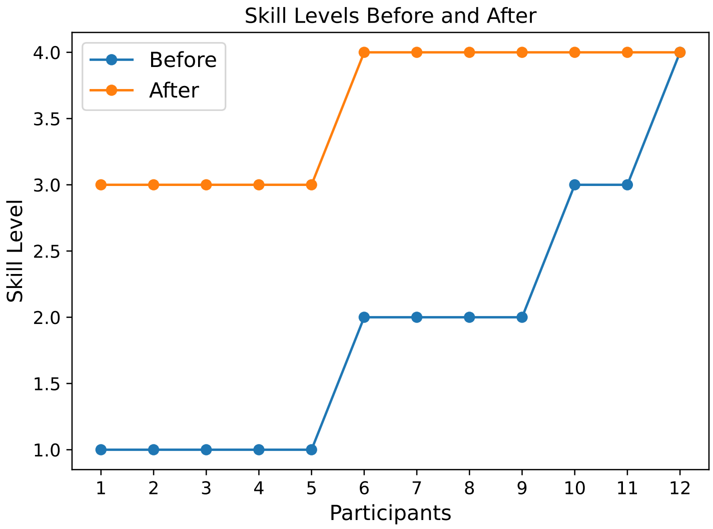


Figure 2.2: Orientation of participants on the topic of ethical hacking before and after the project. Participants were ordered by skill level before the project. Figure taken from [A13].

Authors contribution

Our research focused on integrating real-world Bug Bounty Programs (BBPs) into an academic curriculum as this allows students to work with real systems during their studies. Bug Bounty challenge is a security measure used to increase product security [21]. It is based on the principle of using external entities (ethical hackers) to find vulnerabilities in your system [22]. It usually includes a suitable incentive system involving financial rewards to encourage participation in the program [23].

I was part of a broader project team developing a methodology for a responsible cyber security vulnerability reporting program primarily aimed at SMEs. The knowledge we gained from this project[2] was also used during the preparations of articles in this area.

The main contribution of our work that focused on using bug bounties in education [A13] is the design of the concept of incorporating real-world bug bounties into teaching and practical verification of its effectiveness and pedagogical implications (e.g., providing motivation or how to evaluate students without penalizing them for failing a hunt or how to provide a safe environment for the student from a legislative perspective).

We implemented the proposed solution in a secure coding course at our faculty to evaluate our idea. Students could voluntarily choose a new type of project to solve during the semester, where they participated in a real-world bug bounty. The goal was not a successful solution (that was a bonus) but learning appropriate techniques and describing their implementation during the bug bounty. The students were provided with educational materials and instructed on the requirements to participate in the BBPs. The project evaluation was followed by a questionnaire survey to find out further details. We supplemented the results with information from the project reports.

I was responsible for the entire execution of the experiment. I designed the way to inte-

[2] https://www.muni.cz/vyzkum/projekty/43686

Table 2.1: Personal likings of the project as reported by students. The responses ranged from 1 - 5, where 1 means the worst (negative) and 5 means the best (positive). Symbol σ denotes the standard deviation. Table taken from [A13].

Question	Mean	σ
How much would you like to see this type of project incorporated into regular teaching?	4	0.85
Is the project content beneficial even if the participant has a career path outside of cybersecurity?	3.91	0.79
How do you feel about working in a real environment on real products?	3.92	0.99
How important is the project's social impact to you (helping to improve real safety)	3.42	1.08

grate BB into the subject and project assignment, defined learning areas, created a survey, evaluated the project solutions, and conducted data collection and evaluation.

Primarily, we were interested in the answer to whether students can successfully solve BBPs and what impact this type of project involving a practical scenario has on the student's knowledge.

The main result is that all students were able to successfully solve the assignment and their understanding of ethical hacking increased (see Figure 2.2). In evaluating the projects, we did not observe a significant deviation from other types of projects. Some students even succeeded in finding and reporting a real vulnerability. Upon completion of the project, students reported a good understanding of the topic, a practical understanding of the different phases of penetration testing, an understanding of the attacker's perspective and capabilities, and a working knowledge of ethical hacking tools. Other interesting results include the fact that students found the knowledge they gained beneficial even if they plan to pursue a career outside of cybersecurity, as well as the interest of some to pursue BBPs in their spare time. We also investigated participants' views on questions regarding their personal likings of the project as shown in Table 2.1.

The conducted research is only the first phase of an ongoing longitudinal study aimed at further continuous validation and refinement of the concept.

In our other work [A14], we have elaborated on how the inclusion of other collaborating organizations in our concept can bring further positive effects. This is one of the ways to partially solve the problem of the lack of specialists, which, for example, the state institutions suffer from. They can use students to test their infrastructures and possibly discover existing vulnerabilities, leading to remediation and increased security without the need for further investment. Beneficially, this opportunity gives students the chance to test their knowledge and skills on real infrastructure.

However, using BBP brings with it certain risks, from breaking the law by hackers to data leakage by BBP providers, that must be mitigated. Therefore, we have focused more on the operators of BBPs in our work, and we have shared the lessons learned.

My role in this research was to provide additional information regarding the implementation of the experiment. My task was also to review proposed legal concepts in the context of operational IT to reflect the actual field situation encountered by IT professionals.

To further support this activity, we have conducted an analysis of the legal aspects of cybersecurity vulnerability disclosure [A15], which puts this issue in the context of the European directives such as NIS 2 [24] and provides the legal framework needed to effectively and safely implement a bug bounty program. My task was again to review proposed legal concepts in the context of operational IT to reflect the actual field situation encountered by IT professionals.

2.3 Human recognition of deepfakes

In this section, we focus on the challenges brought by the rapid development of generative AI as managing information security is a continuous process that must consider the development of new technologies. These, in addition to their benefits, also bring new threats. Significant increases in the quality and availability of generative AI models and tools in recent years have enabled the creation of quality synthetic media (voice, images, video), even for people lacking a technological background. Over the last few years, this has led to a significant increase in attacks that use deepfakes - voice, image, or a combination of the two [A16].

Deepfakes are a subset of synthetic media that depict events that never happened and can be used for malicious purposes [A17]. The term itself is a combination of words *deep learning* and *fake*. Deepfakes are created using deep neural networks, depicting events that never happened to entertain, defame individuals, spread fake news, and others [25].

Typical directions of attack are theft and scams, the spread of fake news and hate propaganda, spoofing attacks on biometrics, defamation, and identity thefts. Although the range of attacks can vary widely - from defamation of a single person to fake news spread with high impact potentially influencing geopolitical situations[3], the typical target remains the human being. Recognizing the synthetic medium from the real one is crucial to ensure resistance to these attacks. Another form of defense may be to deploy additional technical means to help detect a deepfake, but this is discussed in greater detail in the next chapter.

Research on humans' ability to recognize deepfakes mostly focuses on video and photos; however, there is also work in the area of audio that covers this area. In image/video domain, studies on human deepfake detection reveal varying success rates based on image or video quality, with images achieving 58-70% accuracy and videos as low as 20% for high-quality deepfakes, increasing to over 80% for lower quality ones [26, 27, 28, 29, 30, 31].

In the voice area, Mai et al. [32] revealed a 73% accuracy rate in identifying deepfake audio. Müller et al. [33] used a game-based approach and the ASVspoof 2019 dataset [34] and reported 80% success rate in human detection. Watson et al. [35] investigated audio

[3] https://theconversation.com/deepfakes-in-warfare-new-concerns-emerge-from-their-use-around-the-russian-invasion-of-ukraine-216393

deepfake perception among college students. Results showed success rate with a varying accuracy of 42% to 90% based on the task.

Authors contribution

In our work [A18, A19], we focus on deepfake recognition by humans in the context of IT security. To better understand how serious the problem is, and to better evaluate its impact, we conducted a series of experiments to test the ability of humans to detect deepfakes without any technological support. We also tackled the problem of lacking a multilingual component; most of the existing research concentrated on English and other major languages, while we were interested in the impact on languages that are less represented in the research such as Czech.

Although most of the related published research has demonstrated a relatively high ability of humans to detect deepfakes, from a security perspective, they had a common drawback in the design of the experiments. Subjects typically worked in a detector-only mode, where they had to distinguish between a bona fide sample and a deepfake. They were thus informed about the nature of the experiment. In real attacks, however, the victim does not have this information. Insufficient emphasis has also been given to the effect of varying the quality of deepfakes (which improves over time).

We explored two principal options for the investigation of the human ability to recognize deepfakes:

1. Uninformed recognition - where the intent of the experiment is hidden from the subject so we can better simulate a real attack.

2. Informed recognition - where subjects are fully informed about the issue of deepfakes

My role in this area of research was different from previous ones. I was already the principal investigator who defined the direction of our research group in this area. I identified the bottlenecks of the previous experiments and proposed the whole concept of measuring uninformed recognition using a cover story. I also defined the methodology used. Together with colleagues, we later significantly extended the set of analyzed properties.

For both options, we used a similar methodology: using publicly available state-of-the-art SW for speech synthesis, we created deepfake samples of the required quality, which we then used in experiments on a selected demographic group. In addition to the results obtained by the direct measuring of responses, we also used a questionnaire survey to obtain additional information.

The initial publication [A18] focused on investigating people's ability to detect deepfakes in casual conversation. A major difference from research papers with a similar focus was the use of a cover story to conceal the nature of the experiment. This allowed us to investigate situations that are comparable to real attack conditions.

According to the cover story, participants evaluated the usability of voice messages via the WhatsApp application by playing a game *Two Truths One Lie* with a figurant.

Table 2.2: Human ability to identify deepfake recording during casual conversation. Table taken from [A19].

Reaction during conversation	
Reacted	0%
Described unnatural things from the conversation	
Poorer audio quality	41.90%
Deepfake sign	3.20%

During the experiment, the figurant was replaced by his deepfake. The experiment aimed to see if participants would notice the change in some form and thus detect the attack. However, the results showed almost zero success rate in this scenario.

None of the participants reacted to the deepfake during the conversation. In follow-up questions before revealing the main idea of the experiment, only one respondent specifically addressed deepfakes. The participants stated that the reason behind the low success rate is their focus on content. The possibility of a fraudulent recording did not occur to them during the interview, which supports our opinion about the need to simulate scenarios close to real attacks. Results are summarized in Table 2.2.

Our experiment showed an extremely high vulnerability of the general population that was not detected by the design of previous experiments. We believe that we have addressed a critical gap in existing research.

In the second area [A19], we used a common experimental setup where participants knew they would be exposed to deepfakes and just decided whether the sample was bona fide or deepfake. In this case, we focused on the effect of deepfake quality on how people recognize them and other attributes that might play a role, such as primary language, gender, or prior experience with deepfakes.

The main output of the second part of the experiment was the development of the quality metric for deepfake speech. Quality can be expected to play a significant role in the success rate of an attack, but it is usually not quantified in relevant research beyond the description of the dataset used. Determining quality value is thus important to make it easier to compare the results of multiple independent experiments. Moreover, it is also relevant from the attacker's point of view, which will focus on three key parameters during design - Speaker Similarity, Perceptual Evaluation of Speech Quality, and Technical Evaluation of Speech Quality. Our quality metric covers all mentioned areas. We used a voice biometric system to measure speaker similarity, Perceptual Evaluation of Speech Quality (PESQ) to measure the people's subjective opinions of synthetic audio samples, and Mel Cepstral Distortion to assess speech quality as it is often used in speech synthesis systems. The measured values were appropriately combined to calculate the final quality.

This metric was then used to evaluate the dataset used in the experiment. We tested 85 participants (48 men, 37 women) over two months using an online survey.

The results revealed that although none of the deepfakes used so far represented a threshold quality beyond which they could no longer be detected, the higher quality made it more challenging to detect deepfakes. Given the rapid advances in technology, it is likely that the results of a more powerful and modern synthesizer will already be different.

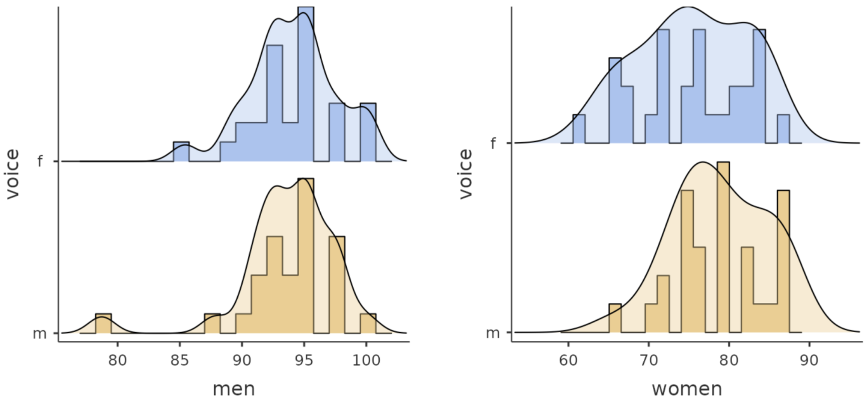


Figure 2.3: Plots depicting the accuracy of deepfake detection by gender: Men's accuracy is shown on the left, and women's on the right. The X-axis indicates the percentage of correctly identified deepfakes, while dual Y-axes show the volume of accurately labeled recordings. The graphs employ orange (m) and blue (f) to distinguish between recordings voiced by male and female speakers, respectively, sharing a common X axis but with separate Y axes for each gender's count of correctly identified recordings. Figure taken from [A19].

Next, we considered what effect the output device, gender, or native language, might have on recognition. As an example of the results, we present here the difference in recognition of deepfakes between men and women (see Figure 2.3). Our findings reveal that men recognized 93.90% of all deepfakes, while women identified 77.20%. Specifically, men detected 94.10% of deepfakes spoken by women and 93.70% spoken by men. Women had a 78.90% accuracy rate for deepfakes voiced by men and 75.50% for those voiced by women, as shown in Figure 2.3.

This part of the experiments confirmed the results of previous studies but further showed a strong dependence of recognition on the quality of the deepfakes used, differences in perception in the population, and also a strong influence of the used output device.

The information gathered is important for further preparation of campaigns to raise user awareness in this area. The fundamental conclusion is that, with the current trend of increasing the quality of synthesizers, people will soon completely lose the ability to recognize a deepfake from a bona fide sample. Thus, educational campaigns should also focus on other areas - explaining attacks and the possibilities of exploiting deepfakes, explaining the need to change internal processes vulnerable to these attacks, or focusing on familiarizing users with the use of detection and other protection tools.

We have further incorporated our findings into the educational campaigns we created for a diverse community to raise awareness about this issue. As most of the available materials are in English, we focused on the Czech Republic. We have co-authored a Czech book for the general population [A20], where we wrote a chapter dedicated to this issue, we have worked with the banking association and the police to create a national education campaign KYBERTEST[4], and we have also delivered over 30 training sessions for business representatives, police, prosecutors, judges, military, and the general population or students.

Contributed papers

This chapter is based on our 8 research articles, parts of which are included in this thesis.

Articles in collection

[A11] Vashek Matyas, Kamil Malinka, Lydia Kraus, Lenka Knapova, and Agata Kruzikova. "Even if users do not read security directives, their behavior is not so catastrophic". *In: Commun. ACM* 65.1 (December 2021), pp. 37–40. ISSN:0001-0782. https://doi.org/10.1145/3471928

I helped with the design of experiments and contributed to text writing. I prepared technical parts necessary for the realization of the experiment such as setting up the data collection environment and implementing the survey. Contribution 25%.

[A13] Kamil Malinka, Anton Firc, Pavel Loutocký, Jakub Vostoupal, Andrej Krištofík, and Frantisek Kasl. "Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report". *In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education* V. 1. ITiCSE 2024. Milan, Italy: Association for Computing Machinery, 2024, pp. 227–233. ISBN: 9798400706004. https://doi.org/10.1145/3649217.3653633.

I led the research, designed the experiments, cooperated on analysis, and wrote a significant part of the text. Contribution 30%.

[A14] Andrej Krištofík, Jakub Vostoupal, Kamil Malinka, František Kasl, and Pavel Loutocký. "Beyond the Bugs: Enhancing Bug Bounty Programs through Academic Partnerships". *In: Proceedings of the 19th International Conference on Availability, Reliability and Security*. ARES '24. Vienna, Austria: Association for Computing Machinery, 2024. ISBN: 9798400717185.
https://doi.org/10.1145/3664476.3670455.

I cooperated on the design and analysis of experiments, on the analysis of legislation and contributed to text writing. Contribution 20%.

[A19] Kamil Malinka, Anton Firc, Milan Šalko, Daniel Prudký, Karolína Radačovská, and Petr Hanáček. "Comprehensive multiparametric analysis of human deepfake speech recognition". *In: EURASIP Journal on Image and Video Processing* 2024.1 (August 2024). ISSN: 1687-5281. http://dx.doi.org/10.1186/s13640-024-00641-4.

[4] https://www.kybertest.cz/

I led the research, proposed the idea of using a cover story for the desired use case, cooperated on the experiment design and analysis of results, and contributed to text writing. Contribution 45%.

Other relevant publications

[A10] Vlasta Stavová, Vashek Matyas, and Kamil Malinka. "The Challenge of Increasing Safe Response of Antivirus Software Users". *In: Mathematical and Engineering Methods in Computer Science*. Ed. by Jan Kofroň and Tomáš Vojnar. Cham: Springer International Publishing, 2016, pp. 133–143. ISBN: 978-3-319-29817-7.

I cooperated on the design of used approaches, helped with the experiment deployment, and contributed to text writing. Contribution 15%

[A15] Jakub Vostoupal, Václav Stupka, Jakub Harašta, František Kasl, Pavel Loutocký, and Kamil Malinka. "The legal aspects of cybersecurity vulnerability disclosure: To the NIS 2 and beyond". *In: Computer Law & Security* Review 53 (2024), p. 105988. ISSN: 0267-3649. https://doi.org/10.1016/j.clsr.2024.105988.

I cooperated on the analysis of legislation and contributed to text writing. Contribution 10%.

[A18] Daniel Prudký, Anton Firc, and Kamil Malinka. "Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation". *In: 2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2023, pp. 1–5. ISBN 978-3-88579-733-3. DOI:10.1109/BIOSIG58226.2023.10346006.

I led the research, proposed the idea of using a cover story for the desired use case, cooperated on the experiment design and analysis of results, and contributed to text writing. Contribution 45%.

[A20] Kamil Malinka and Anton Firc. "Deepfakes: příležitost, nebo hrozba?" Czech. *In: Proč se nebát umělé inteligence?* Praha, CZ: Nakladatelství JOTA, s.r.o., 2024, pp. 271–283. ISBN: 978-80-7689-459-4.

I determined the concept of the whole chapter and wrote a significant part of the text. Contribution 50%.

Chapter 3

Impacts of Deepfakes on Biometric Authentication

In the previous chapter, we discussed the increase in attacks utilizing deepfakes on people. Recently, we have also seen increasing reports of attacks using deepfakes that target technical resources. An example is using morphed images to fool automated gates at airports [36]. There are also multiple real-world reported attacks involving synthetic speech in multiple thefts [25, 37, 38, 39, 40]. The scammers posed as the CEO of an energy company and managed to extort a payment of 250k USD [41]. China has also reported an incident involving a successful deepfake spoofing attack on facial recognition. In early 2021, tax fraudsters used deepfake videos to trick the tax invoicing system into accepting premade deepfake identities to defraud $76.2 million [42].

As shown in the previous chapter, we can't rely on humans' abilities, so we decided to focus on the technical means they use. Since we want to continue focusing on user-related security measures, we decided to investigate the area of user authentication and test their resistance to deepfakes.

Authentication is taken into account while discussing persons or data. The practice of verifying the original data source is known as data authentication. User authentication is binding an identity to a subject [43]. The subject must provide some proof of his identity. In this thesis, we focus on user authentication.

There are usually four ways an entity can provide needed information:

1. What the entity knows (such as passwords or secret information)

2. What the entity has (such as a badge or card)

3. What the entity is (such as fingerprints or retinal characteristics)

4. Where the entity is (such as in front of a particular terminal)

Like any security mechanism, authentication is also vulnerable to various attacks. Therefore, a combination of single-factor approaches is often used to increase its resilience - so-called multi-factor authentication (MFA) [44]. It is also one of the solutions to partially mitigate the human factor, as users very well accept MFA [45]. In the banking sector, MFA is even required for some types of sensitive transactions by legislative standards such as PSD2 [46].

A successful attack on the MFA means a concerted attack on each factor where each part of the attack must succeed. Since only one of the factors is vulnerable to deepfake-based attacks, we, therefore, focused primarily on biometric authentication.

3.1 Deepfake-based attacks on biometric authentication

The concept of attacks utilizing deepfakes was defined more than 10 years ago, and many studies confirmed the vulnerability of verification systems technology to spoof attacks [47]. Also, proof-of-concept on spoofing voice verification was presented at the 2018 Black Hat conference by J. Seymour and A. Aqil [48].

Although the principles of spoofing attacks and possible defenses were defined, these were more general concepts for testing detection methods, which were difficult to apply in practice. Previous works examining the feasibility of deepfake-based spoofing attacks were focused on testing of detection methods, not the whole deployed systems [49, 50]. How individual attacks could be implemented in practice was not sufficiently investigated, nor was their impact clear. We have tried to cover this gap with our research. In our work, we have primarily focused on voice authentication; however, we have research that overlaps with facial authentication because some of the new attacks integrate multiple types of deepfakes.

Authors contribution

First, we take advantage of one of our previous works to set the context for the security of a general authentication method. In our work focusing on e-banking security [A21], we covered authentication mechanisms with a strong focus on the e-banking specifics in greater detail. As the main contributions, we provided a comprehensive overview of authentication schemes and their security evaluation. We also proposed the taxonomy for attacks on e-banking compatible with the general authentication taxonomy by NIST [51], and we discuss security features of authentication schemes in the context of the European directive - Payment Services Directive version 2 (PSD2) [52], which requires satisfaction of various features such as strong authentication. In this research, I defined a taxonomy of attacks, analyzed legislation, and created an overview of current authentication methods and their properties in the context of international standards.

However, the most relevant conclusions of this work for the next part of the thesis are related to biometric authentication. Due to the massive expansion of the use of smartphones, the spread of biometrics has increased significantly. The main reason is the excellent usability and integration directly into the smartphone. Biometrics are used to strengthen the ”Know Your Customer” process (KYC, the process for client identification when opening an account, to be done periodically over time) or for device authorization when using a hardware token, a dynamic password generator, or a secure enclave. One

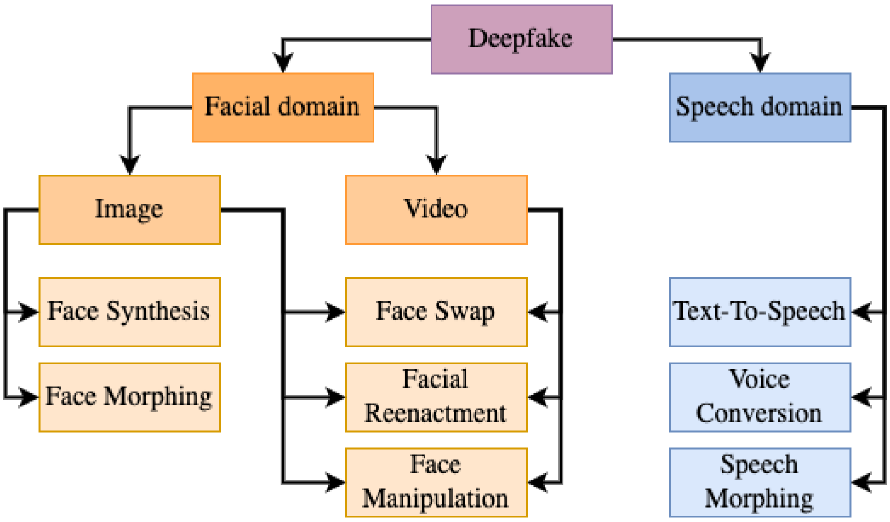


Figure 3.1: Visualization of used united taxonomy of deepfakes. Categories are divided according to the domain and media type. For the facial domain, most of the categories fall under both media types. Figure taken from [A16].

of our conclusions is that biometrics are still vulnerable to, e.g., replay attacks or man-in-the-middle attacks. In addition, another new and much more serious threat has emerged, namely spoofing through deepfakes.

Impacts of deepfakes

In further research, we focused on deepfakes and investigated their impacts on the types of biometric authentication that are vulnerable to them - specifically, voice and face authentication.

In this part, I was again the principal investigator, setting the direction of our research group in this area and working on the topic with my PhD students. I defined the need for detailed research on attacks with a security overlay, determined the appropriate research methodology, and performed the security analysis. The selected attacks (specifically attacks on voice assistants) were entirely under my responsibility - from the design of the focus to the design of the method to the implementation of the experiment and its evaluation.

For a thorough orientation in the field, we have published a survey [A16], which provides a united taxonomy for facial and speech deepfake attacks (see Figure 3.1), defines differences between each category, and provides an overview of deepfake creation tools, available datasets, and detection techniques. We also define attack vectors for each deepfake category. These attack vectors respect the differences in all deepfake categories and show the potential of each category to spoof biometrics systems and their usability in other types of attacks.

We can use face swap as an example. *Face swapping* refers to a technique where a face from *source* photo is transferred onto a face in a *target* photo (see Figure 3.2). Face swapping can be misused, for example, to impersonate someone else on a Zoom call, pornographic material for slander, or to attack facial biometrics. While face swapping

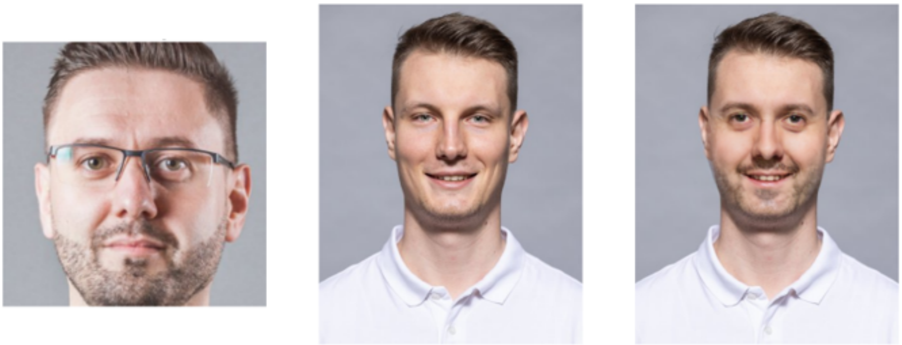


Figure 3.2: Face swap. a) The source image (victim), b) The target image (attacker), c) The result of automatic face swap. Figure taken from [A16].

can be more easily computed and performed in real-time, it does not hide most of the actor's image, so a potential attacker is limited by having to maintain some similarity to the victim in the actor. In contrast, face reenactment hides the actor completely, making the attack easier.

In our work [A16], we summarized the latest developments in each category. However, our main contribution was the analysis of the threats posed by each type of deepfake and their combination, as well as a review of the current tools. This information gives us the baseline for performing a correct risk analysis of the deployed system and correctly assessing the attacker's strengths and capabilities, which, in turn, impacts the proper evaluation of the price and the attack's impact.

Practical performing of attacks

While analyzing the current state-of-the-art research, we found a lack of security overlap. Although the principles of spoofing attacks and possible defenses were defined, these were more general concepts for testing detection methods, which were difficult to apply in practice. How individual attacks could be implemented in practice was not sufficiently investigated, nor was their impact clear. We decided to fill this gap, so we focused on specific types of attacks in more detail.

We contributed to authentication security by conducting multiple attacks on state-of-the-art authentication mechanisms by utilizing relevant deepfake synthetic media and exploring their feasibility. We also performed an impact analysis of successful attacks. Specifically, we focused on voice biometric authentication used in the KYC process and voice assistants, and next, we also targeted facial biometrics. We used commonly available tools for the attacks. Our results showed the practical feasibility of selected attack types and contributed to a better understanding of the whole process. This information is essential for the design of effective defenses.

Typically, we have tried to target the expected types of attackers: the casual to moderately advanced user who can use commonly available models, the advanced attacker who can develop custom models tailored to his purpose, and the strongest type of attacker imaginable in a corporation with unlimited technology. According to the attacker model, we chose synthesizers of appropriate quality and attack vectors.

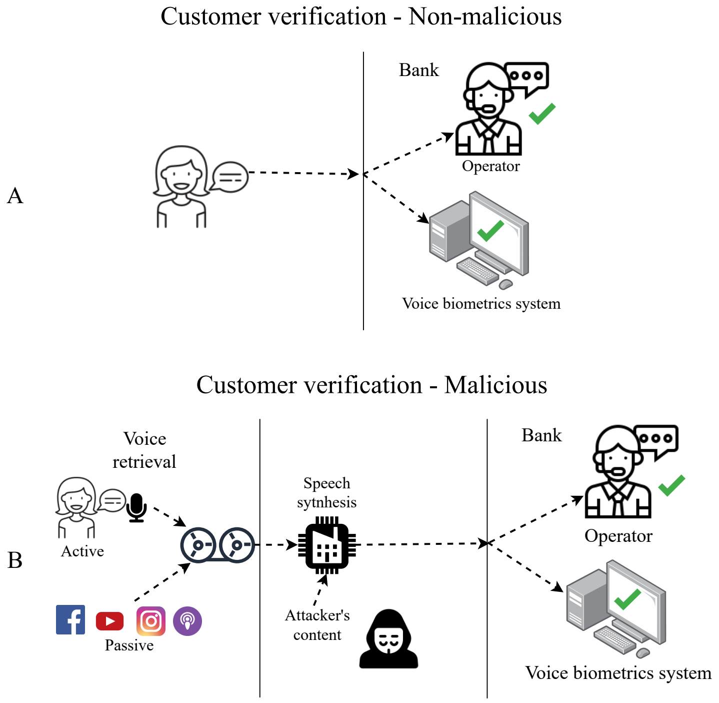


Figure 3.3: Attack scheme. Part A represents non-malicious (genuine) access to customer care call center, Part B represents malicious access with target voice retrieval phase and speech synthesis. Figure taken from [A17].

To verify the feasibility of the attacks, evaluate the necessary tools, and assess the effectiveness, we researched several types of attacks - an attack on the ”Know you customer” KYC process implemented using voice and facial biometrics [A17, A22] and an attack on voice assistants [A23]. We managed to execute all the attacks successfully and proved their high efficiency. For selected attacks, we evaluated the impact of new technology (diffusion models) on their success rate [A24].

Attacking KYC process using voice biometrics

First, we focused on the area of customer verification in companies providing customer care call centers, which is often used also for the ”Know you customer” (KYC) process [A17]. Usually, setup is a combination of human operator and biometrics (see Figure 3.3). While the customers talk to the operator about their request, the voice biometrics system verifies the customer’s identity. After successful authentication, the operator executes the customer’s requested action.

For the practical implementation of the attack, we have chosen the following procedure. First, we created a dataset by using text-to-speech tools for deepfake creation - two commercial tools *Overdub* [53] and *ResembleAI* [54] and one open-source tool *Real Time Voice Cloning* [55]. The created dataset consists of genuine and deepfake speech of 100 English and 60 Czech speakers selected from the Common Voice Corpus [56]. Next, we tested bona fide samples and created deepfake samples on two voice biometrics systems: *Microsoft Speaker Recognition API* [57] and *Phonexia Voice Verify demo* [58].

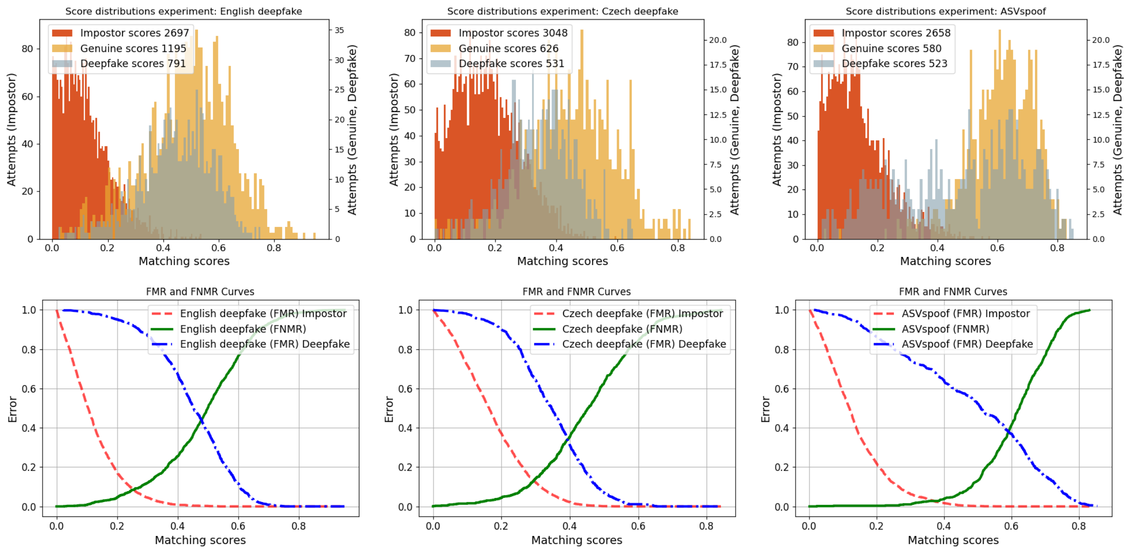


Figure 3.4: Comparison scores distribution graphs (top) and FMR / FNMR graphs (bottom). The left plots represent the created English deepfake dataset. The middle plots represent the created Czech deepfake dataset. The right plots represent the ASVSpoof 2019 challenge dataset [34]. Figure taken from [A17].

We examine the differences by collecting comparison scores of the genuine, impostor, and deepfake attempts. Then we compare collected comparison scores by plotting score distribution plots and false non-match rate (FNMR) and false match rate (FMR) curves.

As Figure 3.4 shows, the deepfake dataset performed very well. The deepfake comparison score distributions almost identically overlay the genuine comparison score distributions, showing that the tested voice biometrics systems could not detect synthetic speech.

The main contribution of the successful practical attack on a state-of-the-art voice biometric system we have demonstrated in this paper is the exploration and confirmation of the technical feasibility of the attack and the provision of information on the range of victim samples needed for a successful full attack. We also investigated the difference between text-dependent and text-independent verification, where text-dependent verification was shown to be more robust to this type of attack. If we combine the results with the results of attacks on humans, we find that this method does not provide sufficient robustness in either part.

Attacking voice assistants using voice biometrics

In our other work, we experimentally demonstrate the vulnerability of four voice assistants (Google Assistant, Siri, Bixby, and Alexa) to attack based on voice deepfakes and replay attacks [A23]. As part of the experiment, we also evaluate the suitability of the selected speech synthesis tools for this type of attack. We also analyzed the proposed scenarios to evaluate the security impacts of demonstrated attacks.

Seventy-two respondents participated in the experiment in a controlled environment. Each participant was enrolled in all voice assistants and performed 30 bona fide authen-

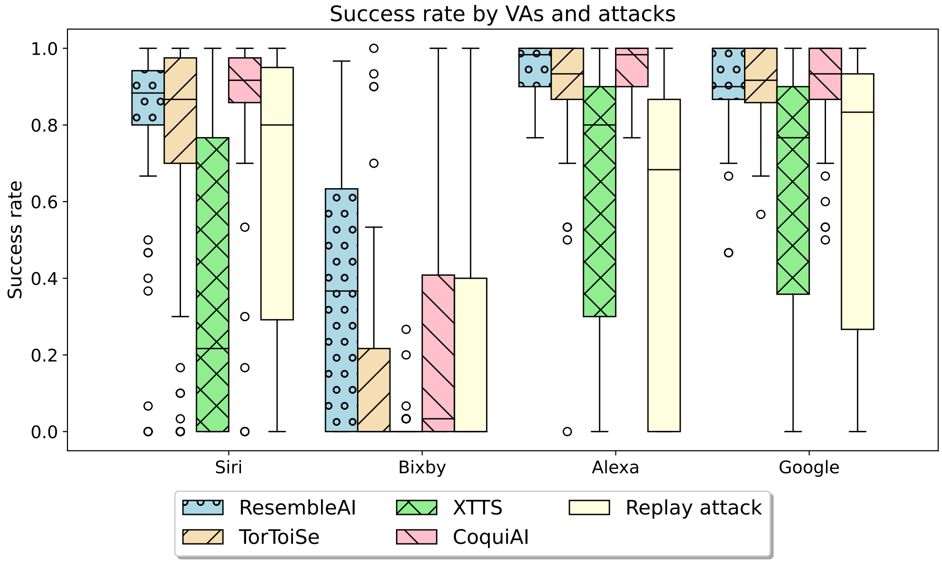


Figure 3.5: Success rates of attacks on voice assistants. Figure taken from [A23].

tication trials with each assistant. Next, we recorded the participant's speech and created a deepfake speech using four state-of-the-art speech synthesizers in a text-to-speech (TTS) setting - *CoquiAI*[1] [59], *ResembleAI* [54], *TorToiSe* [60], and *XTTS* [61]. To perform replay attacks, we replayed the original sentences with wake words for each assistant. Next, we played the wake words synthesized using different tools to all assistants. The success rate was computed to evaluate the efficacy of each verification attempt.

The breakdown of attack success rates is shown in Figure 3.5. The replay attacks succeeded approximately every second time, while some of the deepfakes reproduced the bona fide success rates of more than 90%. We then conducted a threat analysis to discuss a potential attack's impact. The analysis revealed potential privacy breaches and financial damage, but also that assistants do not allow activation of critical functions, such as payments, via voice commands. Overall, we showed the great vulnerability of voice assistants and the importance of choosing the appropriate authentication mechanism for a desired use case.

Attacking KYC process based on face biometrics

Another type of attack we experimentally verified was the attack on face biometrics [A22].

First, we analyzed the attacker model. It is important to understand where this kind of attack makes sense because it is sometimes easier for attackers to use different technical means (see Figure 3.6).

Based on our analysis, we defined three categories of system types. The first category includes systems where implementing an attack is difficult or infeasible. These cases mostly involve access control, often supplemented by human surveillance, where an attacker pointing a tablet at the camera would be suspicious and probably caught very quickly. The second category includes use cases where deepfake spoofing is unnecessary.

[1] Discontinued in 12/2023.

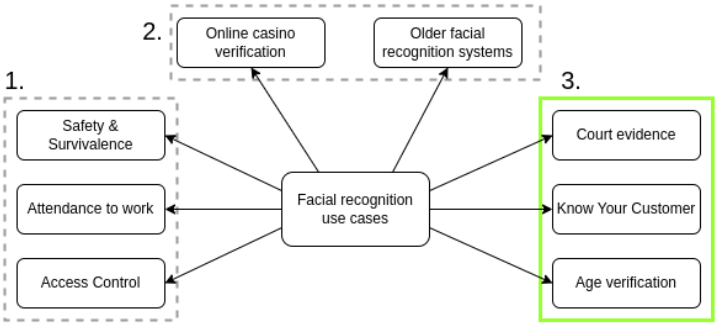


Figure 3.6: Scenarios using facial recognition can be divided into three categories depending on whether it makes sense to attack them using deepfakes. Category 1: It is difficult to attack them with deepfakes. Category 2: It is not worth using deepfakes to attack. This category includes ”Older facial recognition systems”, which refers to all use cases of outdated and poorly designed systems for which more conventional attack methods are sufficient. Category 3: Appropriate use of deepfakes in an attack (green solid line). Figure taken from [A22].

Such use cases typically involve outdated or simple facial recognition systems vulnerable to basic presentation attacks. Thus, the extra effort to create a deepfake is unnecessary from the attacker’s perspective. The third category includes use cases well suited for deepfake spoofing attacks, including current biometric authentication or age verification systems. These use cases allow a meaningful implementation of spoofing attacks because they mostly use video-based input data for verification, and are usually remote (without surveillance).

Next, we experimentally performed a deepfake-based attack on the selected commercial systems (*IFace SDK 3.0* [62] and *Megamatcher* [63]) and evaluated its effectiveness. We tested two scenarios. In the image-to-profile scenario, only a single facial image (frame) was extracted from the input video and compared to the user profile stored in the database. In multiple-image-to-profile scenarios, multiple frames were extracted from the video and compared to the profile. The scores for each frame were averaged. For both phases of the experiment, individual images and image sequences of 58 identities were selected from the Celeb-DF dataset [64] and fed into systems via their API.

Similar to previous attacks, results show insufficient resilience of current systems. The plot clearly illustrates a disturbing phenomenon - the overlap between deepfake and genuine scores (see Figure 3.7). Some of the deepfake samples have reached a sufficient threshold for acceptance by the system. However, modifying the threshold settings to also reject these samples would result in a significant increase in the rejection of eligible samples.

Similar results were obtained for the second, more advanced scenario. Using multiple frames extracted from the video was expected to provide more robustness as it can identify inconsistencies between these frames. However, the comparison scores distribution graphs have the same properties as in the previous scenario. The experimental results thus show the high vulnerability of facial biometrics, where even advanced approaches do not provide good protection as the systems struggled to identify deepfakes even with multiple snapshots.

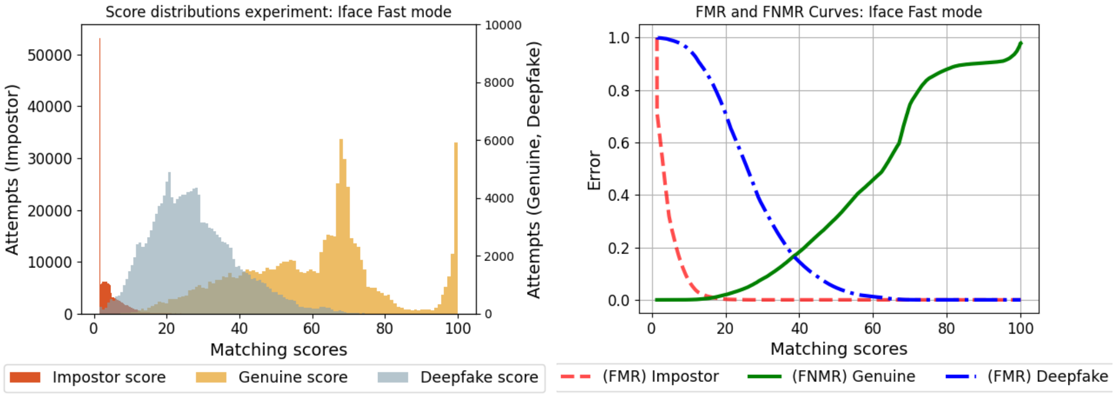


Figure 3.7: Comparison scores distribution graph on the left and right FMR / FNMR graphs for the IFace fast mode for image-to-profile comparisons. Figure taken from [A22].

Impact of diffusion models on speech synthesis

An essential part of deepfake research is the impact of new approaches and models that can improve the quality of a deepfake or the speed of its creation. Improvement has implications on the attacker's power, where, for example, the ability to create a deepfake in real time has significant consequences because it enables new types of attacks. To keep up with the attackers, we explored diffusion models, a novel method for creating a realistic synthetic speech, and their impact on the attacker's strength [A24].

Diffusion models have emerged as a new technique for producing highly realistic synthetic speech [65]. To evaluate their impact on security, we compare diffusion-generated deepfakes with non-diffusion-generated ones in the context of their ability to fool the deepfake speech detectors. We also focus on the quality and characteristics of generated speech to determine if they present a more significant threat.

The experiment aimed to determine whether diffusion-based synthesizers produce deepfakes of better quality. To verify this, we created a dataset consisting of deepfake samples created by diffusion synthesizers and non-diffusion synthesizers. We used representatives of the four basic synthesizer types: diffusion synthesizers with non-diffusion vocoders, diffusion-only synthesizers, diffusion-based vocoders, and non-diffusion synthesizers. Next, the datasets was used to evaluate three state-of-the-art (SOTA) deepfake speech detectors: *LFCC-LCNN* [66], *Wav2vec + GAT* [67] and *IDSD* [A25]. Part of the experiment also involved evaluating the quality of the synthesized speech.

The results show a low impact of the diffusion models as the effectiveness of detection remains consistent across both types of synthesizers. Also, the similarity of the speaker was consistent, and the quality of the generated speech was comparable.

The main contribution is the dataset itself, as we have published the dataset, to enable further analysis in this area. We also plan to further expand it in the future. The second contribution is the finding that diffusion models do not introduce additional negatives and do not increase the strength of an attacker using voice deepfakes. Deepfake samples produced by these models have similar properties to existing deepfakes and are detectable by existing detectors at a comparable level.

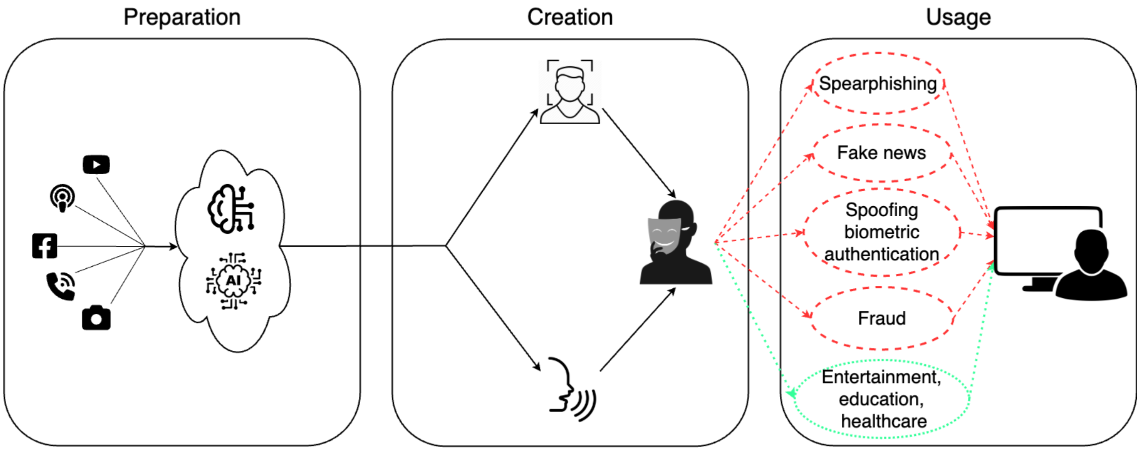

Preparation
Creation
Usage
Spearphishing
Fake news
Spoofing biometric authentication
Fraud
Entertainment, education, healthcare
AI

Figure 3.8: Deepfake lifecycle with areas allowing mitigation of deepfake threats. Malicious usages are visualized using dashed (red) lines, and beneficial usages using dotted (green) lines. Figure taken from [A26].

3.2 Protection against deepfake attacks

In the previous chapters, we have discussed the implementation of attacks using deepfakes, the factors affecting their success, and their impact. The results show that they represent a real threat that needs to be addressed. Thus, we also decided to contribute to the area of protection methods.

A report [68] published by the U.S. Department of Homeland Security in 2021 presented broad-based deepfake mitigation measures. Six sequential areas were defined: Intent, Research, Create, Disseminate, Viewer Response, and Victim Response. These areas map the whole cycle from the malicious actor's first idea to the targeted individual's response. For each phase, stakeholders and potential mitigation measures are identified. However, the report is oriented towards policymakers and the legal side of mitigation measures, primarily focusing on the US. The mitigation measures are rather broad and vague and do not include the required technical details.

In contrast, in our approach, we focus on the technical aspect of deepfake mitigation. For a better understanding of the problem, we will present the deepfake use lifecycle and illustrate other applications of protection mechanisms on top of it [A26]. Most of the discussion that has been devoted to protection mechanisms has focused on deepfake detection. The disadvantage of this approach is that if detection fails, no other protection is standing in the way of a successful attack. Detection is thus an important part of building protection, but it is not the only option.

Deepfake lifecycle

The deepfake lifecycle (see Figure 3.8) is an abstract model describing the *life of a deepfake* from its creation to mis/usage. This lifecycle might be divided into three distinct areas where deepfake threats might be mitigated: *preparation*, *creation*, and *usage*.

The first area in the deepfake life cycle is preparation, which includes the following stages:

- **Data collection**: The first step in preparing for the creation of deepfakes is to collect data that will be used to create them. This may involve collecting images, videos or utterances of the target individual, along with any other relevant data that can be used to create a convincing deepfake.

- **Data transformation**: Once the data has been collected, it may need to be transformed or processed before it can be used for training or inference. This may involve cleaning the data, removing noise or artefacts, or converting it into a format that can be used by the deepfake creation tool.

- **Training the model**: After the data has been collected and transformed, the next step is to train the deepfake model. The training process involves using large labelled datasets.

- **Fine-tuning the model**: Once the model has been trained, it may need to be fine-tuned or adjusted to improve its performance. Fine-tuning is often done to published pre-trained models to adjust them to a specific individual. Much less data is required than for training the whole model.

Overall, the preparation area of the deepfake life cycle involves collecting and preparing the data, models, and tools needed to create deepfakes. By collecting high-quality data, transforming it as needed, training the deepfake model, and fine-tuning it for optimal performance, malicious actors can create convincing and realistic deepfakes.

The second area involves using the prepared data and tools to generate the deepfake persona. The key stages are as follows:

- **Inference**: The first step in creating a deepfake is to use the trained model to generate the fake media. This process involves using the deepfake tool (speech synthesizer, face-swap application, etc.) to generate images, videos, or utterances based on the input data collected during the preparation stage.

- **Image processing**: Once the deepfake media has been generated, it may need to be processed further to improve its quality or enhance its realism. This may involve using image processing techniques such as noise reduction, color correction, or sharpening to improve the visual quality of the deepfake.

- **Audio processing**: Besides processing the visual aspect of the deepfake media, audio processing may also be required to improve the quality of the audio included in the deepfake. This may involve noise reduction, filtering, or equalization techniques to enhance the audio quality and make it more convincing.

- **Integration**: Once the deepfake media has been generated and processed, it may need to be integrated into a larger project or application. This may involve integrating deepfake media into a video editing project, a virtual reality application, or a social media platform.

The creation involves using the prepared data and tools to generate deepfake media. By using the deepfake tool to generate images, videos, or utterances and processing the media to improve its quality and realism, malicious actors can create highly convincing and realistic deepfakes that can be used for various purposes.

The third area involves the distribution and misuse of deepfake media. This stage is critical because it determines the ultimate impact of the deepfake, which can be either positive or negative. We have already covered various types of misusage in previous sections.

Protection methods

Protection against deepfakes usually focuses on detection; however, this is deployed in the last phase of the lifecycle, and failure means a successful attack. Realistically, we could have a much more diverse range of protection measures deployed in other parts of the attack lifecycle: watermarking [69], legal regulations such as AI Act [70] and related legislation (in the preparatory phase) [71], Digital Services Act [72], methods for obstructing deepfake creation [73, 74], forensics analysis [75], methods ensuring proof of authenticity [76, 77], or straightforward removal of vulnerable components.

Digital watermarking: Digital watermarking is a process of imperceptibly altering a piece of data to embed information about the data. Watermarking has been accepted as an effective and practical technique to protect the copyright of digital multimedia [69]. The watermark should not be easily removed or added to the media, but this is difficult to accomplish, so it remains a research problem. The digital watermark might be used in two major ways: *Verifying source* and *Marking synthetic media*. The first uses a watermark to verify the media is genuine [78]. The media capturing device, such as a digital camera, may add this watermark. The second uses a watermark to mark synthetic media. Developers of deepfake creation tools, both commercial and open-source, might be required to include such a watermark in the output of their tools.

Legal regulations: The regulation of deepfakes at both EU and national levels involves a complex framework of hard and soft laws. However, enforcement remains challenging, often failing to protect victims adequately. The involvement of multiple entities in the deepfake lifecycle fragments responsibilities. Priority should be given to defining the obligations of technology providers and hosting platforms, which play central roles in creating and disseminating deepfakes. A recent study by van Huijstee et al. [79] comprehensively analyses possible ways forward. The most effective regulatory framework for the Preparation area seems to be the AI Act [70] and related legislation (in the preparatory phase) [71]. Also, in the Creation area, the regulatory framework for AI allows for the development of labeling guidelines accompanied by a broad obligation to label deepfake. Furthermore, specific applications with clearly negative impacts, such as non-consensual deepfake pornography, should be expressly prohibited. Additionally, the European democracy action plan [80] offers a suitable instrument for banning deepfakes containing political disinformation and manipulative communication. National-level criminal law should be revised to encompass and react to creating specific deepfake tools. The Usage area is where the platforms should play the central role in helping to detect deepfakes, support victims, and identify perpetrators. Digital Services Act [72] can serve as a basis for having deepfake detection and authenticity verification systems in place. Independent oversight and increased transparency should allow for the development and deployment of additional third-party measures and solutions that protect individuals. Furthermore, GDPR [81], in its current version or with partial revision, provides a suitable vehicle for additional regulatory measures, primarily EDPB guidelines[2]. GDPR is highly relevant as voice and facial data should be considered as biometric data and protected as special categories of personal data. On top of that, the trust services introduced by the eIDAS Regulation [82] could be based on further guidelines and explanations specified to be offered.

Obstructing deepfake creation: We can also try to obstruct deepfake creation. Li et al. [73] proposed the Landmark Breaker method that disrupts facial landmark extraction and thus obstructs the usage of facial images for deepfake creation. Khachaturov et al. [83] proposed a process that allows augmenting any arbitrary image so that any attempt to edit it using a specific model will add arbitrary visible information. Such obstruction techniques might be applied directly by a device used to capture the media, manually by the creator, or automatically when sharing the media online.

Proof of authenticity: Proofs of authenticity are common for physical media, such as a certificate of authenticity (COA) given with the purchase of artwork. A similar concept might be transferred into the digital domain to prove the authenticity of digital content. An example of such a system has been presented by Hasan and Salah [76], who use blockchain technology to pose as proof of authenticity by providing credible and secure traceability to the source. Boneh et al. [77] suggest using a cryptographic content-signing key. All media exported from a camera would be signed, meaning that every piece of media would have a digital signature identifying the device used to capture it. The functionality of this kind is available using ProofMode[3].

Deepfake detection: However, current deepfake protection efforts focus mainly on detection. In the area of deepfake face detection methods, there are plenty of methods: detection based on artifact detection [84, 85, 86, 87], based on deep learning [88, 89], or based on physiological features such as eye blinking [90, 91].

In the area of deepfake voice detection, the community has long been involved in developing detection methods. One of the ways is participation in competitions for the most effective detector - e.g., the recent ASVspoof challenge 5 [92]. However, the transfer of obtained results to security practice is problematic and slow. Published methods are mostly not commercially available and not ready for production deployment, leaving users to defend against the increased attacks that have proliferated in recent years.

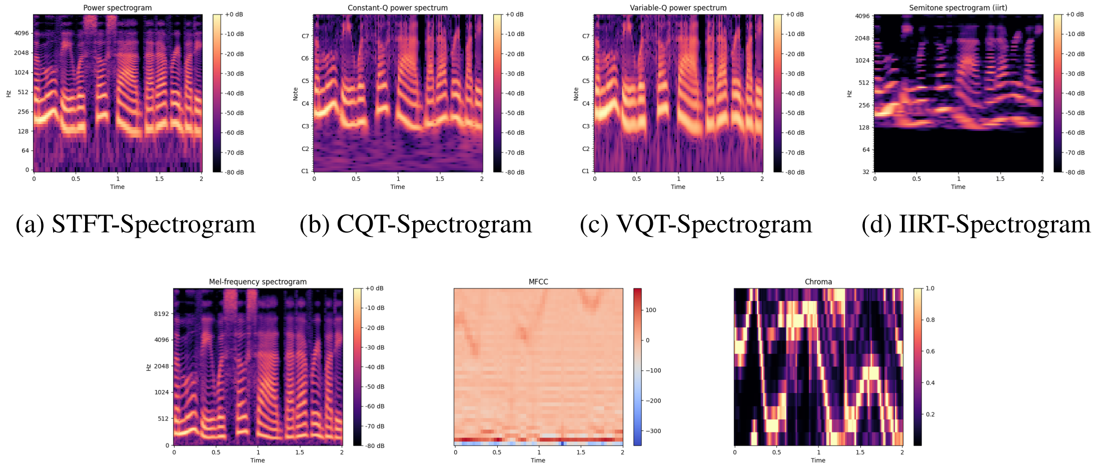
Based on ASVspoof challenge 5 results[4], current detectors have architectures based on deep neural networks [92]. Most systems in the challenge are based on the AASIST framework [93] and pre-trained SSL models, such as Wav2Vec2 [94] or WavLM [95].

In addition, the authors of the detection methods themselves are aware of the lim-

[2] https://edpb.europa.eu/our-work-tools/general-guidance/guidelines-recommendations-best-practices_en

[3] https://github.com/guardianproject/proofmode-android

[4] https://www.asvspoof.org/



(a) STFT-Spectrogram (b) CQT-Spectrogram (c) VQT-Spectrogram (d) IIRT-Spectrogram

(e) Mel-Spectrogram (f) MFCC-Cepstrogram (g) Chromagram

Figure 3.9: Examples of used spectrograms. Each image represents the same recording. Figure taken from [A25].

itations of their approach - to ensure comparability of results, the dataset used in the competition is specially prepared. Despite the authors' efforts to update it as much as possible, it cannot reflect all attacks. It is also often focused on a narrower area. It does not accentuate the problem of generalization, where even the authors of successful models themselves point out that the identical successful architecture does not achieve the same results on other datasets - e.g., the older ASVspoof challenge [96].

Authors contribution

Our goal is primarily to bring the existing field of detection design closer to real-world cybersecurity and to emphasize the need to find effective ways of translating scientific results into practice so that we have the tools to protect us from previously theoretical attacks. We also wanted to contribute to the field, so, in addition to working with methods that already exist, we designed and evaluated a new detection mechanism based on spectrogram analysis.

I was primarily responsible for defining the deepfake lifecycle and finding appropriate measures for each area. I also defined the need to focus on areas other than detection. My role in the case of the detector design was primarily to supervise the methodology. In the field of detector comparison, I defined this research need and participated in the experimental evaluation of the detectors.

Detection based on spectrogram analysis

In our work [A25], we build on Reimao [97], who first proposed image-based deepfake speech detection, and Khochare et al. [98] who later extended the idea and explored the behavior of Temporal Convolutional Networks (TCNs) with Mel-Spectrogram as input.

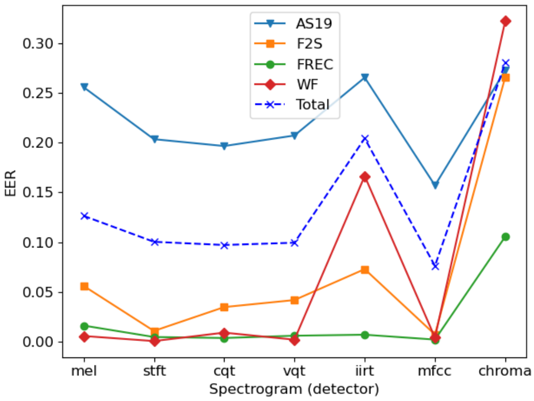


Figure 3.10: Equal Error Rates (EER) for each training dataset and detector. The dashed line shows the total EER for each detector. Figure taken from [A25].

We designed and implemented various detectors utilizing different spectrograms as inputs and evaluated their performance and data requirements on multiple datasets.

In the first part of the experiment, we evaluated the accuracy of selected spectrograms (see Figure 3.9). In the second part, we investigate the data requirements for each spectrogram and compare the spectrograms in terms of storage space consumed, RAM consumed during training, and time required to extract each spectrogram.

Figure 3.10 shows the comparison of EER collected from all detectors. Our other results may be of interest in a context where you have limited resources and are looking for a detector that best handles this limitation. STFT-spectrogram is the most resource-hungry. It is also the fastest one for extraction. However, in scenarios with unlimited resources, the STFT-spectrogram provides the highest accuracy. The accuracy over known data is the best when using the MFCC-spectrogram, while its data requirements are among the lowest. The accuracy over unknown data is the best for Mel-spectrogram and VQT-spectrogram.

As it turns out, there is no single answer to the question of which detector is most effective or which dataset is best. Many parameters can be taken into account, and detector behavior varies significantly based on these parameters. It is, therefore, essential to understand the environment in which such a detector will be used and to set it up to give the best performance under the given circumstances.

Evaluation of existing detection methods

As already mentioned, comparing the performance of individual detectors is difficult. The main reason is the different testing methodologies and test datasets used by the individual authors. Another key challenge in the field of false speech detection is generalization—ensuring good detector performance under different and unprecedented conditions, such as different speakers, recording environments, and false speech generation techniques. While models may excel on their training datasets, they often have problems with real data, which limits their effectiveness. An important activity in this area, the ASVSpoof challenge [34], particularly ASVSpoof5 [92], addresses this problem by evaluating detectors' generalization capabilities to improve their robustness in practical applications.

We decided to address both problems in our next work [A27]. We focused on the problem of a detector comparison and tried to find an answer to how to develop detectors in a way that makes them easier to compare because the existing methods used in the community do not sufficiently reflect the general quality of the detector and its generalization capability. We also aimed to provide guidance on how to test and develop new methods to ease their transition to practice.

We proposed a detailed framework for evaluating and comparing deepfake speech detectors. Our goal was to provide a testing environment that ensures replicability of experiments, comparability of deepfake speech detectors, and easy incorporation of new deepfake speech detectors.

Setting up a proper testing environment is quite complicated, as we can see for example in the ASVSpoof challenge. So it is not advisable to force the detector authors to solve this additional problem and it is useful to provide them with best practices on how to carry out the evaluation process. While they are proficient at model design, it is more complicated to achieve correct verifying processes, which take into account all advancements.

To showcase the usage of our framework and the benefits it can bring, we then used this framework to evaluate 40 state-of-the-art deepfake speech detectors. We performed extensive experiments, where we extended common approaches by testing for previously unobserved forms of manipulated speech. In fact, we extended the testing to include simple modifications that an attacker could use to prevent detection. We also investigated the most appropriate detector architecture concerning accuracy and robustness.

The basic principle of the framework is straightforward: it allows us to compare different detectors on different datasets. It provides us consistent environment, as it allows us to train and evaluate various detectors on the same datasets. This ensures that any differences in performance are due to the design or characteristics of the detectors rather than to inconsistencies in the data.

The framework's flexibility allows easy expansion with new detectors, validation datasets, or training data. While adding a new detector is a simple operation, adding a new validation dataset requires evaluating each detector to obtain valid results. The most demanding part is the change in training datasets, as it requires complete re-training of all detectors and their full re-evaluation, which is time and resource-demanding.

To demonstrate our framework, we conducted experiments comparing state-of-the-art deepfake speech detectors and assessing their robustness, particularly against potential attacks. We selected 14 publicly available implementations for testing. As some implementations contained multiple models, the total number of tested methods increased to 40.

To achieve the goal of testing resistance to methods that can be used by a common attacker, it was necessary to create a modified dataset that contained the application of these modifications to common datasets before the actual testing. We selected modifications to simulate real-world audio distortions, including environmental noise (white, street, bird), compression artifacts (MP3, WMA), frequency reduction, and volume reduction, reflect-

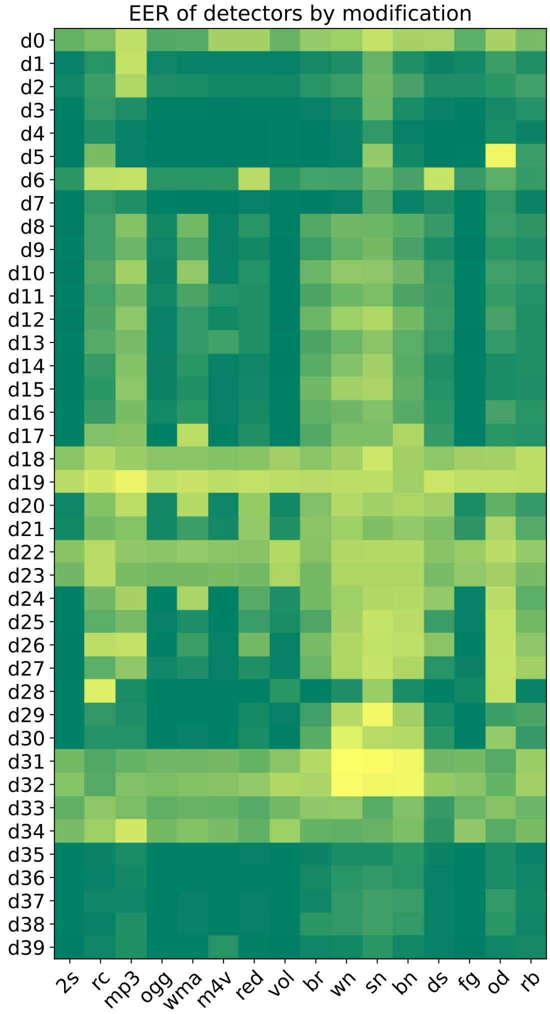


Figure 3.11: Heatmap visualisation - Each row represents a tested detector and each column modification to speech. The visualisation highlights the least robust detectors and the most challenging modifications. Green denotes low EER (good), and yellow denotes high EER (bad). Figure taken from [A27].

ing challenges like lossy transmission, low-quality microphones, and weak signals.

The resulting comparison of detectors yielded several interesting conclusions (see Figure 3.11). Some detectors show inferior performance compared to others and specialized architectures proved to outperform more general architectures. As you can see in the last 5 lines of the figure, the important difference was also played by whether the detectors experienced some form of adversarial sampling during training as they simulated the challenging conditions of evaluation. Furthermore, it turned out that although the resulting architecture is important, the proper training procedure plays a more significant role, as some of the inferior detectors outperformed the better detectors when properly trained on a well-constructed dataset.

We also verified the successful modifications using biometric authentication to see if they succeeded for both control mechanisms. This is because in practice we can expect interconnection between the detector and the biometrics, where rejection of a sample by one of the components means automatic rejection. Results show that most modifications have little to no effect on the speaker recognition system's performance.

We believe that our framework has many uses (and opens up other interesting directions of research): evaluation of new detectors, identification of superior approaches, architectures, and training data, the formal basis for the creation of large-scale deepfake detection evaluation tools, and a prelude to certification of deepfake speech detection systems involving advanced acceptance testing.

Contributed papers

This chapter is based on our 9 research articles, parts of which are included in this thesis.

Articles in collection

[A21] Kamil Malinka, Ondřej Hujňák, Petr Hanáček, and Lukáš Hellebrandt. "E-Banking Security Study—10 Years Later". *In: IEEE Access* 10 (2022), pp. 16681–16699. DOI: 10.1109/ACCESS.2022.3149475.

I led the research, proposed the attack taxonomy, cooperated on the analysis, and significantly contributed to text writing. Contribution 35%.

[A16] Anton Firc, Kamil Malinka, and Petr Hanáček. "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors". *In: Heliyon* 9.4 (2023), e15090. ISSN: 2405-8440.
DOI: https://doi.org/10.1016/j.heliyon.2023.e15090.

I cooperated on the analysis and contributed to text writing. Contribution 45%.

[A17] Anton Firc and Kamil Malinka. "The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems". *In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. SAC '22. Virtual Event: Association for Computing Machinery, 2022, pp. 1646–1655. ISBN: 9781450387132. https://doi.org/10.1145/3477314.3507013.

I cooperated on the design and analysis of experiments and contributed to text writing. Contribution 50%.

[A23] Kamil Malinka, Anton Firc, Petr Kaška, Tomáš Lapšanský, Oskar Šandor, and Ivan Homoliak. "Resilience of Voice Assistants to Synthetic Speech". *In: Computer Security – ESORICS* 2024. Ed. by Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas. Cham: Springer Nature Switzerland, 2024, pp. 66–84. ISBN: 978-3-031-70879-4.

I led the research, proposed the idea of an attack on voice assistants, designed the experiments, cooperated on the performance of the experiments and analysis of results, and contributed to text writing. Contribution 48%.

Other relevant publications

[A22] Milan Šalko, Anton Firc, and Kamil Malinka. “Security Implications of Deepfakes in Face Authentication”. *In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC ’24. Avila, Spain: Association for Computing Machinery, 2024, pp. 1376–1384. ISBN: 9798400702433.
https://doi.org/10.1145/3605098.3635953.

I cooperated on the design and analysis of experiments and contributed to text writing. Contribution 25%.

[A24] Anton Firc, Kamil Malinka, and Petr Hanáček. “Diffuse or Confuse: A Diffusion Deepfake Speech Dataset”. *In: 2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2024, pp. 1–7.
DOI: 10.1109/BIOSIG61931.2024.10786752.

I cooperated on the design and analysis of experiments and contributed to text writing. Contribution 45%.

[A25] Anton Firc, Kamil Malinka, and Petr Hanáček. “Deepfake Speech Detection: A Spectrogram Analysis”. *In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC ’24. Avila, Spain: Association for Computing Machinery, 2024, pp. 1312–1320. ISBN: 9798400702433.
https://doi.org/10.1145/3605098.3635911.

I cooperated on the design and analysis of experiments and contributed to text writing. Contribution 45%.

[A26] Kamil Malinka, Anton Firc, Jakub Reš, Pavel Loutocký, František Kasl, and Vashek Matyáš. Deepfakes – we can do more than just detect. *In: Security Protocols XXIX*. 2025. Lecture Notes in Computer Science, Springer.
(accepted, to be published)

I led the research, proposed the idea of other protection areas, cooperated on the analysis, and wrote a significant part of the text. Contribution 30%.

[A27] Anton Firc, Kamil Malinka, and Petr Hanáček. Evaluation Framework for Deepfake Speech Detection: A Comparative Study of State-of-the-art Deepfake Speech Detectors. Cybersecurity (2025)
(accepted, to be published)

I cooperated on the design and analysis of experiments and contributed to text writing. Contribution 45%.

Chapter 4

Conclusion

In this thesis, we presented the summary of our research and its contributions to the areas of usable security, cybersecurity impacts of deepfakes with a focus on users, attacks on biometric authentication as well as detection methods, and education topics in cybersecurity.

The observed experience shows the current vulnerability of biometric authentication, which has no natural defenses against this type of attack. Furthermore, the lessons learned from the practical implementation of attacks are further applicable e.g., for the development of methods for penetration testing of biometric authentication systems, and should be considered in biometric testing standards and processes. They can also form the basis for risk analysis in developing and deploying authentication tools. The knowledge gained in the field of deepfake detection can help to accelerate the deployment of detectors in practice.

Our results are used in practice in multiple projects where we are involved. For example, we are developing a tool for the police to detect fake voice recordings. Furthermore, our results are used in educational campaigns, and we have incorporated them into teaching.

Future work

The research presented here can be built upon in multiple areas. We expect usable security to continue to grow in importance, especially as IT becomes more integrated into all areas of human activity. Thus, we can expect an increasing number of people working with IT who do not have in-depth knowledge. In contrast, the new generation that has grown up with IT is likely to use it for a much wider range of activities than we have been used to.

However, we expect more development in the area of deepfakes as deepfakes bring new risks to society. Their impact and the number of attacks can be expected to continue to increase - we expect better social engineering methods that take advantage of people not recognizing deepfakes. Extremely challenging issues await biometric authentication, which will likely require adding features to protect it from deepfakes. The use of synthetic media for fake news, manipulation of public opinion, or to carry out and cover up other crimes (hate speech, i.e. hate crimes) will increase. All this with continuous improvement of the quality and availability of tools for synthesis.

We are also approaching the state of real-time deepfakes (or rather, we are already there), so we can expect another significant increase in new types of attacks combining voice and image deepfakes and creation, e.g., filters on MS Teams, allowing attackers to impersonate victims. Moreover, with the development of language models (ChatGPT, etc.), the risk of automating attacks with widespread impact is approaching.

Solving each of the problems mentioned above represents an interesting research area. There is an opportunity for deeper exploration of the human ability to detect deepfakes, e.g., to analyze the decision-making of forensic experts or to investigate whether and how people's decisions are influenced by the provided aid (detectors of different quality).

New methods of raising awareness of deepfake technology and people's resilience can be developed, e.g., based on repeated exposure to deepfake media or inspired by mock phishing campaigns.

We can also focus on developing new detectors with more information - e.g., knowledge of the speaker's context and how they speak. There will also be a need to respond to the creativity of attackers who may start combining real and deepfake media. Thus, there will be a need to design detectors capable of detecting these scenarios.

The involvement of other IT security areas is also an option. Instead of trying to solve the generalization problem, one can also focus on other tools to help mitigate the threats that deepfakes pose. We believe, that our work can help to open a discussion about whether other proven techniques from other areas of IT security and cryptography might be applicable in this area to help cover presented attack vectors.

Bibliography

Authored publications and manuscripts referenced in the thesis

[A1] Kamil Malinka. "Usability of Visual Evoked Potentials as Behavioral Characteristics for Biometric Authentication". In: *The Fourth International Conference on Internet Monitoring and Protection*. Venice, IT: IEEE Computer Society, 2009, p. 6. ISBN: 978-0-7695-3612-5. URL: https://www.fit.vut.cz/research/publication/8961 (page 3).

[A2] Kamil Malinka, Petr Hanáček, and Michal Trzos. "Evaluation of biometric authentication based on visual evoked potentials". In: *2011 Carnahan Conference on Security Technology*. 2011, pp. 1–7. DOI: 10.1109/CCST.2011.6095875 (page 3).

[A3] Kamil Malinka and Petr Hanáček. "Behavioural Patterns and Social Networks in Anonymity Systems". In: *Computational Social Networks: Security and Privacy*. Ed. by Ajith Abraham. London: Springer London, 2012, pp. 311–340. ISBN: 978-1-4471-4051-1. DOI: 10.1007/978-1-4471-4051-1_13. URL: https://doi.org/10.1007/978-1-4471-4051-1_13 (page 3).

[A4] Kamil Malinka, Petr Hanáček, and Daniel Cvrček. "Analyses of Real Email Traffic Properties". In: *Radioengineering* 2009.4 (2009), pp. 644–650. ISSN: 1210-2512. URL: https://www.fit.vut.cz/research/publication/9103 (page 3).

[A5] Kamil Malinka and Jiří Schäfer. "Development of Social Networks in Email Communication". In: *The Fourth International Conference on Internet Monitoring and Protection*. Venice, IT: IEEE Computer Society, 2009, p. 5. ISBN: 978-0-7695-3612-5. URL: https://www.fit.vut.cz/research/publication/8960 (page 3).

[A6] Pavel Loutocký and Kamil Malinka. "Bezpečnost ICT ve vnitřních předpisech a školení zaměstnanců". cze. In: *Revue pro právo a technologie* 7 (2016). ISSN: 1804-5383. URL: https://journals.muni.cz/revue/article/view/6144/pdf (page 3).

[A7] Kamil Malinka and Jakub Harašta. "Právní aspekty sledování využití výpočetní techniky zaměstnancem". cze. In: *Data Security Management* 4 (2016). ISSN: 1211-8737. URL: https://tate.cz/archiv-cisel-dsm/109-dsm-2016-4 (page 3).

[A8] Radim Polčák, Zdeněk Říha, and Kamil Malinka. "Právní aspekty interních směrnic – část I". cze. In: *Data Security Management* 2 (2015). ISSN: 1211-8737. URL: https://tate.cz/archiv-2015/585-dsm-2015-2 (page 3).

[A9] Radim Polčák, Zdeněk Říha, and Kamil Malinka. "Právní aspekty interních směrnic – část II". cze. In: *Data Security Management* 3 (2015). ISSN: 1211-8737. URL: https://tate.cz/archiv-2015/584-dsm-2015-3 (page 3).

[A10] Vlasta Stavova, Vashek Matyas, and Kamil Malinka. "The Challenge of Increasing Safe Response of Antivirus Software Users". In: *Mathematical and Engineering Methods in Computer Science*. Ed. by Jan Kofroň and Tomáš Vojnar. Cham: Springer International Publishing, 2016, pp. 133–143. ISBN: 978-3-319-29817-7 (pages 5, 15, 54).

[A11] Vashek Matyas, Kamil Malinka, Lydia Kraus, Lenka Knapova, and Agata Kruzikova. "Even if users do not read security directives, their behavior is not so catastrophic". In: *Commun. ACM* 65.1 (December 2021), pp. 37–40. ISSN: 0001-0782. DOI: 10.1145/3471928. URL: https://doi.org/10.1145/3471928 (pages 6, 14, 52, 54).

[A12] Kamil Malinka, Martin Perešíni, Anton Firc, Ondřej Hujňák, and Filip Januš. "On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?" In: *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. ITiCSE 2023. Turku, Finland: Association for Computing Machinery, 2023, pp. 47–53. ISBN: 9798400701382. DOI: 10.1145/3587102.3588827. URL: https://doi.org/10.1145/3587102.3588827 (page 7).

[A13] Kamil Malinka, Anton Firc, Pavel Loutocký, Jakub Vostoupal, Andrej Krištofík, and František Kasl. "Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report". In: *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. ITiCSE 2024. Milan, Italy: Association for Computing Machinery, 2024, pp. 227–233. ISBN: 9798400706004. DOI: 10.1145/3649217.3653633. URL: https://doi.org/10.1145/3649217.3653633 (pages 7–9, 14, 52, 54).

[A14] Andrej Krištofík, Jakub Vostoupal, Kamil Malinka, František Kasl, and Pavel Loutocký. "Beyond the Bugs: Enhancing Bug Bounty Programs through Academic Partnerships". In: *Proceedings of the 19th International Conference on Availability, Reliability and Security*. ARES '24. Vienna, Austria: Association for Computing Machinery, 2024. ISBN: 9798400717185. DOI: 10.1145/

3664476.3670455. URL: https://doi.org/10.1145/3664476.3670455 (pages 9, 14, 52, 54).

[A15] Jakub Vostoupal, Václav Stupka, Jakub Harašta, František Kasl, Pavel Loutocký, and Kamil Malinka. "The legal aspects of cybersecurity vulnerability disclosure: To the NIS 2 and beyond". In: *Computer Law & Security Review* 53 (2024), p. 105988. ISSN: 0267-3649. DOI: https://doi.org/10.1016/j.clsr.2024.105988. URL: https://www.sciencedirect.com/science/article/pii/S0267364924000554 (pages 10, 15, 54).

[A16] Anton Firc, Kamil Malinka, and Petr Hanáček. "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors". In: *Heliyon* 9.4 (2023), e15090. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2023.e15090. URL: https://www.sciencedirect.com/science/article/pii/S2405844023022971 (pages 10, 18, 19, 33, 52, 54).

[A17] Anton Firc and Kamil Malinka. "The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems". In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. SAC '22. Virtual Event: Association for Computing Machinery, 2022, pp. 1646–1655. ISBN: 9781450387132. DOI: 10.1145/3477314.3507013. URL: https://doi.org/10.1145/3477314.3507013 (pages 10, 20, 21, 33, 52, 54).

[A18] Daniel Prudký, Anton Firc, and Kamil Malinka. "Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation". In: *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2023, pp. 1–5. DOI: 10.1109/BIOSIG58226.2023.10346006 (pages 11, 15, 54).

[A19] Kamil Malinka, Anton Firc, Milan Šalko, Daniel Prudký, Karolína Radačovská, and Petr Hanáček. "Comprehensive multiparametric analysis of human deepfake speech recognition". In: *EURASIP Journal on Image and Video Processing* 2024.1 (August 2024). ISSN: 1687-5281. DOI: 10.1186/s13640-024-00641-4. URL: http://dx.doi.org/10.1186/s13640-024-00641-4 (pages 11–14, 52, 54).

[A20] Kamil Malinka and Anton Firc. "Deepfakes: příležitost, nebo hrozba?" Czech. In: *Proč se nebát umělé inteligence?* Praha, CZ: Nakladatelství JOTA, s.r.o., 2024, pp. 271–283. ISBN: 978-80-7689-459-4. URL: https://www.fit.vut.cz/research/publication/13287 (pages 13, 15, 54).

[A21] Kamil Malinka, Ondřej Hujňák, Petr Hanáček, and Lukáš Hellebrandt. "E-Banking Security Study—10 Years Later". In: *IEEE Access* 10 (2022), pp. 16681–16699. DOI: 10.1109/ACCESS.2022.3149475 (pages 17, 33, 52, 54).

[A22] Milan Šalko, Anton Firc, and Kamil Malinka. "Security Implications of Deepfakes in Face Authentication". In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC '24. Avila, Spain: Association for Computing Machinery, 2024, pp. 1376–1384. ISBN: 9798400702433. DOI: 10.1145/

3605098.3635953. URL: https://doi.org/10.1145/3605098.3635953 (pages 20, 22–24, 34, 54).

[A23] Kamil Malinka, Anton Firc, Petr Kaška, Tomáš Lapšanský, Oskar Šandor, and Ivan Homoliak. "Resilience of Voice Assistants to Synthetic Speech". In: *Computer Security – ESORICS 2024*. Ed. by Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas. Cham: Springer Nature Switzerland, 2024, pp. 66–84. ISBN: 978-3-031-70879-4 (pages 20–22, 33, 53, 54).

[A24] Anton Firc, Kamil Malinka, and Petr Hanáček. "Diffuse or Confuse: A Diffusion Deepfake Speech Dataset". In: *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2024, pp. 1–7. DOI: 10.1109/BIOSIG61931.2024.10786752 (pages 20, 24, 34, 54).

[A25] Anton Firc, Kamil Malinka, and Petr Hanáček. "Deepfake Speech Detection: A Spectrogram Analysis". In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC '24. Avila, Spain: Association for Computing Machinery, 2024, pp. 1312–1320. ISBN: 9798400702433. DOI: 10.1145/3605098.3635911. URL: https://doi.org/10.1145/3605098.3635911 (pages 24, 29, 30, 34, 54).

[A26] Kamil Malinka, Anton Firc, Jakub Reš, Pavel Loutocký, Fratišek Kasl, and Vashek Matyáš. "Deepfakes – we can do more than just detect." In: *Security Protocols XXIX*. Lecture Notes in Computer Science, Springer, 2025 (pages 25, 34, 54).

[A27] Anton Firc, Kamil Malinka, and Petr Hanáček. "Evaluation Framework for Deepfake Speech Detection: A Comparative Study of State-of-the-art Deepfake Speech Detectors". In: *Cybersecurity* (2025). ISSN: 2523-3246 (pages 31, 32, 34, 54).

Other publications referenced in the thesis

[1] Simson Garfinkel and Heather Richter Lipford. "Introduction". In: *Usable Security: History, Themes, and Challenges*. Cham: Springer International Publishing, 2014, pp. 1–11. ISBN: 978-3-031-01215-0. DOI: 10.1007/978-3-031-02343-9_1. URL: https://doi.org/10.1007/978-3-031-02343-9_1 (Accessed on December 1, 2022) (page 2).

[2] Peter Leo Gorski, Luigi Lo Iacono, and Matthew Smith. "Eight Lightweight Usable Security Principles for Developers". In: *IEEE Security & Privacy* 21.1 (2023), pp. 20–26. DOI: 10.1109/MSEC.2022.3205484 (page 5).

[3] Anne Adams and Martina Angela Sasse. "Users are not the enemy". In: *Commun. ACM* 42.12 (December 1999), pp. 40–46. ISSN: 0001-0782. DOI: 10.1145/322796.322806. URL: https://doi.org/10.1145/322796.322806 (page 5).

[4] Mary Theofanos. “Is Usable Security an Oxymoron?” In: *Computer* 53.2 (2020), pp. 71–74. DOI: 10.1109/MC.2019.2954075 (page 5).

[5] Francesco Di Nocera, Giorgia Tempestini, and Matteo Orsini. “Usable Security: A Systematic Literature Review”. In: *Information* 14.12 (2023). ISSN: 2078-2489. DOI: 10.3390/info14120641. URL: https://www.mdpi.com/2078-2489/14/12/641 (page 5).

[6] Matthew Green and Matthew Smith. “Developers are Not the Enemy! The Need for Usable Security APIs”. In: *IEEE Security and Privacy* 14.5 (September 2016), pp. 40–46. ISSN: 1540-7993. DOI: 10.1109/MSP.2016.111. URL: https://doi.org/10.1109/MSP.2016.111 (page 5).

[7] Felix Fischer, Konstantin Böttinger, Huang Xiao, Christian Stransky, Yasemin Acar, Michael Backes, and Sascha Fahl. “Stack Overflow Considered Harmful? The Impact of Copy&Paste on Android Application Security.” In: *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2017, pp. 121–136. ISBN: 978-1-5090-5533-3. URL: http://dblp.uni-trier.de/db/conf/sp/sp2017.html#FischerBXSA0F17 (page 5).

[8] Sonia Chiasson, Robert Biddle, and Anil Somayaji. “Even Experts Deserve Usable Security: Design guidelines for security management systems”. In: 2007. URL: https://api.semanticscholar.org/CorpusID:6364968 (page 5).

[9] Mutlaq Jalimid Alotaibi, Steven Furnell, and Nathan Clarke. “A framework for reporting and dealing with end-user security policy compliance”. In: *Information & Computer Security* 27.1 (January 2019), pp. 2–25. ISSN: 2056-4961. DOI: 10.1108/ICS-12-2017-0097. URL: https://doi.org/10.1108/ICS-12-2017-0097 (page 5).

[10] Adéle Da Veiga. “Comparing the information security culture of employees who had read the information security policy and those who had not”. In: *Information & Computer Security* 24.2 (January 2016), pp. 139–151. ISSN: 2056-4961. DOI: 10.1108/ICS-12-2015-0048. URL: https://doi.org/10.1108/ICS-12-2015-0048 (page 5).

[11] Teodor Sommestad, Jonas Hallberg, Kristoffer Lundholm, and Johan E. Bengtsson. “Variables influencing information security policy compliance: A systematic review of quantitative studies”. In: *Inf. Manag. Comput. Secur.* 22 (2014), pp. 42–75. URL: https://api.semanticscholar.org/CorpusID:18387933 (page 5).

[12] Simon Parkin, Aad van Moorsel, Philip Inglesant, and M. Angela Sasse. “A stealth approach to usable security: helping IT security managers to identify workable security solutions”. In: *Proceedings of the 2010 New Security Paradigms Workshop*. NSPW ’10. Concord, Massachusetts, USA: Association for Computing Machinery, 2010, pp. 33–50. ISBN: 9781450304153. DOI: 10.1145/1900546.1900553. URL: https://doi.org/10.1145/1900546.1900553 (page 5).

[13] Cormac Herley. “More Is Not the Answer”. In: *IEEE Security & Privacy* 12.1 (2014), pp. 14–19. DOI: 10.1109/MSP.2013.134 (page 5).

[14] Regina Hartley. “Ethical Hacking Pedagogy: An Analysis and Overview of Teaching Students to Hack”. In: *Journal of International Technology and Information Management* 24 (January 2015), pp. 95–104. DOI: 10.58729/1941-6679.1055 (page 7).

[15] Regina Hartley, B. Medlin, and Zach Houlik. “Ethical Hacking: Educating Future Cybersecurity Professionals”. In: October 2017 (page 7).

[16] Zouheir Trabelsi and Margaret McCoey. “Ethical Hacking in Information Security Curricula”. In: *Int. J. Inf. Commun. Technol. Educ.* 12.1 (January 2016), pp. 1–10. ISSN: 1550-1876. DOI: 10.4018/IJICTE.2016010101. URL: https://doi.org/10.4018/IJICTE.2016010101 (page 7).

[17] Brian A. Pashel. “Teaching students to hack: ethical implications in teaching students to hack at the university level”. en. In: *Proceedings of the 3rd annual conference on Information security curriculum development*. Kennesaw Georgia: ACM, September 2006, pp. 197–200. ISBN: 978-1-59593-437-6. DOI: 10.1145/1231047.1231088. URL: https://dl.acm.org/doi/10.1145/1231047.1231088 (Accessed on April 15, 2024) (page 7).

[18] Tim Greene. *Training Ethical Hackers: Training the Enemy?* July 2004. URL: https://defcon.org/html/links/dc_press/archives/12/ebcvg_training_ethical_hackers.htm (Accessed on April 15, 2024) (page 7).

[19] Patricia Y. Logan and Allen Clarkson. “Teaching students to hack: curriculum issues in information security”. en. In: *ACM SIGCSE Bulletin* 37.1 (February 2005), pp. 157–161. ISSN: 0097-8418. DOI: 10.1145/1047124.1047405. URL: https://dl.acm.org/doi/10.1145/1047124.1047405 (Accessed on April 15, 2024) (page 7).

[20] Nicole Radziwill, Jessica Romano, Diane Shorter, and Morgan Benton. “The Ethics of Hacking: Should It Be Taught?” In: *Software Quality Professional* 18.1 (December 2015). URL: https://arxiv.org/abs/1512.02707 (page 7).

[21] Thomas Walshe and Andrew Simpson. “An Empirical Study of Bug Bounty Programs”. In: *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*. London, ON, Canada, February 2020, pp. 35–44. DOI: 10.1109/IBF50092.2020.9034828 (page 8).

[22] Jukka Ruohonen and Luca Allodi. “A Bug Bounty Perspective on the Disclosure of Web Vulnerabilities”. en. In: *17th Annual Workshop on the Economics of Information Security, Innsbruck* (May 2018). arXiv: 1805.09850. URL: http://arxiv.org/abs/1805.09850 (Accessed on March 29, 2022) (page 8).

[23] Jacob Riggs. “I hacked the Dutch government and all I got was this t-shirt”. en. In: *Jacob Riggs* (May 2021). URL: https://jacobriggs.io/blog/posts/i-hacked-the-dutch-government-and-all-i-got-was-this-t-shirt-24.html (Accessed on May 3, 2023) (page 8).

[24] Sandra SCHMITZ and Stefan SCHIFFNER. “Responsible Vulnerability Disclosure under the NIS 2.0 Proposal”. English. In: *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 12.5 (2021). URL: https://www.jipitec.eu/issues/jipitec-12-5-2021/5495 (page 10).

[25] Jon Bateman. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Tech. rep. Carnegie Endowment for International Peace, 2020, pp. i–ii. URL: http://www.jstor.org/stable/resrep25783.1 (pages 10, 16).

[26] Matthew Groh, Ziv Epstein, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. “Human detection of machine-manipulated media”. en. In: *Communications of the ACM* 64.10 (October 2021), pp. 40–47. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3445972. URL: https://dl.acm.org/doi/10.1145/3445972 (Accessed on December 26, 2022) (page 10).

[27] Sankini Rancha Godage, Froy Lovasdaly, Sushma Venkatesh, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. “Analyzing Human Observer Ability in Morphing Attack Detection -Where Do We Stand?” In: *IEEE Transactions on Technology and Society* (2023), pp. 1–1. DOI: 10.1109/tts.2022.3231450. URL: https://doi.org/10.1109/tts.2022.3231450 (page 10).

[28] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. “FaceForensics++: Learning to Detect Manipulated Facial Images”. In: October 2019, pp. 1–11. DOI: 10.1109/ICCV.2019.00009 (page 10).

[29] Pavel Korshunov and Sébastien Marcel. *Deepfake detection: humans vs. machines*. arXiv:2009.03155 [cs, eess]. September 2020. URL: http://arxiv.org/abs/2009.03155 (Accessed on December 26, 2022) (page 10).

[30] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. “Deepfake detection by human crowds, machines, and machine-informed crowds”. In: *Proceedings of the National Academy of Sciences* 119.1 (2022), e2110013119. DOI: 10.1073/pnas.2110013119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2110013119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2110013119 (page 10).

[31] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. “Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Sys-*

tems. ACM, May 2021. DOI: 10.1145/3411764.3445699. URL: https://doi.org/10.1145/3411764.3445699 (page 10).

[32] Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. "Warning: Humans cannot reliably detect speech deepfakes". In: *PLOS ONE* 18.8 (August 2023). Ed. by Yogan Jaya Kumar, e0285333. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0285333. URL: http://dx.doi.org/10.1371/journal.pone.0285333 (page 10).

[33] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. "Human Perception of Audio Deepfakes". In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. DDAM '22. Lisboa, Portugal: Association for Computing Machinery, 2022, pp. 85–91. ISBN: 9781450394963. DOI: 10.1145/3552466.3556531. URL: https://doi.org/10.1145/3552466.3556531 (page 10).

[34] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicolas Evans, Tomi Kinnunen, Kong Aik Lee, Ville Vestman, and Andreas. Nautsch. *ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*. 2019. URL: https://doi.org/10.7488/ds/2555 (pages 10, 21, 30).

[35] Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja. *Audio Deepfake Perceptions in College Going Populations*. 2021. arXiv: 2112.03351 [cs.SD] (page 10).

[36] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. "Face Recognition Systems Under Morphing Attacks: A Survey". In: *IEEE Access* 7 (2019), pp. 23012–23026. DOI: 10.1109/ACCESS.2019.2899367 (page 16).

[37] Thomas Brewster. *Fraudsters cloned company director's voice in $35 million bank heist, police find*. Online: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/. October 2021. URL: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/ (page 16).

[38] Lindsey O'Donnell. *CEO 'Deep fake' swindles company out of $243K*. September 2019. URL: https://threatpost.com/deep-fake-of-ceos-voice-swindles-company-out-of-243k/147982/ (page 16).

[39] Laurie Iacono, Josh Hickman, and Caitlin Muniz. *The rise of vishing and Smishing attacks – the monitor, issue 21*. August 2022. URL: https://www.kroll.com/en/insights/publications/cyber/monitor/vishing-smishing-attacks (page 16).

[40] Tim Ring. “Europol: the AI hacker threat to biometrics”. In: *Biometric Technology Today* 2021.2 (2021), pp. 9–11. ISSN: 0969-4765. DOI: https://doi.org/10.1016/S0969-4765(21)00023-0. URL: https://www.sciencedirect.com/science/article/pii/S0969476521000230 (page 16).

[41] Valencia A. Jones. “Artificial Intelligence Enabled - Deepfake technology The Emerge of a New Threat”. Master thesis. Utica College, 2020 (page 16).

[42] Masha Borak. *Tax scammers hack government-run facial recognition system*. March 2021. URL: https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax (page 16).

[43] M. Bishop. *Introduction to Computer Security*. Addison-Wesley, 2005. ISBN: 9780321247445. URL: https://books.google.cz/books?id=Z-lQAAAAMAAJ (page 16).

[44] Sven Kiljan, Koen Simoens, Danny De Cock, Marko Van Eekelen, and Harald Vranken. “A Survey of Authentication and Communications Security in Online Banking”. In: *ACM Comput. Surv.* 49.4 (December 2016). ISSN: 0360-0300. DOI: 10.1145/3002170 (page 16).

[45] Maha M. Althobaiti and Pam Mayhew. “Security and usability of authenticating process of online banking: User experience study”. In: *2014 International Carnahan Conference on Security Technology (ICCST)*. 2014, pp. 1–6. DOI: 10.1109/CCST.2014.6986978 (page 17).

[46] Fredrik Mennes. *PSD2: Which Strong Authentication and Risk Analysis Solutions comply with the EBA’s Final Draft RTS?* Accessed on Nov. 12, 2020. April 2017. URL: https://frederikmennes.wordpress.com/2017/04/19/psd2-which-strong-authentication-and-risk-analysis-solutions-comply-with-the-ebas-final-draft-rts/ (page 17).

[47] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. “ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge”. In: *Proc. Interspeech* 1 (2015), pp. 2037–2041. DOI: 10.21437/Interspeech.2015-462 (page 17).

[48] John Seymour and Azeem Aqil. *Your Voice is My Passport*. 2018. URL: https://www.blackhat.com/us-18/briefings/schedule/%5C#your-voice-is-my-passport-11395 (page 17).

[49] Shahroz Tariq, Sowon Jeon, and Simon S. Woo. “Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack”. In: *Proceedings of the ACM Web Conference 2022*. WWW ’22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022,

pp. 512–523. ISBN: 9781450390965. DOI: 10.1145/3485447.3512212. URL: https://doi.org/10.1145/3485447.3512212 (page 17).

[50] Pavel Korshunov and Sebastien Marcel. *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*. 2018. arXiv: 1812.08685 [cs.CV] (page 17).

[51] Paul Grassi, Elaine Newton, Ray Perlner, Andrew Regenscheid, Jim Fenton, William Burr, Justin Richer, Naomi Lefkovitz, Jamie Danker, Yee-Yin Choong, Kristin Greene, and Mary Theofanos. *Digital Identity Guidelines: Authentication and Lifecycle Management*. NIST Special Publication 800-63B. Gaithersburg, MD: National Institute for Standards and Technology, June 2017. DOI: 10.6028/NIST.SP.800-63b (page 17).

[52] European Union. "Directive (EU) 2015/2366 of the European Parliament and of the Council". In: *Official Journal of the European Union* L 337 (November 2015), pp. 35–128 (page 17).

[53] Descript. *Overdub*. online. 2021. URL: https://www.descript.com/overdub (page 20).

[54] Resemble AI. *Resemble AI webpage*. online. 2020. URL: https://www.resemble.ai (pages 20, 22).

[55] Jemine Corentin. "Real-time Voice Cloning". Master thesis. Liège, Belgique: Université de Liège, Liège, Belgique, 2019. URL: https://matheo.uliege.be/handle/2268.2/6801?locale=en (page 20).

[56] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. "Common Voice: A Massively-Multilingual Speech Corpus". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.520 (page 20).

[57] Microsoft. *About the Speech SDK*. https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview#speaker-verification. 2020 (page 20).

[58] Phonexia. *Phonexia Voice Verify*. https://www.phonexia.com/en/product/voice-verify/. 2021 (page 20).

[59] CoquiAI. *CoquiAI webpage*. online. URL: https://coqui.ai/ (page 22).

[60] James Betker. *Better speech synthesis through scaling*. 2023. arXiv: 2305.07243 [cs.SD] (page 22).

[61] XTTS. *XTTS webpage*. online. URL: https://github.com/coqui-ai/TTS (page 22).

[62] Innovatrics s.r.o. *Face functions*. J01 2021. URL: https://developers.innovatrics.com/digital-onboarding/docs/functionalities/face/ (page 23).

[63] Neurotechnology. *Megamatcher SDK*. January 2023. URL: https://www.neurotechnology.com/megamatcher.html (page 23).

[64] Yuezun Li. “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics”. In: *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 2020 (page 23).

[65] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. *A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI*. 2023. arXiv: 2303.13336 [cs.SD] (page 24).

[66] Xin Wang and Junichi Yamagishi. “A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection”. In: *Proc. Interspeech 2021*. 2021, pp. 4259–4263. DOI: 10.21437/Interspeech.2021-702 (page 24).

[67] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. *Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation*. 2022 (page 24).

[68] Tina Brooks, Princess G., Jesse Heatley, Jeremy J., Scott Kim, Samantha M., Sara Parks, Maureen Reardon, Harley Rohrbacher, Burak Sahin, Shani S., James S., Oliver T., and Richard V. *Increasing Threat of Deepfake Identities*. Tech. rep. U.S. Department of Homeland Security, 2021 (page 25).

[69] Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. “A comprehensive survey on robust image watermarking”. In: *Neurocomputing* 488 (2022), pp. 226–247. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.02.083 (page 27).

[70] Council of European Union. *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Emerging Certain Union Legislative Acts*. https://artificialintelligenceact.eu. 2021 (page 27).

[71] Council of European Union. *Liability Rules for Artificial Intelligence*. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en. 2022 (page 27).

[72] Council of European Union. *The Digital Services Act: ensuring a safe and accountable online environment*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en. 2022 (pages 27, 28).

[73] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. "Toward the Creation and Obstruction of DeepFakes". In: *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Ed. by Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. Cham: Springer International Publishing, 2022, pp. 71–96. ISBN: 978-3-030-87664-7. DOI: 10.1007/978-3-030-87664-7_4 (pages 27, 28).

[74] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. "AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis". In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. CCS '23. Copenhagen, Denmark: Association for Computing Machinery, 2023, pp. 460–474. ISBN: 9798400700507. DOI: 10.1145/3576915.3623209 (page 27).

[75] Luisa Verdoliva. "Media Forensics and DeepFakes: An Overview". In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932. DOI: 10.1109/JSTSP.2020.3002101 (page 27).

[76] Haya R. Hasan and Khaled Salah. "Combating Deepfake Videos Using Blockchain and Smart Contracts". In: *IEEE Access* 7 (2019), pp. 41596–41606. DOI: 10.1109/ACCESS.2019.2905689 (pages 27, 28).

[77] Dan Boneh, Andrew J. Grotto, Patrick McDaniel, and Nicolas Papernot. "How Relevant Is the Turing Test in the Age of Sophisbots?" In: *IEEE Security & Privacy* 17.6 (2019), pp. 64–71. DOI: 10.1109/MSEC.2019.2934193 (pages 27, 28).

[78] Adnan Alattar, Ravi Sharma, and John Scriven. "A System for Mitigating the Problem of Deepfake News Videos Using Watermarking". In: *Electronic Imaging* 32.4 (January 2020), pp. 117–1–117–10. DOI: 10.2352/issn.2470-1173.2020.4.mwsf-117 (page 27).

[79] Mariëtte van Huijstee, Pieter van Boheemen, Djurre Das, Linda Nierling, Jutta Jahnel, Murat Karaboga, and Martin Fatun. *Tackling deepfakes in European policy*. Publications Office, 2021. DOI: 10.2861/325063 (page 27).

[80] Council of European Union. *European Democracy Action Plan*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en. 2020 (page 27).

[81] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (page 28).

[82] Council of European Union. *electronic IDentification, Authentication and trust Services (eIDAS) Regulation*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en. 2014 (page 28).

[83] David Khachaturov, Ilia Shumailov, Yiren Zhao, Nicolas Papernot, and Ross Anderson. “Markpainting: Adversarial Machine Learning meets Inpainting”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 5409–5419 (page 28).

[84] Yisroel Mirsky and Wenke Lee. “The Creation and Detection of Deepfakes: A Survey”. In: *ACM Comput. Surv.* 54.1 (January 2021). ISSN: 0360-0300. DOI: 10.1145/3425780. URL: https://doi.org/10.1145/3425780 (page 28).

[85] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. “On the Detection of Digital Face Manipulation”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5780–5789. DOI: 10.1109/CVPR42600.2020.00582 (page 28).

[86] Jun Jiang, Bo Wang, Bing Li, and Weiming Hu. “Practical Face Swapping Detection Based on Identity Spatial Constraints”. In: *2021 IEEE International Joint Conference on Biometrics (IJCB)*. 2021, pp. 1–8. DOI: 10.1109/IJCB52358.2021.9484396 (page 28).

[87] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. “Fighting Deepfakes by Detecting GAN DCT Anomalies”. In: *Journal of Imaging* 7.8 (June 2021), p. 128. ISSN: 2313-433X. DOI: 10.3390/jimaging7080128. URL: http://dx.doi.org/10.3390/jimaging7080128 (page 28).

[88] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. “Detection of Face Morphing Attacks by Deep Learning”. In: *Digital Forensics and Watermarking*. Cham: Springer International Publishing, 2017, pp. 107–120. ISBN: 978-3-319-64185-0 (page 28).

[89] Ulrich Scherhag, Christian Rathgeb, and Christoph Busch. “Towards Detection of Morphed Face Images in Electronic Travel Documents”. In: *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. 2018, pp. 187–192. DOI: 10.1109/DAS.2018.11 (page 28).

[90] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. “Fakecatcher: Detection of synthetic portrait videos using biological signals”. In: *IEEE transactions on pattern analysis and machine intelligence* (2020) (page 28).

[91] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. “FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). ISSN: 1939-3539. DOI: 10.1109/tpami.2020.3009287. URL: http://dx.doi.org/10.1109/TPAMI.2020.3009287 (page 28).

[92] Xin Wang et. al. “ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale”. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*. 2024 (pages 28, 30).

[93] Kirill Borodin et. al. “AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge”. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop*. 2024, pp. 48–55. DOI: 10.21437/ASVspoof.2024-8 (page 28).

[94] Yuxiong Xu, Jiafeng Zhong, Sengui Zheng, Zefeng Liu, and Bin Li. “SZU-AFS antispoofing system for the ASVspoof 5 Challenge”. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop*. 2024, pp. 64–71. DOI: 10.21437/ASVspoof.2024-10 (page 28).

[95] Théophile Stourbe, Victor Miara, Theo Lepage, and Reda Dehak. “Exploring WavLM back-ends for speech spoofing and deepfake detection”. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop*. 2024, pp. 72–78. DOI: 10.21437/ASVspoof.2024-11 (page 28).

[96] Pierre Falez and Tony Marteau. “Whispeak speech deepfake detection systems for the ASVspoof5 Challenge”. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*. August 2024, pp. 32–35. DOI: 10.21437/ASVspoof.2024-5 (page 29).

[97] Ricardo Reimao. “Synthetic Speech Detection Using Deep Neural Networks”. Master’s thesis. Toronto, Ontario: York University, 2019 (page 29).

[98] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, and Faruk Kazi. “A Deep Learning Framework for Audio Deepfake Detection”. In: *Arabian Journal for Science and Engineering* (November 2021). ISSN: 2191-4281. DOI: 10.1007/s13369-021-06297-w. URL: https://doi.org/10.1007/s13369-021-06297-w (page 29).

Part II

SELECTED PAPERS

This part contains the 8 research articles (out of more than 50 published total) selected as the representatives of my research contributions into investigated areas. The complete articles are inserted into the corresponding appendix of the printed version of this thesis.

Acknowledgment of the publishers. Due to copyright reasons, we list below the bibliographic information of all attached papers, acknowledging the original source.

Article [A11]: Vashek Matyas, Kamil Malinka, Lydia Kraus, Lenka Knapova, and Agata Kruzikova. "Even if users do not read security directives, their behavior is not so catastrophic". *In: Commun. ACM* 65.1 (December 2021), pp. 37–40. ISSN:0001-0782. https://doi.org/10.1145/3471928

Article [A14]: Andrej Krištofík, Jakub Vostoupal, Kamil Malinka, František Kasl, and Pavel Loutocký. "Beyond the Bugs: Enhancing Bug Bounty Programs through Academic Partnerships". *In: Proceedings of the 19th International Conference on Availability, Reliability and Security*. ARES '24. Vienna, Austria: Association for Computing Machinery, 2024. ISBN: 9798400717185.
https://doi.org/10.1145/3664476.3670455.

Article [A13]: Kamil Malinka, Anton Firc, Pavel Loutocký, Jakub Vostoupal, Andrej Krištofík, and Frantisek Kasl. "Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report". *In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education* V. 1. ITiCSE 2024. Milan, Italy: Association for Computing Machinery, 2024, pp. 227–233. ISBN: 9798400706004. https://doi.org/10.1145/3649217.3653633.

Article [A19]: Kamil Malinka, Anton Firc, Milan Šalko, Daniel Prudký, Karolína Radačov-ská, and Petr Hanáček. "Comprehensive multiparametric analysis of human deepfake speech recognition". *In: EURASIP Journal on Image and Video Processing* 2024.1 (August 2024). ISSN: 1687-5281. http://dx.doi.org/10.1186/s13640-024-00641-4.

Article [A21]: Kamil Malinka, Ondřej Hujňák, Petr Hanáček, and Lukáš Hellebrandt. "E-Banking Security Study—10 Years Later". *In: IEEE Access* 10 (2022), pp. 16681–16699. DOI: 10.1109/ACCESS.2022.3149475.

Article [A16]: Anton Firc, Kamil Malinka, and Petr Hanáček. "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors". *In: Heliyon* 9.4 (2023), e15090. ISSN: 2405-8440.
DOI: https://doi.org/10.1016/j.heliyon.2023.e15090. .

Article [A17]: Anton Firc and Kamil Malinka. "The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems". *In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. SAC '22. Virtual Event: Association for Computing Machinery, 2022, pp. 1646–1655. ISBN: 9781450387132.
https://doi.org/10.1145/3477314.3507013.

Article [A23]: Kamil Malinka, Anton Firc, Petr Kaška, Tomáš Lapšanský, Oskar Šandor, and Ivan Homoliak. "Resilience of Voice Assistants to Synthetic Speech". *In: Computer Security – ESORICS* 2024. Ed. by Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas. Cham: Springer Nature Switzerland, 2024, pp. 66–84. ISBN: 978-3-031-70879-4.

Author's Contribution

In Table 4.1, we describe the author's contributions to the 17 papers used as the basis for this thesis[1]. They include 5 original scientific papers in a scientific journal classified by SCOPUS/SJR in Q1, 1 original scientific paper in a scientific journal classified by SCOPUS/SJR in Q2, 1 original scientific paper in a CORE conference in category A, 5 original scientific papers in a CORE conference in category B and 5 publications in non-ranked category. For assessing the qualitative and quantitative contribution, the standardized metric CRediT (Contributor Roles Taxonomy) [2] is used. We have supplemented this with a precise proportion of the contribution.

Explanation of used CRediT categories

- Conceptualization: Ideas; formulation or evolution of overarching research goals and aims
- Methodology: Development or design of methodology; creation of models
- Software: Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components
- Investigation: Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection
- Writing: Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)

[1] Note that the table contains a sorting of the papers according to the order used in sections Contributed papers (Section 2.3, Section 3.2).

[2] https://credit.niso.org/implementing-credit/

Paper	**Conceptualization**	**Methodology**	**Software**	**Investigation**	**Writing**	**Share**
[A11]	▒	▒	■	▒	▒	25%
[A13]	■	■	■	▒	▒	30%
[A14]	▒	▒	■	▒	▒	20%
[A19]	■	■	▒	▒	▒	45%
[A10]	▒	▒	▒	▒	▒	15%
[A15]		▒	X	▒	▒	10%
[A18]	■	■	▒	▒	▒	45%
[A20]	■	▒	X	▒	▒	50%
[A21]	■	■	X	▒	▒	35%
[A16]	▒	▒	▒	▒	▒	45%
[A17]	▒	▒	▒	▒	▒	50%
[A23]	■	■	▒	■	▒	48%
[A22]	▒	▒			▒	25%
[A24]	▒	▒			▒	45%
[A25]	▒	▒	▒	▒	▒	45%
[A26]	■	■	X	▒	▒	30%
[A27]	▒	▒		▒	▒	45%

Table 4.1: Author's contributions to selected articles related to this work. The degree of contribution is as *lead* (depicted in black), *equal* (depicted in gray), or *supporting* (depicted in white). **X** denotes *not usable*. The papers highlighted in bold are attached to this thesis.

DOI:10.1145/3471928

Vashek Matyas, Kamil Malinka, Lydia Kraus, Lenka Knapova, and Agata Kruzikova

Viewpoint

Even If Users Do Not Read Security Directives, Their Behavior Is Not So Catastrophic

Turning believers into nonbelievers.

IMAGE FROM SHUTTERSTOCK.COM

MORE THAN TWO decades ago, Adams and Sasse in their highly cited seminal work[1] challenged the belief widely held—among IT security professionals—that users are the enemy within an organization—the one who does not care about security and subsequently behaves in a threatening way. While much effort has been undertaken by the research community since then to take the burden from end users and to make security systems more usable,[6] it seems the situation in organizational security has not improved. According to a survey[4] conducted by an online community for IT security professionals—a majority of these professionals still deems "users who are negligent or break the security policy" as "the top data breach risk." Also, as Herley suggests,[7] there can be rational reasons why users do not follow security advice, simply because the cost of following it can be higher than the benefits.

At Masaryk University (MU)—a Czech university with approximately 30,000 students—we wanted to find out more about the current state of affairs from the user perspective: Do users (still not) follow the security policy? At the same time, the fact that our university IT infrastructure management had the intention to redesign the (outdated) security directive, constituted an ideal opportunity for us to deeper investigate the topic.

A security directive (a.k.a. information security policy) is a high-level document that builds the basis for defining, communicating, and enforcing an organization's information security strategy. It describes the principles a user must follow to support the protection of an organization's assets (for example, technical infrastructure or knowledge). Some people believe (and we were of this belief too) that having a usable security directive is the cornerstone for motivating users to behave securely.[2] Similarly, security decision makers have been repeatedly criticized when they ignored usability aspects in security directive design.[9]

We tried to improve our security directive to motivate users to follow it. Yet our faith has been hit hard—as we describe in some detail here, but it was not a wasted effort at all. The data we

obtained as a side effect shows a new perspective on this area.

Improving the Directive

Six years ago, we had the great opportunity to participate in the modernization of security directive at MU and since we were keen to find the truth, we decided then that besides the implementation of modern trends into the directive,[8] we would also design surveys to measure the impact of these changes on the (user self-)reported security behavior. Both legal and IT teams invested quite some effort, as did the university management, hoping the university directive will be easier to follow in practice (its umbrella design is followed by additional documents and measures such as user training, emphasis on a single point of contact to ease communication about incidents, and so forth) and more usable (in terms of accessibility and ease of reading), thus also read by more students and that this would positively contribute to improving their security behavior.

The directive for "Management and Use of Computer Networks of MU" was subsequently modified to the "Use of Information Technology" directive focusing on acceptable use (for work and study tasks, and so forth), behavior during security incidents and protection of authentication data. Within the redesign process, the directive length shrank from 5.5 pages to two pages; moreover, the new directive carried significantly less definitions. Previously mentioned technical issues that did not concern all users (for example, administrator tasks or network hierarchy) were removed, access rights issues shrank to one sentence, and privacy issues were left for a specific directive. Eventually, no sanctions were specified.

Obviously, the directive also concerns the interaction with the MU information system (IS), since students use the IS for critical tasks including registration of courses, exam terms, access study materials, grades, and use the IS email front-end for communication with staff.

Our Study: Better Directive = Better Security?

To find out, we organized a longitudinal study at MU—where we aimed to investigate both security attitudes/behavior and knowledge of the institutional security directive. The study repeatedly ran in years 2015, 2017, and 2018–2019 corresponding to three phases of institutional life: before the release of a (redesigned) directive (that then happened with a delay in September 2017); two months after the release through standard institutional channels; and finally after a campaign on several security issues like password sharing, extent of malware, or access abuse victims, and so forth.

When we eventually started to analyze the obtained data after the third survey round, we discovered surprising results that led to heated discussions among the research team members.

The campaign we coordinated before the third phase took advantage of the university magazine (the only university-wide periodical) with both online and paper versions. We had a front-page attractor in the print version (6,000 copies), presented surprising results from the first phase, with poor password security. We also emphasized the existence of the new security directive. To increase visibility, the article was promoted via three campaigns on different Facebook groups that reached approximately 15,000 people. For the online version, we achieved 966 unique article page views (650 thanks to Facebook). The article was approximately 1,000 words, and with the measured average time spent on the article page five minutes and one second, the article was read in full considering the average reading speed of approximately 200 words per minute.[11]

The survey was conducted in MU computer halls available to all MU students. At a student login, the study questionnaire was displayed to each student, with the possibility to completely skip it (and sometimes to be presented at next login). Students of all nine faculties (schools) of MU were exposed to the study, with the primary aim to avoid focus on students of selected disciplines, for example, computing. We had 613 respondents in 2015, 1,100 in 2017, and 1,309 in 2019. They were females at 52.7%, 62.3% and 63.3% in the respective years, with average age 22.98, 22.24, and 22.11.

When we eventually started to analyze the obtained data after the third survey round, we discovered surprising results that led to heated discussions among the research team members. While the results of the directive-related questions were relatively easy to interpret (as unsatisfying), the opinions of what constitutes "bad" or "good" security behavior naturally widely diverged in a multidisciplinary research team consisting of a psychologist, a sociologist, an engineer, computer scientists, and people in management positions. Although security behavior is not yet ideal, we concluded it is quite reasonable under the given context—the majority of users not having read the directive—as we describe here.

Users Read the Directive Less and Less …

The percentage of users who never read the security directive increased significantly over time (see Figure 1; all following results reported here—if not explicitly noted otherwise—were checked for statistical significance at $p < 0.05$), as did the percentage of those who declared to know nothing of the directive. Please note the answer scales to most of the survey items were grouped and dichotomized to clearly show the differences in behavior and knowledge. The knowledge on matters regulated by the directive also decreased: While 43.6% of respondents (correctly) attested the directive regulates their use of laptops in a dorm network in 2015, the same was attested only by 34.9% in 2017 and 34.1% in 2019. For a private smartphone connected to the university Wi-Fi network, the difference was non-significant but there was a decrease from the first wave (by 4 and 1.8 percentage points, respectively)—31.3% in 2015, 27.3% in 2017, and 29.5% in 2019. These findings negatively surprised us and security decision makers at MU, especially as related studies hint at much lower non-reading rates among employees.[5]

Yet User Security Behavior Is Not that Bad

However, users also reported protecting their computers at levels that we deem quite reasonable and without any significant changes during the period of our study (with the exception of updating applications, where the percentage of users who do not regularly knowingly update increased from 30% in 2015, to 36% in 2017, and 34% in 2019)—only about 30% do not regularly update their OS (or maybe are not aware of this happening), only 12% do not use (or are not aware this being built in their OS) up-to-date malware protection and just 5% report not using a firewall (again, they may not even be aware of this being included in their OS setting). We consider these to be quite positive findings.

Locking a workstation in use when leaving the computer hall is another dimension we investigated. As Figure 2 shows, we discovered improving trends in terms of the percentage of users who lock their screens when they leave their workplace.

Figure 1. The graph reports the share of users that never read the directive and the share of users reporting to not know the content. These trends contrast with the decreasing share of users that shared their MU password.

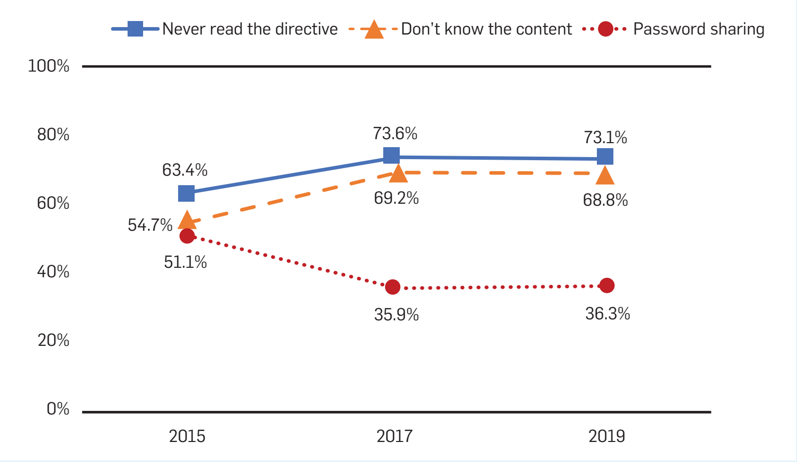


Figure 2. The graph shows trends in computer locking through the three data collections for three different reasons of leaving the computer hall—for lunch (blue), coffee/drinks (orange), and a phone call (grey).

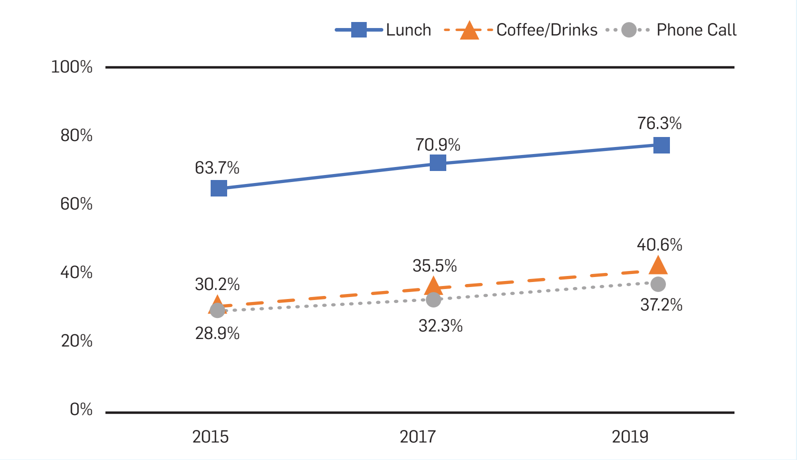
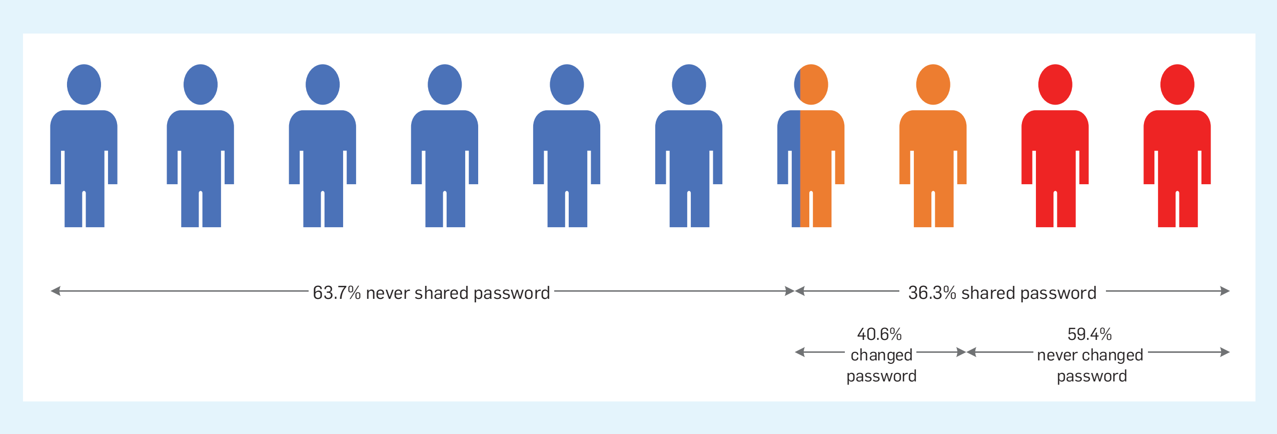


Figure 3. While only 36.3% of our subjects in the last data collection ever shared their IS passwords, 59.4% of them then did not bother changing their password after sharing and only 40.6% did so.



Viewing the same situation from a different viewpoint, we wanted to find out the percentage of users who would do something about somebody else's unlocked computer when passing by, following the principle "If you see something, do something." Here, while the percentage of users who would not do anything at all about another user's unlocked computer decreased, we consider it still unpleasantly high—it was 78.5% in 2015, 73.7% in 2017, and 70.0% in 2019.

User behavior when dealing with passwords comes with both positive and negative observations. Our study showed a very positive trend in the decrease of the proportion of respondents who ever shared their IS password—while 51.1% of them reported ever sharing in 2015, the proportion of sharers reduced to 35.9% in 2017 and 36.3% in 2019. Similarly, approximately one-quarter (26.5%–27.3%) reuses the IS password elsewhere while reusing other passwords between other services even more often (34.8%–37.4%). We consider the latter two numbers a pleasantly surprising finding—as password reuse in the studied age group has been found elsewhere to be as high as 76%.[3] On the negative side, for those who shared their password enabling access to the IS, more than a half (56.1%–59.4%) reported not changing their password afterward, making themselves potentially vulnerable to future impersonation attacks. See Figure 3 and its caption for more details.

User behavior when dealing with passwords comes with both positive and negative observations.

Epilogue: There Is Reason for Hope

We expected our effort to improve the security directive would show a positive impact on students' security behavior. That did not happen—users simply did not read it. However, the results of a longitudinal study on a large group of university students still show a positive trend in self-reported security behavior—despite the small exposure to the directive. While not yet ideal, the protection of endpoint devices and how people handle their passwords is getting to a reasonable level. Whether this is due to exposure to external sources of relevant information (for example, related work[10] hints that only 29.5% learn about secure behavior at work) or to a more naturally increased adoption of technologies remains to be investigated in future work. C

References

1. Adams, A. and Sasse, M.A. Users are not the enemy. *Commun. 42*, 12 (Dec. 1999), 40–46.
2. Alotaibi, M.J., Furnell, S. and Clarke, N. A framework for reporting and dealing with end-user security policy compliance. *Information & Computer Security 27*, 1 (2019), 2–25.
3. CSID. Consumer survey: password habits. A study of password habits among American consumers (2012); https://bit.ly/3Hnfd84
4. Dark Reading. Strategic Security Survey (2019); https://bit.ly/3DgMrn8
5. Da Veiga, A. Comparing the information security culture of employees who had read the information security policy and those who had not. *Information and Computer Security* (2016).
6. Garfinkel, S. and Lipford, H.R. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust 5*, 2 (2014), 1–124.
7. Herley, C. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the 2009 New Security Paradigms Workshop*, (2009); 133–144.
8. Lampson, B. Privacy and security usable security: How to get it. *Commun. ACM. 52* (2009), 25–27; DOI: 10.1145/1592761.1592773.
9. Parkin, S. et al. A stealth approach to usable security: Helping IT security managers to identify workable security solutions. In *Proceedings of the 2010 New Security Paradigms Workshop*, (2010), 33–50.
10. Redmiles, E.M. et al. How I learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, (2016), 666–677.
11. Trauzettel-Klosinski, S. and Dietz, K. Standardized assessment of reading performance: The new international reading speed texts IReST. *Investigative Ophthalmology and Visual Science 53*, 9 (2012), 5452–5461.

Vashek Matyas (matyas@fi.muni.cz) is a professor at Masaryk University in Brno, Czechia.

Kamil Malinka (malinka@ics.muni.cz) is an IT architect at Masaryk University in Brno, Czechia.

Lydia Kraus (lydia.kraus@mail.muni.cz) is a senior researcher at Masaryk University in Brno, Czechia.

Lenka Knapova (knapova@mail.muni.cz) is a Ph.D. candidate at Masaryk University in Brno, Czechia.

Agata Kruzikova (kruzikova@mail.muni.cz) is a Ph.D. candidate at Masaryk University in Brno, Czechia.

The authors want to express their thanks to students and support staff of Masaryk University for their involvement in our experiments, to David Smahel, Lenka Dedkova, and Vlasta Bukacova for cooperation during the years of experiment setting and data collection, to Michal Muzik for assistance with data analyses, and to the reviewers for useful suggestions. Lydia Kraus also thanks the ERDF project CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence (No. CZ.02.1.01/0.0/0.0/16_019/0000822) for support during the preparation of this Viewpoint.

Beyond the Bugs: Enhancing Bug Bounty Programs through Academic Partnerships

Andrej Krištofík
kristofik@mail.muni.cz
CERIT – Faculty of Informatics, and Institute of Law and Technology – Faculty of Law, Masaryk University
Brno, Czechia

Jakub Vostoupal
jakub.vostoupal@law.muni.cz
CERIT – Faculty of Informatics, and Institute of Law and Technology – Faculty of Law, Masaryk University
Brno, Czechia

Kamil Malinka
malinka@ics.muni.cz
Institute of Computer Science and Faculty of Informatics, Masaryk University, Faculty of Informatics, Brno Technological University
Brno, Czechia

František Kasl
frantisek.kasl@muni.cz
CERIT – Faculty of Informatics, and Institute of Law and Technology – Faculty of Law, Masaryk University
Brno, Czechia

Pavel Loutocký
loutocky@muni.cz
CERIT – Faculty of Informatics, and Institute of Law and Technology – Faculty of Law, Masaryk University
Brno, Czechia

ABSTRACT

This paper explores the growing significance of vulnerability disclosure and bug bounty programs within the cybersecurity landscape, driven by regulatory changes in the European Union. The effectiveness of these programs relies heavily on the expertise of participants, presenting a challenge amid a shortage of skilled cybersecurity professionals, particularly in less sought-after sectors. To address this issue, the paper proposes a collaborative approach between academia and bug bounty issuers.

By integrating bug bounty programs into cybersecurity courses, students gain practical skills and soft skills essential for bug hunting and cybersecurity work. The collaboration benefits both issuers, who gain manageable manpower, and students, who receive valuable hands-on experience. A pilot conducted during the current academic year yielded positive results, indicating the potential of this approach to address the demand for skilled cybersecurity professionals. The insights gained from the pilot inform future considerations and advancements in this collaborative model.

CCS CONCEPTS

• **General and reference** → **General conference proceedings**; • **Applied computing** → **Law**; **Education**; • **Security and privacy** → **Human and societal aspects of security and privacy**; • **Social and professional topics** → *Employment issues*; **Codes of ethics**; *Testing, certification and licensing*;

KEYWORDS

Cybersecurity; Bug Bounty; ethical hacking; education; curriculums

ARES 2024, July 30–August 02, 2024, Vienna, Austria
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1718-5/24/07
https://doi.org/10.1145/3664476.3670455

ACM Reference Format:
Andrej Krištofík, Jakub Vostoupal, Kamil Malinka, František Kasl, and Pavel Loutocký. 2024. Beyond the Bugs: Enhancing Bug Bounty Programs through Academic Partnerships. In *The 19th International Conference on Availability, Reliability and Security (ARES 2024), July 30–August 02, 2024, Vienna, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3664476.3670455

1 INTRODUCTION

Whereas the critical nature of cybersecurity is seldom disputed, the conflict between Russia and Ukraine drastically affected the threat landscape and highlighted the dangers of cybersecurity expert shortage, not only in the governmental services and public sector but in the private sector as well [5]. ENISA, in its Threat Landscape 2022 report, underscored the surge in geopolitically motivated cyber-attacks, hacktivism and the escalating capabilities of threat actors, spanning both nation-state and non-state entities [5]. Particularly alarming is the report's spotlight on the utilization of 0-day exploits, supply chain attacks and the uptick in phishing activities leading to data breaches [5]. Combined with the rapid development of AI-based tools and the increase in the offer of Malware as a Service, this trend significantly reduces the proficiency threshold required to execute successful cyberattacks or cybercrime campaigns.

As the frequency and impact of cyber-attacks continue to rise, the imperative to expand not only cybersecurity strategies and technologies but also the pool of cybersecurity experts becomes increasingly evident. Even though the skills shortage in cybersecurity is not a new trend, and the EU implemented several strategies [19] and created a number of projects to "fill" the cybersecurity market (e.g., CyberSec4Europe [15] or SPARTA [2]) through the means of higher education [31], the progress remains slow. According to the Digital Economy and Society Index 2021 report, 55 % of enterprises report difficulties recruiting ICT specialists [3]. Moreover, Microsoft's LinkedIn data analysis shows that *"the demand for cybersecurity expert workforce has grown by an average of 22 % between 2021 and 2022 alone"* [6].

In response to this challenge, companies have turned to crowdsourcing as a potential solution. Among the various crowdsourcing approaches, bug bounty programs and responsible/coordinated vulnerability disclosures stand out as more established yet underutilized methods. However, as previous research has highlighted [22, 23, 33, 41, 43, 47], numerous technical, legal and ethical considerations pose challenges to the widespread adoption of these techniques (e.g., lack of experienced experts, low visibility of certain areas, low transparency of the programs' legal conditions [47]).

These challenges may be best addressed within controlled environments, particularly within cybersecurity education [22, 23]. By incorporating bug bounty activities into the curriculum, students not only acquire the necessary skills to excel as cybersecurity experts (or ethical hackers [20]) but also gain crucial legal and ethical insights essential for navigating the complexities of bug bounty programs. This educational approach not only benefits students but also contributes to making bug bounty programs more appealing and transparent, thus potentially enhancing the overall cybersecurity landscape.

To underscore the benefits of integrating bug bounty education into the cybersecurity curricula, we commence the next section by introducing the fundamental concepts of Bug Bounty and responsible vulnerability disclosure, along with their associated benefits. Following this, we present our proposal on enhancing the efficiency of bug hunting while concurrently advancing cybersecurity education through collaboration between academia and vendors or bug bounty providers. Further, we delve into more granular strategies for academia to integrate real-world bug bounty and vulnerability disclosure programs into the curricula of cybersecurity-oriented courses, addressing the challenges inherent in implementing this proposal and conclude the article with a brief experience report from testing our theories in practice.

2 VULNERABILITY DISCLOSURE AND BUG BOUNTIES – THE JEDI'S WAY TO HACK?

Ensuring a sufficient level of cybersecurity within an organization is a never-ending Sisyphean struggle. Despite the pressing need for robust security solutions, implementation efforts are often hampered by insufficiently skilled or understaffed personnel, compounded by inadequate financial compensation [6, 36]. This situation will only worsen with the upcoming national implementations of the NIS 2 Directive, which will significantly expand the number of entities obligated to comply with relevant cybersecurity laws. In such conditions, many organizations will face heightened difficulties in meeting essential obligations, including vulnerability assessment and handling. That, of course, presents a serious threat, as an unchecked vulnerability remains an open door for a potential attacker [36].

Outsourcing, a popular solution for many entities in the context of cybersecurity, is not a silver-bullet solution in case of vulnerability handling as the red teaming is usually very expensive and thus out of reach for many of the small and medium enterprises [36]. A cheaper and more viable alternative (even though somewhat limited in extent and depth) can be found in "crowdsourcing", specifically vulnerability disclosure and bug bounty programs [14]. These are essentially ways of opening a product or service to a general or specific public to test the security solutions and report any vulnerabilities, altruistically or for a bounty [36]. Ensuring a sufficient level of cybersecurity within an organization is then, at least partially, reduced from "finding and fixing the vulnerability" to just "fixing the vulnerability" [36].[1] Moreover, involving external experts and enthusiasts broadens the capabilities and increases the probability of uncovering different kinds of vulnerabilities.

2.1 How do the Bug Bounty Programs work?

When an individual encounters a cybersecurity vulnerability, whether by chance or through intentional search, they may opt to report it to the responsible entity, allowing them to address and patch the issue [39]. This process of discovering, reporting, and patching vulnerabilities constitutes a vulnerability disclosure procedure, with terms like 'responsible' or 'coordinated' providing a further specification of the procedure itself [39].

General vulnerability disclosure programs primarily serve as a permission and procedural tool for facilitating the reporting process. And while enlisting the help of volunteers driven by purely altruistic motives proves effective in some cases [40, 47, 48], there are instances where an additional incentive becomes necessary. Although financial compensation it among the most traditional rewards [25, 29, 36, 47], hence the analogy with bounty hunting, it isn't the sole or necessarily the most motivating form of reward [21, 30, 37]. In certain scenarios, even reputational rewards alone suffice, as exemplified by the case of Jacob Riggs [35].

Essentially, bug bounty programs are nothing more than incentivised vulnerability disclosure programs, offering rewards for reported vulnerabilities [18].[2] Further elaboration on the mechanics of vulnerability disclosure and bug bounty hunting is not pertinent within this context. For those interested in delving deeper into this topic, we recommend referring to the following articles: [26, 46, 51].

The private sector, particularly tech giants and large corporations, frequently and effectively utilize bug bounty programs [7, 11, 12]. However, as ENISA pointed out [1, 8, 9], the situation differs in the governmental sector and among some SMEs [14, 16]. SMEs often lack awareness of these programs and their benefits, while the governmental sector also faces numerous legal hurdles in establishing adequate vulnerability disclosure procedures, let alone bug bounty programs, particularly within the European Union [16, 51].

Nevertheless, significant changes are underway in this regard, notably with the introduction of the NIS 2 Directive and the Cyber Resilience Act, as these regulations inter alia mandate Member States to develop national coordinated vulnerability disclosure policies and obligate manufacturers of products with digital elements to implement vulnerability disclosure as one of the security solutions [38].[3] Consequently, bug bounties and vulnerability disclosures are poised to become much more pertinent and prevalent.

[1] Issuing bug bounty programs cannot guarantee that all vulnerabilities are found, nor does it absolve any obliged entities of other obligations under cybersecurity laws (and other related regulations).

[2] There are some other differing aspects which are not necessarily relevant to this article.

[3] We have analysed this topic further in another paper, see [46]

3 THE PROBLEMS OF BUG HUNTING AND THE ROLE OF ACADEMIA

With the rising relevance and expected growth in the usage of bug bounty programs, it's crucial to address inherent problems in the current system before their widespread introduction to the European Union on a massive scale. A primary limitation of bug bounty programs is that they are only as effective as their participants. Often, the issue lies not only in the lack of skilled participants but more in the absence of any participants altogether, particularly in areas considered less attractive among specialists and students [17, 27, 32, 47].[4]

Although bug bounty programs in some of the less attractive areas are no less interesting than their more prestigious counterparts, it is necessary to admit that in a competitive race for skilled bug hunters among program providers [24], some public sector agencies, such as postal services, may be at a disadvantage [13] compared to private entities offering significant rewards (e.g., [7]) or even more attractive public sector bodies such as the military (e.g., [4]). This creates a kind of vicious circle for these sectors, where their bug bounty programs are not popular, resulting in minimal attention from hunters and, consequently, even organizations. Thus, the issuing organizations may not respond to the reports they receive or update rewards to be more attractive, further diminishing the programs' attractiveness. As a result, these sectors become even less appealing [13].

Generating sufficient interest from bug hunters is often hindered by a lack of knowledge about and experience with bug bounty programs, as well as concerns about their legal implications. Bug hunters may be wary of participating due to the risk of criminal sanctions. Additionally, the educational system, particularly in regions like the Czech Republic, is deficient in addressing the ethical aspects of hacking [22, 23]. Without proper education on the ethical and legal considerations surrounding hacking, bug hunters may not be familiar with the correct procedures, thereby increasing the risk of incurring criminal penalties and undermining their goodwill.

For these reasons, we advocate for a more active role of Academia, respectively the higher education as a whole, in addressing these issues. Educating future cybersecurity and IT experts about the benefits of ethical hacking, instilling a sense of altruism, and familiarizing them with bug bounty programs are crucial aspects for fostering a more secure cyber environment. Specifically, we propose integrating bug hunting into study courses to introduce students to bug bounty programs during their academic studies. However, this approach presents several challenges, such as motivating and assessing students without penalizing them for unsuccessful bug hunts, as bug hunting is indeed a challenging task. We delve into these challenges further in the text.

Nonetheless, it's important to emphasize that, in our view (supported by the conclusions of Hartley's [22, 23] and Trabelski's teams [42–44]), the benefits outweigh the challenges. Moreover, we believe that potential cooperation between academia, private and public sectors, and bug bounty providers is mutually beneficial. Such collaboration can facilitate knowledge exchange, skill development, and the identification and resolution of cybersecurity vulnerabilities, ultimately enhancing the overall cyber resilience of organizations and society as a whole.

4 THE FORM OF THE COOPERATION AND ITS BENEFITS

The form of students' involvement in bug hunting can vary [22, 23, 43]; however, we consider two models as the most "cost-effective" - for simplicity's sake, we refer to them as one-sided and two-sided involvement. The first model is straightforward to administer but may be somewhat limited in terms of experiences, personalization, and liability. In this model, students are introduced to bug-hunting topics, ethical and legal aspects, reporting procedures, and various testing procedures within a specific course or as part of a separate course, depending on the study program. Subsequently, students are assigned to one or a set of public bug bounty programs, preferably from less attractive sectors, and tasked with hunting bugs under these programs for the duration of the course. They document their progress and report throughout. Even this basic scheme contains a variety of issues and decisions that need to be taken into account by the academics in charge of such courses. In the next section, we discuss the general challenges that can be faced during the implementation process.

The second model is slightly more challenging to prepare but offers greater benefits. It entails establishing a partnership with bug bounty providers from disadvantaged sectors. Even if not for the purpose of ongoing cooperation, it is highly advisable to contact the provider of the selected bug bounty program *a priori*, as relying solely on a single mandatory program for a whole course may place a higher and potentially damaging strain on the given service or product. Moreover, early and effective communication may open up further opportunities for deeper collaboration. This may include creating a specific, focused bug hunt for the class, adjusting the terms and conditions of the bug bounty to better suit the class's needs (including specific liability regulations), or modifying the bounty to offer internships for successful hunters. Additionally, it could involve providing internal feedback to successful hunters and other variations to the standard program.

Despite potential disadvantages (e.g., educating future attackers), we contend that the benefits of incorporating bug hunting into courses outweigh them. Education in computer science degrees differs significantly from many other academic disciplines due to its strong emphasis on practicality [22, 23, 34]. Cybersecurity degrees or at least courses focused on cybersecurity, present an opportunity for a fruitful symbiosis with bug bounty programs. Furthermore, academia stands to gain significantly from more innovative, interactive, and practical courses. These not only attract students but also offer collaboration opportunities with the private and public sectors, along with the potential for talent acquisition. Additionally, such courses can greatly enhance the performance of graduates.

We would also like to emphasize that universities are the ideal candidates for this cooperation, given their vast reservoir of intellectual potential.[5] And as we firmly believe that any competent

[4] This situation is exacerbated by a lack of transparency in the terms and conditions of bug bounty programs, leading to legally uncertain situations where testers understandably choose not to report or participate in bug hunting.

[5] It's worth noting that the scope of participation doesn't necessarily need to be limited to cybersecurity students; including students from general "coding" study programs can also be beneficial.

cybersecurity expert should be capable of thinking like an attacker, experiencing the perspective of their possible "opponent" can help students better incorporate *security-by-default* thinking and develop better coding practices. Moreover, universities possess the necessary infrastructure to support such initiatives, and the university environment provides an excellent (and sometimes also last) opportunity to instil the principles of ethical hacking [22, 23, 43].

The benefits on the "bug bounty" side are twofold. Firstly, there is the obvious advantage of having more "manpower"[6] for programs that may not usually receive much attention. Secondly, there is the more abstract benefit of potentially increasing the popularity of bug hunting in the long run.

Hata, Guo, and Babar interviewed 2,504 bug bounty contributors to understand the motivations behind bug bounty contributions. They found that among "non-specific" contributors, i.e., those not specifically attached to a particular bug bounty program, the main reasons for participation were a) they use the product, and b) they like the company [24]. This sheds light on several aspects: firstly, it indicates that monetary rewards are not as crucial as often assumed,[7] which is promising for public sector bug bounties.[8] However, it also highlights the relatively unfocused nature of these bug hunts.

Therefore, we believe that fostering a better educational environment is essential. Not only should students be informed about the existence of bug bounty programs, but they should also be equipped with all the necessary knowledge.

5 COLLABORATION - THE PILOT RUN

To assess the viability of our proposed collaboration, we conducted a pilot run within a cybersecurity course focused on safe coding practices during the autumn semester of the current academic year. We introduced bug-hunting projects as voluntary alternatives to traditional study tasks because we were initially unsure about the difficulty of these projects and students' interest in such activities. A list of selected programs (based on their (un)popularity and technical level) was provided to the students. However, they were allowed to choose programs outside of this list in order to encourage them to choose programs close to their interests. Despite our initial uncertainties, the response exceeded our expectations. Out of 31 active students in the course, 13 opted for the bug bounty projects despite the increased time commitments. Among them, 3 found a vulnerability, and one made a full-scale report. Every student had to submit a report that detailed their process, used tools as well as overall strategy. This report was then based on their final evaluation, as their success in bug finding did not play a role in the final grade as long as they had chosen an appropriate strategy and set of tools. This approach solves the issue of different technical levels of different programs as well as provides all of the participating students with valuable feedback that could benefit them in their future bug-hunting endeavours.

Following the pilot program's conclusion and subsequent course grading, we solicited feedback from participating students through self-assessment surveys. The majority of participants regarded the bug-hunting project as a valuable, engaging, and educational experience. Many reported a significant increase in awareness regarding ethical hacking and bug bounty topics as well as in coding and hacking skills. Whilst not being rigorously objective, the self-assessment that informed our conclusion (Table 1) as to the educational value for students still provides an interesting insight and shows the increased awareness of the student about Bug Bounty and relevant tools and issues at the end of the pilot run.

We utilized a simple five-grade scale to quantify students' experience (with 1 representing minimal value and 5 representing maximum value) and also required a textual explanation of their evaluations to better understand their answers. The excerpt of students' experience is shown in the next table (due to the constraints of this paper, we have simplified the data):

	Average
Hours spent on Preparation	23.2
Hours spent on Search	27,4
Topic Awareness before (1-5)	1,9
Topic Awareness after (1-5)	3,6
Skills improvement (1-5)	3,4
Do you find this project beneficial? (1 - No, 5 - Yes)	4,7
How do you feel about working in a real environment on real products? (1-Not beneficial, 5-Very beneficial)	3,9
Was the ethicality of the project an important aspect for you? (1-5)	3,4

Table 1: Self-assesment results

Furthermore, 7 out of the 13 participating students expressed their intention to pursue further involvement in ethical hacking and bug bounty activities. For those interested in delving deeper into the specifics of this pilot run, we direct them to our other paper, where we provide a more comprehensive description of the experience and relevant data (see [28]).

6 LESSONS LEARNED

While the pilot run could be deemed successful, despite its inherent limitations and the absence of an objective assessment methodology, it also provided invaluable insights into various aspects and issues that academics overseeing such courses must take into account. These issues span from practical considerations to pedagogical ones.

Regarding practical aspects, it's crucial to identify suitable bug bounty programs. We still argue that less attractive programs should be prioritized. In addition to the benefits mentioned above, these

[6] Furthermore, this manpower is relatively easily manageable, and the testing process can be easily modified.

[7] This finding aligns with broader psychological research into the motivation of volunteers. Notably, monetary rewards can have a negative effect on those volunteers motivated purely by altruism or a love for the brand or society [30, 37, 49].

[8] Besides this one, a different question in this research was as well aimed at the motivation with regards to bounties, where the respondents showed that the bounty is as important as making the users safe and helping the developers in their list of values.

also offer much more undiscovered "easy" bugs that students are more likely to find; hence these programs may also serve to motivate students to pursue this topic further.

We advise against selecting a single bug bounty program, as various factors could negatively impact individual students' performance if they cannot choose according to their skills and preferences. These factors may include the platform used in the bug bounty program or the programming language of the tested solution. Therefore, the list of "offered" bug bounty programs should be sufficiently diverse to cover all of the most frequently used technologies, ensuring that each student has a chance of success.

However, determining the exact disadvantaged sectors may pose challenges initially, as comprehensive data throughout the EU is currently lacking. The broad implementation of vulnerability disclosure programs and strategies has yet to come, with the implementation of the NIS 2 Directive and the adoption of the Cyber Resilience Act, so the uncertainty is still high.

To address this issue, we propose a possible solution for universities: open cooperation with the governmental sector, establish a reputation, and raise general awareness about such collaboration. Subsequently, universities can motivate individual vendors to initiate further cooperation. Vendors could leverage their participation as part of their promotion strategy, positioning themselves as more responsible, transparent, and potentially secure.

Another viable option is identifying underused sectors in the bug-hunting context through cooperation with a strategic coordinator of vulnerability disclosures and bug bounty programs. A coordinator typically simplifies the process by coordinating activities between vendors and vulnerability reporters [10]. Examples of such coordinators include HackerOne and BugCrowd. Creating a national coordinator is expected during the implementation of the NIS 2 Directive and could be instrumental in addressing this problem.

This mutual cooperation could also be instrumental in addressing another practical issue, which is partially pedagogy-oriented: assessing the students' participation in the bug bounty program. While evaluating them solely based on their success may be tempting, this approach may not be the most appropriate, as successful bug hunting is inherently demanding. This situation highlights the importance of focusing on the learning process rather than the outcome [41–43, 50].

One ideal, albeit unlikely, scenario is for the teacher to have access to the provider's bug bounty platform through mutual cooperation, enabling them to monitor the activities of individual students. A more realistic solution is to require students to record the process[9] and submit it for assessment along with a detailed bug report. Alternatively, students could submit the bug report for assessment along with an essay describing not only their process but also the reasoning behind their chosen approach. This approach emphasizes the importance of understanding the process rather than achieving a specific outcome.

In practice, it's possible to combine these methods into a fully interactive presentation of the results, where students share their thought processes, dead ends, and final solutions with their colleagues. However, timing the school project, the platform's response time and the actual grading can present challenges. Finding the right balance between these factors is crucial for ensuring a fair and effective assessment process.

As described earlier, one of the benefits of bug bounty programs is the transfer of non-technical skills and knowledge necessary for participating in bug bounties. This includes understanding legal issues, which is a critical aspect that teachers must address before "releasing" students onto real-world bug bounty programs.

Students should not only have a comprehensive understanding of potential legal issues and how to navigate them but also confidence that their actions will not lead to any legal consequences. This aspect needs to be emphasized in the course to ensure students understand the extent to which legal issues are covered by rules and regulations. Additionally, defining the allowed scope, methods, and other parameters will provide legal protection for both the teacher and the university.

Choosing suitable bug bounty programs (assuming enough vulnerable systems are available) and aligning curriculums pose challenges. Therefore, the preparation phase for teachers must not be underestimated.[10] Proper preparation is crucial to ensure that students are well-equipped to participate in bug bounty programs safely and effectively.

This brings us to the issue of bug reports not only being part of the assessment but also the foundational component of bug bounties themselves. In cybersecurity education, it's crucial to provide students not only with technical knowledge but also with soft skills and other relevant abilities to navigate the professional world of cybersecurity effectively [23, 43, 47]. This includes understanding legal issues, which is a critical aspect that teachers must address before "releasing" students onto real-world bug bounty programs. Students should not only possess a comprehensive understanding of potential legal issues and how to navigate them but also confidence that their actions will not lead to any legal consequences. This includes the ability to navigate the intricacies of a particular program's rules. Additionally, teachers should properly define allowed and banned scope, methods, and other parameters that will provide legal protection for the teacher, students, and the university.[11]

In this context, it is also noteworthy that the preparation phase for teachers is rather time-demanding and must not be underestimated. Proper preparation is crucial to ensure that students are well-equipped to participate in bug bounty programs safely and effectively.

Apart from the technical and legal know-how, essential soft skills directly tied to bug hunting itself have been emphasized by Malladi and Subrama [29]. Therefore, it is highly advisable not only to thoroughly go through the steps of bug hunting and emphasize when the process should stop but also to review bug reports and the process of writing and submitting them. Additionally, delving deeper into the ethical aspects and benefits, as highlighted by Hartley [22], is crucial.

Ideally, students should complete a mock report in the preparatory part of the course and focus on bug hunting with technical

[9] Either in a written document or via a screen capture.

[10] As mentioned above, we suggest *a priori* agreement with a public entity as the most suitable solution for the beginnings of the more thorough course

[11] This should be done in an even stricter manner than the allowed techniques according to the rules of the bug bounty program, as some testing activities may be more dangerous when used by inexperienced students.

case-based troubleshooting in the second part of the course. The mock report should be accompanied by proper feedback to prepare students for submitting actual reports and enable them to describe relevant technical aspects of the vulnerability meaningfully. This approach ensures that students develop the necessary skills to participate effectively in bug bounty programs.

Last but not least, and definitely one of the more controversial aspects of this proposal that cannot be overlooked, is teaching students ethical hacking. There has been much debate on this topic over the last few years, with experts still not reaching a consensus [22, 23, 34, 45]. However, from our perspective, a proficient cybersecurity expert needs to be able to think like an attacker and use offensive techniques to anticipate and analyze risks sufficiently.

Of course, there is always a possibility that students could misuse these skills; however, we argue that the skillset of an ethical hacker (and bug hunter) is crucial for cybersecurity experts, and it is better to try to teach students the ways of ethicality and influence them accordingly. We also cannot ignore that anyone motivated to be a black-hat hacker can easily acquire the necessary skillset on the internet while getting into ethical hacking (including garnering the necessary motivation) without external help is more demanding.

We also emphasize the importance of how incorporating bug hunting into a curriculum would impact non-cybersecurity-oriented IT students. Encouraging them to think not only about "how to make it work" but also "how to crack it" could help them implement the standard of security by design and default into their mindsets, furthering cybersecurity as a whole.

Based on the findings from our pilot run, we hold that this option should be incorporated more into relevant curricula, as it proved to raise awareness, and willingness of students, to participate in Bug Bounty Programs, which could, in the long run, help with the current personnel problem of the (European) cybersecurity. While the scenario we have described, and the issues and their solutions, should scale to bigger classes and programs directly, two further points should be considered for larger classes. Firstly, in order to truly provide students with a benefit (and confidence) for their future bug hunting, the final report requires thorough and time-intensive feedback by the lecturer(s) and having large classes could degrade the quality of such feedback. Secondly, proper distribution of students to Bug Bounty programs should be ensured, for example, by publicly "claiming" a given program so that there is no overload of a single service or a program. Other than that, we hold that the suggested scenario could prove beneficial even for larger classes or study programs.

For further discussion on this matter, we recommend referring to the article written by Dr. Hartley, as we almost exclusively agree with the presented solutions, suggestions, analyses, and conclusions [22].

7 CONCLUSION

As a result of regulatory shifts within the European Union, vulnerability disclosure and bug bounty programs are poised to assume heightened significance within the cybersecurity landscape. However, the efficacy of these initiatives is inherently contingent upon the expertise of participating individuals. Consequently, the burgeoning demand for skilled cybersecurity professionals, particularly in less sought-after sectors, poses a substantial challenge for the market.

In response to this pressing issue, this article proposes a proactive approach in the form of a collaboration between academia and both private and public bug bounty issuers. The proposed collaboration is beneficial not only to the issuers, who get relatively easily manageable manpower (which allows for easier management of the program itself and is therefore suitable even for less experienced issuers), but also to students. By adopting bug bounty programs into cybersecurity courses, students can better prepare for real-world scenarios and transition theoretical knowledge into practical skills, aligning with computer science education trends. This integration also offers a platform for students to develop soft skills essential for bug hunting and cybersecurity work in general. Leveraging the educational environment to train future professionals in bug hunting not only addresses the immediate need for skilled individuals but also fosters a long-term solution by nurturing a talent pipeline.

During the autumn semester of the current academic year, we conducted a pilot of the proposed collaboration within a cybersecurity course focused on safe coding practices. Initially offered as a voluntary activity due to our uncertainties about task difficulty and students' interest, the bug-hunting initiative exceeded our expectations. Many students enthusiastically embraced this alternative, dedicating significant time not only to their projects but also to voluntary skill enhancements, such as through Portswigger Academy.

This trial yielded invaluable insights, which we presented in the last section of this article.

ACKNOWLEDGMENTS

This article was supported by the Ministry of the Interior of the Czech Republic project "Capacity Building in Cybersecurity (CaBuCy)", no. VJ03030052.

REFERENCES

[1] 2016. Good Practice Guide on Vulnerability Disclosure. From challenges to recommendations. https://www.enisa.europa.eu/publications/vulnerability-disclosure

[2] 2019. Strategic Programs for Advanced Research and Technology in Europe - SPARTA. https://cordis.europa.eu/project/id/830892

[3] 2021. Digital Economy and Society Index 2021: Overall progress in digital transition but need for new EU-wide efforts. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_5481

[4] 2022. Hack The Army. https://www.arcyber.army.mil/Resources/Fact-Sheets/Article/3106335/hack-the-army/https%3A%2F%2Fwww.arcyber.army.mil%2FResources%2FFact-Sheets%2FArticle%2F3106335%2Fhack-the-army%2F

[5] 2022. Threat Landscape 2022. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

[6] 2022. The urgency of tackling Europe's cybersecurity skills shortage. https://blogs.microsoft.com/eupolicy/2022/03/23/the-urgency-of-tackling-europes-cybersecurity-skills-shortage/

[7] 2023. Apple Security Bounty. https://security.apple.com/bounty/

[8] 2023. Coordinated Vulnerability Disclosure: Towards a Common EU Approach. https://www.enisa.europa.eu/news/coordinated-vulnerability-disclosure-towards-a-common-eu-approach

[9] 2023. Developing National Vulnerabilities Programmes. https://www.enisa.europa.eu/publications/developing-national-vulnerabilities-programmes

[10] 2023. HackerOne - About us. https://www.hackerone.com/

[11] 2023. Meta Bug Bounty Information. https://www.facebook.com/whitehat/info

[12] 2023. Rewards Program. https://security.samsungmobile.com/rewardsProgram.smsb

[13] Omer Akgul, Taha Eghtesad, Amit Elazari, Omprakash Gnawali, Jens Grossklags, Michelle L. Mazurek, Daniel Votipka, and Aron Laszka. 2023. Bug Hunters' Perspectives on the Challenges and Benefits of the Bug Bounty Ecosystem. https://doi.org/10.48550/ARXIV.2301.04781

[14] Jasmine Arooni. 2021. Debugging the System: Reforming Vulnerability Disclosure Programs in the Private Sector. *Federal Communications Law Journal* 73, 3 (2021), 443–466.

[15] Cyber Security for Europe. 2020. Addressing the Shortage of Cybersecurity Skills in Europe. https://cybersec4europe.eu/addressing-the-shortage-of-cybersecurity-skills-in-europe/

[16] Cristina Del-Real and María José Rodriguez Mesa. 2023. From Black to White: The Regulation of Ethical Hacking in Spain. *Information & Communications Technology Law* 32, 2 (May 2023), 207–239. https://doi.org/10.1080/13600834.2022.2132595 Publisher: Routledge _eprint: https://doi.org/10.1080/13600834.2022.2132595.

[17] Amit Elazari. 2019. Private Ordering Shaping Cybersecurity Policy: The Case of Bug Bounties. In *Rewired: Cybersecurity Governance*, Ryan Ellis and Vivek Mohan (Eds.). Rochester, NY, 231–246.

[18] Ryan Ellis and Yuan Stevens. 2022. Bounty Everything: Hackers and the Making of the Global Bug Marketplace. https://doi.org/10.2139/ssrn.4009275

[19] Miguel González-Sancho. 2019. Four EU pilot projects launched to prepare the European Cybersecurity Competence Network. https://ec.europa.eu/digital-single-market/en/news/four-eu-pilot-projects-launched-prepare-european-cybersecurity-competence-network 00000.

[20] Tim Greene. 2004. Training Ethical Hackers: Training the Enemy? https://defcon.org/html/links/dc_press/archives/12/ebcvg_training_ethical_hackers.htm

[21] Stefan Tomas Güntert, Isabel Theresia Strubel, Elisabeth Kals, and Theo Wehner. 2016. The quality of volunteers' motives: Integrating the functional approach and self-determination theory. *The Journal of Social Psychology* 156, 3 (May 2016), 310–327. https://doi.org/10.1080/00224545.2015.1135864

[22] Regina Hartley. 2015. Ethical Hacking Pedagogy: An Analysis and Overview of Teaching Students to Hack. *Journal of International Technology and Information Management* 24 (Jan. 2015), 95–104. https://doi.org/10.58729/1941-6679.1055

[23] Regina Hartley, Dawn Medlin, and Zach Houlik. 2017. Ethical Hacking: Educating Future Cybersecurity Professionals. *Information Systems & Computing Academic Professionals: Proceedings of the EDSIG Conference* (2017), 1–10.

[24] Hideaki Hata, Mingyu Guo, and Muhammad Ali Babar. 2017. Understanding the Heterogeneity of Contributors in Bug Bounty Programs. (Sept. 2017).

[25] Keman Huang, Jia Zhang, Wei Tan, and Zhiyong Feng. 2015. An Empirical Analysis of Contemporary Android Mobile Vulnerability Market. In *2015 IEEE International Conference on Mobile Services*. IEEE, New York City, NY, USA, 182–189. https://doi.org/10.1109/MobServ.2015.34

[26] Aron Laszka, Mingyi Zhao, Akash Malbari, and Jens Grossklags. 2018. The Rules of Engagement for Bug Bounty Programs. In *Financial Cryptography and Data Security (Lecture Notes in Computer Science)*, Sarah Meiklejohn and Kazue Sako (Eds.). Springer, Berlin, Heidelberg, 138–159. https://doi.org/10.1007/978-3-662-58387-6_8

[27] Zhen Li and Qi Liao. 2018. Economic solutions to improve cybersecurity of governments and smart cities via vulnerability markets. *Government Information Quarterly* 35, 1 (Jan. 2018), 151–160. https://doi.org/10.1016/j.giq.2017.10.006

[28] Kamil Malinka, Anton Firc, Pavel Loutocký, Jakub Vostoupal, Andrej Krištofík, and František Kasl. 2024. Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report. (April 2024). https://doi.org/10.1145/3649217.3653633 arXiv:2404.12043 [cs].

[29] Suresh S. Malladi and Hemang C. Subramanian. 2020. Bug Bounty Programs for Cybersecurity: Practices, Issues, and Recommendations. *IEEE Software* 37, 1 (Jan. 2020), 31–39. https://doi.org/10.1109/MS.2018.2880508 Conference Name: IEEE Software.

[30] Vanessa Mertins and Christian Walter. 2021. In absence of money: a field experiment on volunteer work motivation. *Experimental Economics* 24, 3 (Sept. 2021), 952–984. https://doi.org/10/gmcztj

[31] Jason R. C. Nurse, Konstantinos Adamos, Athanasios Grammatopoulos, and Fabio Di Franco. 2021. Addressing Skills Shortage and Gap Through Higher Education. https://www.enisa.europa.eu/publications/addressing-skills-shortage-and-gap-through-higher-education

[32] Taiwo A. Oriola. 2011. Bugs for sale: Legal and ethical proprietaries of the market in software vulnerabilities. *John Marshall Journal of Computer & Information Law* 28, 4 (2011). https://repository.law.uic.edu/jitpl/vol28/iss4/1

[33] Brian A. Pashel. 2006. Teaching students to hack: ethical implications in teaching students to hack at the university level. In *Proceedings of the 3rd annual conference on Information security curriculum development*. ACM, Kennesaw Georgia, 197–200. https://doi.org/10.1145/1231047.1231088

[34] Nicole Radziwill, Jessica Romano, Diane Shorter, and Morgan Benton. 2015. The Ethics of Hacking: Should It Be Taught? *ArXiv* (Dec. 2015).

[35] Jacob Riggs. 2021. I hacked the Dutch government and all I got was this t-shirt. https://jacobriggs.io/blog/posts/i-hacked-the-dutch-government-and-all-i-got-was-this-t-shirt-24.html

[36] Jukka Ruohonen and Luca Allodi. 2018. A Bug Bounty Perspective on the Disclosure of Web Vulnerabilities. *17th Annual Workshop on the Economics of Information Security, Innsbruck* (May 2018). http://arxiv.org/abs/1805.09850 arXiv: 1805.09850.

[37] Richard M Ryan and Edward L Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55, 1 (2000), 68–78. https://doi.org/10/c48g8h

[38] Sandra Schmitz and Stefan Schiffner. 2021. Responsible Vulnerability Disclosure under the NIS 2.0 Proposal. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 12, 5 (2021), 448–457. https://heinonline.org/HOL/P?h=hein.journals/jipitec12&i=646

[39] Stephen Shepherd. 2021. How do we define Responsible Disclosure?

[40] Jantje Silomon, Mischa Hansel, and Fabiola Schwartz. 2022. Bug Bounties: Between New Regulations and Geopolitical Dynamics. *International Conference on Cyber Warfare and Security* 17, 1 (March 2022), 298–305. https://doi.org/10.34190/iccws.17.1.21

[41] Zouheir Trabelsi. 2011. Hands-on lab exercises implementation of DoS and MiM attacks using ARP cache poisoning. In *Proceedings of the 2011 Information Security Curriculum Development Conference*. ACM, Kennesaw Georgia, 74–83. https://doi.org/10.1145/2047456.2047468

[42] Zouheir Trabelsi. 2012. Switch's CAM table poisoning attack. In *Computing Education 2012 - Proceedings of the 14th Australasian Computing Education Conference (Conferences in Research and Practice in Information Technology Series)*, Michael de Raadt and Angela Carbone (Eds.). Australian Computer Society, Melbourne, Australia, 113–120. http://www.scopus.com/inward/record.url?scp=85014905333&partnerID=8YFLogxK Publisher: Australian Computer Society.

[43] Zouheir Trabelsi. 2014. Enhancing the comprehension of network sniffing attack in information security education using a hands-on lab approach. In *Proceedings of the 15th Annual Conference on Information technology education*. ACM, Atlanta Georgia USA, 39–44. https://doi.org/10.1145/2656450.2656462

[44] Zouheir Trabelsi and Latifa Alketbi. 2013. Using network packet generators and snort rules for teaching denial of service attacks. In *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*. ACM, Canterbury England, UK, 285–290. https://doi.org/10.1145/2462476.2465580

[45] Zouheir Trabelsi and Margaret McCoey. 2016. Ethical Hacking in Information Security Curricula. *International Journal of Information and Communication Technology Education* 12, 1 (Jan. 2016), 1–10. https://doi.org/10.4018/IJICTE.2016010101

[46] Jakub Vostoupal, Václav Stupka, Jakub Harašta, František Kasl, Pavel Loutocký, and Kamil Malinka. 2023. The Legal Aspects of Cybersecurity Vulnerability Disclosure: To the Nis 2 and Beyond. https://doi.org/10.2139/ssrn.4640775

[47] Thomas Walshe and Andrew Simpson. 2020. An Empirical Study of Bug Bounty Programs. In *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*.

London, ON, Canada, 35–44. https://doi.org/10.1109/IBF50092.2020.9034828

[48] T. Walshe and A. C. Simpson. 2022. Coordinated Vulnerability Disclosure programme effectiveness: Issues and recommendations. *Computers & Security* 123 (Dec. 2022). https://doi.org/10.1016/j.cose.2022.102936

[49] Dale R. Wright, Les G. Underhill, Matt Keene, and Andrew T. Knight. 2015. Understanding the Motivations and Satisfactions of Volunteers to Improve the Effectiveness of Citizen Science Programs. *Society & Natural Resources* 28, 9 (Sept. 2015), 1013–1029. https://doi.org/10.1080/08941920.2015.1054976 Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/08941920.2015.1054976.

[50] Mantz Yorke. 2003. Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education* 45, 4 (2003), 477–501. https://doi.org/10.1023/A:1023967026413

[51] Uldis Ķinis. 2018. From Responsible Disclosure Policy (RDP) towards State Regulated Responsible Vulnerability Disclosure Procedure (RVDP): The Latvian approach. *Computer Law & Security Review* 34, 3 (June 2018), 508–522. https://doi.org/10.1016/j.clsr.2017.11.003

Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report

Kamil Malinka
Brno University of Technology
Brno, Czech Republic
malinka@fit.vut.cz

Anton Firc
Brno University of Technology
Brno, Czech Republic
ifirc@fit.vut.cz

Pavel Loutocký
Masaryk University
Brno, Czech Republic
Pavel.Loutocky@law.muni.cz

Jakub Vostoupal
Masaryk University
Brno, Czech Republic
Jakub.Vostoupal@law.muni.cz

Andrej Krištofík
Masaryk University
Brno, Czech Republic
Andrej.Kristofik@law.muni.cz

František Kasl
Masaryk University
Brno, Czech Republic
Frantisek.Kasl@muni.cz

ABSTRACT

To keep up with the growing number of cyber-attacks and associated threats, there is an ever-increasing demand for cybersecurity professionals and new methods and technologies. Training new cybersecurity professionals is a challenging task due to the broad scope of the area. One particular field where there is a shortage of experts is Ethical Hacking. Due to its complexity, it often faces educational constraints. Recognizing these challenges, we propose a solution: integrating a real-world bug bounty programme into the cybersecurity curriculum. This innovative approach aims to fill the practical cybersecurity education gap and brings additional positive benefits.

To evaluate our idea, we include the proposed solution to a secure coding course for IT-oriented faculty. We let students choose to participate in a bug bounty programme as an option for the semester assignment in a secure coding course. We then collected responses from the students to evaluate the outcomes (improved skills, reported vulnerabilities, a better relationship with security, etc.). Evaluation of the assignment showed that students enjoyed solving such real-world problems, could find real vulnerabilities, and that it helped raise their skills and cybersecurity awareness. Participation in real bug bounty programmes also positively affects the security level of the tested products. We also discuss the potential risks of this approach and how to mitigate them.

CCS CONCEPTS

• **Security and privacy** → Vulnerability management; *Penetration testing*; • **Social and professional topics** → **Computing education**; Computing education programs;

KEYWORDS

University Education, Bug Bounty, Cybersecurity Specialization, Secure coding, Course Assignment, Experience Report

ITiCSE 2024, July 8–10, 2024, Milan, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0600-4/24/07
https://doi.org/10.1145/3649217.3653633

ACM Reference Format:
Kamil Malinka, Anton Firc, Pavel Loutocký, Jakub Vostoupal, Andrej Krištofík, and František Kasl. 2024. Using Real-world Bug Bounty Programs in Secure Coding Course: Experience Report. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8–10, 2024, Milan, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649217.3653633

1 INTRODUCTION

The IT threat landscape continuously increases, resulting in a vastly increased demand for cybersecurity experts. However, the lack of cybersecurity experts is a long-known problem. One of the solutions to how we tried to tackle this problem is to use crowdsourcing solutions to share intelligence. An example of such efforts is bug bounty programs (BBPs).

Bug Bounty Programs should be considered indispensable tools promoting responsible vulnerability disclosure [5]. These programs do not rely only on altruistic and randomly encountered ethical hackers. They incentivize them with rewards for reporting relevant cybersecurity vulnerabilities [5]. Vendors typically announce these specific competitions to external stakeholders, ranging from the general public to researchers and security-oriented companies [11].

In the case of a bug bounty program, which is one of the branches of so-called ethical hacking, they are coordinated to some extent by the entity that has an interest in discovering vulnerabilities in a given system (not necessarily its own). We find it essential to coordinate related activities within these approaches, not only in relation to the vulnerability testing itself but especially apriori at the level of education within specialized study programs to implement and introduce study projects that would lead to the use of bug bounty-related approaches as part of student motivation.

Thus, cybersecurity education can help address the shortage of cybersecurity professionals by introducing students to ethical hacking and allowing them to participate in real-world BBPs.

This cooperation between academia and real-world business yields several key benefits for both parties, fostering much-needed synergy. This includes endowing students with valuable real-life experiences that test their technical knowledge and cultivate non-technical skills, such as effective communication and report filing [9]. Moreover, these experiences benefit students' personal development and enhance their employability, leveraging the prestige associated with successful bounty hunts [5].

We further argue that the positive effects of introducing BBPs into the cybersecurity curriculum are not limited only to the development of hard and soft skills but hold a much more profitable potential by instilling in students the principles of ethical hacking and cultivating an appreciation for the benefits of both BBPs and ethical hacking as a whole, even outside the course work [3].

In the context of this paper, we evaluate specific approaches and experiences with implementing bug bounties in the curriculum, the practical benefits for students, and the impact on their knowledge and skill base. It is also crucial to evaluate whether students of a particular IT program can solve specific tasks and requirements of bug bounty programs. Based on our teaching experience, we know that students like real-world problems and appreciate real-world examples. However, we were interested to see how they would cope with this type of problem, as it represents quite a big leap from conventional school problems.

To experimentally verify our proposal, we let students voluntarily select participating in a BBP as a semester project in a secure coding course. The success or failure of the search for the vulnerability was not reflected in the final grade. We evaluated only the process, used tools, and final report. We also organized an extra lecture for those interested in this project area to give students a basic orientation in the field of ethical hacking.

After completing the assignments, we surveyed the students to find out how they liked the possibility of being involved in solving real-world problems. We focus on three areas of questions: technical parameters, student self-evaluation, and project evaluation. Additionally, we examined how many vulnerabilities will be reported by BBP participants with no prior knowledge (students) and how the students perceive such an assignment (entertainment, skills, risks).

Contributions. The main contributions of this paper may be stated as follows:

- We proposed an innovative way to improve the learning of IT and cybersecurity professionals that leverage real bug bounty programs in the curriculum.
- We have experimentally implemented and evaluated the proposed idea in one run of the university course.
- We discuss the pedagogical implications of the proposed approach and have shown, among other things, a positive impact on learning and that students can solve this type of task successfully.

2 MOTIVATION

No IT solution is free of bugs and vulnerabilities [19], and finding and patching vulnerabilities is an iterative process that is both financially and human capital intensive [20]. Letting external entities (i.e. cybersecurity researchers, testers, enthusiasts or hackers) search for and responsibly disclose cybersecurity vulnerabilities can thus be an effective security tool to mitigate such vulnerabilities.

Such external help can be very welcome (or even needed) from the vendor's side [11]. A specific approach to gathering these external entities is the introduction of the already mentioned BBPs, where the system or device is subjected to "planned" attacks at the request of the vendor and the ethical hackers are then rewarded for reporting found vulnerabilities [10]. It is then essential to ensure that the relevant activities are legal not only from the perspective of the vendor but also from the perspective of the ethical hackers themselves so that they do not have to worry about unwanted sanctions (which would limit or even eliminate the motivation for conducting such actions [18]). Nevertheless, it is essential to emphasize that many actors involved in the BBPs' procedures (the notifying ethical hacker, the vendor, or third parties and coordinators) often have incompatible goals and interests [7]. The potential conflicts between these actors may seriously hinder the motivational aspect, and it is, therefore, crucial to mitigate this risk by thorough setup of the conditions for the announcement (and the whole of the procedure) under a bug bounty program [17].

However, it has to be stated that participation and hands-on experiences in the context of these programs are not often encountered by students (not only) of higher education programs, although they can be of great benefit to them [3, 9, 16]. In our experience, which we present within this paper, this is a benefit not only in terms of gaining actual awareness and concrete, hands-on experience with such cyber security vulnerability detection but especially in terms of gaining superior related skills that appropriately further shape the specialized profile of the students themselves.

In such regard, the cooperation between the educational institutions and the vendors has its own significance, as it may satisfy the vendors' need for such expert labour, which is currently lacking in the European market [1]. On the other hand, this cooperation allows students to gain hands-on experience and help educators better prepare them for real scenarios. Making the courses more interesting by utilising a real-world exercise is also an element that should not be overlooked [3, 9, 16]. Gaining actual awareness and concrete hands-on experience with such cyber security vulnerability detection, and especially gaining superior related skills [3] help further shape the students' specialised profile.

We must also consider another factor - the time allotment structure of the curriculum. The curriculum of security-oriented courses and specializations is often broad and covers only the basics of all the areas concerned. Students are often taught the basics as a part of general security-oriented courses, but it is undesirable to neglect other areas such as cryptography, authentication, or IT security management in favour of ethical hacking. The education in the area of ethical hacking is thus often minimal and left to the students themselves. The situation is even more problematic in the area of IT students not focused on security. Although they should also have the basics of cybersecurity, they have even less teaching space. Also, incorporating a real-world problem into teaching brings with it a number of challenges in addition to the benefits mentioned above - greater difficulty may prevent students from successfully solving, ensuring assignment consistency for a fair assessment, or emphasis on practical knowledge of a wide range of tools.

3 RELATED WORK

The use of bug bounty programs for controlled vulnerability discovery is very common in the case of large companies (e.g. Apple[1], Samsung[2] or Microsoft[3]). Such programs usually have a graduated range of rewards according to the vulnerability discovered in the

[1] https://security.apple.com/bounty/

[2] https://security.samsungmobile.com/rewardsProgram.smsb

[3] https://www.microsoft.com/en-us/msrc/bounty

general terms of the program. It is thus evident from practice that the introduction of bug bounty programmes brings substantial benefits (both financial and security), which are all the more evident in conjunction with specialised service providers, which usually act as intermediaries in the processing of the programme. These include companies such as BugCrowd[4] and HackerONE[5]. Programs offered by these services were used in the 2020 quantitative study conducted by Walshe and Simpson [17] that has demonstrated how a well-deployed program could, in financial terms, substitute for two full-time experts.

In the context of education, however, the problematic implementation of practical knowledge related to bug bounty programmes into university teaching is evident [3, 9, 16][6]. There is a generally noticeable orientation towards broader educational focuses that focus on, for example, cybersecurity specialists; thus, a specific targeting on bug bounty-related skills is sporadic[7], even though it may be deemed a crucial part of the cybersecurity professionals' skillset [8]. In the context of cybersecurity education, this was emphasized already by Greene [2] and also in the research of Logan and Clarkson [6] and Pashel [8], and studied further by Trabelski (and others) [12–16] and Hartley (and others) [3, 4].

We aim to showcase the benefits and insights gained from integrating bug bounty programs into our curriculum, echoing Hartley's findings on the significant impact of hands-on experience in cybersecurity education [3].

We advocate that a proper curriculum design should give the students hands-on experience and provide them with the necessary soft skills and knowledge outside of the tech domain. This part of the curriculum should also focus on the legal and ethical aspects of hacking, as Trablesi et al. [16] have shown in their research that there is a potential for malicious activity done by the students.

4 PROPOSED SOLUTION

This section presents a possible solution to the above-mentioned problems by combining education and real-world ethical hacking. Our innovative approach aims to fill the practical cybersecurity education gap and bring other positive benefits.

To evaluate our idea, we implemented the proposed solution to a course for IT-oriented faculty. We let students choose to participate in a bug bounty programme as an option for the semester project in a secure coding course. We then collected responses from the students to evaluate the outcomes. In addition to the technical question of whether students without prior expertise in computer security would even be able to successfully solve a BBP and what procedures and tools they would use, we also investigated how students evaluate this type of project and what the implications are for teaching this topic.

It has to be said that there are several risks in letting the students participate in BBP. Firstly, the students may not be able to identify any vulnerability. While this may be a problem for the BBP itself, the project focuses on the process and educational benefits. Shortly said, even failure to identify any vulnerability is considered positive if the students improve their skills and broaden their horizons. Secondly, the students may violate one of the rules of the BBP.

To mitigate the risks, we provide a proper introduction to ethical hacking and BBPs in the form of a lecture delivered by an expert in the field. An integral component of the course involves teaching students the necessary skills of the ethical hacker [3]. These encompass technical proficiency and a fundamental understanding of the relevant legal framework and the ability to navigate the legal specifics, rules, and methodologies of the particular BBP [4].

This ensures that their actions are conducted within legal boundaries, minimizing the risk of causing undue harm and facing subsequent consequences [4].

We propose specifically dedicating one of the introductory lessons to the legal issues, where students are acquainted with locating the specific rules of a given BBP and the potential consequences of overstepping these rules or targeting services that fall outside the defined boundaries. This approach aims to establish a secure foundation for both students and educators, fostering a safe learning environment. Additionally, it assists in orienting them toward ethical hacking methodologies and ensures a responsible and lawful engagement with bug bounty initiatives [16].

5 COURSE PROJECT DESCRIPTION

The experiment was conducted as part of an optional university course focused on secure coding. The course is regularly taught in computer security specialization at the IT-oriented faculty. The course is designed for Master's students, introduces the basic principles of secure coding, and explains the general principles of vulnerabilities and defences against them. It covers multiple areas, such as basic vulnerabilities of compiled languages, memory protection mechanisms, input validation, static and dynamic analysis, and more, but in general, the course is not focused on ethical hacking.

An important part of the course is the semester project, for which students can get almost half of all points counted in the final grade. Students have two months to finish the project. Students are expected to work independently on a selected topic that falls within the course's content area. They are expected to study the relevant materials, research the chosen area, and possibly implement selected solutions.

The result is a technical article in the selected area of at least six pages in double-column IEEE format (for implementation, the output may differ by agreement). An integral part of the solution is the oral presentation of the whole work, which takes place in the course seminars. The goal is for students to demonstrate both hard and soft skills, as this combination is expected of future cybersecurity professionals.

Students choose their own topic, but the course teacher must always approve it. They choose from 3 types of projects: *tutorial*, *HW or SW implementation*, and *original work*.

In the *Tutorial*, students have to study a selected topic in depth and write a short tutorial or overview study with the structure of a scientific paper. Their own opinion and analysis are welcome. In the *Implementation*, they choose an algorithm and a platform on

[4] https://www.bugcrowd.com/products/platform/

[5] https://www.hackerone.com/

[6] On the other hand, there is a relatively large number of specifically targeted courses, especially in the online environment. See e.g. here: https://securitytrails.com/blog/popular-bug-bounty-courses or here: https://www.classcentral.com/report/best-bug-bounty-courses/

[7] This is also evident from the report within the SPARTA project - https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5d212c432&appId=PPGMS, e.g.. p. 53 or 56.

which they then implement the algorithm with respect to a specific set of goals (speed, security, and/or performance). The goal is to produce a working code and write a paper explaining the implementation. Within the last area, *Original work*, they can present their own activities if they work on an interesting, relevant topic from the secure coding area: a new method, algorithm, implementation technique, optimization of existing solutions, etc. The form of the solution is again a paper describing the work.

As part of the experiment, we have added a new voluntary area - *Bug Bounty challenge*. The challenge expected active participation in the selected bug bounty program and an attempt to find a real vulnerability. A list of suitable BBPs was provided, but self-selection was also allowed to enable students to pick the most technologically suitable BBP. The success or failure was not reflected in the final grade. We evaluated only the process, used tools, and final report. Thus, if one failed to find something specific, it did not compromise the passing of the course in any way. In case of success, correct reporting of the vulnerability according to the processes of the selected BBP was mandatory. Students were repeatedly reminded of the need to strictly follow the competition's rules.

For those interested in ethical hacking, we organized an additional lecture to provide a basic orientation in the field. Delivered by an expert with over five years of experience, the lecture covered the motivations for ethical hacking and its appropriate applications. It introduced the primary methodologies and tools for web, infrastructure, and mobile environments. We also guided students through relevant learning platforms and Bug Bounty Programs (BBPs). Finally, we detailed how to report vulnerabilities, comply with BBP guidelines, and understand the legal framework.

After completing all the preparatory phases, students had 7 weeks to complete the entire project. After submitting and evaluating the entire project, students were contacted and asked to complete a questionnaire. In addition to the questionnaires, the evaluation included an analysis of the submitted project reports.

The course's general learning outcomes are designed so that students can gain knowledge in the chosen security area, acquire the ability to write a professional text and acquire the skills to present professional content. The specific learning outcomes for a BBS-type project are designed to enable students to:

- understand the skills and qualifications to be an ethical hacker,
- demonstrate the knowledge of information gathering, testing, and ethically hacking a system,
- learn about the different tools and techniques that hackers—including ethical hackers—employ,
- use selected tools in hacking,
- understand cybersecurity laws and the consequences for breaking those laws.

The course strengthens the following working life skills: presentation skills, creativity, problem-solving skills, and information and communication technology skills.

6 EXPERIMENT DESIGN

The experiment began with the preparation of an extended project assignment and the preparation of questionnaires. As part of the publication of the assignment, students were informed in advance that this project is part of an experiment and were asked to consent to the processing of information about the experiment, including their use for the final publication.

After we published the assignment, students could attend a bonus lecture and work throughout the project.

To evaluate the experiment, we used a combination of quantitative and qualitative methods. After the project was handed in, we sent a questionnaire to all participants, asking them additional questions. Further information was obtained through a detailed analysis of the submitted project reports, which included a description of the project solution, technical details, approach description, etc. Analysis of the results was carried out by team members who are also responsible for teaching and assessing the course.

As part of the analysis of the project reports, we focused on collecting information relevant to the responsible reporting process. However, we also discuss other interesting observations.

Within the questionnaires, we focused on 3 main areas. The results are intended to provide a better understanding of the impact of each project parameter. The first batch of questions was focused on the actual work done on the project. These questions were intended to discover the student's previous experience, the influence of his/her knowledge on choosing a particular BBP, the vulnerability search strategy, the tools used, the methodology, and many others.

The second area focused on self-evaluation, where we asked students to evaluate their skills and knowledge before and after the project. The last area concerned the students' evaluation of the project, where we were interested in the fun and usefulness of the project, the perception of risks, etc.

Data were collected and processed anonymously based on the consent of the participants in the experiment. According to the university's internal rules, the course supervisor and department head approved the whole experiment.

7 RESULTS

A total of 38 students signed up for the course project, and a total of 31 students successfully completed it (which is comparable to past years). 19 students chose the experimental bug bounty challenge (BBCh). 13 of whom had BBCh assignments successfully completed the project (12 of them filled out the follow-up survey).

We did not observe a significant deviation from other project types in the project evaluation. The average project score of other types was 39 points out of 49 possible. The average BBCh project score was 40 points. Thus, in terms of difficulty, we rate all types of assignments as comparable.

7.1 Student's work on the assignment

In the first step, we were interested in the BBP choice. Five students chose the T-Mobile bug bounty program because they wanted to pursue the program available in their native language. Three chose HackerOne - specifically Shopify's Bug Bounty, Epic Games, and Boozt Fashion AB. Other programs were covered by only one of the students: TryHackMe, HackTheBox, Hacker101, PicoCTF, Coinbase, Moneta, Microsoft, and Hacktrophy.

Table 1: The overview of used learning resources.

Resource	*Times reported*
PortSwigger academy	8
Videos	5
TryHackMe	3
OWASP	2
Online blogs	2
Hacker101	2
School lectures	1
Scientific literature	1

Some students tried to participate in multiple programs. One of the students even used the obtained knowledge and did a back check of his code and found many errors.

The previous experience did not play any significant role in selecting this assignment. On a scale of 1 - 5 (no impact - high impact), the average score was 3.16 with a standard deviation σ of 1.4. In most cases, the reason for selecting a specific BBP was prior experience with and knowledge of the given system. In addition, the majority of the selected BBPs were focused on the security of web applications. Some students even reported that they perceived the web applications as the easiest, thus suitable for beginners.

The selection of BBP based on these factors is expected. Moreover, interacting with a familiar system should make it easier for inexperienced participants to find new vulnerabilities.

In most cases, the strategy for identifying the vulnerabilities involved following checklists, guides, or tutorials (such as OWASP WSTG[8]), testing for known vulnerabilities, or searching for vulnerabilities based on prior personal experience.

Time-wise, most students have spent 15-30 hours with theoretical preparation. Three students have spent less than 10 hours, with a minimum of two. Three students have spent over 30 hours, with a maximum of 80. The vulnerability identification took 15-30 hours for most of the students. Only one student took less than 15 hours, and two more than 30 hours. Thus, most students were within the recommended time frame set for this project (40-60 hours).

The majority of students utilised the materials we suggested for their additional learning. The most used education resource was PortSwigger Academy, as reported by 8 students, closely followed by video tutorials, reported by 5 students. The summary of used learning resources is shown in Table 1.

Students used many tools to tackle the project, including BurpSuite, Wappalyzer, Kali Linux, Nmap, ffuf, Metasploit, and others. This reflects the variability of the BBP, as different implementations of the tested systems require different analytical tools.

We were most interested in seeing if any real vulnerabilities were found. Two of the students have found some vulnerabilities. The first one discovered an Insecure direct object reference (IDOR) vulnerability but did not report it, as the tested platform did not have a proper BBP. The second student discovered two vulnerabilities (CSRF and incorrect validation of redirect link after login) that he reported to the BBP, but at the time of submitting the paper, no response had yet been received. Furthermore, students found a small number of cases of non-standard behaviour, e.g., the existence and availability of non-actual pages, error messages available to users, etc. It is unclear whether or not these instances could be used to mount a successful attack. However, they still violate best practices and should potentially be fixed.

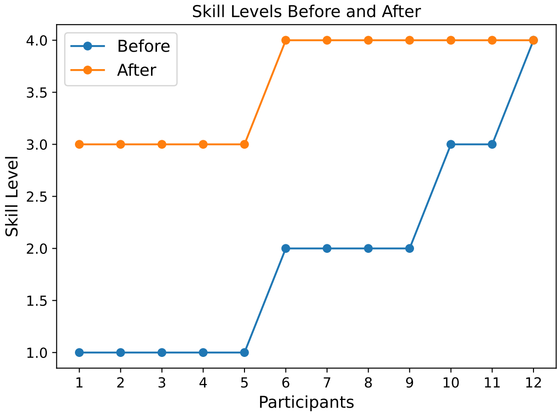


Figure 1: Orientation of participants on the topic of ethical hacking before and after the project. Participants were ordered by skill level before the project.

7.2 Self-evaluation of educational impacts

Before the project, participants in the experiment rated their understanding of ethical hacking as 1.92 on average; this increased to 3.58 at the end. Figure 1 shows that in most cases, the students increased their skill in ethical hacking by two points.

We also assessed participants' prior knowledge of ethical hacking before the project. Nearly half had only theoretical understanding, while the rest had minimal practical experience from another school project. One participant had completed multiple TryHackMe courses and attended sessions on specific vulnerabilities at Burp Suite Academy.

After the end of the project, students reported good orientation in the topic, practical understanding of the different phases of penetration testing, understanding of the attacker's point of view and abilities, and practical knowledge of tools for ethical hacking.

41,7 % of the students reported that they see their future career in cybersecurity before starting the assignment. After completing the assignment, it changed to 50%.

We also wondered whether the students planned to participate in other bug bounty projects after the end of this project. Only two (17%) do not plan to do so, six students (50%) are considering it if there is enough time, and four (33%) firmly plan to continue.

Ultimately, students found the knowledge gained beneficial, even if they pursued careers outside of cybersecurity.

7.3 Self-assessment of the project

Most students rated the experimental project as very beneficial, primarily due to the great practical overlap and the ability to work

[8] https://owasp.org/www-project-web-security-testing-guide/

Table 2: Personal likings of the project as reported by students. The responses ranged from 1 - 5, where 1 means the worst (negative) and 5 means the best (positive). σ denotes the standard deviation.

Question	Mean	σ
How much would you like to see this type of project incorporated into regular teaching?	4	0.85
Is the project content beneficial even if the participant has a career path outside of cybersecurity?	3.91	0.79
How do you feel about working in a real environment on real products?	3.92	0.99
How important is the project's social impact to you (helping to improve real safety)	3.42	1.08

on real problems from practice. They also positively evaluated the relatively high degree of freedom of the project and, paradoxically, the need for self-study to a greater than usual extent, as the project allowed them to get an assessment in an area that interested them. One student summed it up this way: "*If this project is chosen by a person who is interested and fond of this field, it is the most useful and interesting project at the school at all.*"

Next, we asked about the enjoyability of the project as a whole and how it compares to conventional projects with a fixed specification. On a scale of 1 - 5 (not enjoyable at all - it's extremely enjoyable), the average score was 3.81 with a standard deviation σ of 0.87. On a scale of 1 - 5 (significantly more boring than a regular project - significantly more entertaining than a regular project), the average score was 4.16 with a standard deviation σ of 0.72.

When explaining the reasons for the assessment, students primarily mentioned learning new skills, having the freedom to decide how to complete the assignment, and their involvement in trying to solve real-world problems. Despite the high values, almost a third of respondents mentioned time pressure due to the demands of other courses as well as the freedom of the course.

We were also interested in participants' views on questions regarding their personal likings of the project as shown in Table 2.

Students generally rated the project as having a high practicality and educational impact. Negatives included the stress of breaking BBP rules or the frustration of not finding any vulnerabilities.

8 DISCUSSION AND LIMITATIONS

The findings of this study have provided several noteworthy insights regarding the implementation and acceptance of our project-oriented approach to cybersecurity education. Firstly, to our surprise, students could identify vulnerabilities, which unequivocally endorses the efficacy of our pedagogical strategy.

Secondly, the positive feedback received from students about this type of real-world project learning is particularly encouraging. Their enthusiasm underscores the relevance and engagement of hands-on, project-based learning in the conditions of the real world.

Moreover, we consider it of paramount importance that the project was deemed beneficial by those participants who do not intend to pursue a career in cybersecurity. The fact that students outside the field perceive the adversarial perspective as advantageous suggests that the skills and thought processes developed through this project have a wide-reaching impact, extending beyond the immediate domain of cybersecurity. Ethical considerations in teaching hacking techniques within an educational setting have been a topic of some debate. However, various scholarly works have discussed this issue, and our stance aligns with the perspective that such education is legitimate [3]. By introducing IT students to the concept of thinking like an attacker, we foster a critical mindset for developing more secure systems. This cognitive shift is essential for producing programmers who are adept at anticipating and mitigating potential security threats.

Finally, we would like to summarize the main lessons learned. It proved essential to allow freedom in the students' choice of technology by selecting a suitable BBS. Ethical hacking assumes a detailed knowledge of the technology, and it is usually not within the capabilities of the course to deliver this knowledge. Thus, students have to use their existing knowledge, which can be varied and diverse. To reduce the time commitment and increase the chances of success, it is thus advisable to let students work in a familiar environment. We are also very positive about the feasibility of our approach, which was also helped by focusing the assessment on progression rather than finding vulnerabilities. According to the feedback, we are considering how to appropriately integrate existing online courses on ethical hacking, which students widely used to gain detailed knowledge.

We also considered the varying technical skills of students. Some needed more time than the average recommendation due to larger skill gaps, particularly for introductory education. However, this was accommodated by allocating sufficient credits for the project.

8.1 Limitations

We consider the limited dataset as a limitation, as the experiment was only conducted within one run of one course. However, we believe the results are sufficient for the initial opening of the debate and the data-driven presentation of our idea. We are also planning further extensions and repeated runs of experiments.

9 CONCLUSIONS

Based on the positive feedback from students and valuable educational impacts, we plan to include BBP in future courses as well. We plan to shift the orientation towards government agencies, as such agencies often lack the resources to run their own programs or contract ethical hackers. Such a connection would not only be beneficial for students but also increase the security of the public sector.

ACKNOWLEDGMENTS

This work was supported by the Brno University of Technology internal project FIT-S-23-8151.

REFERENCES

[1] 2022. The Urgency of Tackling Europe's Cybersecurity Skills Shortage. https://blogs.microsoft.com/eupolicy/2022/03/23/the-urgency-of-tackling-europes-cybersecurity-skills-shortage/

[2] Tim Greene. 2004. Training Ethical Hackers: Training the Enemy? https://defcon.org/html/links/dc_press/archives/12/ebcvg_training_ethical_hackers.htm

[3] Regina Hartley. 2015. Ethical Hacking Pedagogy: An Analysis and Overview of Teaching Students to Hack. *Journal of International Technology and Information Management* 24 (Jan. 2015), 95–104. https://doi.org/10.58729/1941-6679.1055

[4] Regina Hartley, Dawn Medlin, and Zach Houlik. 2017. Ethical Hacking: Educating Future Cybersecurity Professionals. *Information Systems & Computing Academic Professionals: Proceedings of the EDSIG Conference* (2017), 1–10.

[5] Hideaki Hata, Mingyu Guo, and Muhammad Ali Babar. 2017. Understanding the Heterogeneity of Contributors in Bug Bounty Programs. In *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. Toronto, ON, Canada, 223–228. https://doi.org/10.1109/ESEM.2017.34

[6] Patricia Y. Logan and Allen Clarkson. 2005. Teaching students to hack: curriculum issues in information security. *ACM SIGCSE Bulletin* 37, 1 (Feb. 2005), 157–161. https://doi.org/10.1145/1047124.1047405

[7] Suresh S. Malladi and Hemang C. Subramanian. 2020. Bug Bounty Programs for Cybersecurity: Practices, Issues, and Recommendations. *IEEE Software* 37, 1 (Jan. 2020), 31–39. https://doi.org/10.1109/MS.2018.2880508 Conference Name: IEEE Software.

[8] Brian A. Pashel. 2006. Teaching students to hack: ethical implications in teaching students to hack at the university level. In *Proceedings of the 3rd annual conference on Information security curriculum development*. ACM, Kennesaw Georgia, 197–200. https://doi.org/10.1145/1231047.1231088

[9] Nicole Radziwill, Jessica Romano, Diane Shorter, and Morgan Benton. 2015. The Ethics of Hacking: Should It Be Taught? *Software Quality Professional* 18, 1 (Dec. 2015). https://arxiv.org/abs/1512.02707

[10] Jacob Riggs. 2021. I hacked the Dutch government and all I got was this t-shirt. https://jacobriggs.io/blog/posts/i-hacked-the-dutch-government-and-all-i-got-was-this-t-shirt-24.html

[11] Jukka Ruohonen and Luca Allodi. 2018. A Bug Bounty Perspective on the Disclosure of Web Vulnerabilities. *17th Annual Workshop on the Economics of Information Security, Innsbruck* (May 2018). http://arxiv.org/abs/1805.09850 arXiv: 1805.09850.

[12] Zouheir Trabelsi. 2011. Hands-on lab exercises implementation of DoS and MiM attacks using ARP cache poisoning. In *Proceedings of the 2011 Information Security Curriculum Development Conference*. ACM, Kennesaw Georgia, 74–83. https://doi.org/10.1145/2047456.2047468

[13] Zouheir Trabelsi. 2012. Switch's CAM table poisoning attack. In *Computing Education 2012 - Proceedings of the 14th Australasian Computing Education Conference (Conferences in Research and Practice in Information Technology Series)*, Michael de Raadt and Angela Carbone (Eds.). Australian Computer Society, Melbourne, Australia, 113–120. http://www.scopus.com/inward/record.url?scp=85014905333&partnerID=8YFLogxK Publisher: Australian Computer Society.

[14] Zouheir Trabelsi. 2014. Enhancing the comprehension of network sniffing attack in information security education using a hands-on lab approach. In *Proceedings of the 15th Annual Conference on Information technology education*. ACM, Atlanta Georgia USA, 39–44. https://doi.org/10.1145/2656450.2656462

[15] Zouheir Trabelsi and Latifa Alketbi. 2013. Using network packet generators and snort rules for teaching denial of service attacks. In *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*. ACM, Canterbury England, UK, 285–290. https://doi.org/10.1145/2462476.2465580

[16] Zouheir Trabelsi and Margaret McCoey. 2016. Ethical Hacking in Information Security Curricula. *International Journal of Information and Communication Technology Education* 12, 1 (Jan. 2016), 1–10. https://doi.org/10.4018/IJICTE.2016010101

[17] Thomas Walshe and Andrew Simpson. 2020. An Empirical Study of Bug Bounty Programs. In *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*. London, ON, Canada, 35–44. https://doi.org/10.1109/IBF50092.2020.9034828

[18] Marleen Weulen Kranenbarg, Thomas J. Holt, and Jeroen van der Ham. 2018. Don't shoot the messenger! A criminological and computer science perspective on coordinated vulnerability disclosure. *Crime Science* 7, 1 (Dec. 2018), 16. https://doi.org/10.1186/s40163-018-0090-8

[19] Kim Zetter. 2001. Information Security News: Three Minutes With Microsoft's Scott Culp. https://seclists.org/isn/2001/Oct/37

[20] Uldis Ķinis. 2018. From Responsible Disclosure Policy (RDP) towards State Regulated Responsible Vulnerability Disclosure Procedure (RVDP): The Latvian approach. *Computer Law & Security Review* 34, 3 (June 2018), 508–522. https://doi.org/10.1016/j.clsr.2017.11.003

EURASIP Journal on Image
and Video Processing

RESEARCH Open Access

Comprehensive multiparametric analysis of human deepfake speech recognition

Kamil Malinka[1], Anton Firc[1*], Milan Šalko[1], Daniel Prudký[1], Karolína Radačovská[1] and Petr Hanáček[1]

*Correspondence:
ifirc@fit.vut.cz

[1] Faculty of Information Technology, Brno University of Technology, Božetěchova 2, 61200 Brno, CZ, Czech Republic

Abstract

In this paper, we undertake a novel two-pronged investigation into the human recognition of deepfake speech, addressing critical gaps in existing research. First, we pioneer an evaluation of the impact of prior information on deepfake recognition, setting our work apart by simulating real-world attack scenarios where individuals are not informed in advance of deepfake exposure. This approach simulates the unpredictability of real-world deepfake attacks, providing unprecedented insights into human vulnerability under realistic conditions. Second, we introduce a novel metric to evaluate the quality of deepfake audio. This metric facilitates a deeper exploration into how the quality of deepfake speech influences human detection accuracy. By examining both the effect of prior knowledge about deepfakes and the role of deepfake speech quality, our research reveals the importance of these factors, contributes to understanding human vulnerability to deepfakes, and suggests measures to enhance human detection skills.

Keywords: Deepfake, Synthetic speech, Deepfake detection, Human perception, Speech quality, Cybersecurity

1 Introduction

Deepfakes are digitally manipulated media, typically video or audio recordings, created using advanced artificial intelligence and machine learning techniques. These technologies allow for the alteration or synthesis of human likenesses and voices, making it possible to generate convincingly realistic content that portrays individuals saying or doing things they never actually did [1].

Deepfake technology creates an entirely new threat landscape in IT security. Recent studies show that face and voice biometrics systems are vulnerable to deepfake spoofing attacks [2, 3]. These vulnerabilities motivate the development of protection techniques, such as deepfake detectors [4, 5, 47].

Moreover, the increasing number of deepfake-related headlines in the news documents this technology's malicious impacts on us—humans [6–12]. One of the very recent cases involves the theft of $ 25 million [6]. During a video conference, a finance worker at a multinational firm in Hong Kong was deceived into transferring company funds to scammers using deepfake technology to impersonate the company's CFO. The scam involved deepfake renderings of several staff members. Despite initial suspicions raised by an unusual message from the supposed CFO about a secret transaction, the worker was convinced by the realistic appearance and voices of the colleagues in the video call.

Motivated by the increasing frequency of deepfake attacks on individuals, our research evaluates how well humans can recognise deepfake speech. Previous studies [13, 14] only involved participants who were aware that they would be exposed to deepfakes, with the explicit task of distinguishing between genuine and deepfake speech. This scenario, however, is not representative of real-world situations where individuals are unexpectedly confronted with deepfakes, in critical moments when their ability to detect these spoofs is vital. The key difference in real-world attacks is the absence of forewarning; targets are not pre-alerted to scrutinise the authenticity of the media they encounter. This study addresses this gap by simulating authentic conditions testing individuals' ability to recognise deepfakes without prior notice of exposure.

In addition, existing research has not fully explored how the quality of deepfake speech affects detection capabilities. The former studies only record success or failure to detect the utterance but omit quality information. To fill this gap, we designed a second experiment focusing on the role of speech quality in human deepfake recognition. To complement the obtained knowledge, we also investigate additional factors such as language, sex, or playback devices (speakers, headphones). Exploration of these additional factors is essential to set a baseline for detection and to guide further education about deepfakes better.

In our first experiment, participants were unknowingly exposed to deepfake audio during a "*Two Truths One Lie*" game involving voice messages about countries, one of which was synthetically generated. This setup tested their capacity to spot the deepfake without any prior indication of its presence. Afterwards, a questionnaire unveiled the experiment's real intent and inquired about their detection ability before and after learning about the deepfake, thus comparing their detection skills with and without prior knowledge.

In our second experiment, we investigated whether the quality of deepfake speech notably affects the ability to recognise deepfakes. For this purpose, we created a novel quality metric for deepfake speech, used it to categorise deepfake audio, and then conducted a survey to see how well individuals could differentiate between authentic and deepfake speech, focusing on how speech quality influences their judgments. This approach allowed us to probe for a possible quality threshold at which deepfakes become undetectable to the human ear, as the trends in speech synthesis clearly show continual increases in the quality of the synthesised speech [1].

In addition to the ability to recognise, we map the public awareness of deepfake technology. We ask if respondents ever encountered deepfakes and where. We use this knowledge to better understand public perception of deepfakes and examine a link between public awareness and deepfake recognition accuracy.

The ultimate goal of this paper is to understand the impact of AI-based attacks and scams on humans. This understanding helps to design and employ proper protection mechanisms. Unfortunately, as we demonstrate, the outcomes of the former research do not provide the complete picture of the area, where most of the results claim that the human ability to recognise deepfakes ranges around 70–80%. As our results show, it is essential to consider the impact of languages, demographics or playback devices.

This study thus evaluates how the prior information and quality of deepfake speech influence the human recognition of deepfakes, and it extends our previously published work [15]. Our research protocol was presented to our institution's alternative to an ethics board, and we were advised that no further actions were necessitated on our part.

Contributions

The main contributions of this paper can be summarised as follows:

- We assess the human ability to recognise deepfakes in the Czech and Slovak languages.
- We show that the human ability to recognise deepfakes is affected by the prior information of deepfake exposure and the quality of deepfake recordings.
- We explore the impact of the gender of both the speaker and the listener, the language used, and the playback device on the ability of humans to recognise deepfake recordings.
- We propose a quality measurement for deepfake speech.
- We discuss possible measures to strengthen the human ability to recognise deepfakes.

2 Related work

Related work may be split into two distinct areas: recognition of faces (image and video) and recognition of speech (audio).

Audio

Using unary and binary selection methods, Mai et al. [14] explored speech deepfake detection among 529 participants across English and Mandarin. Their findings revealed a 73% accuracy rate in identifying deepfake audio without a significant difference between languages, showing minimal improvement in detection through awareness efforts.

Wang et al. [16] examined the human ability to distinguish between human and synthetic speech in a simulated commercial bank scenario. Participants evaluated utterances across three categories (bonafide, irrelevant, deepfake) and assigned confidence scores. This study demonstrated a reasonable capability to recognise deepfakes, although exact success rates were not specified.

Müller et al. [13] focused on comparing human and AI detection of voice deepfakes using a game-based approach and the ASVspoof 2019 dataset. They reported an 80% success rate in human detection, noting better performance against TTS-generated deepfakes, particularly among native speakers, with rapid learning observed initially but stabilising at 80% success.

Table 1 Comparison of experiments on the human ability to recognise audio deepfakes

Study	Year	Prior information	Respondents	Accuracy [%]
Wang et al. [16]	2020	Yes	1145	N/A
Watson et al. [17]	2021	Yes	53	42–90
Müller et al. [13]	2022	Yes	410	80
Mai et al. [14]	2023	Yes	529	73
Ours	2024	Yes	85	67–94
Ours	2024	*No*	31	3.20

Watson et al. [17] investigated audio deepfake perception among college students, focusing on English speakers and the impact of grammar complexity. Their study found no significant difference in detection accuracy between senior and junior students, with a varying accuracy of 42% to 90% across different tasks, indicating that complex and shorter sentences were more likely to be identified as synthetic.

Image and video

Studies on deepfake detection reveal varying success rates based on image or video quality, with images achieving 58–70% accuracy and videos as low as 20% for high-quality deepfakes, increasing to over 80% for lower quality ones [18–23]. Training programs have improved detection rates by 33% [23], indicating an average success rate of 60–65%.

Research by M. Groh et al. [24] on recognising deepfake political speeches showed enhanced detection when participants were familiar with the content or speaker's voice. Jilani et al. [25] found that novices could outperform experts in identifying deepfake videos, highlighting the challenge deepfakes pose to forensic analysis.

Bray et al. [26] evaluated human capability to distinguish StyleGAN2 deepfakes, with participants' accuracy around 62%, barely above chance, despite interventions. Similarly, Somoray et al. [27]'s study saw an average detection accuracy of 60.70% without significant improvement from training on visual cues.

Mohammad et al. [28] investigated whether exposure to deepfake videos could enhance detection skills, suggesting potential for awareness to combat deepfake challenges.

Summary

In previous studies, participants were aware they were interacting with deepfakes, which could have influenced their responses. As highlighted in Table 1, our research diverges significantly in this aspect. A key distinction of this study is that it was conducted in Czech and Slovak languages. In addition, we explore how the quality of the deepfake audio, the gender of both the speaker and the listener, and the language used affect the ability of humans to identify deepfake recordings.

3 Experiment design

This study builds on previous research regarding the human ability to recognize deepfake speech. Unlike earlier studies, which informed respondents about deepfakes before testing their recognition skills, we chose not to notify respondents about their exposure to deepfakes. This approach aims to replicate real-world scenarios where such attacks occur without prior warning. In vishing attacks, victims are not pre-informed that an attack is underway or that they should scrutinize the speech for deepfakes. In addition, it remains uncertain how the quality of deepfake speech impacts the human ability to detect it.

The experimental part, thus, consists of two parts. The first part evaluates the human ability to recognise deepfakes in an ordinary conversation (without prior information). The second part examines how the quality of deepfake speech influences the human ability to recognise deepfakes (with prior information). The experiments thus aim to bring new knowledge on the influence of the prior information and quality of deepfake recordings on the human ability to recognise deepfakes and to validate that the results of the former studies are still relevant.

3.1 Experiment one: influence of the prior information

The design of the experiment is inspired by Matyáš et al. [29], who propose using a cover story to hide the true nature of an experiment. Unlike other works, respondents do not know their deepfake detection abilities are being tested. Thus, our goal is to create a realistic attack scenario in which we change a real voice, which respondents know and do not consider suspicious, to a deepfake and try to see if they notice this change.

The experiment took place in the Czech Republic, and as a result, all interactions were conducted in Czech. This included the creation of deepfake speech in the Czech language. Given that most models and tools are designed for English, our work demonstrates the potential for adapting speech synthesis models to other languages. This adaptation necessitates tailored approaches for both training and utilising these models.

The whole experiment is hidden behind a cover story of testing the usability of voice messaging. This approach helps to obscure the true objective of the study, thereby reducing potential bias in respondent behaviour. Respondents play the game *Two Truths One Lie*. They receive five voice messages from the narrator, each containing three facts about a selected country. One of these facts is incorrect, and the respondent's task is to identify the incorrect fact and report it back (using the voice message). This setup simulates communication using voice messages only.

The usage of a cover story shifts the focus of the respondents from carefully examining the recordings to their *normal* mode of operation, where the primary focus is given towards the communication and its content rather than scrutinising the technical aspects of the voice messages. By engaging respondents in a familiar and straightforward game, the cover story encourages natural interaction, ensuring that any observations or feedback provided reflect genuine reactions rather than responses influenced by an awareness of the study's true purpose. In addition, the interactive nature of the game maintains the respondents' engagement and helps to gather more reliable data on their communication patterns and their ability to detect anomalies in the voice messages.

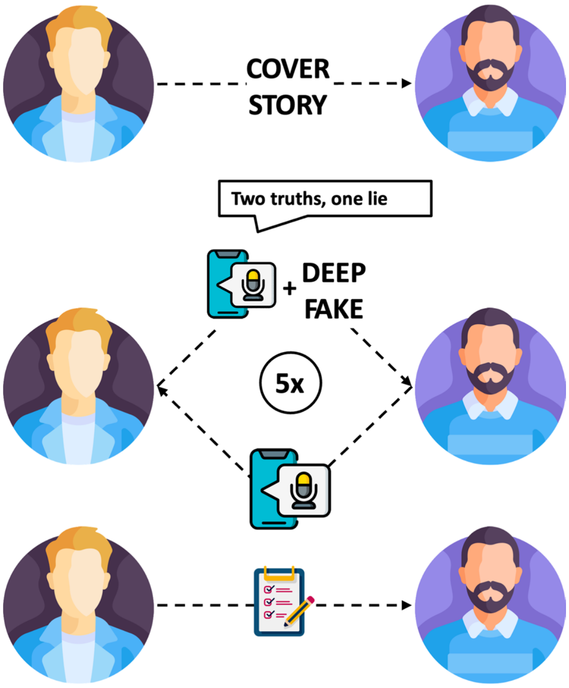


Fig. 1 Flowchart describing the course of the experiment

One of these sets was pre-prepared as a deepfake recording of the narrator's voice. At the end of the experiment, each respondent was sent a questionnaire asking about their knowledge of and attitude towards deepfakes, if they observed anything unusual during the conversation, and ultimately revealed the true nature of the experiment and asked if they could now identify the deepfake set. The flow of the experiment is visualised in Fig. 1. The work described in this experiment results from a previously completed bachelor‘s thesis [30].

3.1.1 Research questions

For the first experiment, we have identified three main research questions:

RQ1: Are humans able to identify deepfake recording during casual conversation?

RQ2: Are humans able to detect a deepfake recording among genuine ones?

RQ3: What is people's awareness of deepfake technology?

3.1.2 Round setup

The experiment was hidden behind a cover story. Participants were presented with simple facts about countries in the form of the *Two Truths One Lie* game. All communication took place within the WhatsApp chat, using voice messages.

Each conversation begins with a brief introduction presenting the pre-prepared cover story, explaining the rules of the experiment, explaining the rules of the game and reminding the respondents that whenever they encounter anything unordinary, they should report it. This is important for our experiment because we need them to report any concerns (mainly about the deepfake set). It is also crucial for us to get used to the narrator's voice and to listen to it. We then gradually send them voice messages containing the sets of facts for the game. The sets include simple facts about world countries, such as:

> *Set: India*
>
> 1. *India is the second most populous country in the world.*
> 2. *The capital is Mumbai.*
> 3. *The most widespread religion is Hinduism.*
>
> **The second fact is a lie. The correct version is: The capital is New Delhi.*

The respondents listened to these sets and replied with voice messages as well. This way, we send five sets (voice messages), including one pre-prepared deepfake set. We refer respondents directly to the questionnaire if they raise any suspicions or questions about the deepfake set. Otherwise, after completing all five sets, we send the respondent a link to the final questionnaire to complete. This questionnaire first collects information about the attitude and knowledge of deepfakes and whether the respondent noticed anything unusual during the experiment (detected the deepfake set). Finally, the questionnaire discloses the true nature of the experiment and that one of the sets is a deepfake and asks the respondents to identify it. The final questionnaire was carefully designed not to reveal the true nature of our experiments in advance, as described in subsubsection 3.1.4.

3.1.3 Synthesizing deepfake set

To synthesise the deepfake set, we use YourTTS [31] tool with provided pretrained models in the voice conversion setting. This decision was motivated mainly by the ease of use and satisfactory multilingual capabilities of the pretrained models. The conversion has been done in a challenging female-to-male setting.[1] After synthesis, we improved the quality of the deepfake set using post-processing. We removed the noise added during creation using *Noise Reducer*[2] tool and smoothed out the frayed phonemes by cutting out the part of the recording where the phonemes resonated using *Audacity*.[3] We also adjusted the pitch of the voice. The test run revealed a significant difference in background noise between bonafide (directly spoken) and deepfake (played by speakers) utterances. To diminish this difference and force the participants to focus on the spoken content instead of the background noise, we played brown noise as the background for all the utterances.

Quality evaluation

The evaluation is inspired by the *Mean Opinion Score (MOS)* subjective listening test method described by Loizou [32]. We played the recording to 12 experts working with deepfakes regularly. Therefore, we expect their knowledge about the qualities of deepfake recordings. Each expert rated the quality on a scale of 1 (poor) to 5 (excellent). The

[1] The source speech was female, the target voice was male, resulting in male deepfake speech.

[2] https://noisereducer.media.io/speech-enhancement.

[3] https://www.audacityteam.org.

final mean score was 3.0; therefore, the recording qualitatively corresponds to the rating "*Fair*".

3.1.4 Questionnaire

In designing the survey, it was crucial to disguise our experiment with a cover story to prevent the sequence of questions from influencing subsequent responses. We aimed to gradually lead up to the most critical questions, ensuring the survey, which comes after the experiment, was not overly lengthy. Consequently, we organised the survey into six distinct sections:

1. *Respondent Profile:* This section gathers basic personal information from participants, such as age, sex, professional field, and contact number. The contact number is used to verify the authenticity of the responses related to the experiment.
2. *Usability:* To avoid directly addressing deepfakes at the beginning, we chose a preliminary question regarding the usability of voice messages, which could be relevant for assessment purposes.
3. *Recordings:* Participants were asked about their impressions of the recordings, specifically if they noticed anything unusual or unnatural, and if so, what it was. This question is critical for our research.
4. *Deepfakes:* At this juncture, we introduced the concept of deepfakes to participants, inquiring if they had previously encountered them and in which contexts. We also assessed their confidence in identifying a deepfake, referencing research on Americans' ability to recognise computer-generated voices pretending to be human [33].
5. *Real Experiment:* We disclosed the full details of our experiment here, unveiled the cover story, and acknowledged sending a deepfake during our interaction. We then checked if participants could identify the deepfakes, knowing at least one was included.
6. *Conclusion:* In the final section, we disclosed which recording was inauthentic and gauged participants' reactions to the quality of the voice deepfakes. We also evaluated whether their confidence in recognising deepfakes changed after this experience and the revelation of the experiment's true purpose.

At the survey's conclusion, we provided links for participants to learn more about deepfakes. Supplementary material contains a comprehensive list of all survey questions.

3.2 Experiment two: influence of deepfake speech quality

The second experiment investigates how the quality of deepfake recordings affects people's ability to identify them. Similar to the first experiment, the tests are conducted in Czech and Slovak. These Slavic languages sound very similar but differ in grammar and pronunciation. They are mutually intelligible, meaning that a speaker of one language can understand the other without studying it. The participants will be asked to recognise deepfakes in these languages. In addition, each deepfake recording will be given a quality score, which will later be used to determine if there is a threshold above which it is no longer possible to identify deepfakes correctly.

The study was conducted via an online survey, which gathered demographic information from the participants. Following this, participants were presented with pairs of audio recordings for evaluation. Each pair contained a genuine audio sample and its corresponding deepfake version, featuring the same speaker delivering the same content. To ensure a diverse and inclusive dataset, we randomly assigned 14 recording pairs to each participant. We carefully balanced the representation of male and female speakers across the two languages featured to cover all pairs from the created dataset.

In addition, the order in which these pairs were presented was randomised to mitigate potential bias. This approach was critical as we anticipated that not all participants would complete the survey in its entirety; by randomising the sequence, we aimed to prevent the latter pairs from being disproportionately overlooked. The task for participants was straightforward: identify the deepfake recording in each pair.

The demographic focus was on young individuals, particularly students and those heavily involved with technology and social media. This group's familiarity with digital media, including potential exposure to deepfake content, suggests a higher proficiency in recognising deepfakes than older generations, making them the experiment's primary audience.

For data analysis, we applied the Student's paired t test, suitable for our data's normal distribution pattern, with a significance level set at $\alpha = 0.05$. Jamovi[4] was used for this analysis to validate our research questions and hypotheses.

The work described in this experiment results from a previously completed bachelor‘s thesis [34].

3.2.1 Hypotheses and research questions

For the second experiment, formulated the following hypotheses:

H1: Women are more likely to detect voice deepfakes than men.

H2: Women, compared to men, are more likely to detect deepfakes spoken by women.

H3: Men, compared to women, are more likely to detect deepfakes spoken by men.

H4: People are likelier to detect deepfakes in their native language.

H5: Headphones increase the human capability to detect deepfakes compared to device speakers.

H6: People who are aware of deepfakes are more likely to detect them than people who have never heard of deepfakes.

H7: People who believe they can detect deepfakes are likelier to detect deepfakes than people without this belief.

In addition to the hypotheses, we formulated the following research questions:

RQ4: Is there a threshold in the deepfake quality rating score beyond which it is no longer possible to recognise deepfakes?

RQ5: Are people more likely to detect deepfakes with the lower score assigned using the proposed quality rating system?

RQ6: Are people able to detect voice deepfakes?

RQ7: How many people with previous knowledge of deepfakes can recognise deepfakes?

RQ8: Does the audio device impact the human ability to recognise deepfakes?

[4] https://www.jamovi.org/.

3.2.2 Speech quality measurement

To investigate the relationship between the quality of deepfake recordings and the human ability to detect them, a system is necessary to measure the quality of these recordings. Since no existing system meets this need, we have undertaken the task of developing one. We approach the quality assessment from the attacker's point of view. The hallmark of an ideal deepfake speech recording for a potential attacker is that it perfectly mimics the voice of the person being imitated, is free from any background noise or artefacts, and delivers clear and easily understood content. With these criteria in mind, we have designed a quality measurement system for deepfake speech that evaluates recordings based on three key factors:

Speaker Similarity of the speaker in deepfake recording with the recording (voiceprint) of the imitated speaker is calculated using the Phonexia Voice Biometrics.[5] The system creates a voiceprint for each user, and the verification is done by comparing at least seven seconds of speech to this voiceprint. The similarity of speakers is expressed as log-likelihood ratio (LLR).

The Perceptual Evaluation of Speech Quality (PESQ) is a measurement designed to predict the mean opinion score (MOS)—the people's subjective opinions of synthetic audio samples. PESQ is the objective quality measure recommended by ITU-T for speech quality of narrow-band telephone networks and speech codecs [32] implementation is available online, as published by Wang et al. [35]. The result PESQ score represents the MOS–LQO, which stands for Mean Opinion Score–Listening Quality Objective. It combines the objective measurements of various parameters (e.g., delay, packet loss) and subjective listening tests to model the relationship between the objective parameters and the perceived quality of the audio. The values lie within the range of 1.0 and 5.0; the higher the score, the better the quality.

Finally, *Mel Cepstral Distortion* (MCD) is a widely used measure to differentiate two mel cepstral coefficient sequences. It is often used in speech synthesis systems to assess speech quality. The smaller result means less distortion between the signals and a better match [36]. Implementation[6] initial step involves generating mel cepstral coefficients (MCCs), a process tailored to the project's specific requirements. This project adopted an approach that necessitates the creation of *.mgc* files due to the original implementation's inability to directly process waveform audio for feature extraction. The *.mgc* files store pre-extracted acoustic features, including the MCCs, with additional support from external helper repositories for *.mgc* file generation.[78] The extraction of these coefficients is performed using the World Vocoder [37]. The fundamental frequency is then identified, logarithmically scaled, and transformed into *.mgc* format via the Speech Signal Processing Toolkit (SPTK).[9] The resulting *.mgc* files, enriched with MCCs, are prepared for subsequent MCD computation. Using the Dynamic Time Warping technique, the MCD calculation is enhanced to account for potential timing discrepancies between

[5] https://www.phonexia.com/product/voice-biometrics/.

[6] https://github.com/MattShannon/mcd.

[7] https://github.com/Lukelluke/MCD-MEL-CEPSTRAL-DISTANCE-MCD-application.

[8] https://github.com/CSTR-Edinburgh/merlin.

[9] https://sp-tk.sourceforge.net/.

Table 2 Table of quality ranges in each cluster

Cluster	Range [%]
1	[20.05, 34.67]
2	[38.29, 52.58]
3	[53.08, 67.77]
4	[72.48, 84.81]

The numbers are rounded to two decimal points. The clusters are left as defined by the clustering algorithm, resulting in gaps between the intervals

sequences, ensuring accurate alignment. The desired outcome of MCD values falls within the 4.0–8.0 range, indicative of the quality of speech synthesis.

Computing final quality

The numerical values of these factors were adjusted to fit within a range of 0 to 1 using min–max normalisation. Typically, we would consider the proposed metrics equally important when evaluating the overall quality. However, PESQ assesses speech quality based on how listeners perceive it, whereas MCD measures how similar two recordings are. In the context of deepfakes, exact similarity to the original (bonafide) recording is less critical for a deepfake to be effective in deceptive scenarios. Therefore, we adjusted the significance of these metrics, reducing the MCD's weight in our evaluation.

The rationale behind the chosen weights is based on this study's specific context and objectives. PESQ and Speaker Similarity were each given a significant weight (40%) because the perceptual quality of speech and the resemblance to the target speaker's voice are crucial for producing convincing and natural-sounding deepfake speech. MCD was assigned a lower weight (20%) as the primary goal is to create a convincing imitation rather than a replica.

The formula used to calculate the quality of deepfake speech is as follows:

$$Q_s = 0.4 * SpeakerSimilarity + 0.4 * PESQ + 0.2 * MCD$$

The final quality score Q_s lies between 0 and 100%. Higher values signalise better quality of deepfake speech. Finally, the parametrisation (weights) may be changed to better suit different use cases. For instance, in applications where exact similarity to the original recording is more critical, the weight for MCD can be increased accordingly. This flexibility ensures that our approach remains generalisable and adaptable to various contexts, maintaining relevance to the specific objectives of different research or practical scenarios.

3.2.3 Data set

A custom data set has been created for this experiment, as no publicly available deepfake datasets contain paired recordings (bonafide–deepfake) with the same content in the Czech or Slovak language. The dataset thus contains pairs of audio clips containing bonafide and deepfake voices. These audio clip pairs are spoken by the same speaker, meaning the deepfake's target voice is the voice from the bonafide clip. The bonafide audio

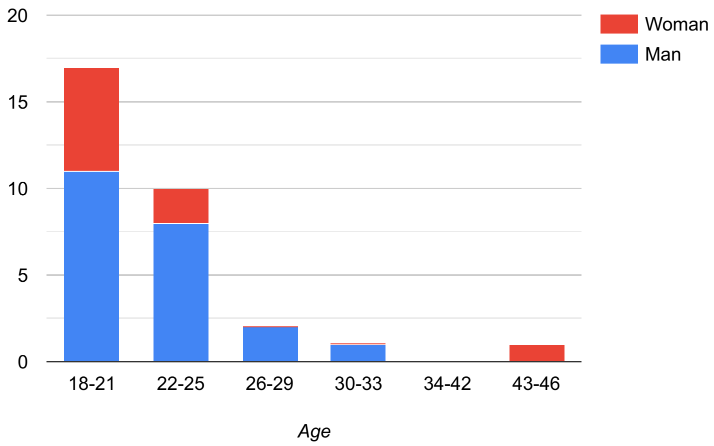


Fig. 2 Age of respondents with a look at the gender ratio in five age groups

clips are taken from the Common Voice Corpus [38] version 12.0.[10] We chose Common Voice because it provides a broad range of audio samples in many languages, including Slovak and Czech, which are essential for this project. These original recordings had to be concatenated to fulfil Phonexia Voice Biometrics' requirements about the length of the audio samples (min. 15 s for enrollment and 7 s for verification). The minimum length of pure speech contained in one enrollment recording was 15 s. All samples were thus gradually concatenated with the following ones to fulfil this requirement. These concatenated original clips were used as input for the voice conversion method to create their deepfake pair. We used Coqui deep learning toolkit[11] with custom YourTTS [31] models for Czech and Slovak languages[12] trained using the Common Voice corpus version 12.0. The resulting deepfake samples have a lot of noise and distortions; however, this is intentional as we need to introduce a quality system rating, dividing the dataset into several groups of recordings sorted according to their assigned quality.

Recordings were assigned quality using the proposed quality measurement (subsubsection 3.3.1) and sorted into quality groups using the k-means clustering algorithm.[13] We chose a one-dimensional array k-means input to sort the recordings into four groups. The quality score ranges of the clusters are displayed in Table 2. The rationale for clustering the recordings into four groups was based on the distribution of the quality scores. The quality scores were not evenly distributed, making it challenging to manually define clear and distinct ranges. To achieve the best possible separation and ensure each group represented a distinct quality level, we utilised k-means clustering. This method provided a more data-driven and objective approach to categorising the recordings into meaningful quality groups.

The final data set consists of twelve speakers.[14] They are divided into six Slovak speakers and six Czech speakers; for each language, there are three male and three female

[10] https://commonvoice.mozilla.org/sk/datasets.

[11] https://github.com/coqui-ai/TTS.

[12] Download links in the Declarations section.

[13] https://pypi.org/project/kmeans1d/.

[14] Download links in the Declarations section.

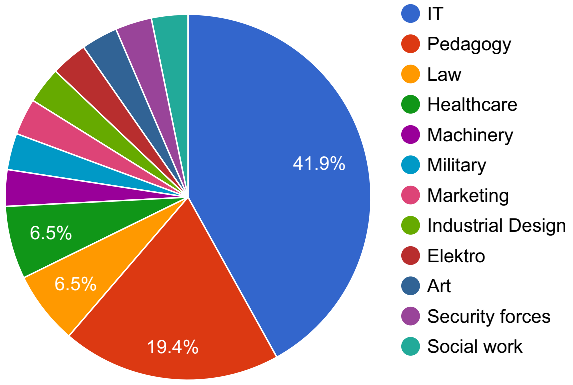


Fig. 3 Proportions of fields in which respondents work

Table 3 RQ1 summary

Reaction during conversation	
Reacted	0%
Described unnatural things from the conversation	
Poorer audio quality	41.90%
Deepfake sign	3.20%

speakers. Each language includes three women and three men. Every cluster has its text file with a table representing every file in the group, its particular quality measure evaluations, and the final score.

4 Experiments and results

Following the experiment design, we executed both experiments with different participant groups.

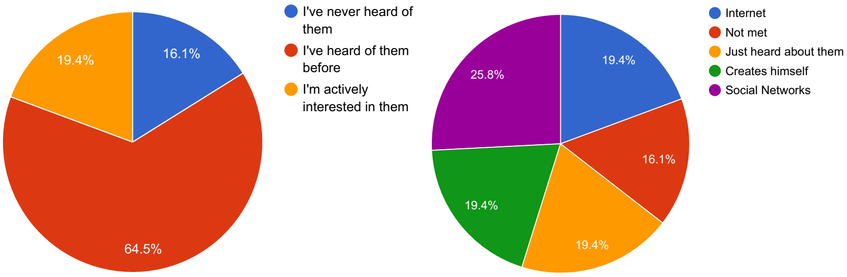
4.1 Experiment one: influence of the prior information

During the first experiment, we collected 31 responses. In terms of sex, 71% of respondents were male and 29% were female. The age of the respondents ranges from 18 to 46, but 80% of the values are less or equal to 23, and the average age is about 22.39 years, as shown in Fig. 2. In focus on the field of work, IT has the highest representation, with 41.90% of respondents. The following common field is education with 19.40%, law and healthcare with 6.50%, and other fields like machinery, marketing, military, art, etc., as shown in Fig. 3.

Participants were recruited through a convenience sampling method, whereby we randomly selected individuals from our personal and professional networks. We approached and invited a larger pool of individuals to participate, but only a subset of them chose to take part in the study. This method ensured a diverse but accessible pool of respondents, leveraging existing contacts to gather a broad spectrum of data efficiently.

Table 4 RQ2 summary

Identify deepfake set	
Marked	96.80%
Correctly identify	83.90%
Justification for identification	
Different from the others	54.80%
Lower quality than others	29%
Deepfake sign	22.60%



(a) Proportion of deepfake knowledge groups. (b) Proportion of deepfake knowledge sources.

Fig. 4 Awareness of deepfake technology of the participants

Table 5 RQ3 summary

Heard of deepfakes	
Heard of them	64.50%
Actively interested	19.40%
Never heard of them	16.10%
Where they heard about them	
Social media	25.80%
Internet	19.40%
Not specify	19.40%
Create them themselves	19.40%
Never heard of them	16.10%

All of the research questions have been answered:

RQ1: Are humans able to identify deepfake recording during casual conversation?

No one reacted to the deepfake at all during the conversation. One respondent even asked to repeat this set, yet he continued and answered the question as the others did without noticing.

Only one respondent mentioned anything specific about deepfakes before the true nature of the experiment was revealed. This gives us a deepfake detection success rate of 3.20%. 13 respondents mentioned a lower quality of this recording; however, we cannot consider this a successful identification of the deepfake set.

Finally, a third of the respondents told us after the experiment or in their text responses in the questionnaire that the possibility of a fraudulent recording did not occur to them during the interview, and they focused primarily on the content and the correct answer,

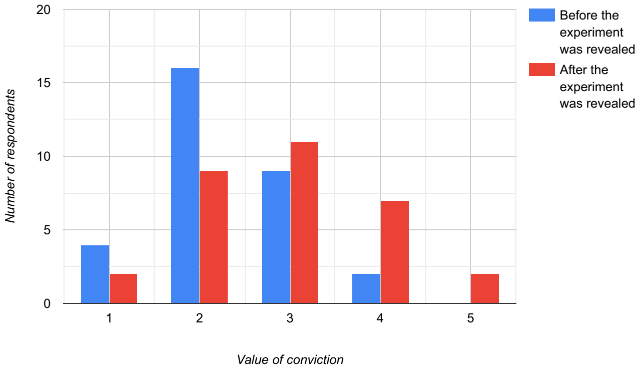


Fig. 5 Responses to the question of how confident respondents are in detecting a deepfake, quantified by a number of respondents

stating that they considered the lower quality to be expected. These results are summarised in Table 3.

RQ2: Are humans able to detect a deepfake recording among genuine ones?

After revealing that one of the sets is a deepfake, 83.90% of all respondents correctly identified this set. Respondents who marked the deepfake set and other options are not counted as successful. Counting these responses as successful would result in 96.80% respondents identifying the deepfake set. Five out of the respondents (23.80%) incorrectly identified at least one genuine (bonafide) audio set as a deepfake. In addition, the only participant who did not identify the actual deepfake set incorrectly labelled the bonafide set as a deepfake.

54.80% of respondents justify selecting the deepfake set because it was different to others. The second most-stated reason was the lower quality compared to bonafide recordings, as mentioned by 29% of respondents. Finally, the third most-stated reason is the presence of typical deepfake artefacts, mentioned by 22.60% of respondents. These artefacts included slight distortion and glitches in the last word of the sentence. Some respondents gave a combination of stated reasons. These results are summarised in Table 4.

RQ3: What is people's awareness of deepfake technology?

Respondents had a choice of three options: 16.10% of respondents answered, "*I've never heard of deepfakes*", 64.50% answered, "*I've heard of deepfakes before*", and 19.40% answered, "*I'm actively interested in deepfakes*" as shown in Fig. 4a. Where they heard about deepfakes is variable but can still be classified into several groups, and more than a quarter of people (25.80%) said that they heard about deepfakes on social media, mainly in some informative videos, articles, etc. One respondent said they had encountered deepfake videos of politicians on TikTok. Consistently, 19.40% of people wrote that they heard about them on the internet, nothing more specific, or that they heard about them and did not specify where or tried to create them themselves, which were mainly people in the IT environment. The reported sources of deepfake awareness are shown in Fig. 4b. In summary, 83.90% of the participants have at least heard of deepfakes, mainly from social media and informative videos. The responses are detailed in Table 5.

Table 6 Results on confirmed hypotheses

Hypothesis		Mean [%]	Median [%]	SD [%]	*p*-value	Effect Size
H3	Men	93.70	94.50	4.98	< 0.001	2.08
	Women	78.90	78.70	5.76		
H5	Headphones	91.50	92.10	3.82	< 0.001	1.79
	Speakers	81.40	80.70	4.99		
H6	Deepfake awareness	91.80	92.20	3.39	< 0.001	2.63
	No deepfake awaereness	67.00	66.70	8.95		
H7	Believe	89.20	89.10	3.82	< 0.001	1.09
	Don't believe	82.80	82.10	5.41		

Table 7 Results on rejected hypotheses

Hypothesis		Mean [%]	Median [%]	SD [%]	*p*-value	Effect Size
H1	Women	77.20	76.30	6.49	< 0.001	– 2.18
	Men	93.90	94.40	3.70		
H2	Women	75.50	75.70	6.82	< 0.001	– 2.37
	Men	94.10	94.40	3.46		
H4	Native Czech - Czech	91.30	92.30	5.28	< 0.001	1.22
	ative Slovak - Czech	84.30	84.40	3.93		
H4	Native Slovak - Slovak	83.70	83.70	4.42	< 0.001	– 1.05
	Native Czech - Slovak	91.70	92.30	5.65		

Respondents were also asked before and after the experiment how confident they were that they would detect voice deepfakes. They were asked to express this confidence on a scale of 1 (not confident) to 5 (extremely confident). The mean before the experiment was 2.29, and 2.94 after. A total of 51.60% of respondents increased this value, while 45.20% did not change it, and only 3.20% decreased it, as shown in Fig. 5. Younger respondents mainly increased the value of their certainty. This may be due to their familiarity with technology and digital manipulation, a steeper learning curve, and the educational experience provided by the experiment. In addition, successfully identifying deepfakes during the experiment likely boosted their confidence, leading them to believe that detecting deepfakes will be easier in the future.

In addition, after completing the experiment, 74.20% of the respondents said they were surprised by the quality of today's voice deepfake in the Czech language.

4.2 Experiment two: influence of deepfake speech quality

The survey was conducted over two months, during which 85 participants (48 men, 37 women) completed it. The majority of participants were university students specialising in technical fields. An online survey was employed for participant recruitment and disseminated through our colleagues, friends, families, and faculty members. In addition, leveraging the student union facilitated broader reach, as one of the authors was a student then. While a larger pool of individuals was invited to participate, 85 respondents ultimately completed the survey. This recruitment strategy ensured a wide distribution and maximised engagement within our accessible networks.

Table 8 Quality ranges in each cluster

Cluster	Range [%]	Deepfake recognition accuracy [%]
1	[20.05, 34.67]	88.20
2	[38.29, 52.58]	87.90
3	[53.08, 67.77]	86.50
4	[72.48, 84.81]	85.00

The numbers are rounded to two decimal points. The clusters are left as defined by the clustering algorithm, resulting in gaps between the intervals

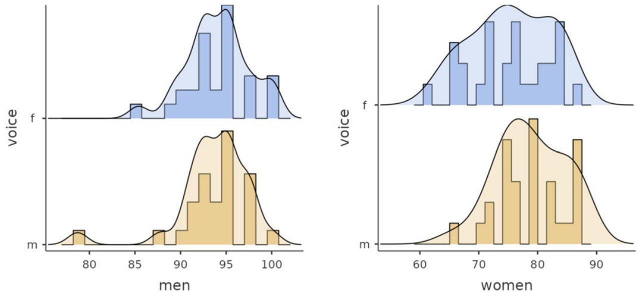


Fig. 6 Plots depicting the accuracy of deepfake detection by gender: Men's accuracy is shown on the left, and women's on the right. The X-axis indicates the percentage of correctly identified deepfakes, while dual Y-axis show the volume of accurately labelled recordings. The graphs employ orange (m) and blue (f) to distinguish between recordings voiced by male and female speakers, respectively, sharing a common X axis but with separate Y axes for each gender's count of correctly identified recordings

However, it is important to note that not every respondent reviewed each pair of recordings presented in the survey. We analysed the gathered data using the students' T test. The analysis enabled us to confirm several hypotheses, as detailed in the results presented in Table 6.

H3: Men, compared to women, are more likely to detect deepfakes spoken by men.

H5: Headphones increase the human capability to detect deepfakes in comparison to device speakers.

H6: People who are aware of deepfakes are more likely to detect them than people who have never heard of deepfakes.

H7: People who think they can detect deepfakes are more likely to detect deepfakes than people who do not think they can detect deepfakes.

As shown in Table 7, the following hypotheses were rejected as there is insufficient significant evidence to support them according to the Student's t test:

H1: Women are more likely to detect voice deepfakes than men.

H2: Women, compared to men, are more likely to detect deepfakes spoken by women.

H4: People are more likely to detect deepfakes in their native language.

Finally, we were able to answer all the research questions:

RQ4: Is there a threshold in the deepfake quality rating score beyond which it is no longer possible to recognise deepfakes?

The results have shown that there seems to be no such threshold in the deepfake quality rating score. Every deepfake recording was correctly recognised at least once.

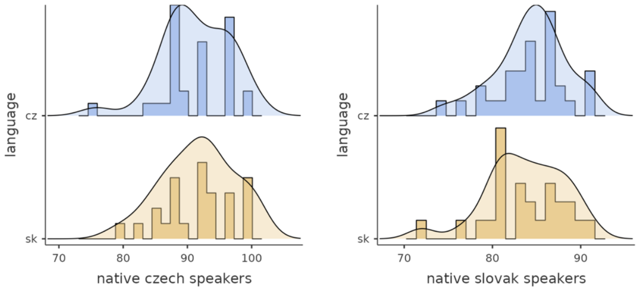


Fig. 7 Plots illustrating the proficiency of native Czech and Slovak speakers in identifying deepfakes, with Czech speakers' results on the left and Slovak speakers' on the right. The X-axis quantifies the percentage of recordings correctly identified. Two distinct colours, blue for Czech (cz) and orange for Slovak (sk) recordings, indicate the language of the recordings. Though these graphs share a common *X* axis, they feature separate *Y* axes to display the count of recordings correctly identified in each language by the respective groups of native Czech and Slovak speakers

Therefore, no deepfake would present the boundary quality beyond which it was impossible to recognise. However, this observation is closely tied to the synthesiser and experimental conditions used. Given the rapid advancements in technology since these experiments were conducted, it is likely that results would differ with a more powerful, state-of-the-art synthesiser.

RQ5: Are people more likely to detect deepfakes with lower score assigned using proposed quality rating system?

As Table 8 shows, the quality of deepfake recordings is inversely proportional to the deepfake recognition accuracy. The higher the quality, the more challenging it is to recognise a deepfake.

RQ6: Are people able to detect voice deepfakes?

The results might be categorised into two main parts: one focusing on gender differences and the other on the impact of native language in deepfake recognition.

Our findings reveal that men are more proficient in identifying deepfakes than women. In the survey, 48 men (56%) and 37 women (44%) participated. Men recognised 93.90% of all deepfakes, while women identified 77.20%. Specifically, men detected 94.10% of deepfakes spoken by women and 93.70% spoken by men. Women had a 78.90% accuracy rate for deepfakes voiced by men and 75.50% for those voiced by women, as shown in figure Fig. 6.

Regarding native language, Czech speakers were more successful at detecting deepfakes than Slovak speakers. The survey included 51 Slovak native speakers and 34 Czech native speakers, with an additional two participants reporting other native languages, accounting for 60% Slovak and 40% Czech speakers, respectively. Czech natives demonstrated a 91.50% accuracy in deepfake detection, compared to the 84% accuracy of Slovak speakers. When evaluating deepfakes by the language spoken (Czech or Slovak), Czech natives showed 91.30% accuracy for Czech-voiced and 91.70% for Slovak-voiced deepfakes. Slovak speakers had an accuracy of 83.70% for Slovak-voiced and 84.30% for Czech-voiced deepfakes. These findings support the hypothesis that Czech native speakers are more adept at detecting deepfakes in both languages, as illustrated in Fig. 7.

RQ7: How many people with previous knowledge of deepfakes can recognise deepfakes?

People who have already heard about deepfakes were more likely to detect deepfakes. Sixty-nine people claimed that they have heard about deepfakes before, which represents 81.18% of all people. The 16 people, representing 18.82%, claimed they had never heard about deepfakes. The correctness of labelling the deepfakes by people who have heard about deepfakes is 91.80%. Conversely, the correctness of labelling the deepfakes by people who have not heard about deepfakes is 67%.

RQ8: Does the audio device impact human's ability to recognise deepfakes?

The results suggest that the audio playback device impacts humans' ability to recognise deepfakes. Of all people, more than 52% were using headphones while listening to the recordings, 47% used a device's speakers, and 1% (one person) used another, unspecified device. The accuracy of proper deepfake detection by people who used headphones is 91.50%. The accuracy of deepfake detection by people who used speakers is 80.70%.

5 Discussion

Related work evaluating human ability often reports more than 60% success rate. The success rate of deepfake detection in the first experiment is 3.20%, which is quite different. It is thus important to say that our approach is fundamentally different from the other works. Considering the case where respondents knew they were presented with deepfakes, the success rate of around 80% for both experiments confirms the related studies' outcomes.

The results of this study revealed several intriguing insights. Notably, none of the participants reacted to the deepfake audio during casual listening. However, when explicitly prompted to pinpoint the deepfake set, nearly all respondents successfully identified it. Many participants confessed that they hadn't detected any anomalies upon first listening. This fundamental discovery has profound implications for educating the public. It suggests that the security risks associated with deepfakes are more extensive than initially anticipated, indicating significant vulnerabilities within modern society. Yet, when participants listened for a second time with the specific intent of identifying the deepfake, they could confidently discern the computer-generated voice. There may be several reasons for this, but we lean towards something similar to a psychological phenomenon called *The Monkey Business Illusion* [39], which states that if people focus on one thing, they are more prone to overlook another, in their opinion, less important things. In our case, it was the answers to the questions and the sound quality. People focused on the correct answers and ignored the difference in the voice recordings. However, they detected it easily when we told them to focus on quality and find the deepfake. These results thus demonstrate the crucial role the knowledge of deepfakes plays in their correct identification and that the education of the broad public on this topic is inevitable.

Moreover, we observe that our ability to recognise deepfakes is connected to the quality of consumed recordings. This goes hand in hand with the used playback device. The increasing quality of the playback device seems to boost our capacity to identify deepfake recordings. In the most favourable cases, we would have the information about possible deepfake exposure and proper playback devices to analyse the recording and make a decision. These findings directly apply to designing protection measures or internal processes to mitigate the possible damage.

The prior experience with deepfakes is similar within both tested groups, meaning that the younger population of the Czech Republic has solid knowledge of deepfake technology. Moreover, we can estimate that the awareness will drop with increasing age [40]. It is thus essential to directly educate these vulnerable groups, such as older people, as vishing attacks or scams often target them. From the collected results, it is evident that prior experience plays a role in the ability to recognise deepfakes, which is also confirmed by other studies [28, 41]. Even though identifying factors that contribute to the correct identification of deepfake recordings led only to the differences in quality and deepfake-specific artefacts, it is evident that raising awareness is a reliable indirect means to improve the ability of the general public to recognise deepfakes.

It is also important to understand to what extent the general public understands deepfakes. As the results from the first experiments suggest, more than 75% of respondents were surprised by the current quality of deepfake speech. Out of the respondents who have at least heard of deepfakes, more than 58% were surprised by the quality. Finally, from the 16% of the respondents actively interested in deepfakes, 40% reported they were surprised by the quality. Moreover, these results align with our personal experience from lectures and demonstrations about deepfakes. Even people with previous knowledge of deepfakes are often surprised by the quality and capability of state-of-the-art models. Awareness is thus a severe issue because knowing that deepfakes exist is very different from understanding their full potential. And without understanding their full potential, people may not expect to encounter them in the increasingly frequent attacks.

This study's findings indicate notable differences in the ability to detect deepfake utterances between genders, with women facing more challenges in this area than men. This observation opens up avenues for further research into how demographic factors influence the recognition of deepfakes and which demographic groups might be more susceptible to such deceptive practices. Understanding these dynamics could lead to more effective strategies for safeguarding vulnerable populations.

In addition, our analysis revealed a discrepancy in deepfake detection abilities between Czech and Slovak speakers, suggesting that Czech speakers were more adept at identifying deepfakes. This difference prompts a broader hypothesis that specific linguistic communities may possess varying levels of resilience or susceptibility to deepfake attacks. For instance, the French language, known for its rigorous pronunciation rules, might present a significant challenge for deepfake creators, as native French speakers may struggle to comprehend speech from non-native speakers [42, 43]. Conversely, languages that are more lenient in pronunciation or have numerous dialects might be more susceptible to convincing deepfake impersonations. This aspect of our research highlights the potential impact of linguistic characteristics on the effectiveness of deepfake technologies and underscores the importance of tailored protective measures for different language communities. Given these preliminary findings, further research is required to deepen our understanding of these phenomena and to develop more nuanced approaches to countering deepfake misinformation across diverse linguistic and demographic landscapes.

5.1 Limitations

The primary issue with the first experiment was the quality of the deepfake recordings, which were attributed to background noise. Despite minimal noise and the recordings being understandable when played on an iPhone 11, many participants reported that the noise significantly compromised the quality. This discrepancy in audio quality perception likely stems from the variability in noise reduction capabilities across different playback devices. The most commonly reported problems by participants were related to the poor quality and presence of noise, with 13 respondents specifically mentioning reduced quality. This observation does not substantially limit the findings of our results but rather shows how deep the problem actually is. Using state-of-the-art models that are currently able to suppress these artefacts would make the results much less favourable for us humans.

In recent months, the field of speech synthesis has seen rapid advancements, significantly improving the quality of synthesised speech. If cutting-edge technology were employed currently, we anticipate the findings would be notably more concerning.

Regarding the second experiment, including a more extensive and diverse group of participants would have been advantageous. Most participants were young individuals with a background in IT, a demographic presumably more adept at identifying deepfakes. Consequently, the performance of this group could be considered the upper bound of deepfake recognition capabilities, suggesting that outcomes from a more varied sample might be even more concerning. Despite this, the comparison with other studies indicates that our participant sample was sufficiently representative, affirming the validity of our observations concerning the quality of deepfake speech.

6 Improving human ability to detect deepfakes

The limited capability of humans to detect deepfakes accurately highlights the critical need to enhance this skill. In light of this, we propose several strategies grounded in existing research and our findings to bolster the ability of individuals to discern deepfakes.

Westerlund [44] cites computer scientist Hao Li, who remarks, *“This is developing more rapidly than I thought. Soon, it is going to get to the point where there is no way that we can actually detect [deepfakes] anymore, so we have to look at other types of solutions.“*

Supporting this, evidence from prior studies and our research indicates that exposure to deepfakes can enhance the human capacity to identify them [28, 41]. Raising public awareness emerges as a broad yet impactful strategy to improve general proficiency in recognising deepfakes, with even basic demonstration materials proving beneficial.

However, it is important to acknowledge that not all studies agree on the impact of prior exposure to deepfakes on detection performance. For instance, Bray et al. [26] and Mai et al. [14] found that previous exposure to deepfakes did not significantly improve detection abilities. This discrepancy in findings highlights the issue’s complexity. It suggests that the effectiveness of exposure may depend on various factors, such as the type and quality of deepfakes, the context of exposure, and individual differences in perceptual and cognitive abilities.

In addition, the concept of super-recognizers, individuals who excel in face recognition, suggests that detection abilities can vary significantly within the population [45]. Auditory perception, abstraction skills, and overall perceptual and cognitive abilities also play a crucial role in recognizing deepfakes. Therefore, while exposure and awareness-raising are beneficial, the varying capabilities among individuals must be considered in strategies aimed at improving deepfake detection.

Given these mixed results, further research is necessary to understand the conditions under which exposure to deepfakes can enhance detection performance. It may be that certain types of training or exposure are more effective than others or that individual differences play a significant role in the ability to detect deepfakes. Thus, while public awareness and exposure remain promising strategies, they should be implemented thoughtfully, considering the nuances highlighted by conflicting research findings.

Tahir et al. [23] significantly improved detection abilities by educating participants through illustrated deepfake videos, emphasising key points and analytical techniques. Transferring this educational approach to audio deepfakes requires identifying specific audio deepfake artefacts and instructing people on these markers using concrete examples. However, the challenge with audio media is notable; internet videos are generally of high quality, while audio media, such as phone calls or voice messages, often experience quality degradation due to transmission or recording methods, which could mistakenly be perceived as signs of deepfakes.

Our experiment revealed that participants initially focused on content, overlooking sound artefacts, and failed to detect the deepfake. Upon a second listening, with attention shifted to audio qualities, most could identify the deepfake. This suggests a dual-listening strategy for deepfake detection: the first for content and the second for audio analysis.

Furthermore, we advocate for training in verification and caution. Given the increasing sophistication of deepfakes, as noted by the FBI [46], adopting the SIFT method—Stop, Investigate the source, Find trusted coverage, and Trace original content—can effectively counter disinformation. This strategy, coupled with scepticism towards online personas and the use of multi-factor authentication, enhances protection against deepfakes. Implementing simple validation steps, such as double authentication for sensitive transactions, can prevent spoofing attempts.

Considering each piece of information as potentially false until verified could also serve as a proactive defence against misinformation. This approach, akin to scepticism towards improbable claims from strangers, could reverse the current trend of credulity in online information.

Detection tools, as shown by Groh et al. [22], can aid in identifying fraudulent media. However, accessible, non-commercial tools for verifying media remain scarce.

To consolidate these strategies, we propose the creation of an educational platform offering:

- Demonstrations of deepfake technologies, misuse examples, vulnerabilities, and defensive measures.
- Interactive training for detecting synthetic media.
- Guidance on information verification and cautious engagement.
- An overview of detection tools, including usage tutorials.
- Resources and links for individuals impacted by deepfakes, such as www.napisnam.cz in the Czech Republic.

A publicly accessible web application where users can explore tutorials, interact with deepfake technology, and learn about its implications could significantly bolster public resilience to these deceptions.

7 Conclusions

This work has shown that the human ability to recognise voice deepfakes is not at a level we can trust. We have pointed out crucial factors that influence the human ability to recognise deepfakes, which significantly change the threat landscape and impacts of deepfake speech. The prior information about deepfake exposure substantially influences the recognition abilities. It is thus challenging for people to distinguish between real and fake voices if they are not expecting them. The human ability to detect deepfakes is influenced mainly by the fact that people don't think about the voice they are listening to, are used to poor-quality audio conversations, and focus primarily on the content of the message.

It is evident that people without knowledge of deepfakes cannot reliably identify deepfake recordings in conversation. Combined with the Czech and Slovak languages, we show this problem is general and poses a significant threat to society. Even less popular languages are threatened, as synthesising speech is no longer limited to English. Moreover, after revealing the presence of a deepfake set, most respondents could identify it. However, this identification was caused by a difference in audio quality or muffled sound compared to the bonafide sets. It is thus essential to address these imperfections in future and assess what role the audio quality plays in the detection process.

As suggested, the second factor influencing the human recognition of deepfakes is the quality of deepfake recording. It is apparent that our ability to distinguish bonafide from deepfake recordings degrades with increasing quality of deepfake speech.

Our results show that awareness of deepfake technology increases individuals' ability to recognise deepfake recordings. It is thus vital to continuously raise public awareness and educate the broad public on the dangers of deepfake technology.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13640-024-00641-4.

Supplementary Material 1.

Acknowledgements

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Author contributions

KM oversaw the design and execution of the experimental work, helped write the manuscript, and proofread it. AF oversaw the design and execution of the experimental work and was a significant contributor to writing the manuscript. MŠ helped write the article and proofread it. DP designed, executed and evaluated the first experiment. KR designed, executed and evaluated the second experiment. PH oversaw the execution of the experimental work and proofread the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the national project NABOSO: Tools To Combat Voice DeepFakes" (with code VB02000060) funded by the Ministry of the Interior of the Czech Republic and the Brno University of Technology internal project FIT-S-23-8151.

Availability of data and materials

The data generated and used during the first experiment are not publicly available as they contain the speech of one of the authors but may be provided upon reasonable request to the corresponding author. The datasets generated and analysed during the second experiment are publicly available at: https://nextcloud.fit.vutbr.cz/s/iwKpdJa4tMYggPe/download/dataset.zip. The models used to synthesise audio in the second experiment are publicly available at: https://nextcloud.fit.vutbr.cz/s/3ENB2rdzzTYp7Qe/download/YourTTS_CZSK.zip.

Declarations

Competing interests

The authors declare that they have no Conflict of interest.

Received: 29 February 2024 Accepted: 19 August 2024
Published online: 30 August 2024

References

1. A. Firc, K. Malinka, P. Hanáček, Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. Heliyon **9**(4), 15090 (2023). https://doi.org/10.1016/j.heliyon.2023.e15090
2. A. Firc, K. Malinka, The Dawn of a Text-dependent Society: Deepfakes as a Threat to Speech Verification Systems, pp. 1646–1655 (2022). https://doi.org/10.1145/3477314.3507013
3. M. Šalko, A. Firc, K. Malinka, Security Implications of Deepfakes in Face Authentication. (2024). https://doi.org/10.1145/3605098.3635953
4. M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, Deepfake detection: A systematic literature review. IEEE Access **10**, 25494–25513 (2022). https://doi.org/10.1109/ACCESS.2022.3154404
5. Y. Mirsky, W. Lee, The creation and detection of deepfakes: A survey. ACM Comput. Surv. **54**(1) (2021) https://doi.org/10.1145/3425780
6. H. Chen, K. Magramo, Finance worker pays out \$25 million after video call with Deepfake "chief financial officer". Cable News Network (2024). https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
7. T. Brewster, Fraudsters cloned company director's voice in \$35 million bank heist, police find. Forbes Magazine (2021). https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
8. M. Bajtler, Falešné videohovory Jsou Tu. Kolegovi Zavolal Måj Deepfake, říká Zakladatel Gymbeamu. Forbes (2023). https://forbes.cz/falesne-videohovory-jsou-tu-kolegovi-zavolal-muj-deepfake-rika-zakladatel-gymbeamu/
9. L. O'Donnell, CEO 'Deep fake' swindles company out of \$243K (2019). https://threatpost.com/deep-fake-of-ceos-voice-swindles-company-out-of-243k/147982/
10. P. Oltermann, European politicians duped into deepfake video calls with mayor of Kyiv. Guardian News and Media (2022). https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko
11. J. Wakefield, Deepfake presidents used in Russia-ukraine war. BBC (2022). https://www.bbc.com/news/technology-60780142
12. S.M. Kelly, Explicit, ai-generated Taylor Swift images spread quickly on social media. CNN (2024). https://www.cnn.com/2024/01/25/tech/taylor-swift-ai-generated-images/index.html
13. N.M. Müller, K. Pizzi, J. Williams, Human perception of audio deepfakes. In: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. DDAM '22, pp. 85–91. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3552466.3556531
14. K.T. Mai, S. Bray, T. Davies, L.D. Griffin, Warning: Humans cannot reliably detect speech deepfakes. PLoS ONE **18**(8), 0285333 (2023). https://doi.org/10.1371/journal.pone.0285333
15. D. Prudký, A. Firc, K. Malinka, Assessing the human ability to recognize synthetic speech in ordinary conversation. In: 2023 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5 (2023). https://doi.org/10.1109/BIOSIG58226.2023.10346006
16. X. Wang, J. Yamagishi, M.Todisco, H.Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S.L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y.Jia, K. Onuma, K. Mushika, T.Kaneda, Y.Jiang, L.-J. Liu, Y.-C. Wu, W.-C.Huang, T.Toda, K.Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S.Ronanki, J.-X. Zhang, Z.-H. Ling, Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language 64, 101114 (2020)https://doi.org/10.1016/j.csl.2020.101114
17. G. Watson, Z. Khanjani, V.P. Janeja, Audio Deepfake Perceptions in College Going Populations (2021)
18. M. Groh, Z. Epstein, N. Obradovich, M. Cebrian, I. Rahwan, Human detection of machine-manipulated media. Communications of the ACM **64**(10), 40–47 (2021). https://doi.org/10.1145/3445972. Accessed 2022-12-26
19. S.R. Godage, F. Lovasdaly, S. Venkatesh, K. Raja, R. Ramachandra, C. Busch, Analyzing human observer ability in morphing attack detection -where do we stand? IEEE Transactions on Technology and Society, 1–1 (2023) https://doi.org/10.1109/tts.2022.3231450
20. A, Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv. arXiv:1901.08971 [cs] (2019). http://arxiv.org/abs/1901.08971 Accessed 2022-12-26
21. P. Korshunov, S. Marcel, Deepfake detection: humans vs. machines. arXiv. arXiv:2009.03155 [cs, eess] (2020). http://arxiv.org/abs/2009.03155 Accessed 2022-12-26
22. M. Groh, Z. Epstein, C. Firestone, R. Picard, Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences **119**(1), 2110013119 (2022) https://doi.org/10.1073/pnas.2110013119https://www.pnas.org/doi/pdf/10.1073/pnas.2110013119

23. R. Tahir, B. Batool, H. Jamshed, M. Jameel, M. Anwar, F. Ahmed, M.A. Zaffar, M.F. Zaffar, Seeing is believing: Exploring perceptual differences in DeepFake videos. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, ??? (2021). https://doi.org/10.1145/3411764.3445699
24. M. Groh, A. Sankaranarayanan, N. Singh, D.Y. Kim, A. Lippman, R. Picard, Human Detection of Political Speech Deepfakes across Transcripts, Audio, and Video (2024)
25. S.K. Jilani, Z. Geradts, A. Abubakar, Decoding deception: Understanding human discrimination ability in differentiating authentic faces from deepfake deceits, in *Image Analysis and Processing - ICIAP 2023 Workshops*. ed. by G.L. Foresti, A. Fusiello, E. Hancock (Springer, Cham, 2024), pp.470–481
26. S.D. Bray, S.D. Johnson, B. Kleinberg, Testing human ability to detect 'deepfake' images of human faces. Journal of Cybersecurity **9**(1) (2023) https://doi.org/10.1093/cybsec/tyad011
27. K. Somoray, D.J. Miller, Providing detection strategies to improve human detection of deepfakes: An experimental study. Computers in Human Behavior 149, 107917 (2023) https://doi.org/10.1016/j.chb.2023.107917
28. M.F.B. Ahmed, M.S.U. Miah, A. Bhowmik, J.B. Sulaiman, Awareness to deepfake: A resistance mechanism to deepfake. In: 2021 International Congress of Advanced Technology and Engineering (ICOTEN), pp. 1–5 (2021). https://doi.org/10.1109/ICOTEN52080.2021.9493549
29. V. Matyas, J. Krhovjak, M. Kumpost, D. Cvrcek, Authorizing card payments with pins. Computer 41, 64–68 (2008) https://doi.org/10.1109/MC.2008.40
30. D. Prudký, Assessing the human ability to recognize synthetic speech. Bachelor's thesis, Brno University of Technology, Brno, Czech republic (2023). https://www.vut.cz/en/students/final-thesis/detail/140541
31. E. Casanova, J. Weber, C.D. Shulby, A.C. Junior, E. Gölge, M.A. Ponti, YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2709–2720. PMLR, ??? (2022). https://proceedings.mlr.press/v162/casanova22a.html
32. P.C. Loizou, Speech quality assessment. Multimedia analysis, processing and communications, 623–654 (2011)
33. K. Martin, New ID R &D research finds over 1 in 3 Americans confident they could detect a computer-generated voice pretending to be a human voice (2020). https://www.idrnd.ai/voice-deepfake-survey/
34. K. Radačovská, Deepfake dataset for evaluation of human capability on deepfake recognition. Bachelor's thesis, Brno University of Technology, Brno, Czech republic (2023). https://www.vut.cz/studenti/zav-prace/detail/140539
35. M. Wang, C. Boeddeker, R.G. Dantas, A. Seelan, ludlows/python-pesq: supporting for multiprocessing features. Zenodo (2022) https://doi.org/10.5281/ZENODO.6549559. https://zenodo.org/record/6549559
36. M. Shannon, mcd. GitHub (2017)
37. M. MORISE, F. YOKOMORI, K. OZAWA, World: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems E99.D(7), 1877–1884 (2016) https://doi.org/10.1587/transinf.2015EDP7457
38. R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, G. Weber, Common Voice: A Massively-Multilingual Speech Corpus (2020)
39. D.J. Simons, C.F. Chabris, The monkey business illusion. Cognition **119**(1), 23–32 (2010)
40. A. Firc, Applicability of deepfakes in the field of cyber security. Master's thesis, Brno University of Technology, Faculty of Information Technology, Brno (2021). Supervisor Mgr. Kamil Malinka, Ph.D
41. S.R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra, C. Busch, Analyzing human observer ability in morphing attack detection-where do we stand? IEEE Transactions on Technology and Society **4**(2), 125–145 (2023). https://doi.org/10.1109/TTS.2022.3231450
42. ThoughtCo: These French pronunciation mistakes are toughest for new speakers. ThoughtCo (2019). https://www.thoughtco.com/french-pronunciation-mistakes-and-difficulties-1364615
43. D. Liakin, W. Cardoso, N. Liakina, Learning l2 pronunciation with a mobile speech recognizer: French /y/. CALICO Journal 32(1), 1–25 (2015). Accessed 2024-06-10
44. M. Westerlund, The emergence of deepfake technology: A review. Technology Innovation Management Review 9, 40–53 (2019) https://doi.org/10.22215/timreview/1282 . Chap. 40
45. R. Russell, B. Duchaine, K. Nakayama, Super-recognizers: People with extraordinary face recognition ability. Psychonomic Bulletin & Review **16**(2), 252–257 (2009). https://doi.org/10.3758/pbr.16.2.252
46. Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations. publisher: FBI (2021). https://www.aha.org/system/files/media/file/2021/03/fbi-tlp-white-pin-malicious-actors-almost-certainly-will-leverage-synthetic-content-for-cyber-and-foreign-influence-operations-3-10-21.pdf Accessed 2023-04-24
47. A. Firc, K. Malinka, P. Hanáček, Deepfake speech detection: A spectrogram analysis, pp. 1312–1320 (2024). https://doi.org/10.1145/3605098.3635911

Publisher's Note

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2022, accepted January 21, 2022, date of publication February 7, 2022, date of current version February 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149475

E-Banking Security Study–10 Years Later

KAMIL MALINKA, ONDŘEJ HUJŇÁK, PETR HANÁČEK, AND LUKÁŠ HELLEBRANDT
Faculty of Information Technology, Brno University of Technology, 612 00 Brno, Czech Republic

Corresponding author: Kamil Malinka (malinka@fit.vutbr.cz)

This work was supported by the Internal Project of Brno University of Technology (BUT) under Grant IGA FIT-S-20-6427.

ABSTRACT ICT security in the banking area is going through rapid changes. It is ten years since we covered the state of e-banking security, and both authentication schemes and legislation has evolved. With the Payment Services Directive (PSD2) for European Union coming into force, we believe it is a good time to update our findings. PSD2 brings new requirements for multi-factor authentication, thus it is necessary to revise compliance of currently used schemes. This work's main contribution is an overview of current authentication methods, their properties with respect to international standards, and their resistance against attacks. We further discuss the multi-factor authentication schemes composed of those methods and their compliance with the PSD2 requirements. In order to present the overview, we introduced the e-banking attacks taxonomy, which is compatible with authenticator threats from NIST Digital Identity Guidelines but has an increased level of detail with respect to the e-banking area. The available sources in this area are usually either very broad, targeted on the business executive, or focus on one particular issue or attack in greater detail. We believe our article can bridge such diverse sources by providing a comprehensive and complex tool to help with orientation in the area.

INDEX TERMS Online banking, PSD2, authentication, multi-factor, cybersecurity, secure hardware.

I. INTRODUCTION

Ten years ago, we published a comparative study focused on the security of e-banking [1], where we summarised basic forms of electronic banking and widely used authentication and authorisation methods. Given the drastic evolution of the situation over the years, shift to mobile banking and the emergence of new European directives that affect this area, we believe it is an ideal time to update our findings. Changes in user behaviour and used equipment directly impact the security of the whole environment. For example, by integrating smart banking into the smartphone, we are losing a secure second channel used for SMS verification, as opposed to the traditional web application e-banking performed through a PC (i.e., second channel). Also, the ''smartness'' of the devices brings new vectors of attacks as they can be targeted by malware.

The goal of this paper is to present an overview of current authentication methods, their relation to the most common attacks on electronic banking and the level of protection they can provide. With new requirements on two-factor authentication brought by the new European directive PSD2, we also discuss possible combinations of authentication methods and evaluate their usability and security properties.

A. PARADIGM SHIFT

The banking sector keeps going through continuous digital evolution as the paradigms in the finance sector are shifting. Ten years ago, we witnessed the transfer from in-person banking to online e-banking, and this trend continues towards mobile banking. Moreover, with the increase of accessibility and digitisation of services, fully digital banking emerged and, in such an environment, physical contact with the customer is completely dropped in favour of digital means.

SwissFinanceCouncil estimates that nearly 60% of retail banking transactions worldwide go through mobile and online channels [2]. This corresponds with reports from other countries such as Brazil [3] as shown in Figure 1, where digital channels carry out 63% of banking transactions (either mobile or internet banking) and the share of mobile banking is increasing every year. The shift towards mobile solutions is apparent also in the Deloitte GMCS report, which claims that according to the UK survey, a smartphone is a device of preference when using banking services for the majority of people [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Weizhi Meng.

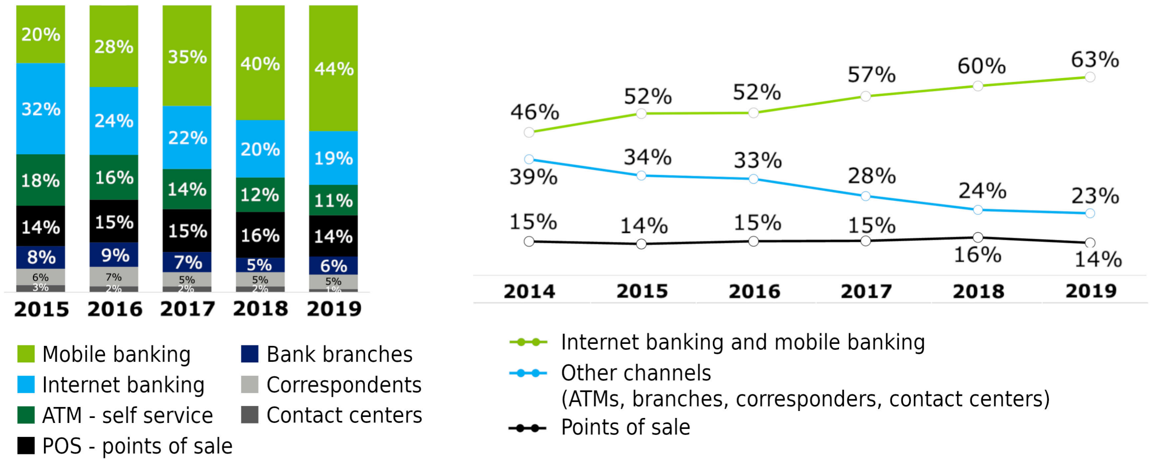


FIGURE 1. Composition of banking transactions in Brasil (in %) [3].

The popularity of mobile banking corresponds with the adoption of smartphones and, as noted by Pew Research Centre, "today, most people who own a mobile phone own a smartphone" [5]. The users of basic cell phones, which were targeted as 2nd factor for e-banking ten years ago, are rapidly decreasing. Also, in advanced economies, only 18% of people own a basic cell phone, as seen in Figure 2.

These changes bring not only increased comfort to the users but impose new security challenges on the banks. The worldwide spread of FinTech services as observed by EY [6] brings further focus on using emerging technologies to provide financial operations that rapidly expand the attack surface. One of the most eminent new challenges is identity verification in fully digital banking.

When facing these challenges, financial institutions were forced to develop new mitigation techniques and update the legacy ones. The identity and operation verification shifted to enforce multi-factor authentication with various factors. While the passwords and PIN codes are still broadly used, authentication calculators were superseded by SMS codes, which are now replaced by newly established factors such as digital tokens and biometrics. For payment card operations, 3D Secure protocol [7] was introduced.

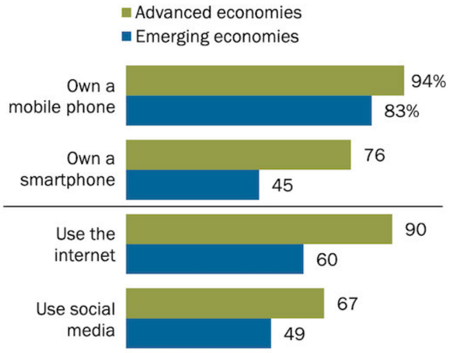


FIGURE 2. Mobile technology, internet, and social media use [% of adults] [5].

New legislation frameworks were developed to establish interoperability and security across the financial sector and updated to address emerging concerns. The international standard addressing security is Payment Card Industry Data Security Standard (PCI DSS [8]), while the European Union introduced a significant change in approach with its Payment Services Directive version 2 (PSD2) [9]. PSD2 forces financial institutions to publish interface to their data, effectively allowing FinTech companies to work with it. To ensure security, PSD2 dictates strong authentication and multi-factor authentication.

B. CONTRIBUTIONS

The main contribution of this paper lies in the security evaluation of authentication schemes composed of viable combinations of authentication methods concerning the concurrent standards - mostly the PSD2 directive of the European Union. In order to perform this evaluation, we propose a specific e-banking attacks taxonomy and define authentication primitives and their security features.

All contributions of the paper can be summarised as follows:

- We propose e-banking attacks taxonomy compatible with authenticator threats from NIST Digital Identity Guidelines, but with an increased level of detail with respect to the e-banking area. In taxonomy, we describe relevant attacks and trends in their occurrence. We look in more detail at the two most common types of attacks – phishing and malware.
- We present an overview of current authentication methods and their properties in the context of international standards. We also benchmark their resistance against attacks from taxonomy.

- We introduce possible combinations of the multi-factor authentication schemes composed of current authentication methods and evaluate their compliance with the newest standard available – the PSD2 directive of the European Union.
- We address future trends and present open gaps.
- We provide up to date and comprehensive view of e-banking security, enabling the reader to get an overview of available options and their advantages and disadvantages in the context of current international regulations.

C. ORGANIZATION

The rest of the paper is organised as follows: In Section II, we present e-banking attacks taxonomy with the description of attacks, their impact and historical context. In greater detail, we focus on malware and phishing, which are currently the most common. In Section III, we explain elementary authentication primitives and their properties with respect to the requirements from NIST and PSD2. To increase the detail for the e-banking area, we have expanded the former categories into multiple subcategories and discuss their resistance to attacks from our taxonomy. We discuss the usability of multiple factors combination in the context of new PSD2 requirements in Section IV. Then, in Section V we summarize related work. Finally, in Section VI, we conclude the paper.

II. ATTACKS ON E-BANKING SYSTEMS

This section presents an overview of the current attacks on e-banking and their trends. We identify which attacks are declining in their usage and which are emerging thanks to the new technologies used nowadays and a shift in user behaviour.

A. PROPOSED E-BANKING ATTACK TAXONOMY

E-banking security is a well-known research topic, and many scientific papers and studies can be found (eg. [10]–[13] [14], [15]). The main issue of these sources is that they often focus on very specific attacks (mostly scientific papers) or broadly cover the topic for business executives. Thus, we decided to create a reference taxonomy, which can be used to ease orientation in the broad spectrum of e-banking attacks.

The proposed e-banking attack taxonomy, shown in Figure 3, defines four most common categories:

- ***Authentication and authorisation attacks*** – Goal of attacker is to obtain valid user credentials (such as passwords, PINs, certificates etc.) for a specific service.
- ***Identity theft (identity stealing)*** – In these attacks, the adversary attempts to misuse or take over someone's identity to enable various malicious behaviour.
- ***Communication attacks*** – Attacks that passively listen to communication between two parties or actively transparently participate in it.
- ***Attacks focused on bank*** – For the sake of comprehensive overview, we add the category of attacks directly targeting bank infrastructure and bank employees as opposed to previous categories focusing on users (bank clients) and their communication channels.

  These attacks represent a high risk, high value for attackers as successful attacks of this category can cause financial gain in hundreds of millions of dollars. However, they are hard to carry out because banking infrastructure protection usually uses state of the art technologies. A deeper study of this category is beyond the scope of this article.

The category list we propose is not exclusive by nature. As some attack types may fall into various categories depending on the point of view used, the defined categories blend seamlessly. We have subdivided every category into specific attack techniques that fulfil the malicious goal of the category. The full proposed taxonomy is displayed in Figure 3, and the description of individual elements follows:

A-1 Authentication and Authorisation Attacks

The attacker's goal is to obtain valid user credentials (such as passwords, PINs, certificates, etc.) for a specific service.

A-1.1 Password Guessing

Password guessing is an attack based on systematic guessing of a user password. There are multiple approaches, such as a dictionary-based attack, brute-force, or even attacks based on neural network usage [16].

There are three main approaches used for e-banking attacks: first mentioned above, second uses guessing of one high-quality password which is then used on a large number of clients (because it is unfeasible to test a large number of passwords on one user account). The third approach uses a password obtained from other sources of leaked passwords.

A-1.2 Exhaustive Search

Known also as a brute force attack is based on trying a large number (all) possible passwords or secret values. This approach is not commonly used in the banking environment due to the limited attempts and difficulty of obtaining an encrypted password file. However, it is viable to crack passwords from leaked credential databases or other sources and use those against bank authentication. Distributed high-performance computing can be used to increase attack speed [17].

A-1.3 Phishing

Phishing tries to deceive the user by fraudulent e-mail/webpage to steal credentials or other personal data [18]. Targeting the user as the weakest link has proved to be a very dangerous and successful technique.

A-1.4 Pharming

This technique is similar to phishing but usually requires the assistance of malware or DNS spoofing attacks which redirect users to fake sites with

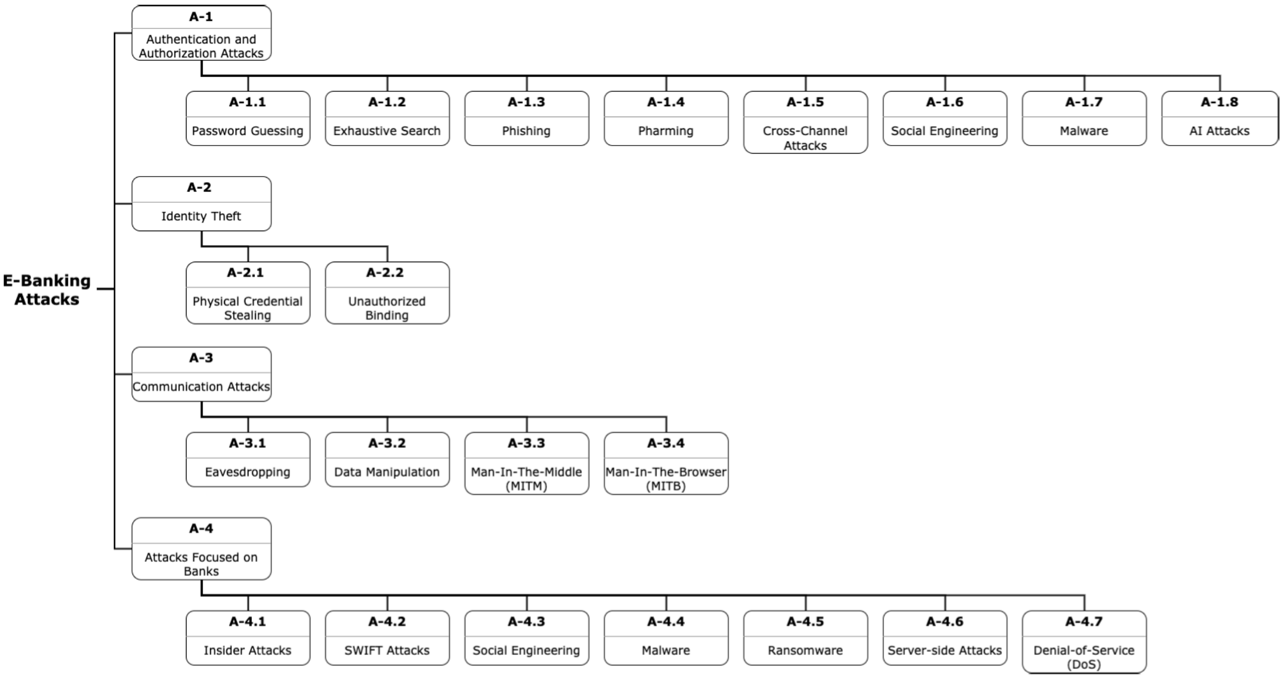

E-Banking Attacks
A-1 Authentication and Authorization Attacks
A-1.1 Password Guessing
A-1.2 Exhaustive Search
A-1.3 Phishing
A-1.4 Pharming
A-1.5 Cross-Channel Attacks
A-1.6 Social Engineering
A-1.7 Malware
A-1.8 AI Attacks
A-2 Identity Theft
A-2.1 Physical Credential Stealing
A-2.2 Unauthorized Binding
A-3 Communication Attacks
A-3.1 Eavesdropping
A-3.2 Data Manipulation
A-3.3 Man-In-The-Middle (MITM)
A-3.4 Man-In-The-Browser (MITB)
A-4 Attacks Focused on Banks
A-4.1 Insider Attacks
A-4.2 SWIFT Attacks
A-4.3 Social Engineering
A-4.4 Malware
A-4.5 Ransomware
A-4.6 Server-side Attacks
A-4.7 Denial-of-Service (DoS)

FIGURE 3. E-banking attacks taxonomy.

a similar appearance as original pages (bank pages etc.) [19], [20].

A-1.5 Cross-Channel Attacks
These attacks usually target systems that use multi-factor authentication (2FA), where the adversary is forced to attack multiple channels simultaneously, e.g., a simultaneous attack on internet connection and SMS messages. Attacking multiple channels usually requires various methods, such as a combination of social engineering and hacking.

A-1.6 Social Engineering
Social engineering is a way of manipulating people, so they give up confidential information, which includes passwords, bank information, or access to a computer to install malicious software secretly.

A-1.7 Malware
Specialized malware that is designed for credential stealing [21]. The most common class is the banking trojan which advertises itself as a useful application while scraping credentials or supporting other attacks in the background. Modern malware is multi-purpose, and there is malware for both computer and mobile OS's.

A-1.8 AI Attacks
Attacks specialise exclusively on (biometric) authentication, usually in the context of the "Know Your Customer" (KYC) process, such as creating fake samples to deceive voice and face recognition systems. Attackers can utilise an algorithm class called generative adversarial network (GAN), which is a class of machine learning algorithms designed to generate artificial data with the same statistics as the training set [22]. An example of the exact use of GAN is false eye image creation [23].

A-2 Identity Theft
The goal of the attacker is to misuse or take over someone's identity to enable various malicious behaviours.

A-2.1 Physical Credential Stealing
Real-world theft of ID cards or creating counterfeit documents with high value to criminals such as passports, driver licenses, credit cards, bank statements, tax statements, medicare cards and utility bills.

A-2.2 Unauthorised Binding
Unauthorised Binding is a class of attacks aiming to bind an adversarial authentication device (such as HW token, SIM card or private key) to the victim's account. Because of the common usage of SMS codes as an out-of-band (oob) second factor, the most relevant attacks are aimed at telecommunication operators (telco). We provide examples of two such attacks – SIM Swapping and SS7 attacks. SIM Swapping is a technique for diverting telco services (including calls and SMS) from victims' mobile carrier account to a new SIM card controlled by an adversary [24]. Similarly, SS7 attacks divert telco services to an attacker, but those attacks

follow the Man-in-the-middle (MITM) scheme and abuse directly vulnerabilities in Signalling System 7 (SS7) protocols used for public telephone calls [25].

A-3 Communication Attacks
Attacks that passively listen to communication between two parties or actively and transparently participate in it.

A-3.1 Eavesdropping
Passive listening to some communication that is happening on a network of any kind without generating any activity – e.g., running a piece of software on a network device, which is merely saving all the data that has passed through it. The collected data, either encrypted or unencrypted, is later analysed and can be further used by the attacker [26]–[29].

A-3.2 Data Manipulation
In Data Manipulation, the adversary not only listens to communication but also actively modifies the messages [30]–[32].

A-3.3 Man-In-The-Middle (MITM)
In Man-in-the-middle, the adversary hijacks the whole communication channel and positions themself in the middle of the communication in order to gain access to the data the communicating parties wouldn't reveal voluntarily [33], [34]. In encrypted communication, the adversary creates encrypted channels to both communicating parties and decrypts and re-encrypts all the messages [35].

A-3.4 Man-In-The-Browser (MITB)
The Man-in-the-Browser attack corresponds to a Man-in-the-middle attack, but in this case, the attacking malware is embedded within a web browser [36].

A-4 Attacks Focused on Banks
Attacks targeting banking organisations with various goals such as theft, data breach, disruption, and espionage.

A-4.1 Insider Attacks
An insider threat is a security risk that originates within the targeted organisation. This doesn't mean that the actor must be a current employee or officer in the organisation. It can be an outsider who positioned himself within an organisation infrastructure [37].

A-4.2 SWIFT Attacks
SWIFT (Society for Worldwide Interbank Financial Telecommunications) is a bank-to-bank electronic messaging system that is the primary means of communication for international wire transfers. These attacks exploit vulnerabilities in the SWIFT interface system allowing the attackers to gain control of the banks' legitimate SWIFT credentials or endpoint device [11]. This leads to sending fake SWIFT funds transfer requests to other banks [38].

A-4.3 Social Engineering
As mentioned above (A-1.6), but with respect to a different environment – focused on bank employees to disrupt trustworthy bank operations. The goal is to convince a bank employee to perform an illegitimate action, e.g. create a forged transaction, reveal private information.

A-4.4 Malware
Specialised malware targeting the bank IT infrastructure [39]. Because banks are usually well protected, this type of malware is usually a custom made and highly sophisticated product of an organised group.

A-4.5 Ransomware
Ransomware is a category of malware able to encrypt the user data and prevent the user from accessing it, thus attacking the availability and causing DoS [40]. The attacker then demands a ransom from the victim to restore access to the data upon payment.

A-4.6 Server-side Attacks
Attacks launched directly from an attacker (the client) to a listening service. Web application attacks are dominant these days, as described by OWASP [41] such as SQL injection, cross-site scripting, broken authentication and session management etc. [42].

A-4.7 Denial-of-Service (DoS)
The DoS attack will send multiple requests to the attacked web resource to exceed the website's capacity to handle multiple requests and prevent the website from functioning correctly [43], [44].

The attack grouping in our taxonomy is compatible with NIST "Digital Identity Guidelines" [45] and provides a deeper focus on the e-banking area. Some NIST threat groups are directly corresponding with attacks we have identified - *Social Engineering*, *Eavesdropping* and *Unauthorised Binding*. For a deeper understanding, we name the password attacks differently than NIST - we use the term *Password Guessing* for *Online Guessing* and *Exhaustive Search* for *Offline Cracking*. We consider *Duplication* a special case of *Theft* and call the group *Physical credential stealing*. Because *Phishing* and *Pharming* attacks have different implications as pharming needs additional supporting malware, we have decided to split them. We consider the groups *Assertion Manufacture or Modification* and *Endpoint Compromise* too general and divide them into more specific groups – *Data manipulation*, *Man-in-the-middle*, *Man-in-the-browser* and *Malware*. A special case is our group *Cross-channel attacks*, which may correspond with both *Eavesdropping* and *Endpoint Compromise* NIST groups depending on the execution of the attack. We define a new group *AI attacks* (Artificial Intelligence attacks) for novel attacks utilising AI to bypass security defences (focused on biometric systems). Lastly, we omit the *Side Channel Attacks* because the extraction of secrets from authenticator is for e-banking security in principle the same as *Physical credential stealing*.

B. OVERALL TRENDS OF ATTACKS

To provide additional value and insight, we attempted to include a recent trend for every attack identified, but this was proven tricky due to the lack of consistent long-term data. Available data from reports usually focus only on the most relevant attacks in the selected time period. Thus, it is not possible to create such trends for a wider range of attacks over the years. Instead, we decided to present an excerpt of findings, which we consider the most important:

- "*Financial services firms fall victim to cybersecurity attacks 300 times more frequently than businesses in other industries*" [10].
- "*Attacks on this sector accounted for 17 percent of all attacks in the top 10 attacked industries*" [47].
- "*Number of security incidents in this sector has tripled in the past five years*" [10].
- "*Social engineering remains the number one threat in breaching security defences, regardless of the maturity and frequency of security awareness campaigns*" [48], [49].
- "*Denial of service, social engineering, drive-by downloads and phishing to disseminate banking Trojans, and malicious insiders remain the most prevalent attack strategies*" [10].

One of the reasons for the increasing number of attacks is their availability and accessibility, even for people without deep knowledge. Many attack tools, especially malware [50] (even zero-day exploits) and phishing kits [51] are available for purchase on dark web marketplaces [52].

Some attacks are very stable in time, and their evolution follows the development of countermeasures, such as phishing and Man-In-the-Middle (MITM). On the other hand, some are brand new, often enabled by new technologies, e.g., mobile malware or Man-In-The-Browser (MITB). Brand new possibilities for attacks are introduced by the increased use of Artificial Intelligence (AI) both in production systems and attack tools. Even though AI attack surfaces are just emerging, Accenture warns that security strategies have to focus on strengthening their critical AI models. Those models are becoming more and more complex, which increases the risk of an adversary discovering a particular behaviour of the model leading to its exploitation [53].

In a historical context, we can also state some additional trends. Because the risk of phishing and other forms of social engineering is too high, the standalone password-based authentication disappeared. The old generation of One-Time Password (OTP) hardware tokens is quickly diminishing, and because of PSD2, there is a trend in the decreasing number of areas where SMS codes are still applicable for authentication, because in some cases (e.g., banking app and SMS on the same smartphone) SMS no longer provides a secure external channel. PKI has changed its role (user software for endpoint authenticators was abandoned and replaced by usage of TLS and some token authentication schemes). We are also awaiting the spread of a new generation of authentication tokens (based on Universal 2nd Factor – U2F, using Trusted Execution Environment – TEE and wireless communication via Bluetooth or NFC).

In the subsections II-C and II-D we focus on two main attack areas, which we consider the most significant – phishing and malware. The significance of these areas is supported by IBM Threat Intelligence Index [47] and Organization of American States (see Figure 4). Even though both are not new in principle and have been used a decade ago, they have both undergone a big evolution and are still the most serious threat in the sector.

C. PHISHING STILL ON THE RISE

Since the first description of this concept in 1987, phishing has become a well-known social engineering technique for user data exfiltration. Even though the principle hasn't changed, phishing campaigns nowadays are more sophisticated and subliminal than people only ten years ago could imagine. In 2001 phishing campaigns started targeting online payment systems, and its share has increased since. The graph in Figure 5 describes the share of financial (and bank) phishing out of all phishing e-mails detected by Kaspersky Labs. The global reach of phishing attacks is also shown in a longitudinal study by Thomas *et al.* [21].

The main development in financial phishing lies in the use of advanced techniques to imitate official correspondence and exploit user gullibility. Modern phishing e-mails are almost indistinguishable from original e-mails thanks to flawless translations, convincing information and carefully crafted landing pages with nearly identical URLs, often with valid HTTPS certificates. A special category of phishing called "spear phishing" keeps growing. Spear phishing campaigns are strictly targeted at a single user or company, allowing attackers even deeper impersonation of valid correspondence. Spear phishing is used to target companies and their financial departments or as a part of the deployment phase of Advanced Persistent Threats [59].

D. MALWARE NEVER DISAPPEARED

Malware (compound of words *malicious* and *software*) is a general term describing software developed with the intent of harming and exploiting the user and his resources. Since the first malware in 1971, it has undergone significant evolution and remains a constant threat even in the financial market.

Computer malware is a broad set of harmful software targeting operating systems and desktop computer users. Traditional categories are *virus*, *worm*, *trojan*, *rootkit* and *spyware*, but modern malware blurs the differences between them, so novel taxonomies based on behaviour have been created [12], [13], [60].

Modern malware is usually a complex piece of software able to launch or support multiple attacks, including

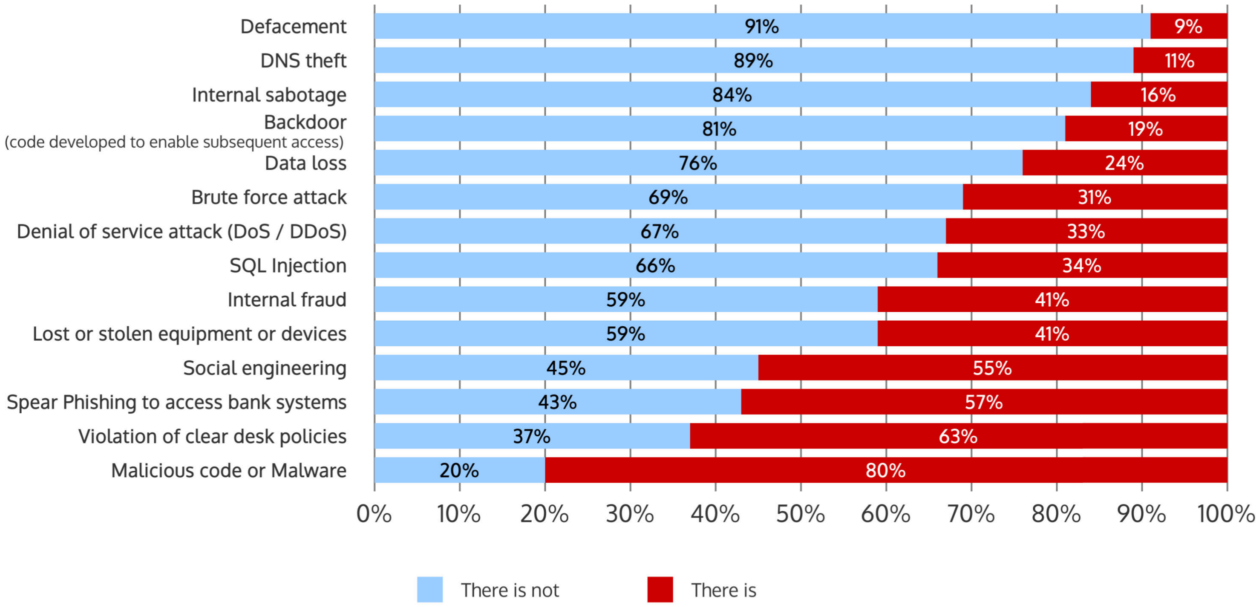


FIGURE 4. Digital security events against banking entities in 2018 in Latin America and the Caribbean [46].

attacks on bank infrastructure, full user simulation or turning the infected system into a remote-controlled bot. Moreover, concurrent malware is often able to deliver additional malware as payload [14]. The development of novel malware never ceases, as is demonstrated in Figure 6 and its detection and mitigation is a never-ending process. We can also notice that recently banking trojans and ransomware have gained popularity, and thus the development of malware from those categories is on the rise.

With the increased versatility and capabilities of mobile phones emerged malware tailored for mobile devices. At first, its use was limited to sending or intercepting messages, but with mobile operating systems (Android, iOS) mobile malware skyrocketed with capabilities similar to computer malware [61]. To describe differences in mobile financial malware more in-depth, Kadir *et al.* suggest a taxonomy of financial malware attacks [15].

Because of the popularity and openness of the Android OS, it became a major target of mobile malware. Android banking trojans are usually injected into a phone by malicious SMS, URL or third-party app stores and installed as APK. Despite all implemented countermeasures, they can also appear in the official app store (Google Play). After installation, they set themselves as a default SMS app or use Accessibility Services to intercept messages (2FA bypass), display phishing screen overlays for banking apps or extract information.

Even though iOS uses a closed ecosystem and APIs, which limit the attack vector for malware at the cost of accessibility and auditability, it is not free of malware as well [62], [63]. But the share of iOS malware is marginal compared to Android because of the difficult spread (tightly controlled App Store) and smaller user base, especially in countries most affected by mobile malware.

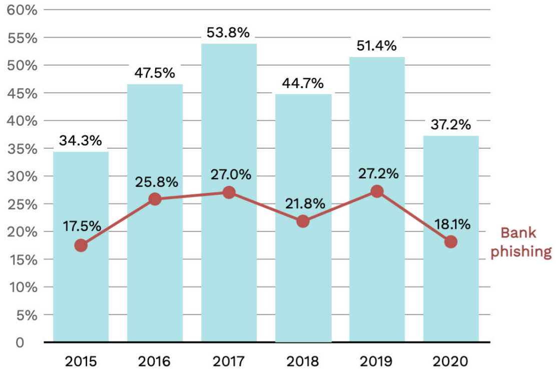


FIGURE 5. Financial phishing share [54]–[58].

The evolution of banking malware led from personal computers targeted malware in 2010 to mobile malware nowadays, thanks to the spread of smartphones and the increasing use of mobile banking. Over the years, the number of people affected by mobile banking malware has skyrocketed and is now catching up with (and sometimes surpassing) computer malware (see graph in Figure 7), so it must be taken into consideration. Malware in general also evolved from single-purpose tools to versatile pieces of code capable of multiple attack scenarios and downloading additional modules or different malware.

In 2018 users affected by Android malware spiked, which was caused mainly by three banking trojan families – Asacub, Agent and Svpeng [56]. The probable reason for this spread is the novel use of DNS hijacking in Android attacks and misuse of Accessibility Services giving the malware superior possibilities.

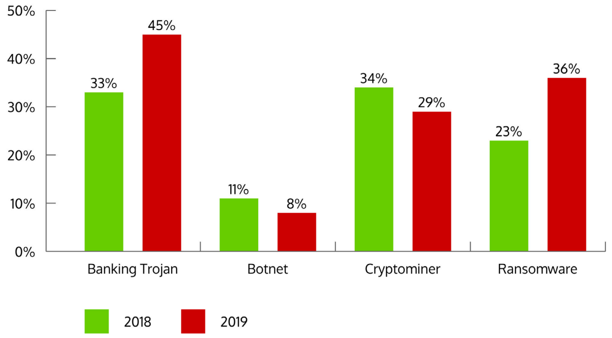


FIGURE 6. Percentage of new (previously unobserved) code by category [47].

III. AUTHENTICATION METHODS

This section contains an overview of elementary authentication primitives and their features. The defined primitives can be used to build any authentication scheme and are compatible with NIST Digital Identity Guidelines [45]. The main contribution of this section lies in the overview of the susceptibility of authentication methods to attacks from the presented taxonomy. Later we take a closer look at secure hardware because of its importance for concurrent e-banking authentication schemes.

The reader should gain a solid overview of authentication primitives whose knowledge is required for a proper understanding of the following section in which we discuss properties of multi-factor authentication and thus combinations of those primitives.

A. AUTHENTICATION PRIMITIVES

In this subsection, we describe the authentication primitives used for user verification in ICT systems. Every authentication scheme can be seen as a combination or specific use of those primitives. Nowadays, direct implementations of those primitives (methods) are not considered inherently secure and to achieve satisfiable security, modern authentication schemes combine multiple methods.

In Table 1, for eight identified authentication primitives, we observe four key characteristics (*replay*, *MITM* and *impersonation* resistances and *dynamic linking*). These characteristics were selected from requirements for authentication methods by NIST [45] based on their relevance to the e-banking area. *Replay resistance* is a fundamental feature preventing an adversary from recording the authentication process and replaying it at a later time, granting him unauthorised access. *Man-In-The-Middle resistance* prevents an adversary from positioning himself in the middle of the authentication process and manipulating the data flow. Under *impersonation resistance* we understand the verifiable identity of the authenticating user and non-repudiation of the authentication process. Lastly, we included *dynamic linking* as one of the key requirements of the PSD2 standard, which will be described later. It describes the ability to use this method to authenticate individual transactions (as opposed to authentication of the session).

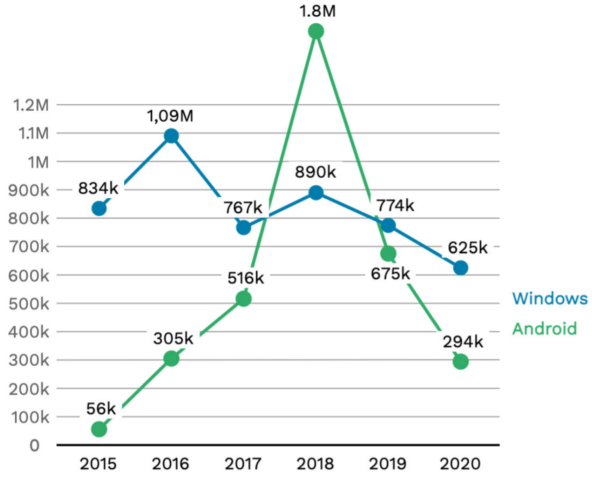


FIGURE 7. Users attacked by banking malware [54]–[58].

The classic approach is utilisation of basic *memorized secret* (**P-MSC**) such as static password or PIN code. Nowadays, we assume the usage of TLS for transport encryption, but despite this, the method shows the weakest resistance against various attacks and alternative authentication options are actively researched [64]. The resistance was greatly improved by facilitating some form of dynamic passwords, and their first usage was in the form of *look-up secrets* (**P-LUS**) represented by grid cards. A grid card contains a matrix of random combinations of alphanumeric characters, and the authentication is based on XY coordinate look-up system [65]. Because of the limited usability as the user was required to look up and construct the code himself, this primitive was practically replaced by HW tokens and SMS codes, where SMS codes are the most commonly used *out-of-band* method (**P-OOB**).

Similarly, the direct *cryptographic authentication* (**P-SCR** and **P-MCR**) based on public key infrastructure (PKI) never spread even though it has superior security features. Mostly because of the usability limits as it imposes the burden of managing private keys on the user. The difference between single-factor (**P-SCR**) and multi-factor (**P-MCR**) lies in the former one keeping the key on the device accessing the e-banking, while the latter keeps the key on a separate device, such as a smart card, which improves not only security but also usability as the key management is usually ensured by the device vendor. The implementations of P-MCR, which we find feasible in the near future, are payment cards, government ID cards, U2F authenticators, and cryptographic smart cards with asymmetric cryptography.

TABLE 1. Overview of authentication primitives and their features.

	Primitive	Replay Resistance	MITM Resistance	Impersonation Resistance	Dynamic Linking
P-MSC	Memorized secret	✗	✗	✗	✗
P-LUS	Look-up secret	✓	✗	✓	✗
P-SCR	Single-factor cryptographic authentication	✓	✓	✓	✓
P-MCR	Multi-factor cryptographic authentication	✓	✓	✓	✓
P-MFO	Multi-factor OTP device	✓	✓	✓	Opt.
P-OOB	Out-of-band authentication	✓	✗	✗	Opt.
P-SFO	Single-factor OTP device	✓	✓	✓	✓
P-BIO	Biometric authentication	✗	✗	✗	✗

Opt. (Optional) - depends on the implementation

The first dynamic password primitive that penetrated the market was *multi-factor OTP device* (**P-MFO**) consisting of various HW tokens generating one-time passwords (OTP). The first generation of those OTP devices was HW tokens with a display (often called ''calculator'') generating passwords that the user had to retype into e-banking. The following generations usually include some interface to ease the password transfer, such as USB or wireless connections (Bluetooth, NFC) or standard chip payment card interface (EMV). The EMV standard for the user and transaction authentication in e-banking is used in MasterCard as the Chip Authentication Program (CAP), while in VISA it is known as Dynamic Passcode Authentication (DPA) [66]. Some tokens lack the display, therefore the ability to show the details of an operation undergoing, and such tokens provide authentication based only on its presence.

The modern trend in OTPs is the usage of mobile applications as dynamic password generator (DPG). This approach simulates HW tokens in SW and often implements a challenge-response protocol. In case this DPG is run as a part of an e-banking application, we talk about a *single-factor OTP device* (**P-SFO**). Much more frequented is a case where this DPG is packed as a separate application, and if we accept that the mobile operating system securely isolates those applications, we can consider it a *multi-factor OTP* despite running on one device. A special case of a multi-factor OTP on one device is the utilisation of a secure enclave – a special cryptographic chip dedicated to key management and isolated from the rest of the system and thus even the e-banking application. We describe the secure enclave in detail in Section III-B.

With the spread of biometric sensors (such as fingerprint readers, face, voice and retina recognition) the inherent features of the user started to be used for authentication (**P-BIO**) as well. The biggest problem of biometrics is uncertainty as the biometrics compares the sensor data with a stored model and returns binary result – verified / unverified – depending on whether the comparison exceeds a given threshold. Because pure biometrics is not resistant against attacks such as replay or MITM, it is usually used bundled with secure key storage or as additional protection of other methods. The main issue of using biometrics as the primary authentication method is that biometric authentication gives local information about successful authentication, but it is difficult to transfer this information to the remote server securely. Existing attempts to achieve secure transfer (called crypto biometrics) are still not mature enough.

After we described the primitives, we take a deeper look at the feasibility of attacks described against selected most frequently used implementations of each authentication primitive (Table 2), which we call authentication methods. It turned out that the NIST requirements are not detailed enough for the assessment of authentication methods in the banking industry. Thus, we have to use more categories (linked to the attacks categories in Figure 3) to describe authentication mechanisms in full detail. It is essential to define the features of individual primitives as later in Section IV we discuss multi-factor authentication, and these features have a direct impact on compounds. If a method is resistant to an attack, then the multi-factor containing this method is resistant as well. E.g., a multi-factor scheme has replay resistance if at least one of the used factors has replay resistance.

In Table 2, we have examples of authentication methods (and their primitives) in the rows and selected attacks in the columns. The attack list is not complete because we omitted the irrelevant attacks (such as the whole category A–4 Attacks Focused on Banks) and merged attacks with the same features (such as Social Engineering based attacks).

We put the check mark if the attack is viable and cross mark for methods resistant to the attack. In case there are some constraints affecting the attack feasibility, we evaluated the most commonly used option and put the result in brackets.

B. SECURE HARDWARE AND ENCLAVE

Secure Hardware (also a trusted device) is a hardware module equipped with a microprocessor containing some security relevant data (keys) and algorithms for manipulating them (see Figure 8). This specialised hardware ensures both logical security by isolating such data from the system and hardware security as these modules are designed to be tamper-proof. The features of secure hardware can be utilised in several ways:

1) Storage of data, which can be manipulated only in a specific way (e.g., counter with only decrease operation allowed).
2) Storage of cryptographic private key allowing only selected operations such as encryption and never revealing the key itself.
3) Operating System integrity, where thanks to secure boot mechanism, secure hardware provides trust that the OS (and its security functions like process separation) hasn't been tampered with.

Ten years ago, in our previous article [1] we stated: "*Most widely use of the trusted device is a smart card. Smart cards offer very cheap implementation of one of the security concepts, and this concept is called tamper-resistant hardware.*" This statement is, obviously, not valid anymore. Nowadays, secure hardware is present in personal computers as Trusted Platform Module (TPM) chips and even in mobile devices as Apple Secure Enclave or Android Keystore System.

With respect to online services, FIDO (Fast IDentity Online) Alliance is the main driving force in the adoption of these technologies. Their protocol FIDO2 [67] consisting of W3C (World Wide Web Consortium) open web standard WebAuthn [68] and CTAP2 is becoming the de facto standard for using secure hardware for authentication in a web environment as it is implemented in all major browsers. While older FIDO standards enabled either local biometrics (FIDO UAF [69]) or Hardware Authentication Module (FIDO U2F [70]) to be used as a second factor for web services, FIDO2 includes support for both second factors and can even be used as a single authentication factor for passwordless authentication.

In general, the secure hardware can be implemented in multiple ways, where some lost the status over time:

Personal Computer (PC) In the infancy of information technology, PCs were considered a trusted device, which protects the interests of the user. The massive spread of malware and cyber-attacks destroyed this principle.

Smartcard Smartcard (electronic payment card or electronic ID card) appears to be an almost ideal implementation of secure hardware. It is cheap, security-hardened, provides cryptographic operations and is easy to manipulate. The main issue is the lack of a user interface requiring a special connector (either hardware or radio) to utilise its features. An example of current usage is CAP/DPA.

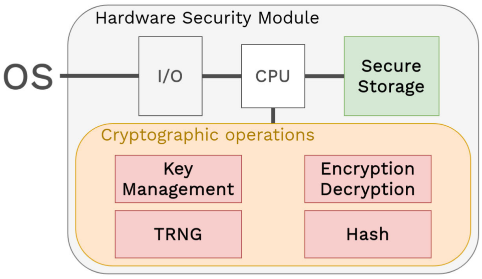


FIGURE 8. High level hardware security module architecture.

Cell phone Traditionally, the cell phone was used as a secure device by utilising the SIM card either through SIM Toolkit or SMS authentication. But SIM Toolkit is scarcely used because of the limitations of SIM card applications and SMS authentication does not comply with new PSD2 factor independence requirements and thus cannot be seen as the primary authentication mechanism in the future.

Hardware Authentication Token Traditionally, these devices take the form of an "authentication calculator". Despite the initial high price and lack of standardisation, they overcame the problems and became the widespread secure device. They declined in favour of SMS authentication but might see a comeback with a new generation featuring wired or wireless interfaces.

Hardware Security Module (HSM) HSM is a separate chip satisfying secure device requirements. It can be added to a system to bring secure device features such as key storage, cryptographic operations and true random number generator. In PCs, it is usually represented by TPM (although it serves mostly as the root of trust and does not offer full HSM possibilities), and in smartphones there are options like Apple Secure Enclave and Android Keystore System, which are discussed in the next subsection.

The modern addition to the secure devices is **Secure Enclave** which implements a trusted execution environment (TEE) concept, where an application is being run isolated from the operating system and protected from outside threats. A secure enclave guarantees confidentiality, integrity, and security for the application running within it [71]. TEE extends the concept of secure devices by allowing us to run arbitrary operations (opposed to a very specific set of operations in HSM) within the device while maintaining a high level of trust and security. Examples of secure enclave technology are Intel® SGX and ARM TrustZone.

1) SECURE HARDWARE IN SMARTPHONES

Because modern banking trends focus on smartphones and mobile banking, we describe the possibilities of smartphones

TABLE 2. Viable attacks on authentication methods.

	Password Guessing	Exhaustive Search	Social engineering (Phishing)	Pharming	Cross-channel attacks	Malware	AI Attacks	Physical credential stealing	Unauthorised Binding	Eavesdropping	Data manipulation	MITM
	A-1.1	A-1.2	A-1.3, A-1.6	A-1.4	A-1.5	A-1.7	A-1.8	A-2.1	A-2.2	A-3.1	A-3.2	A-3.3, A-3.4
Static password, PIN (P-MSC)	✓	✓	✓	✓	N/A	✓	✗	✗	(✓)[1]	(✓)[2]	✓	✓
Grid card (P-LUS)	✗	✓	✓	✓	N/A	✗	✗	✓	(✓)[1]	✗	✓	✓
Dynamic password generator (P-SFO)	✗	✗	✗	(✗)[3]	(✓)[4]	✓	✗	✓	✓	✗	(✗)[3]	(✗)[3]
PKI (P-SCR)	✗	✗	✗	✗	N/A	✓	✗	✗	✓	✗	✗	✗
PKI token (P-MCR)	✗	✗	✗	✗	N/A	✓	✗	✓	✓	✗	✗	✗
SMS Code (P-OOB)	✗	(✓)[5]	✗	✓	✓	✓	✗	✓	✓	✓	(✗)[3]	(✗)[3]
HW tokens with display (P-MFO)	✗	(✓)[5]	✗	(✗)[3]	N/A	✗	✗	✓	✓	✗	(✗)[3]	(✗)[3]
HW tokens without display (P-MFO)	✗	(✓)[5]	✗	✓	N/A	✓	✗	✓	✓	✗	(✗)[3]	(✗)[3]
Secure Enclave (P-MFO)	(✓)[6]	✓	✗	✗	N/A	✓	✗	✓	✓	✗	✗	✗
Biometric authentication (P-BIO)	✗	✗	✗	✓	N/A	✓	✓	✗	✓	(✓)[7]	✓	✓

[1] In case the user can reset the authentication method (password reset, resend grid card) the attack is possible.
[2] If there is no HSTS header set, an adversary might force unencrypted connection.
[3] The method is vulnerable in case that it does not support Dynamic Linking as defined in PSD2 directive.
[4] Applies only if the DPG is running on different device than e-banking (2da – see section IV-B).
[5] In case the length of the one-time cryptogram is not sufficient, and attempts are not limited.
[6] If the enclave is unlocked by a password, an attacker may use password guessing.
[7] Depends on the implementation parameters.

in detail. As we have mentioned in the previous section, SIM card use as a root of trust is becoming obsolete with SIM Toolkit long gone and SMS being replaced by other means, especially HSM seem to be a very promising candidate. Both major smartphone OS's include HSM support.

Apple Secure Enclave (ASE) is Apple's implementation of secure hardware, which (since the iPhone 5) is tightly integrated into iOS due to Apple developing both the iOS operating system and the iPhone hardware. It provides standard HSM features such as secure storage, random number generator and cryptographic operations (key generation, encryption/decryption, and hashing). ASE extends those features with person-to-device authentication using fingerprints, face image or password, as well as secure communication with corresponding sensors [72].

Android Keystore System (AKS) covers the secure hardware in the Android operating system from Google and, unlike ASE, is not bound to a specific HSM as different HW vendors include various modules [73]. Depending on the module provided, AKS supports different security assurance levels starting with *Secure Element*, which can only store cryptographic keys and is represented usually by SIM card or EMV chip. More advanced phones offer true HSM chips not only for key storage but they also include cryptographic operations and random number generators. The most advanced approach offers full TEE able to run custom applications in a secure environment independent from the OS.

IV. MULTI-FACTOR AUTHENTICATION

Multi-factor authentication is a common technique to increase the strength of the authentication process by combining multiple factors (methods). The resulting scheme inherits their properties (such as resilience to specific types of attacks). The usual perception of these combinations must be revised for the e-banking sector due to the new European directives, which bring additional requirements. Thus, some common combinations cease to be viable. The main contribution of this section is our classification of authentication schemes, which is compatible with both the NIST Digital Identity Guidelines [45] and the EU PSD2 regulations [9].

A. PSD2 DIRECTIVE MOTIVATION

The security of authentication in internet banking applications is being pushed forward also by EU activities. The big step came on September 14th, 2019, when the Regulatory Technical Standards (RTS) of the Revised Payment Service Directive (PSD2) started to be mandatory for the EU banks [74]. The end of the migration period for PSD2 Strong Customer Authentication is December 31th, 2020, but every country can temporarily mitigate the effects by delaying the enforcement.

The concepts enforced by PSD2 to the area of client authentication are *two factor authentication* (with requested *factor independence*), *strong customer authentication* (SCA) and the *dynamic linking* of the authentication code to the transaction's beneficiary and amount. The next requirement is *cloning protection* – the ability to withstand memory cloning attacks.

For the sake of completeness, it should be added that as far as regulatory technical standards for strong client authentication and common and secure open communication standards are concerned, PSD2 is further complemented by Commission Delegated Regulation (EU) 2018/389 [75]. In addition to security requirements, it specifies further technical details such as auditability requirements, technology neutrality in the implementation of authentication codes, interface quality requirements, the use of open standards, but also defines exceptions to strong authentication. The exceptions are usually determined by the type of transaction and the presence of additional risk mitigation measures such as low-value contactless payments at the point of sale, which also take into account the maximum number of consecutive transactions.

From our point of view, the two main PSD2 requirements are the factor independence and SCA that together form the requirements for multi-factor authentication (covered later in this section). But it is necessary to say that the SCA requirement is stronger than the general requirement for multi-factor authentication because SCA requires at least two methods from the exact list of primitive categories and not an arbitrary combination of methods. Cloning protection is an additional security requirement, and dynamic linking only allows usage for other purposes than pure authentication.

1) STRONG CUSTOMER AUTHENTICATION (SCA)

The term Strong Customer Authentication is defined in the document *Directive (EU) 2015/2366 of the European Parliament and of the Council* [9]. The definition in *Article 4 – Definitions* paragraph 30 states: "*"strong customer authentication" means an authentication based on the use of two or more elements categorised as knowledge (something only the user knows), possession (something only the user possesses) and inherence (something the user is) that are independent, in that the breach of one does not compromise the reliability of the others, and is designed in such a way as to protect the confidentiality of the authentication data.*"

2) FACTOR INDEPENDENCE

Furthermore, *Article 9 - Independence of the elements* of *Commission Delegated Regulation (EU) 2018/389* [75] supplementing the same regulation defines that:

1. *Payment service providers shall ensure that the use of the elements of strong customer authentication referred to in Articles 6, 7 and 8 is subject to measures which ensure that, in terms of technology, algorithms and parameters, the breach of one of the elements does not compromise the reliability of the other elements.*
2. *Payment service providers shall adopt security measures, where any of the elements of strong customer authentication or the authentication code itself is used through a multi-purpose device, to mitigate the risk which would result from that multi-purpose device being compromised.*

3. *For the purposes of paragraph 2, the mitigating measures shall include each of the following:*
   (a) *the use of separated secure execution environments through the software installed inside the multi-purpose device;*
   (b) *mechanisms to ensure that the software or device has not been altered by the payer or by a third party;*
   (c) *where alterations have taken place, mechanisms to mitigate the consequences thereof.*

3) CLONING PROTECTION

The additional requirements for authentication defined in Article 7 - *Requirements of the elements categorised as possession* of the same supplement are particularly relevant for mobile devices. This article says that " *The use by the payer of those elements shall be subject to measures designed to prevent replication of the elements.* [75]"

4) DYNAMIC LINKING

The term dynamic linking is defined in Article 5 of the supplement as well, and it states that payment transaction details should be protected and tied to authentication:

1. *Where payment service providers apply strong customer authentication in accordance with Article 97(2) of Directive (EU) 2015/2366, in addition to the requirements of Article 4 of this Regulation, they shall also adopt security measures that meet each of the following requirements:*
   (a) *the payer is made aware of the amount of the payment transaction and of the payee;*
   (b) *the authentication code generated is specific to the amount of the payment transaction and the payee agreed to by the payer when initiating the transaction;*
   (c) *the authentication code accepted by the payment service provider corresponds to the original specific amount of the payment transaction and to the identity of the payee agreed to by the payer;*
   (d) *any change to the amount or the payee results in the invalidation of the authentication code generated.*

The PSD2 requirements for Strong Customer Authentication and Factor independence make an urgent demand for the new group of authentication mechanisms called multi-factor authentication mechanisms. Thus, in the following section, we will prospect viable combinations of authentication methods and how these combinations (called authentication schemes) will meet the PSD2 requirements.

B. MULTI-FACTOR ASSESSMENT

In Table 3, we describe features of common combinations of authentication methods. The reader's main takeaway is the quick overview of the possibilities and their features with respect to the PSD2 requirements, which we consider the most advanced in the e-banking area. The table uses the same notation as previous tables, where checkmark denotes that the combination of methods satisfies the requirement and cross mark that it doesn't. If the mark is in brackets, the assessment is not unambiguous, in which case we used the more common rating and added the condition in the footnote. In case the feature can be enabled possible by some additional adjustment of the basic scheme, we use 'Opt.' as in the optional feature.

For further ease of understanding, we state the category of the given scheme as defined by Frederik Mennes in his SCA requirement analysis [76]. There are four categories based on the segregation of the factors:

1aa (one-app-authentication) describes the e-banking apps with built-in authenticators
2aa (two-app-authentication) means both authentication and e-banking apps are separate apps
2da (two-device-authentication) extracts authentication to a separate device
oob (out-of-band) uses a third party (such as telco service) for authentication

C. TRENDS IN MULTI-FACTOR AUTHENTICATION

What we consider interesting is the continuous evolution of a typical e-banking system, especially in the authentication and authorisation area. It moved from *password-based* authentication, over HW tokens and SMS codes, to currently used *Dynamic passwords* generated by mobile applications.

We can essentially divide authentication schemes into four categories based on their viability. The first category, we call it *Schemes on retreat*, contains an ever-growing group of deprecated schemes used from the beginning of the e-banking era. The schemes that declined in use but with some adaptation or under special conditions can be used again belong to the second category - *Reincarnating schemes*. Then, described in *Schemes still here with us*, we have a group of robust schemes which keep their properties over the span of time. The last category contains the schemes based on novel approaches, which are emerging in the e-banking area, and we call them *Schemes on the rise*.

1) SCHEMES ON RETREAT

The classic method used in early authentication schemes was *password* authentication. Because of the lack of resistance against phishing and replay attacks, it was soon replaced. The combination with *PIN* did not bring much improvement as both are static, memorable secrets, but the combination with *Grid cards* brings replay resistance and force the adversary to recover (usually by phishing) a substantial part of the grid card in order to perform an attack.

Later the usage of *SMS codes* used telecommunication services for OTP delivery, which improved usability and shifted the security from banks to the telco providers (attacks such as SIM Swapping and SS7 attacks became relevant). The decline of this scheme was brought by the spread of mobile banking and the factor independence requirement of PSD2. If the user

TABLE 3. PSD2 features of authentication methods and their combinations.

Method combination	Cloning protection	Factor independence	Dynamic linking	SCA	Comment
Password	✗	✗	✗	✗	
Password + PIN	✗	✗	✗	✗	
Password + Grid card	✗	✓	✗	(✓)[1]	2da
PKI (private key) protected by password	✓	✗	✓	✗	
HW Token (MCR)	✓	✓	✓	✓	2da
HW Token protected by PIN	✓	✓	✓	✓	2da
HW Token protected by BIO	✓	✓	✓	✓	2da
SMS	(✓)[2]	✗	Opt.	✗	oob
Password + SMS	(✓)[2]	✓	Opt.	(✓)[3]	oob
Integrated DPG protected by password (SFO)	(✗)[4]	(✗)[5]	✓	(✗)[5]	1aa
Separated DPG protected by password (MFO)	(✗)[4]	(✓)[4]	✓	(✓)[4]	2aa
Dynamic password + BIO	(✗)[4]	✓	✓	(✓)[4]	2aa
BIO	✗	✗	✗	✗	
BIO + Password	✗	✓	✗	✓	
BIO + SMS	(✓)[2]	✓	Opt.	✓	oob
PKI (private key) protected by BIO	(✗)[4]	✗	✓	✗	
Secure Enclave protected by BIO	✓	✓	✓	✓	
Secure Enclave protected by password	✓	✓	✓	✓	

[1] Technically could be considered valid, but in reality is not used as such.
[2] Indirectly possible by attacks on telecommunication links such as SIM Swapping and SS7 attacks.
[3] To fulfil SCA requirements, the SMS receiving device have to be independent of the other one where the password is entered. Nowadays, this is difficult to achieve.
[4] Depends on OS capabilities, in case of rooted OS (called jailbreak in iOS) cannot be ensured [78]. E.g. biometrics is cloneable by design, and protection depends on second factor properties.
[5] The satisfiability of factor independence is not decided yet; if it does not satisfy factor independence, it isn't SCA.

receives the SMS code on the same device where the mobile banking is running, the malware can compromise both at once, breaking the requirement.

These schemes by themselves cannot satisfy the SCA requirements, but methods used in these schemes can be combined to create more resilient schemes. Such an example is the combination of SMS with a password which would comply with the SCA.

2) REINCARNATING SCHEMES

A few years ago, the *Hardware Tokens* were the most spread method for authentication. They represent the first true OTP systems, and as every code is unique and generated by a specially crafted single-purpose device, it provides very high security. The tokens started declining because of the usability constraints as they required users to carry an extra device and manually transfer the generated code into the e-banking system. Despite the fact that they were superseded by *SMS codes*, the situation changes because, unlike SMS, they easily satisfy the SCA requirements of PSD2, and the burden of manual code transfer is overcome by the new generation of HW Tokens with NFC/Bluetooth technologies. Furthermore, the HW tokens are usually protected by passwords or biometrics, which mitigate the risk of theft.

3) SCHEMES STILL HERE WITH US

The schemes based on *PKI* method use the full power of modern cryptography directly, which make them very resilient. The attempts of using those schemes have been present since the beginning of e-banking, but overhead and lack of usability for a common user prevented their spread. Nevertheless, the PKI found its use in HW Tokens and smart cards. The private key is usually further protected by encryption and some other factors such as passwords or biometrics.

The schemes based on PKI can be SCA compliant, but they are strongly dependent on the usage because if used incorrectly, they do not satisfy the factor independence.

4) SCHEMES ON THE RISE

The advances in technology enabled new methods to emerge. The most apparent is the spread of *Biometrics* for authentication, which prevailed mostly in smartphones despite the fact that standalone biometrics is not mature enough yet and can thus be used only in combination with another method. The common use of biometrics is strengthening the KYC process for remote customer verification or device authorisation when using a hardware token, a dynamic password generator or secure enclave.

Secure Enclave is a relatively recent addition to the authentication methods, which was briefly covered in Section III-B. A secure enclave is usually protected by other authentication methods (password, PIN or biometrics) and, apart from providing key storage and cryptographic operations, is used as a root of trust and checks the integrity of operating systems and applications. These integrity features are used to defend the factor independence of standalone e-banking and authentication apps.

The typical e-banking system nowadays utilises a mobile application generating *Dynamic passwords*, which is usually protected by either user *password* or *biometrics* if the smartphone supports it. In case this application is distinct from the mobile banking application, and we trust the mobile operating system to isolate contexts of different applications properly, this approach satisfies factor independence condition and is thus preferred by EU banks.

FIDO2 is the concurrent standard for including biometrics and hardware tokens for online authentication. However, despite its spread in fintech and other web services, the adoption in e-banking is still in its infancy. Examples of banks already including it are Bank of America[1] (member of FIDO Alliance) or Boursorama Banque.[2]

5) ADDITIONAL CLARIFICATION

In our work, we have described only two-factor combinations and not the general multi-factor. We decided to narrow this area for the sake of clarity and because multiple factors are usually used chained, where one factor is strengthened by another such as HW Token as a second factor further protected by a PIN. Furthermore, multi-factor combinations of third and higher order usually do not satisfy the factor independence, and their security features require deeper analysis.

The class *Integrated DPG protected by password* is the subject of research and discussions as it is unclear whether a DPG included within mobile banking application satisfies the SCA requirements [77]. To be considered SCA compliant, such an application has to meet multiple conditions such as *secure device boot*, *secure checking of application integrity* to ensure the application hasn't been tampered with and strong protection against attacks from other applications. In the now prevalent mobile environment, it is usually required that the device is not rooted (or jailbroken), which would allow full access to all application contents breaking cloning protection and factor independence at once. The solution for the future is protecting the application against attacks from the operating system by utilising secure hardware.

D. BROADER PERSPECTIVE

Our primary focus is on authentication methods and schemes. However, there are many other methods to increase security or, specifically in the e-banking area, to reduce the risk of fraud. The output of these methods is usually a score representing the level of the risk (which can be calculated by combining multiple sources), so we consider them as risk management tools. However, they cannot be used for authentication on their own, nor in combination with another factor. One of the used principles is continuous authentication, which constantly monitors selected parameters. For example, it can enhance biometric authentication by continuous sample evaluation (face is constantly scanned by a camera) or analyse behavioural metrics such as user behavioural patterns (e.g. payer/payee location, spending habits). In some cases (such as low-risk transactions), an appropriate combination of score sources could be sufficient to replace SCA in e-banking [75].

In a wider context, with the standardisation of authentication requirements under PSD2, the focus shifts on setting up and consolidating e-identity (eID) systems because they are seen as an important element of future payment systems. It remains an open question what role banks should play within these systems. In Europe, two different approaches can

[1] https://www.bankofamerica.com/security-center/online-mobile-banking-privacy/usb-security-key/

[2] https://www.boursorama.com/aide-en-ligne/mon-espace-client/identifiant-et-mot-de-passe/question/en-quoi-consiste-la-connexion-par-cle-de-securite-sur-internet-5165516

be distinguished – bank-driven eID, where banks are identity providers (e.g. Norway, Sweden or the Czech Republic), and government-driven eID, where banks are mere consumers (e.g. Belgium and Estonia) [78]. Given the potential consolidation of eIDs, PSD2 SCA requirements (as well as other requirements) may impact other e-services, so a thorough evaluation of authentication methods will have wider implications.

V. RELATED WORK

While there are many manuscripts dealing with e-banking security, they usually either broadly cover the topic for business executives [10], [79], focus on a narrow area (e.g. specific types of systems [11], malware [12], [14], usability [80]) or they are location specific (e.g. Switzerland [2], Brazil [3], India [81], Nigeria [82]). On the other hand, in the area of authentication, most of the literature cover general problems [83]–[85]. This fragmentation makes it difficult to develop a coherent view of the current state of the art.

We have also found that a number of papers are outdated and need to be revised to take into account new approaches and regulations such as PSD2.

In related work, we focus on two main areas: attack taxonomies and authentication methods used in e-banking. We also add a third area where we briefly mention usability aspects.

A. ATTACK TAXONOMY

Many works have addressed the classification of attacks in the online environment. There are general overviews, such as Authenticator Threat/Attack from NIST [45], but more common are more narrowly focused taxonomies. Examples of such could be Android financial malware attack taxonomy by Kadir *et al.* [15], banking trojans taxonomy by Kiwia *et al.* [13], or phishing attacks classification by Gupta *et al.* [86].

In our work, we introduce a high-level attack taxonomy tailored specifically for e-banking (as opposed to NIST) while covering the entire domain. We believe that this taxonomy will provide a better understanding of the problem and will comprehensively link existing narrowly focused classifications in the e-banking domain.

B. AUTHENTICATION METHODS IN E-BANKING

The available research in authentication methods in e-banking lately consists of analysing solutions deployed by selected banks and their features. We can see the great variability of the area as virtually every bank develops its authentication solutions independently.

Chaimaa *et al.* [87] recently published an overview of e-banking services, summarising the available research, based mostly on outdated papers (5 years old and more). The grasp of this article is very superficial, while in contrast, our article provides a more comprehensive and deeper insight.

Kiljan *et al.* conducted a survey about the usage of authentication schemes in online banking [88] by actively researching 80 banks across the world. They found that most banks use passwords for single-factor authentication, while PIN is used in multi-factor schemes. SMS has gained popularity, but they already argue about the security of connected (online) possession factors. They mention behaviour anomaly detection as a form of biometrics, but as this factor is fully implemented in the backend, they have not been able to verify its use. The conclusion is that the adoption of multi-factor authentication has increased in all regions except North America and that authentication and transaction authorisation need to be unified across the banking area.

One of the main works in the e-banking authentication area is a survey published by Sinigaglia *et al.* [89], which reviews the EU regulations and strong authentication mechanism implemented by 26 major EU and non-EU banks for their online payment systems. The analysis is based on available public bank documentation. One of their key findings is a diversity of implementations which opens a large attack surface and observation that mobile devices became privileged targets and single point of failure.

Sinigaglia *et al.* further extend their work with a stronger focus on multi-factor authentication [90] while distinguishing between the internet and mobile payments. Again they took a sample of EU and non-EU banks and evaluated their multi-factor authentication with respect to the existing regulations and best practices, security, and complexity. They conduct a comparison of available documents in terms of requirements and analyse the applicability of seven attacker models to identified authenticators. For all included banks, they describe an overview of deployed authentication protocols, their compliance, susceptibility to attacks, and complexity.

In comparison with our work, due to frequent changes, we neglect specific technical solutions and focus on general authentication principles and their compliance with regulations. Additionally, we also propose a complex e-banking attack taxonomy and a more detailed analysis of authentications mechanisms vulnerabilities.

C. USABILITY

The usability of authentication methods and other e-banking solutions is an important parameter for user adoption. Although we do not address this topic in our paper, we consider it important for a comprehensive understanding of the context. We selected two papers to demonstrate the acceptability of more complex multi-factor authentication solutions by the public.

MFA authentication has become more accepted by users, as shown, for example, by Althobaiti and Mayhew [91]. They conducted a survey among 302 e-banking users and concluded that users feel confident using tokens, which they perceive as more secure.

Lyastani *et al.* [92] show that FIDO has a great potential for web user authentication due to its high usability, which users consider more acceptable than password-based authentication.

VI. CONCLUSION

In the paper, we cover the state of the current e-banking authentication area to enable the reader to have an easy orientation in the problem. The existing materials usually cover the general authentication mechanisms and do not focus on the e-banking specifics, or, on the other hand, are too specific and detailed to provide a comprehensive overview. Our paper suggests a taxonomy for attacks on e-banking compatible with general authentication taxonomy by NIST [45] and a comprehensive overview of authentication schemes and their resistance against those attack classes.

Because the *Payment Services Directive 2 (PSD2)* by European Union brings important security requirements for the banking area, which we find the most advanced in the world, we discuss security features of authentication schemes in the context of this standard. For every scheme, we discuss the satisfaction of Strong Customer Authentication (SCA) as well as other essential features as described in PSD2.

Moreover, we provide the reader with an informed discussion about trends in multi-factor authentication schemes and conveniently group them into four classes depending on their current usage and future prospects. We point out unresolved issues, especially in the area of the feasibility of mobile devices as a secure element.

The main contribution of the article is a comprehensive overview of authentication schemes and their security evaluation. We emphasize viable combinations of authentication methods concerning the concurrent standards, mainly PSD2.

REFERENCES

[1] P. Hanacek, K. Malinka, and J. Schafer, "E-banking security—A comparative study," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 1, pp. 29–34, Jan. 2010.

[2] Swiss Finance Council. (2020). *Getting Ready for the '20s-Technology and the Future of Global Banking*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.swissfinancecouncil.org/images/SFC_Discussion_Paper_2020.pdf

[3] Deloitte. (Aug. 2020). *FEBRABAN Banking Technology Survey*. Accessed: Jan. 20, 2021. [Online]. Available: https: //www2.deloitte.com/content/dam/Deloitte/br/Documents/financial-services/2020%20FEBRABAN%20Banking%20Technology%20Survey.pdf

[4] Deloitte. (2019). *Global Mobile Consumer Survey: UK Cut*. Accessed: Feb. 15, 2021. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-plateauing-at-the-peak-the-state-of-the-smartphone.pdf

[5] Pew Research Center. (Feb. 2019). *Smartphone Ownership is Growing Rapidly Around the World, but Not Always Equally*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf

[6] EY. (2019). *Global FinTech Adoption Index 2019*. Accessed: Feb. 15, 2021. [Online]. Available: https://assets.ey.com/content/dam/ey-sites/eycom/en_gl/topics/financial-services/ey-global-fintech-adoption-index-2019.pdf

[7] EMVCo. (Sep. 2021). *EMV 3-D Secure Protocol and Core Functions Specification*. Accessed: Dec. 6, 2021. [Online]. Available: https://www.emvco.com/emv-technologies/3d-secure/

[8] Security Standards Council. (May 2018). *Payment Card Industry (PCI) Data Security Standard*. Accessed: Nov. 27, 2020. [Online]. Available: https://www.pcisecuritystandards.org/documents/PCI_DSS_v3-2-1.pdf

[9] E. Union, "Directive (EU) 2015/2366 of the European parliament and of the council," *Off. J. Eur. Union*, vol. 337, pp. 35–127, Nov. 2015.

[10] L. Pascu. (2018). *Top Security Challenges for the Financial Services Industry in 2018*. Bitdefender. Accessed: Feb. 15, 2021. [Online]. Available: https://www.bitdefender.com/files/News/CaseStudies/study/240/Bitdefender-Top-Security-Challenges-for-the-Financial-Whitepaper-EN-interactive.pdf

[11] F-Secure. *Threat Analysis: SWIFT Systems and the SWIFT Customer Security Program*. Accessed: Nov. 13, 2020. [Online]. Available: https://www.f-secure.com/content/dam/f-secure/en/business/common/collaterals/f-secure-threat-analysis-swift.pdf

[12] M. A. Kazi, S. Woodhead, and D. Gan, "A contemporary taxonomy of banking malware," in *Proc. Int. Conf. Sci. Comput. Cryptogr.*, Dec. 2018, p. 7. [Online]. Available: https://www.researchgate.net/publication/344017237_A_Contempory_Taxonomy_of_Banking_Malware

[13] D. Kiwia, A. Dehghantanha, K.-K.-R. Choo, and J. Slaughter, "A cyber kill chain based taxonomy of banking Trojans for evolutionary computational intelligence," *J. Comput. Sci.*, vol. 27, pp. 394–409, Jul. 2018.

[14] P. Black, I. Gondal, and R. Layton, "A survey of similarities in banking malware behaviours," *Comput. Secur.*, vol. 77, pp. 756–772, Aug. 2018.

[15] A. F. A. Kadir, N. Stakhanova, and A. A. Ghorbani, "Understanding Android financial malware attacks: Taxonomy, characterization, and challenges," *J. Cyber Secur. Mobility*, vol. 7, no. 3, pp. 1–52, Jul. 2018.

[16] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," in *Proc. 25th USENIX Conf. Secur. Symp.* Berkeley, CA, USA: USENIX Association, 2016, pp. 175–191.

[17] R Hranický, L Zobal, O Ryšavý, and D Kolár, "Distributed password cracking with BOINC and hashcat," *Digit. Invest.*, vol. 30, pp. 161–172, Sep. 2019.

[18] K. D. Nguyen, H. Rosoff, and R. S. John, "Valuing information security from a phishing attack," *J. Cybersecur.*, vol. 3, no. 3, pp. 159–171, Nov. 2017.

[19] G. Ollmann. (Jul. 2005). *The Pharming Guide*. Whitepaper. Accessed: Feb. 15, 2021. [Online]. Available: https://research.nccgroup.com/wp-content/uploads/2020/07/thepharmingguide.pdf

[20] C. Karlof, U. Shankar, J. D. Tygar, and D. Wagner, "Dynamic pharming attacks and locked same-origin policies for web browsers," in *Proc. 14th ACM Conf. Comput. Commun. Secur. (CCS)*. New York, NY, USA: ACM, 2007, pp. 58–71.

[21] K. Thomas, F. Li, A. Zand, J. Barrett, J. Ranieri, L. Invernizzi, Y. Markov, O. Comanescu, V. Eranti, A. Moscicki, D. Margolis, V. Paxson, and E. Bursztein, "Data breaches, phishing, or malware? Understanding the risks of stolen credentials," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: ACM, Oct. 2017, pp. 1421–1434.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[23] J. E. Tapia and C. Arellano, "Soft-biometrics encoding conditional GAN for synthesis of NIR periocular images," *Future Gener. Comput. Syst.*, vol. 97, pp. 503–511, Aug. 2019.

[24] K. Lee, B. Kaiser, J. Mayer, and A. Narayanan, "An empirical study of wireless carrier authentication for SIM swaps," in *Proc. 16th Symp. Usable Privacy Secur.* Berkeley, CA, USA: USENIX Association, Aug. 2020, pp. 61–79.

[25] K. Ullah, I. Rashid, H. Afzal, M. M. W. Iqbal, Y. A. Bangash, and H. Abbas, "SS7 vulnerabilities—A survey and implementation of machine learning vs rule based filtering for detection of SS7 network attacks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1337–1371, 2nd Quart., 2020.

[26] W. Yang, Z. Zheng, G. Chen, Y. Tang, and X. Wang, "Security analysis of a distributed networked system under eavesdropping attacks," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 7, pp. 1254–1258, Jul. 2020.

[27] E. Cronin, M. Sherr, and M. A. Blaze, "On the reliability of current generation network eavesdropping tools," *Int. Fed. Inf. Process.*, vol. 222, pp. 199–214, Jan. 2006.

[28] Y. Zeng and R. Zhang, "Wireless information surveillance via proactive eavesdropping with spoofing relay," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 8, pp. 1449–1461, Dec. 2016.

[29] D. Li, H. Zhou, and W. Yang, "Privacy-preserving consensus over a distributed network against eavesdropping attacks," *Electronics*, vol. 8, no. 9, p. 966, Aug. 2019.

[30] J. Fuller, B. Ramsey, J. Pecarina, and M. Rice, "Wireless intrusion detection of covert channel attacks in ITU-T G.9959-based networks," in *Proc. 11th Int. Conf. Cyber Warfare Secur. (ICCWS)*, 2016, pp. 137–145.

[31] Tetra Defense. (Jan. 2019). *Data Manipulation: A Rising Trend in Cyberattacks, and How to Address it*. Accessed: Dec. 3, 2020. [Online]. Available: https://www.tetradefense.com/incident-response-services/data-manipulation-a-rising-trend-in-cyberattacks-and-how-to-address-it/

[32] S. Sridhar and G. Manimaran, "Data integrity attacks and their impacts on SCADA control system," in *Proc. IEEE PES Gen. Meeting*, Jul. 2010, pp. 1–6.

[33] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2027–2051, 3rd Quart., 2016.

[34] M. Knežević, S. Tomović, and M. J. Mihaljević, "Man-in-the-middle attack against certain authentication protocols revisited: Insights into the approach and performances re-evaluation," *Electronics*, vol. 9, no. 8, p. 1296, Aug. 2020.

[35] F. Callegati, W. Cerroni, and M. Ramilli, "Man-in-the-middle attack to the HTTPS protocol," *IEEE Secur. Privacy*, vol. 7, no. 1, pp. 78–81, Jan./Feb. 2009.

[36] The OWASP® Foundation. (2006). *Man-in-the-Browser Attack*. Accessed: Nov. 24, 2020. [Online]. Available: https://owasp.org/www-community/attacks/Man-in-the-browser_attack

[37] J. Petters. (Sep. 2020). *What is an Insider Threat? Definition and Examples*. Accessed: Nov. 12, 2020. [Online]. Available: https://www.varonis.com/blog/insider-threats/

[38] J. A. Hill, "SWIFT bank heists and article 4A," *J. Consum. Commercial Law*, vol. 22, no. 1, pp. 1–7, 2018.

[39] R. Cohen and D. Walkowski. (Aug. 2019). *Banking Trojans: A Reference Guide to the Malware Family Tree*. Accessed: Oct. 20, 2020. [Online]. Available: https://www.f5.com/labs/articles/education/banking-trojans-a-reference-guide-to-the-malware-family-tree

[40] A. Mauraya, N. Kumar, A. Agrawal, and R. Khan, "Ransomware: Evolution, target and safety measures," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 80–85, 2017.

[41] The OWASP® Foundation. (2017). *OWASP Top Ten*. Accessed: Nov. 25, 2020. [Online]. Available: https://owasp.org/www-project-top-ten/2017/

[42] D. Watson, "Web application attacks," *Netw. Secur.*, vol. 2007, no. 10, pp. 10–14, Oct. 2007.

[43] Kaspersky. (2017). *What is a DDoS Attack?-DDoS Meaning*. Accessed: Nov. 9, 2020. [Online]. Available: https://www.kaspersky.com/resource-center/threats/ddos-attacks

[44] T. Mahjabin, Y. Xiao, G. Sun, and W. Jiang, "A survey of distributed denial-of-service attack, prevention, and mitigation techniques," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 12, Dec. 2017, Art. no. 155014771774146.

[45] P. A. Grassi, E. M. Newton, R. Perlner, A. Regenscheid, J. Fenton, W. Burr, J. Richer, N. Lefkovitz, J. Danker, Y.-Y. Choong, K. Greene, and M. Theofanos, "Digital identity guidelines: Authentication and lifecycle management," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, NIST Special Publication, Tech. Rep., 800-63B, Jun. 2017.

[46] Organization of American States. (Sep. 2018). *State of Cybersecurity in the Banking Sectorin Latin America and the Caribbean*. Accessed: Feb. 15, 2021. [Online]. Available: http://www.oas.org/es/sms/cicte/sectorbancarioeng.pdf

[47] IBM. (2020). *IBM X-Force Threat Intelligence Index*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.ibm.com/account/reg/signup?formid=urx-42703

[48] Accenture. (2019). *Future Cyber Threats*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.accenture.com/_acnmedia/pdf-100/accenture_fs_threat-report_approved.pdf

[49] Accenture. (2018). *Phishing as a Service: The Phishing Landscape*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.accenture.com/t00010101T000000Z_w_/gb-en/_acnmedia/PDF-71/Accenture-Phishing-As-Service.pdf

[50] E. Mikalauskas. (Sep. 2020). *Report: Buying Your Own Malware has Never Been Easier*. Accessed: Feb. 15, 2021. [Online]. Available: https://cybernews.com/security/buying-your-own-malware-has-never-been-easier/

[51] H. Poston. (2020). *Cybercrime at Scale: Dissecting a Dark Web Phishing Kit*. Infosec. Accessed: Feb. 15, 2021. [Online]. Available: https://resources.infosecinstitute.com/cybercrime-at-scale-dissecting-adark-web-phishing-kit/

[52] A. Lakhani. (Jul. 2020). *How Threat Researchers Leverage the Darknet to Stay Ahead of Cyber Threats*. Fortinet. Accessed: Feb. 15, 2021. [Online]. Available: https://www.fortinet.com/blog/threat-research/howthreat-researchers-leverage-darknet-to-stay-ahead-of-cyber-threats

[53] Accenture. (2019). *Know Your Threat: AI is the New Attack Surface*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.accenture.com/_acnmedia/Accenture/Redesign-Assets/DotCom/Documents/Global/1/Accenture-Trustworthy-AI-POV-Updated.pdf

[54] Kaspersky. (Feb. 2017). *Financial Cyberthreats in 2016*. Accessed: Sep. 8, 2020. [Online]. Available: https://securelist.com/financial-cyberthreats-in-2016/

[55] Kaspersky. (Feb. 2018). *Financial Cyberthreats in 2017*. Accessed: Sep. 8, 2020. [Online]. Available: https://securelist.com/financial-cyberthreats-in-2017/

[56] Kaspersky. (Mar. 2019). *Financial Cyberthreats in 2018*. Accessed: Sep. 8, 2020. [Online]. Available: https://securelist.com/financial-cyberthreats-in-2018/

[57] Kaspersky. (Apr. 2020). *Financial Cyberthreats in 2019*. Accessed: Sep. 8, 2020. [Online]. Available: https://securelist.com/financial-cyberthreats-in-2019/

[58] Kaspersky. (Mar. 2021). *Financial Cyberthreats in 2020*. Accessed: Jan. 5, 2021. [Online]. Available: https://securelist.com/financial-cyberthreats-in-2020/

[59] Z. Bederna and T. Szadeczky, "Cyber espionage through Botnets," *Secur. J.*, vol. 33,s pp. 43–62, Sep. 2019.

[60] A. R. A. Grégio, V. M. Afonso, D. S. F. Filho, P. L. D. Geus, and M. Jino, "Toward a taxonomy of malware behaviors," *Comput. J.*, vol. 58, no. 10, pp. 2758–2777, Oct. 2015.

[61] A. Qamar, A. Karim, and V. Chang, "Mobile malware attacks: Review, taxonomy & future directions," *Future Gener. Comput. Syst.*, vol. 97, pp. 887–909, Aug. 2019.

[62] Kaspersky. (Jul. 2020). *Mobile Security: Android vs iOS-Which One is safer?*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.kaspersky.com/resource-center/threats/android-vs-iphone-mobile-security

[63] D. Morán. (Oct. 2019). *Analyzing the Risk of Banking Malware in Android vs. iOS*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.buguroo.com/en/labs/analyzing-the-risk-of-banking-malware-in-android-vs-ios

[64] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 553–567.

[65] B. B. Balilo, Jr., B. D. Gerardo, and R. P. Medina, "A comparative analysis and review of OTP grid authentication scheme: Development of new scheme," *Int. J. Sci. Res. Publications*, vol. 7, no. 11, pp. 1–5, 2017.

[66] J. van den Breekel, D. A. Ortiz-Yepes, E. Poll, and J. de Ruiter, "EMV in a nutshell," Radboud Univ. Nijmegen, Nijmegen, The Netherlands, Tech. Rep., 2016. [Online]. Available: http://www.cs.ru.nl/~erikpoll/publications/EMVtechreport.pdf

[67] FIDO Alliance. (2019). *FIDO2: WebAuthn & CTAP*. Accessed: Nov. 26, 2021. [Online]. Available: https://fidoalliance.org/fido2/

[68] World Wide Web Consortium. (2021). *Web Authentication: An API for Accessing Public Key Credentials Level 2*. Accessed: Dec. 2, 2021. [Online]. Available: https://www.w3.org/TR/webauthn-2/

[69] FIDO Alliance. (2020). *FIDO UAF Architectural Overview*. Accessed: Nov. 26, 2021. [Online]. Available: https://fidoalliance.org/specs/fido-uaf-v1.2-ps-20201020/fido-uaf-overview-v1.2-ps-20201020.html

[70] FIDO Alliance. (2017). *Universal 2nd Factor (U2F) Overview*. Accessed: Nov. 26, 2021. [Online]. Available: https://fidoalliance.org/specs/fidou2f-v1.2-ps-20170411/fido-u2f-overview-v1.2-ps-20170411.html

[71] N. Vaish. (Jul. 2019). *Why are Enclaves Taking Over the Security World?*. Accessed: Dec. 7, 2020. [Online]. Available: https://fortanix.com/blog/2019/07/why-are-enclaves-taking-over-security-world/

[72] Apple Inc. (2017). *iOS Security*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.apple.com/kr/business-docs/iOS_Security_Guide.pdf

[73] Google. *Android Keystore System*. Accessed: Sep. 21, 2021. [Online]. Available: https://android-doc.github.io/training/articles/keystore.html

[74] European Banking Authority. (Oct. 2019). *Opinion of the European Banking Authority on the Deadline for the Migration to SCA for E-Commerce Card-Basedpayment Transactions*. Accessed: Dec. 13, 2020. [Online]. Available: https://eba.europa.eu/sites/default/documents/files/documents/10180/2622242/e8b3ec84-c1c6-4e9a-96ea-3575361dc230/Opinion%20on%20the%20dead line%20for%20the%20migration%20to%20SCA.pdf

[75] E. Union, "Commission delegated regulation (EU) 2018/389," *Off. J. Eur. Union*, vol. 69, pp. 23–43, Nov. 2017.

[76] F. Mennes. (Apr. 2017) *PSD2: Which Strong Authentication and Risk Analysis Solutions Comply With the E's Final Draft RTS?*. Accessed: Nov. 12, 2020. [Online]. Available: https://frederikmennes.wordpress.com/2017/04/19/psd2-which-strong-authentication-and-risk-analysis-solutions-comply-with-the-ebas-final-draft-rts/

[77] V. Haupert and T. Müller, "On app-based matrix code authentication in online banking," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, 2018, pp. 149–160.

[78] ENISA. (Mar. 2020). *eIDAS Compliant eID Solutions*. [Online]. Available: https://www.enisa.europa.eu/publications/eidas-compliant-eid-solutions/@@download/fullReport

[79] I. Pollari, C. Bekker, and C. Jowell. (2019). *The Future of Digital Banking: Banking in 2030*. KPMG. Accessed: Jul. 1, 2020. [Online]. Available: https://home.kpmg/au/en/home/insights/2019/07/future-of-digital-banking-in-2030.html

[80] K. Reese, T. Smith, J. Dutson, J. Armknecht, J. Cameron, and K. Seamons, "A usability study of five Two-Factor authentication methods," in *Proc. 15th Symp. Usable Privacy Secur. (SOUPS)*. Santa Clara, CA, USA: USENIX Association, Aug. 2019, pp. 357–370. [Online]. Available: https://www.usenix.org/conference/soups2019/presentation/reese

[81] M. Kumar and S. Gupta, "Security perception of e-banking users in India: An analytical hierarchy process," *Banks Bank Syst.*, vol. 15, no. 1, pp. 11–20, Feb. 2020, doi: 10.21511%2Fbbs.15%281%29.2020.02.

[82] O. Sarjiyus, N. D. Oye, and B. Y. Baha, "Improved online security framework for e-banking services in Nigeria: A real world perspective," *J. Sci. Res. Rep.*, vol. 6, pp. 1–14, Apr. 2019.

[83] A. Ometov, S. Bezzateev, N. Mäkitalo, S. Andreev, T. Mikkonen, and Y. Koucheryavy, "Multi-factor authentication: A survey," *Cryptography*, vol. 2, no. 1, pp. 1–31, 2018. [Online]. Available: https://www.mdpi.com/2410-387X/2/1/1

[84] C. Jacomme and S. Kremer, "An extensive formal analysis of multi-factor authentication protocols," *ACM Trans. Privacy Secur.*, vol. 24, no. 2, pp. 1–34, Jan. 2021, doi: 10.1145/3440712.

[85] I. Velásquez, A. Caro, and A. Rodríguez, "Authentication schemes and methods: A systematic literature review," *Inf. Softw. Technol.*, vol. 94, pp. 30–37, Feb. 2017.

[86] B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: Taxonomy of methods, current issues and future directions," *Telecommun. Syst.*, vol. 67, no. 2, pp. 247–267, Feb. 2018.

[87] B. Chaimaa, E. Najib, and H. Rachid, "E-banking overview: Concepts, challenges and solutions," *Wireless Pers. Commun.*, vol. 117, no. 2, pp. 1059–1078, Mar. 2021.

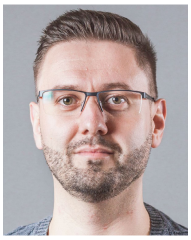
[88] S. Kiljan, K. Simoens, D. D. Cock, M. V. Eekelen, and H. Vranken, "A survey of authentication and communications security in online banking," *ACM Comput. Surveys*, vol. 49, no. 4, pp. 1–35, Dec. 2017.

[89] F. Sinigaglia, R. Carbone, and G. Costa, "Strong authentication for e-banking: A survey on European regulations and implementations," in *Proc. SECRYPT*, 2017, pp. 1–6.

[90] F. Sinigaglia, R. Carbone, G. Costa, and N. Zannone, "A survey on multi-factor authentication for online banking in the wild," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101745.

[91] M. M. Althobaiti and P. Mayhew, "Security and usability of authenticating process of online banking: User experience study," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2014, pp. 1–6.

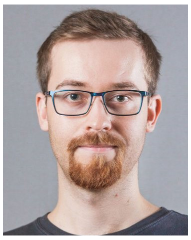
[92] S. G. Lyastani, M. Schilling, M. Neumayr, M. Backes, and S. Bugiel, "Is FIDO2 the kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 268–285.



KAMIL MALINKA received the M.S. degree in applied informatics from Masaryk University (MU), in 2005, and the Ph.D. degree in biometrics and anonymity systems from the Faculty of Information Technology, Brno University of Technology, Czech Republic, in 2010. Currently, he is working as an Assistant Professor with the Faculty of Information Technology, Brno University of Technology. He is also working as an IT Architect and a Researcher at MU and a member of the Local Security Research Group Security, Faculty of Information Technology, where he is focusing on computer and network security.

ONDŘEJ HUJŇÁK received the M.S. degree in information technology security from the Brno University of Technology, in 2016, where he is currently pursuing the Ph.D. degree in computer science and engineering. He is a member of the Research Group Security, Faculty of Information Technology, where he is focusing on computer and network security. His research interests include the security of IoT networks and devices, privacy-enhancing technologies, and cyber-physical systems.

PETR HANÁČEK received the M.S. degree in computer engineering and the Ph.D. and Habilitation degrees in computer science from the Brno University of Technology, Czech Republic, in 1988, 1997, and 2003, respectively. He leads the Local Security Research Group Security, Faculty of Information Technology, where he is focusing on computer and network security. From 1987 to 2001, he worked at the Department of Computer Science, Faculty of Electrical Engineering and Computer Science, Brno University of Technology. Since 2002, he works with the Faculty of Information Technology, Brno University of Technology. He is currently the Head with the Department of Intelligent Systems and an Associate Professor with the Faculty of Information Technology, Brno University of Technology.



LUKÁŠ HELLEBRANDT received the M.S. degree in information technology security from the Faculty of Information Technology, Brno University of Technology, in 2016, where he is currently pursuing the Ph.D. degree. He takes part in the security at the Faculty of Information Technology, Information Technology Security Research Group. He works as a Senior Quality Engineer at Red Hat. His research interests include anonymity networks and privacy-enhancing technologies.

Heliyon 9 (2023) e15090

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors

Anton Firc [*], Kamil Malinka, Petr Hanáček

Brno University of Technology, Božetěchova 2, Brno, 612 00, Czech Republic

ARTICLE INFO

Keywords:
Face deepfakes
Speech deepfakes
Biometrics systems
Facial recognition
Speaker recognition
Deepfake detection
Cybersecurity

ABSTRACT

Deepfakes present an emerging threat in cyberspace. Recent developments in machine learning make deepfakes highly believable, and very difficult to differentiate between what is real and what is fake. Not only humans but also machines struggle to identify deepfakes. Current speaker and facial recognition systems might be easily fooled by carefully prepared synthetic media – deepfakes. We provide a detailed overview of the state-of-the-art deepfake creation and detection methods for selected visual and audio domains. In contrast to other deepfake surveys, we focus on the threats that deepfakes represent to biometrics systems (e.g., spoofing). We discuss both facial and speech deepfakes, and for each domain, we define deepfake categories and their differences. For each deepfake category, we provide an overview of available tools for creation, datasets, and detection methods. Our main contribution is a definition of attack vectors concerning the differences between categories and reported real-world attacks to evaluate each category's threats to selected categories of biometrics systems.

1. Introduction

Deepfake is a term that denotes a subset of synthetic media. The term itself is a combination of words *deep learning* and *fake*. Deepfakes are created using deep neural networks, and they depict events that never happened to entertain, defame individuals, spread fake news, and others [18].

The constant advancements in machine learning make deepfake creation available for a broader spectrum of users.

The simplest tools even feature a graphical user interface that lets inexperienced users create deepfakes [1–4].

In addition, deepfakes still grow in popularity. The publication counts significantly increased in the past years (see Fig. 1), as well as the popularity of the *deepfake* keyword in Google searches (see Fig. 2). The immense popularity and easy accessibility of deepfakes urge investigating what threats and impacts deepfakes might have on cyberspace, as society is still unsure.

The currently known usages might be divided into two major categories by the usage intention: *beneficial* and *malicious* [72]. The beneficial usages find a place in entertainment, education, or even healthcare. For example, the visitors of the Dalí Museum in St. Petersburg, Florida, may interact with a deepfake persona of the painter himself [158]. Deepfake technology also made Val Kilmer speak again for a special appearance in the latest Top Gun movie [17].

The malicious usages, in contrast, include fake news, fraud, or identity theft. One of the latest examples that could have had a tremendous impact is the impersonation of the Ukrainian president in an appeal to surrender [103]. Another example involves fake social media profiles spreading fake news about the Belgian government plans [85].

[*] Corresponding author. Brno University of Technology, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic.
E-mail address: ifirc@fit.vut.cz (A. Firc).

https://doi.org/10.1016/j.heliyon.2023.e15090
Received 1 November 2022; Received in revised form 16 March 2023; Accepted 27 March 2023
Available online 3 April 2023

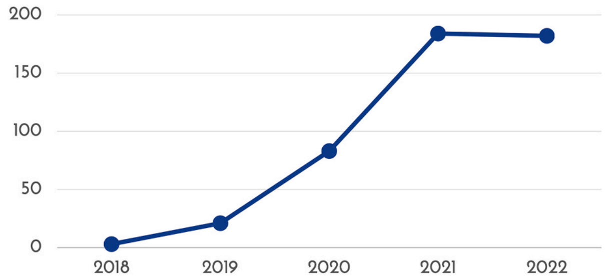


Fig. 1. Publication count by last years for keyword "deepfake" according to WoS. Source: https://www.webofscience.com/wos/woscc/summary/894c369c-e777-4199-89f3-dea9078a35bd-6a94f062/relevance/1.

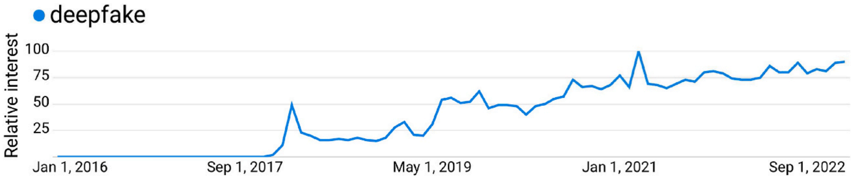


Fig. 2. Trend in Google searches for deepfake queries. Retrieved from Google Trends: .https://trends.google.com/trends/explore?date=2016-01-01%202023-01-12&q=deepfake

Finally, deepfakes might be used to spoof biometrics systems [73,244]. Both voice recognition and facial recognition systems are prone to be spoofed by synthetic media. Moreover, not every deepfake type might have the potential to be used against biometrics systems.

Our survey focuses on the latest development of deepfakes in facial and speech domains. We divide facial deepfakes into categories according to the level of manipulation needed, and as an extra category, we incorporate face morphing. We divide speech deepfakes into categories according to voice transfer technology. For each category, technologies, tools, datasets, attack vectors, and detection methods are provided. The attack vectors propose how each of the presented deepfake categories might be misused. Additionally, we focus on identifying deepfake types that might be used to spoof face or voice biometrics systems.

1.1. Motivation

The primary motivation for creating this survey is the constant advancements in deepfake technology. The currently available surveys are becoming outdated. It is crucial for security-related researchers, developers, or operators to understand the attacker model: how deepfakes might be misused for illegitimate purposes. A second important factor is the quality of current deepfakes. Rapid developments change the power of the attacker and the used technology from day to day. These facts emphasize the need to provide an update on technology. Additionally, access to the most up-to-date deepfake creation tools is essential for creating detection methods that can respond to state-of-the-art threats. Another important fact to consider is the effortless access to media suitable for deepfake creation. Social networks and similar platforms contain tremendous amounts of facial images, videos, or recordings. This empowers the attacker and significantly expands the possible attack vector scope. A more detailed discussion on media retrieval and the attack. vectors is further provided in Section 5.2.

We also consider it essential to interconnect the facial and speech deepfake areas. Attacks exploiting combinations of these areas are starting to be executed, and the combination of fake video and audio makes the attack more powerful. It is thus viable to grasp how these combinations might be used and prevent this usage.

Motivated by the stated facts, we publish this survey updating the existing surveys with the latest publications on the creation and detection of deepfakes. Moreover, we connect the latest publications with corresponding tools (implementations). Because of many existing works, it is often hard to connect current research results with existing tools implementing proposed methods.

1.2. Literature collection and selection criteria

This survey reviews existing research papers that focus on techniques for creating and detecting deepfake media in face and speech domains. A more detailed description of the approach and protocols employed for the review is given in Table 1.

Table 1
Literature collection protocol.

Preparation protocol	Description
Purpose	• To provide an update of existing surveys on the latest face and speech deepfake creation and detection technologies. • To demonstrate the current trends in deepfake creation and detection areas. • To connect face and speech deepfake areas and set a united taxonomy. • To identify the threats posed by deepfake media. • To connect existing academic publications with their implementations.
Sources	Google Scholar, IEEE explore, ACM Digital Library, Springer Link, and online sources for incident reports
Query	Following queries were used on the data sources above for the collection of publications: deepfakes/face synthesis/image synthesis/face morphing/face swap/facial reenactment/reenactment/face manipulation/text to speech/voice conversion/speech morphing/voice morphing/deepfake detection/face synthesis detection/image synthesis detection/face morphing detection/morphing detection/face swap detection/reenactment detection/face reenactment detection/face manipulation detection/deepfake speech detection
Method	Literature was categorized as follows: • Deepfake creation (including tools) methods based on the proposed taxonomy. • Deepfake detection methods based on the proposed deepfake taxonomy. • Deepfake creation tools (GitHub (GitLab) repositories, online tools). • Incident reports and online news reporting usage of deepfakes for malicious purposes. • Discussion on future developments, challenges, and limitations of the deepfake area.
Inclusions and Exclusions	Preference was given to peer-reviewed journal papers and conference proceedings articles published between and including 2021 and 2023. In addition, articles from the archive literature were also taken into account to demonstrate evolution in discussed fields or settle common facts.

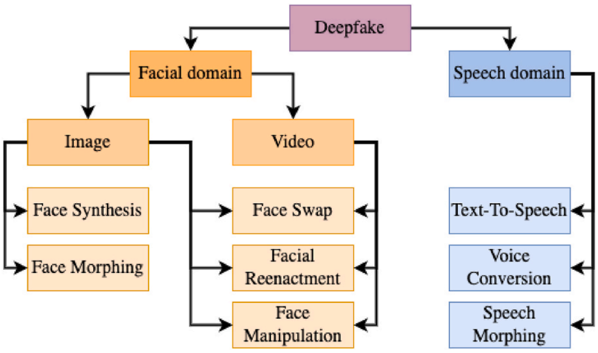


Figure 3. Visualization of used united taxonomy of deepfakes. Categories are divided according to the domain and media type. For the facial domain, most of the categories fall under both media types.

1.3. Contributions

The main contributions of this article might be summarized as follows.

- We provide a united taxonomy for facial and speech deepfakes (see Fig. 3) and define differences between each category. The facial deepfakes are categorized according to the level of manipulation needed, and the speech deepfakes are according to the voice transfer technology.
- We provide an overview of deepfake creation tools. We connect these tools with relevant research results, which usually come from different authors.
- We provide an overview of the latest detection techniques.
- We define attack vectors for each deepfake category. These attack vectors respect the differences in all deepfake categories and show the potential of each category to spoof biometrics systems.

1.4. Document structure

Related work is discussed in Section 2. The face deepfakes are discussed in Section 3. The speech deepfakes are discussed in Section 4. Section 5 provides a discussion on the future of deepfakes and Section 6 summarizes all of the stated knowledge.

2. Related work

In this work, we overview the deepfake area from the security point of view. We use existing deepfake surveys through all domains, unite them, and supplement their results with the latest trends in each area. The style and content of this survey aim to provide a unified overview and trends in the past years of the deepfake problems to security-related researchers and developers. To perform a security analysis focused on the deepfake resilience of biometrics authentication systems, it is necessary to have orientation in all deepfake types, their properties, tools for their creations, what threats they present, and methods for their detection.

Tolosana et al. [267] provide a detailed overview of facial manipulation techniques and corresponding detection methods. The authors divided facial manipulations into categories according to the level of manipulation needed. These categories are well-established by the research community and have received the most attention in the last few years; we thus continue to use this categorization.

The remaining facial deepfake categories are covered in surveys by Li et al. [168], Nguyen et al. [203], Kietzmann et al. [140], Verdoliva et al. [277], Malik et al. [185] or Rana et al. [230]. We especially point out Mirsky et al. [189], who gives a very detailed and technical description of the latest technologies used for deepfake creation and detection. These techniques are explained in-depth using visual aids such as graphs and schemes. The authors focus on the facial and human reenactment deepfakes. Moreover, the authors discuss the future development of deepfake technology and the possible impacts on human society.

As a special category, facial morphing is covered in a survey by Venkatesh et al. [274]. The authors provide a detailed overview of face morphing techniques and detection mechanisms. Various aspects of the creation of morphed facial images are described and illustrated by the authors. Special attention is also given to state-of-the-art detection.

Methods, emphasizing the reproducibility of the method's benchmarks. Finally, the authors discuss future challenges and research directions in face morphing.

Mohmmadi et al. [192], Sisman et al. [252], Machado et al. [183], and Yannis et al. [255] provides an overview of Voice Conversion (VC) systems and the principles behind VC. The authors present the mappings between source and target speakers, prominent evaluation approaches for VC performance, and finally, review different applications that use VC methods.

Ultimately, Tabett et al. [256] provide an overview of former Text-to-Speech (TTS) techniques. Unfortunately, this survey is more than ten years old. The deepfakes in the facial domain are well discussed by various surveys. However, developments in the last two years remain unaddressed. Our survey thus provides an update of the latest developments in facial deepfakes to existing surveys. The speech domain seems to be even less explored and discussed. We thus provide an overview of the most recent publications and tools for all deepfake speech categories to address this issue. This survey poses as an update to the existing ones, where we summarize the latest developments in each category. Moreover, in contrast to the stated deepfake creation and detection surveys, our work examines the usability of different deepfake types to threaten society and biometrics systems. To the best of our knowledge, no other works provide an overview of deepfakes regarding these threats. In addition to other published surveys, we provide an overview of visual and speech deepfakes, discuss the potential misuse of each deepfake category, and provide an exhaustive list of tools for each category connected with corresponding academic publications. We consider the combination of facial and speech deepfakes as essential knowledge for future security practitioners or researchers. This combination increases the power of an attack, and we expect to see more such attacks soon.

3. Face deepfakes

This section discusses available deepfake creation techniques and their security impacts on facial recognition systems. We divide face deepfakes into five categories depending on the level of manipulation needed to create such a deepfake. From the highest level of manipulation to the lowest: face synthesis, face morphing, face swap, face reenactment and face manipulation. For each category, we provide a general overview of the latest technologies, available tools, and publicly accessible datasets, discuss potential attack vectors, and provide a detailed overview of detection methods and techniques.

3.1. Face synthesis

We define face synthesis as a process of synthesizing non-existing faces based on learned high-level attributes, such as pose or identity [136]. There are different applications of face synthesis. They range from synthesizing virtual characters in the film industry to providing a human-looking representation of computer agents to interact with their users. Moreover, face synthesis might be beneficial for face recognition applications to generate needed training data [314]. An example of an entirely fictional face synthesized by the StyleGAN2 model is shown in Fig. 4.

3.1.1. Technologies

The utter majority of algorithms and techniques for face synthesis utilize Generative Adversarial Networks (GANs) [295]. GAN is a generative framework first described by Goodfellow [84]. The network is composed of two networks that work against each other. GANs are successfully deployed to handle various image syntheses tasks, such as style transfer, image-to-image translation, and representation learning [62].

Numerous variations and tweaks exist to the used GAN backbones [106,162]. One of the most influential author groups in face synthesis is Karras et al. A series of publications propose and improve very well-known architecture StyleGAN [133–136]. In their latest work [134], they battle aliases that often leak to the generator network, which ultimately causes details in images to be locked to a specific image coordinate rather than the object surface. The authors thus suggest a change in architecture representing all signals as continuous. Since 2021, most approaches have used controllable GANs that allow the manipulation of the latent space, thus controlling the look of the final output. Liu et al. [174] propose to learn a linear sub-space that mimics the distribution of a target dataset in the latent space. This allows the generation of similar but new images to a chosen dataset. Additionally, Nguyen et al. [202] propose the

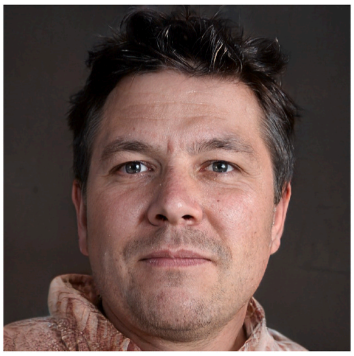


Fig. 4. Face synthesized using StyleGAN2 model [136]. Image retrieved from https://thispersondoesnotexist.com.

implementation of a quality code to the layers of StyleGAN2, which ultimately allows controlling the quality of synthesized faces, as in some scenarios, it is desirable to lower the final quality.

The latest publications seem to focus on synthesizing faces based on sketches [301,309]. For example, Yadav et al. [301] combine attention for improved performance with cyclic-synthesize loss that currently provides the best image-to-image translation results. In contrast, Yoshikawa et al. [309] use a three-stage framework that utilizes auto-encoders. This setting allows the generation of more diverse images from one input sketch.

3.1.2. Tools

Face synthesis is quite a widespread utilization of GANs. Many tools from this area have been developed and published in the past three years (see Table 2). Most tools are simple, while some even feature an intuitive user interface. In addition, the remaining open-source tools are almost always supplied with pre-trained models that provide outputs of outstanding quality. In general, a meager knowledge is required to use the listed tools. This allows the broad public to access these tools and uses synthesized faces for malicious or beneficial purposes.

3.1.3. Datasets

The amount of available datasets containing synthetic faces is quite limited. However, existing datasets consist of a large number of images. Moreover, the easily accessible tools for face synthesis allow for the simple creation of novel datasets with outstanding quality and quantity of synthesized images. The publicly available datasets are listed in Table 3.

3.1.4. Attack vector

Bateman [18] suggests that synthesized facial images of non-existent people might be misused for creating synthetic social botnets. The synthesized image used as a profile picture improves the stealth of such a profile as it cannot be easily detected as a duplicate fake account by automated detectors. A report by Graphika Team [85] uncovers a cluster of inauthentic Twitter accounts that amplified, and sometimes created, articles that attacked the Belgian government's recent plans to limit the access of "high-risk" suppliers to its 5G network. The plans were reportedly designed to limit the influence of Chinese firms. Similar misuses of synthetic images have been reported [32]. The other use of synthetically generated faces might be to provide anonymity to a user who is required to upload her or his photo into any system. These faces might also be used to make adverse identity claims, where the attacker claims that she or he is

Table 2
Face synthesis tools. Each line represents a different tool, the first column denotes the name and a publication, the second column link to the corresponding tool, and the last column main features of the linked tool.

Tool	Link	Main features
CFSM [174]	https://github.com/liuf1990/CFSM	Allows generation of additional dataset samples.
QC-StyleGAN [202]	https://github.com/VinAIResearch/QC-StyleGAN	Control over final image quality.
clip2latent [216]	https://github.com/justinpinkney/clip2latent	Synthesizing images from text.
StyleKD [292]	https://github.com/xuguodong03/stylekd	Reduces StyleGAN computational requirements.
StyleGAN3 [134]	https://github.com/NVlabs/stylegan3	Vastly improved quality.
ProGAN [133]	https://github.com/akanimax/pro_gan_pytorch	Faster training and better quality.
AdvFaces [62]	https://github.com/ronny3050/AdvFaces	Automatic generation of imperceptible perturbations.
GLO [27]	https://github.com/clvrai/Generative-Latent-Optimization-Tensorflow	No unstable adversarial training dynamics.
GAN Zoo	https://github.com/facebookresearch/pytorch_GAN_zoo	Standalone GAN toolbox.
MMGeneration	https://github.com/open-mmlab/mmgeneration	A powerful toolkit curated by the community.
EigenGAN [98]	https://github.com/LynnHo/EigenGAN-Tensorflow	Manipulation of output features.
generated.photos	https://generated.photos/face-generator	Online, interactive and simple to use tool.
thispersondoesnotexist [136]	https://thispersondoesnotexist.com	Online, simple tool with impressive quality.

Table 3
Face synthesis datasets. Each line represents a different dataset, and footnotes contain links for datasets where the link is not explicitly provided in the publication.

Name	Fake Images
non-curated images[a]	NA
PGGAN [133]	80,000
iFakeFaceDB [200]	87,000
TPDNE[b]	150,000
generated.photos[c]	2,683,964
TrueFace [26]	210,000
PerceptionSyntheticFaces [156]	300
FaceSynthetics [284]	100,000
SFHQ [19]	425,000
SPRITZ-PS [71]	1600

[a] https://drive.google.com/drive/folders/1j6uZ_a6zci0HyKZdpDq9kSa8VihtEPCp[b]https://www.kaggle.com/potatohd404/tpdne-60k-128x128/version/2.

[c] https://generated.photos/datasets.

not enrolled in the system but should be. This might, for example, happen when enrolling for government social support. We estimate the threat posed to biometrics systems as low.

3.1.5. *Detection*

The latest synthetic face detection methods seem to exploit discrepancies in synthesized persons' biological traits primarily. Guo et al. [91] propose a simple method that analyzes pupil shape. It thus seems that the GAN-generated persons have irregular pupil shapes. Guo et al. [92] also employ deep learning models to look for inconsistent eye components. Additionally, Hu et al. [107] look for inconsistent corneal specular highlights between two eyes. Simply put, the authors examine the position and orientation of light source reflections in both eyes as they tend to be inconsistent in synthesized images. Finally, Xue et al. [300] propose extracting physiological properties such as iris or pupils using deep learning. Similar to previous works, eye properties are used to detect synthesized faces.

Current research seems to focus on more generalizable solutions as the detectors are model-specific, and the performance deteriorates with different datasets or modifications to synthesized images. [92,300].

Despite the statements that the artifact-based methods will soon become ineffective because of the GAN developments that will remove artifacts from synthesized images [188,311], we still see several artifact-based methods published in the last year (2022) [39,76,280].

3.2. *Face morphing*

Morphing is a special effect in motion pictures or animations that changes one image into another using a seamless transition. Morphing is often used to depict one person turning into another [69]. For the scope of this work, we understand morphing as a method to produce a facial image that is very similar to the face of one subject but also contains facial features of the second subject, as shown in Fig. 5. This process might be summarized into the following steps as proposed by Ferrara et al. [69].

1. Both facial images are placed as separate layers in the same image and then are manually positioned to superimpose the eyes.
2. Important facial features are marked on both faces.
3. Sequence of frames showing the transition from one face to another is automatically generated.
4. Final frame is selected from the animation based on the similarity score.
5. Selected frame is retouched to look more realistic.

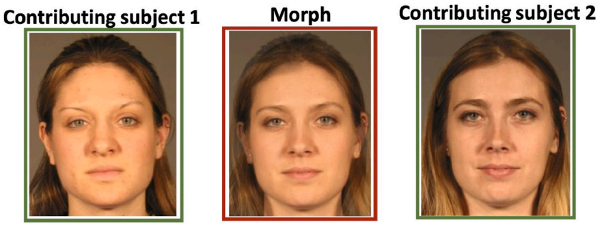


Fig. 5. An example of face morphing [274]. Left and right-most images show the original subjects, often referred to as con artists and accomplices. The middle image shows the result of morphing both subjects.

Table 4
Face morphing tools overview. Each line represents a different morphing tool; the first column denotes the name, the second column links to the tool, and the last column is an interesting tool feature.

Name	Link	Feature
Sarkar et al. [240]	https://gitlab.idiap.ch/bob/bob.paper.icassp2022_morph_generate	Multiple morphing tools in one.
MIPGAN [315]	https://github.com/ZHYYYYYYYYYYYY/MIPGAN-face-morphing-algorithm	High-quality morphs generation.
MorphSG2	https://gitlab.idiap.ch/bob/bob.morph.sg2	StyleGAN2 based generator.
FaceMorpher	http://alyssaq.github.io/face_morpher/index.html	Available as a Python package.
3Dthis	https://3dthis.com/morph.htm	Simple online interactive tool.
Face-Morphing	https://github.com/Azmarie/Face-Morphing	Automatic face morphing.
Face Morphing	https://github.com/cirbuk/face-morphing	Video face morphing.
WebMorp	https://webmorph.org/	Free online morphing tool.
JPsychoMorph	https://cherry.dcs.aber.ac.uk/trac/wiki/jpsychomorph	Averaging, blending or exaggerating differences between facial images.
MorphAnalyser	https://cherry.dcs.aber.ac.uk/trac/wiki/MorphAnalyser	3D facial modelling.
FantaMorph	https://www.fantamorph.com/	Commercially developed software.
FotoMorph	https://fotomorph.informer.com/	Free standalone application.

3.2.1. *Technologies*

Face morphing might be divided into two categories as proposed by Venkatesh et al. [274]: landmark based and deeplearning based. In recent years, primarily deep-learning-based solutions have been published. This approach utilizes Generative Adversarial Networks (GANs). The GAN synthesizes the morphed face more accurately by sampling two facial images in latent space. One of the latest standalone morph creation solutions is MIPGAN [315], which is a modification to StyleGAN with a reworked loss function and added identity factor. Finally, Moser et al. [194] propose using AutoEncoder architecture with multiple decoders for each identity to generate morphs. In addition, the most recent work that reports morph creation uses a landmark-based approach with the OpenCV library [115].

From the literature overview, it is apparent that the current research is focused on developing novel detection methods rather than methods for morph creation.

3.2.2. *Tools*

Recently, numerous tools were published for face morphing. Table 4 shows the combination of open-source and freeware tools. Similarly to the face synthesis tools, a wide range of easy-to-use tools exist that allow non-experienced users to use them. The morphing process thus essentially consists of only uploading (providing) two suitable facial images. Additionally, photo or video editing software, such as GNU Image Manipulation Program (GIMP) with GIMP Animation Package extension or Adobe After Effects, allows for the manual creation of morphed images. However, such manual approaches cannot be considered deepfakes. We list them only for the sake of completeness.

3.2.3. *Datasets*

An extensive amount of datasets containing morphed facial images exist. Unfortunately, the majority of datasets are not publicly available. Most recently, the focus on dataset creation seems to be given to different modifications and perturbations of morphed images to better assess new detection methods' robustness. Table 5 provides an overview of existing datasets.

Table 5
Face morphing datasets. Each line represents a different dataset; the first column denotes the author, the second size (if available), and the last if the dataset is publicly available.

Author	Size	Public
Ferrara et al. [69]	14	no
Kramer et al. [152]	60	upon request
Ferrara et al. [70]	80	no
Scherhag et al. [241]	231	no
Makrushin et al. [184]	326	no
Raghavendra et al. [224]	450	no
Venkatesh et al. [275]	1500	no
Raghavendra et al. [225]	2518	no
Raja et al. [226]	5748	yes
Seibold et al. [242]	9000	no
Venkatesh et al. [273]	14,305	no
Peng et al. [213]	59,344	yes
Damer et al. [56]	80,000	yes
Sarkar et al. [240]	N/A	yes
Dunstone et al. [65]	N/A	upon request

3.2.4. Attack vector

Face morphing presents a real and serious threat to face recognition systems. The most severe threat is posed to the airports' Automatic Border Control systems. Since 2002, the face has been selected as the primary globally interoperable biometric trait for machine-operated identity verification in electronically Machine Readable Travel Documents. The facial image for such a document might be provided in two ways depending on the issuer country: a) image is captured live with a high-end camera at the enrollment station, b) image is provided by the citizen printed on paper. The second way creates an opportunity to hand in a morphed facial image and to enroll that image into an official genuine identification document. Enrolling a morphed face into a document of this kind then allows both of the persons whose faces were used to create the morphed image to use that exact document [69,223,242,243].

Incidents involving morphed facial images in documents were already reported by Slovenian police [34]. These attacks were first observed in 2018. In 2021, the Slovenian police reported more than 40 incidents, which were provided as part of professional service to issue Slovenian passports to Albanians to allow them to travel to Canada.

A recent study by Sarkar et al. [240] reveals that landmark-based morphs seem the most capable of fooling facial recognition systems. Moreover, it is apparent that the better the facial recognition system, the more it is prone to morphing attacks.

3.2.5. Detection

Detection methods are referred to as Morphing Attack Detection (MAD). The detection techniques are either single or differential image-based [274].

The single image-based methods detect morphing attacks based only on a single presented image. Hamza et al. [94] developed a method that uses deep learning to extract special features that are then used as input for the Support Vector Machine classifier. Venkatesh [276] proposes to combine multiple deep-CNNs for feature extraction and then use feature fusion to classify morphed images.

Le-Bing et al. [317] propose to use Auto Encoders to denoise the images and then VGG19 for classification. Raja et al. [228] propose an end-to-end architecture based on the encoder-decoder network that examines the residuals of the morphing process.

Long et al. [179] develop a lightweight solution that extracts patches from faces and classifies them. The probabilities are combined in the end into the ultimate result. Similar methods exploiting local features are proposed by Qin et al. [221].

Neto et al. [199] develop a novel feature that considers the existent identity information and uses ResNet-18 architecture for classification. Finally, Aghadie et al. [10] propose to exploit inconsistencies in the frequency content of morphed images.

The differential image-based methods require the images to be captured in a trusted environment. These situations mainly refer to the border crossing scenario, where a morphed image is provided on the travel document, and a live image is captured simultaneously. Ramachandra and Li [229] propose to exploit the inconsistencies in the color scale-space of morphed and genuine images. Peng et al. [213] propose a method that identifies the attacker. A watchlist containing the biometric reference is used to identify morphed images and retrieve the original identities. Singh et al. [251] propose a fusion of features from two deep CNNs that are later classified using pair of Support Vector Machines.

Generalization is predominantly a number one challenge in most of the publications [221,228,317]. To this extent, Spreeuwers et al. [254] examines the robustness of different approaches on different datasets and to modifications such as Gaussian noise. Raja et al. [227] propose further recommendations to improve generalization abilities.

Additionally, in 2022, a MAD challenge was conducted to evaluate the advancements in the field [114]. The best detection model [118] was a single image-based approach using Convolutional Neural Networks with special tweaks to classification heads.

In contrast to machine-based morph detection, we can find trials evaluating human abilities to differentiate between real and morphed images [83]. This is a critical and timely issue, as morphed images circulate in travel documents. Human detection abilities thus should be developed in addition to machine-based detection to improve security.

3.3. Face swap

Face swapping refers to a technique where a face from source photo is transferred onto a face in a target photo (Fig. 6 (a)). The result is desired to look realistic and unedited. For the scope of this work, we only refer to the one-to-one face-swapping paradigm. Moreover, we focus on the automatic face swapping (Fig. 6 (b)), instead of the manual one (Fig. 6 (c)) as the manual process is outside the deepfake categorization.

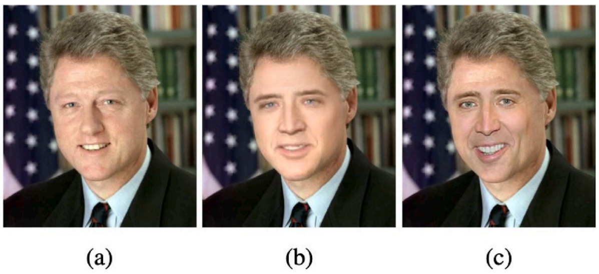


Fig. 6. Face swap. a) The input image, b) The result of automatic face swap, c) The result of manual face swap.

Blanz et al. [24] brought up the first mention of face swapping in 2004. A few years after, automated techniques were described by Bitouk et al. [23]. The original purpose of those methods was to be used for privacy preservation. Face swapping can be used instead of blurring or pixelating faces in graphical materials. In contrast to the original intended usage, face-swapping is nowadays primarily used for entertainment purposes [207].

3.3.1. Technologies

Former face-swapping methods utilized operations over 3D models [24] or 2D image-composition [23]. Current face-swapping methods utilize deep learning algorithms. In particular, GANs, Variational Auto-Encoders (VAEs), or CNNs are used to transfer facial expressions and motion patterns between persons in videos [249].

We can define a general pipeline for face swapping by generalizing the pipeline presented by Refs. [150,214].

1. Extraction: Faces are extracted from the source and target data and then processed by face detection, face alignment, and face segmentation algorithms.
2. Training: A model is trained to learn the correct way of transposing the source person's face onto the target person's head.
3. Conversion: The source face gets transposed onto the target head, and the result is retouched to fit the target seamlessly.

Various publications currently exist incorporating different neural network architectures and face-swapping process automation levels [60,78,260,328].

One of the most influential works in face-swapping was published by Perov et al. [214], who published a complete face-swapping framework utilizing the Encoder-Decoder architecture. Xu et al. [294] propose face swapping in a local-global manner. The method uses a Local Facial Region-Aware branch that augments local identity-relevant features and a Global Source Feature Adaptive branch that complements global identity-relevant cues. Li et al. [161] propose a special Attribute-Conditioned network that can preserve identity attributes even in low-resolution media. Kim et al. [143] develop a method that allows smooth identity embedding by employing a custom loss and promoting a smoother latent space.

Numerous zero-shot and few-shot methods have been published [166,168,206,248]. This development allows for much easier access to face-swapping to the broad public than ever because of removing the requirements to train the network before using it.

In addition, lightweight mobile face-swapping tools are being developed that address the demand for face-swapping applications in smartphones [296,310]. These methods are designed to work with limited resources with minimal impact on the final quality.

A portion of the recently published works also examines the possibilities of disentangling the latent space to swap faces [166,294]

Ultimately, modifications to GAN architectures, such as StyleGAN2, were proposed for face-swapping [175,305].

3.3.2. Tools

The immersive popularity of face swapping sprung the development of many tools. One of the most influential tools is DeepFaceLab [214]. This tool implements a complete face-swapping pipeline, from processing input data to polishing the output visuals and extensive community, providing support and pre-trained models. An overview of currently available face-swapping tools is provided in Table 6. The general population searches for this technology. This may be seen on social media, where face-swapping apps have gone viral. The primary reason behind this popularity is entertainment – plain fun [186]. In contrast to other categories, face-swapping contains the largest number of smartphone/online and paid tools. However, many of these tools try to mitigate the potential misuse by design. For example, watermarking is used, or the users can only use a developer-managed set of photos and videos (mostly movie scenes).

3.3.3. Datasets

A vast amount of datasets containing face swap images or videos exist. Moreover, the majority of datasets are publicly available. However, a more detailed look at the contents often reveals a low quality of the face-swapped videos. Such media do not represent the current quality and power of the face-swapping tools. Thus, it would be beneficial.

To review the existing datasets and propose an update to reflect the current possibilities better. Table 7 provides an overview of existing face swap datasets and their size.

3.3.4. Attack vector

Usage of face swapping to spoof biometrics systems overlaps with facial reenactment (Section 3.4). Face swapping provides resources for identity theft. An attacker can prepare videos or images of a selected individual and impersonate her or him [18,20,167]. This way, a facial biometrics system might be spoofed, even when interacting with the user is required, such as turning the head or looking in different directions. However, this ability is limited compared to facial reenactment, as the media must be prepared beforehand. Face swapping, thus, better serves the purpose of defaming individuals or manipulating evidence rather than spoofing biometrics systems.

3.3.5. Detection

Face-swapping videos might be easier to detect than single images, as they contain temporal information. One of the possibilities is to exploit physical and physiological signals. These signals are not well captured in deepfake videos and may include spontaneous and involuntary physiological activities such as breathing, pulse, eye movement, or eye blinking. As these signals are often overlooked in the process of deepfake video creation, they are suitable to be used as indicators for detection [167]. The temporal information might

Table 6
Face swap tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding tool, and the last column is an interesting feature.

Tool	Link	Features
DeepFaceLab [214]	https://github.com/iperov/DeepFaceLab	Most advanced pipeline, large community of users providing help and models.
MobileFSGAN [310]	https://github.com/HoiM/MobileFSGAN	Lightweight architecture for use in mobile devices.
MobileFaceSwap [296]	https://github.com/Seanseattle/MobileFaceSwap	Real-time swapping and deployment to edge devices.
FaceSwapper [166]	https://github.com/liqi-casia/faceswapper	One-shot face swapping.
FSLD-HiRes [294]	https://github.com/cnnlstm/fslsd_hires	High resolution face-swapping.
GHOST [87]	https://github.com/ai-forever/ghost	High-quality one-shot video or image face-swapping.
Face Swap Live	http://faceswaplive.com	Real-time, for mobile devices.
FaceSwapper	https://faceswapper.ai/	Free online tool.
FaceApp	https://www.faceapp.com	AI manipulation of facial images.
Zao App	https://zaodownload.com	For mobile devices.
Reface	https://hey.reface.ai	For mobile devices, closed set of curated target videos and images.
Deepfakes Web	https://deepfakesweb.com	Online tool, "Responsible Deepfake Technology"
Realistic-Neural-Talking- Head-Models [313]	https://github.com/vincent-thevenin//Realistic-Neural-Talking-Head-Models	Few-shot learning model.
One-Shot Face Swapping on Megapixels [328]	https://github.com/zyainfal/One-Shot-Face-Swapping-on-Megapixels	1024 × 1024 resolution of result images.
SimSwap [43]	https://github.com/neuralchen/SimSwap	Arbitrary source to arbitrary target swap.
FaceShifter [164]	https://github.com/mindslab-ai/faceshifter	High fidelity and occlusion aware framework.
FSGAN [205,206]	https://github.com/YuvalNirkin/fsgan	Subject agnostic, can be applied to a pair of faces unseen during training.
Deepfakes	https://github.com/deepfakes/faceswap	Community and forums.
FaceSwap	https://github.com/MarekKowalski/FaceSwap/	Transposes faces to head captured by the device camera.
Deep Alignment Net [151]	https://github.com/MarekKowalski/DeepAlignmentNetwork	Uses entire face images, which allows handling faces with large variation in head pose.

be used in other methods as proposed by Das et al. [58]. The face-swapped image detection approaches might be shared with the image-based video approaches or employ different methods for extracting artifacts or visual inconsistencies [20,82,120]. For example, Guan et al. [90] propose to extract a 3D mask of the individual and then look for the inconsistency of 3D facial shape and facial appearance created by the face-swapping tools.

Hassani and Malik [96] propose to examine the camera-specific noiseprint to detect deepfake videos. This method is reported to deliver outstanding results while consuming a few resources. In addition, Hassani et al. [97] propose to exploit alternation traces photo-response profile by comparing challenge images with enrolment images.

In contrast to the previously discussed methods, Zhao et al. [322] believe deepfake detection should not be modeled as a vanilla binary classification task but rather a fine-grained classification task. The difference between real and fake images is often subtle and local, and the binary classification cannot fully capture this difference. To this extent, the authors propose a multi-attention network architecture for deepfake detection.

Ultimately, even in the face-swap area, the human ability to detect deepfakes begins to be evaluated. Nichols et al. [204] propose a methodological approach incorporating test procedures from sensory science for visual detection. The findings suggest that human detection ability is limited.

3.4. Facial reenactment

Facial reenactment is a photo-realistic facial re-animation of a target video with expressions of a source actor. This method was formerly proposed to provide the missing visual channel used in a scenario of a digital assistant. It is essential that these methods not only generate audio and visual information, but the audio has to be synced to the motions of generated human visual [263].

The usage of this technology is not limited to providing a human look for digital assistants. It can be used to create virtual avatars in teleconferencing or video production for dubbing [263].

The critical difference between facial reenactment and face-swapping (Section 3.3) is that facial reenactment generates an entirely new visual representation of an individual. However, the identity of the depicted individual remains intact. Face swapping modifies only the existing visual representation of an individual by replacing his face with a different person's face. An example of reenactment is shown in Fig. 7.

3.4.1. Technologies

Facial reenactment technologies might be divided into three main categories, depending on the input and desired output [263]: Video-Driven Facial Reenactment, Audio-Driven Facial Animation and Audio-Driven Facial Reenactment.

Video-Driven Facial Reenactment methods mainly rely on reconstructing a source and target face using a parametric face model. The target face is reenacted by replacing the expression parameters with ones from the source face [102,104,172]. Agrawal et al. [7] improve the reenactment results by using face mesh and face segmentation mask as priors for generation and using the audio

Table 7
Face swap datasets. Each line represents a different dataset. The first column denotes the dataset name or author, and the second column size of the dataset.

Name	Content
DSI-1 [61]	25 deepfake images
DSO-1 [61]	100 forged images
Zhou et al. [323]	1005 deepfake images
FakeAVCeleb [138]	500 deepfake videos
Korshunov and Marcel [149]	620 deepfake videos
DFGC-21 [212]	17,000 deepfake images
MFC Datasets [89]	50,000 deepfake images and 500 deepfake videos
WildDeepfake [330]	707 deepfake videos
VideoForensicsHQ [74]	1737 deepfake videos
FaceForensics++ [235]	1.8 million deepfake images from 4000 videos
DFDC Dataset [64]	124,000 deepfake videos
Celeb-DF [169]	5369 deepfake videos
CelebV-HQ [327]	35,666 deepfake videos
DeeperForensics-1.0 [121]	11,000 deepfake videos
GBDF [196]	10,000 deepfake videos
ZoomDF [210]	400 deepfake videos
FFIW [324]	10,000 deepfake videos
KODF [155]	175,776 deepfake videos

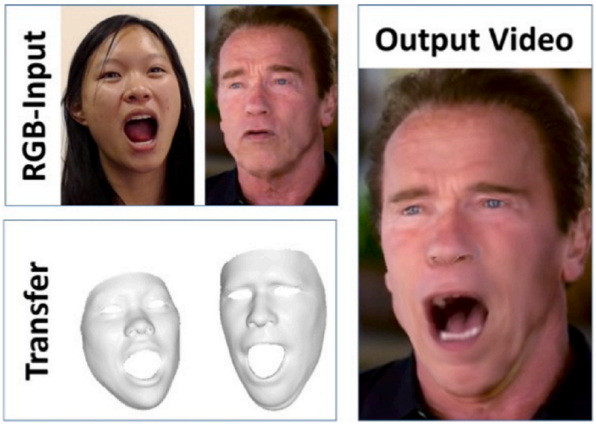


Fig. 7. An example of facial reenactment [264]. RGB-Input shows the input to the reenactment method (driving actor and target identity). The transfer shows 3D masks used for expression transfer for each identity of the input actors. The output video shows the output of the reenactment using the input identities shown in the example.

information to more properly sync the lip movement. Bounareli et al. [29] decided to use existing and pre-trained GAN architecture and approach the reenactment by disentangling the latent code. This approach allows for quality reenactment in a one-shot setting. Xue et al. [297] propose to modify the face representation of GANs by using Projected Normalized Coordinate Code. This novel representation allows high-fidelity generation and identity preservation. Kong et al. [148] propose to use the prior information on facial motion to decompose the movement into two parts: pose and expression. This approach achieves more flexible movement control and does not suffer from losing identity information.

One of the current challenges of facial reenactment tools is the proper representation of mouth opening. Predicting the inside of the mouth and teeth is challenging, so most available methods produce inconsistent results. This issue is addressed by Fu et al. [77], who propose to use a special geometry-aware encoder that extracts the teeth structure from the driving video.

Audio-Driven Facial Animation does not focus on photo-realistic results but the prediction of facial motions [22,291]. Zhang et al. [318] propose a combination of networks to predict motion, pose, and expression, which is then mapped to a 3D space for the final render.

Audio-Driven Facial Reenactment generates realistic videos in sync with the input audio stream [180,316]. Tripathy et al. [269] propose a self-supervised approach that extracts paired feature points from source and driving media and uses them to predict the movement.

Finally, it is important to mention image-to-image reenactment. This category is posed aside from the previously mentioned, as it does not produce video results but still images. Hsu et al. [101] construct a model that uses two generator networks, one for shape-preserving and the second for reenactment. For the same task, Hu et al. [105] propose a modified GAN architecture that uses generative landmark coordinates to estimate reenacted landmark coordinates for the driving image, excluding the original identity. This allows for better identity transfer.

Table 8
Facial reenactment tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding implementation, and the last column special feature of the tool.

Name	Link	Feature
AVFR [7]	http://cvit.iiit.ac.in/research/projects/cvit-projects/avfr	Real-time interactive demo.
FDGLS [29]	https://github.com/StelaBou/stylegan_directions_face_reenactment	One-shot reeanctment.
StyleMask [30]	https://github.com/StelaBou/StyleMask	High-quality results even in extreme poses.
NeuralVoicePuppetry [263]	https://github.com/miu200521358/NeuralVoicePuppetryMMD	Short target video sequence (2–3 min).
Face2Face [264]	https://github.com/datitran/face2face-demo	Real-time using webcam.
ATVGnet [40]	https://github.com/lelechen63/ATVGnet	Novel approach generating sharper and well-synchronized image.
You said that? [52]	https://github.com/joonson/yousaidthat	Generation of videos with arbitrary person from the arbitrary audio input.
Speech-Driven Animation [278]	https://github.com/DinoMan/speech-driven-animation	Generates final video only from a still image of the target person.
First Order Model [249]	https://github.com/AliaksandrSiarohin/first-order-model	Unsupervised learning, not limited to face swapping.
Articulated Animation [250]	https://github.com/snap-research/articulated-animation	Unsupervised learning, more animation types.

3.4.2. *Tools*

The popularity of facial reenactment tools is smaller in comparison to face-swapping. This reflects in the availability of tools. The facial reenactment tools are listed in Table 8. All the tools exist just as open-source implementations with no special user interface provided. This places higher knowledge requirements on the users. However, with pre-trained models and extensive tutorials available, this technology remains accessible to tech-savvy individuals.

3.4.3. *Datasets*

One dataset dedicated solely to facial reenactment deepfakes exists – DeepFake MNIST+ [110]. It consists of 10,000 facial animation videos in ten different actions. Additionally, a portion of the FaceForensics dataset [236] contains reenacted videos (FaceForensics++ dataset also contains reenacted videos; however, their content is the same). The dataset consists of 1004 unique YouTube videos modified by Face2Face [264]. Ultimately, the quality of available datasets is the same as for face-swapping (Section 3.3.3).

3.4.4. *Attack vector*

Reenacted videos might be primarily used for identity theft. The malicious video can portray an individual saying things she or he never did. This way, fake news can be spread, the reputation of an individual ruined, or fraud committed. We consider reenactment deepfakes to be the most harmful type as they allow complete impersonation using only one image of the victim. In combination with deepfake speech, extremely realistic media can be created. These.

Deepfakes might be spread as fake news or in real-time scenarios such as videoconferencing. The possibilities and impacts of combining facial and speech deepfakes are further discussed in Section 5.2.

This form of deepfakes might also pose severe threats to the Know Your Customer (KYC) process [233]. We need to undergo the KYC process in various applications developed for smartphones where our identity needs to be verified, i.e., gambling portals, insurance portals, or banking applications. This might be done by taking a picture of a customer's ID and then capturing the customer's face with a smartphone camera. While the customer's face is captured, they must move their head, follow a dot on the screen, or others. We can find numerous solutions implementing this scenario.[1–5] This approach prevents an attacker from using a still image to get verified as another person. However, facial reenactment techniques might allow the attacker to overcome this kind of identity verification. Unlike face-swapping, facial reenactment might currently be used in real time. An incident of this kind has already been reported from China, where two individuals used stolen facial images to create deepfake videos [28]. They used a special phone with a hijacked camera to trick the tax invoice system into accepting these premade deepfake identities.

Another possibility of how to misuse reenactment is in child predator threat scenarios [32]. The predator hides his identity behind a virtual avatar that does not even need to represent any existing individual. Only a child's face and voice are important.

We might say that the reenactment is a suitable form of deepfakes to spoof face biometrics systems implementing challenge-response liveness detection mechanisms in real-time.

[1] https://getid.com/solutions/aml-and-kyc-compliance/
[2] https://kyc-chain.com/id-verification/
[3] https://get.cognitohq.com/kyc-know-your-customer/
[4] https://www.mobbeel.com/en/mobbscan-onboarding/
[5] https://www.bioid.com/identity-proofing-photoverify/

3.4.5. Detection

There are many similarities in detecting face-swapped media and facial reenacted ones. We thus discuss only the papers focused mainly on reenactment detection.

As previously mentioned, one of the possibilities is to exploit visual artifacts. Only Kumar et al. [153] propose a multi-stream network that learns regional artifacts solely for reenactment detection. We et al. [286] further extend this architecture by adding global information to each part responsible for a local part.

In contrast, Demir et al. [63] exploit biological signals by examining the eye and gaze features. Agrawal et al. [9] learn the movement and position of the head from an original video and use them to detect spoofs. Mittal et al. [190] exploit the emotional cues. Moreover, methods examining inconsistencies between mouth movement a spoken phoneme exist [8,326].

3.5. Face manipulation

Face manipulation is a technique used to modify a specific part of a target's face in an image or video. The identity of the target remains unchanged. The attacker can either add or remove features like facial hair, glasses, and others or change/transfer the head's expressions, lighting, or pose [271]. The expression and pose transfers overlap with the facial reenactment (Section 3.4). Because of this overlap, we will primarily focus on the techniques for altering the appearance of an individual, such as a change of hair color, rather than the expression or the pose.

3.5.1. Technologies

The most used technologies are GANs or Variational Auto-Encoders (VAEs), which are tweaked on various aspects [141,177,246, 247]. For example, Liu et al. [177] propose a modification to the GAN framework that allows the manipulation of faces in 3D space. StyleGAN is used to synthesize the manipulated faces. Similarly, Kwak et al. [154] propose a 3D-aware GAN architecture. Hou et al. [100] propose to use a knowledge network that controls the StyleGAN-based generator. This combination controls various facial attributes, including smiling, eyeglasses, gender, mustache, or hair color.

In addition, Zhu et al. [329] propose a method for manipulation based on free-form text input from the user. The method combines the generative model space of StyleGAN and the text embedding space of CLIP [222]. Mapping the latent spaces of both tools allows manipulation based on a text prompt.

3.5.2. Tools

As mentioned in this section, we focus solely on the tools for appearance altering. The situation for face manipulation is similar to facial reenactment. The available tools are in the form of experimental implementations available as open-source tools. Again, no special tools with an intuitive user interface exist, increasing the entry knowledge needed. However, the availability of pre-trained models and tutorials enables tech-savvy individuals to access these tools. Table 9 provides an overview of available tools.

3.5.3. Datasets

There is an overlap between facial reenactment and face manipulation datasets. Most of the datasets containing facial reenactment deepfakes also contain manipulated faces. In addition, there is no distinct line between facial reenactment and face manipulation in terms of their exact definition. Table 10 provides an overview of available datasets and their sizes.

3.5.4. Attack vector

An attacker can manipulate specific attributes of her/himself and present the modified face to the biometrics system so that she/he is misclassified as someone else. This might be either achieved by tricking the system into classifying the attacker as a specific another person (impersonation) or arbitrary another person (dodging) [245]. This might be done by changing the hair color, adding facial hair, or removing tattoos from the individual's face. We thus evaluate the threat posed to biometrics systems as moderate – higher than face synthesis but lower than face swapping or reenactment.

Table 9
Face manipulation tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding implementation, and the last column special feature of the tool.

Tool	Link	Feature
InterFaceGan [246, 247]	https://github.com/genforce/interfacegan	Novel framework for semantic face editing by interpreting the latent semantics learned by GANs.
SkinDeep	https://github.com/vijishmadhavan/SkinDeep	Removal of tattoos from images.
StyleMapGAN [141]	https://github.com/naver-ai/StyleMapGAN	State-of-the-art results regarding local editing and image interpolation.
GAIA [237]	https://github.com/timsainb/GAIA	Utilization of pixel-wise error function to minimalize blurriness.
ELEGANT [290]	https://github.com/Prinsphield/ELEGANT	Transferring multiple face attributes by exchanging latent encodings.
MGPE [88]	https://github.com/cientgu/Mask_Guided_Portrait_Editing	Local manipulation, face and hair swapping.
UCLT [67]	https://github.com/endo-yuki-t/UserControllableLT	Allows multiple manipulations based on mouse input.
SURF-GAN [154]	https://github.com/jgkwak95/surf-gan	Users can control different facial and camera parameters.

Table 10
Face manipulation datasets. Each line represents a different dataset. The first column denotes the dataset name or author, and the second column denotes the dataset size.

Name	Size
Zhou et al. [323]	1005 deepfake images
Dang et al. [57]	240,336 deepfake images
Face-Forensics++ [235]	1.8 million deepfake images

Such manipulated media might also be submitted to the court as evidence. This way, the attacker might disguise her/himself in video or image evidence to evade justice. Moreover, the attacker might tamper with the evidence to accuse an innocent person falsely.

3.5.5. Detection

Facial manipulation detection might be generally divided into two classes: data driven methods and handcrafted feature-based methods [293].

Data-driven methods use the classification abilities of neural networks. Afchar et al. [6] utilize CNNs with a small number of layers and focus on the mesoscopic properties of images. Rössler et al. [235] propose custom network architecture XceptionNet. Guo et al. [93] propose an adaptive residual extraction network that exploits the tampering artifacts.

Handcrafted feature-based methods aim to detect the affected regions of the manipulated faces [57,178,323]. Nataraj et al. [197] use the color co-occurrence matrix as input for a neural network. Bappy et al. [16] propose using joint pixel-wise segmentation of manipulated regions as input for CNN, and Li et al. [165] propose using novel image representation called face X-ray.

4. Speech deepfakes

There are two main methods for creating deepfake speech: text-to-speech synthesis (TTS) and voice conversion (VC) [282]. The main difference is in the input data. As the name suggests, TTS consumes written text as input and produces synthesized speech that sounds like a particular individual. In contrast, VC consumes a source voice saying desired phrase and a target voice and outputs the source phrase spoken by the target voice [198].

The following sections discuss mentioned speech synthesis approaches. In addition, speech morphing is discussed as a relatively unknown means of deepfake speech creation. An overview of currently used technologies, tools, available datasets, and attack vectors for each category is discussed. Finally, we provide an overview of deepfake speech detection methods. We provide this overview as a whole, not for each category individually, as there seems to be no special dedication of detection methods to the specific deepfake speech types.

4.1. Text-to-speech synthesis

Text-to-speech (TTS) synthesis is a process of generating speech from written text [262]. This process aims to synthesize speech that is not only easily understandable but also indistinguishable from the speech spoken by humans [256]. This technique finds a place in providing computer-human interfaces for smart assistants or navigation systems.

4.1.1. Technologies

The most used approach to text-to-speech synthesis nowadays is concatenative synthesis. Speech is generated by concatenating small, prerecorded speech units into the final utterance. The concatenative approaches are often called corpus-based speech synthesis [181,256]. Moreover, most recently published TTS architectures are non-autoregressive models [66,232,332]. However, one-shot (few-shot) multi-speaker architectures seemed to dominate in 2022 [112,287,298]. These architectures require only a short embedding recording (a few seconds) of the target speaker to synthesize speech in its voice. Zhao et al. [321] use a speaker-guided conditional variational autoencoder to extract the speaker-specific information from embedding recording and use it to condition the synthesis process further. Choi et al. [51] propose a new learning method for zero-shot TTS. The proposed method first generates an additional speech of a query speaker using the external untranscribed datasets at each training iteration. Then, the model learns to consistently generate the speech sample of the same speaker as the corresponding speaker embedding vector by.

Employing an adversarial learning scheme. Another trend in TTS is end-to-end synthesis. End-to-end is a type of system that can be trained on (text, audio) pairs without phoneme duration annotation. This vastly simplifies the training process and speeds up the data-preparation phase. Numerous architectures have been published recently by Kim et al. [142], Cho et al. [50], and others [95].

Additionally, Riberio et al. [239] focus on synthesizing speech with expressions. To this extent, voice conversion is first used to generate data from the set of expressive speakers; then, the expressive data is pooled with the natural data of the target speaker. This combination is ultimately used to train a single-speaker TTS. Expressional TTS approaches were also published by Monge Alvarez et al. [193], or Huang et al. [111].

One of the limitations of the expressional TTS seems to be the single-speaker setting, which limits the synthesis of arbitrary speech or language without additional training.

Table 11
Text-to-speech synthesis tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding implementation, and the last column special feature of the tool.

Name	Link	Feature
MozillaTTS	https://github.com/mozilla/TTS	High-performance models, large community.
Real-Time-Voice-Cloning [54]	https://github.com/CorentinJ/Real-Time-Voice-Cloning	Very short embedding for synthesis (5s).
Overdub	https://www.descript.com/overdub	Commercial tool for high quality speech synthesis.
ResembleAI	https://www.resemble.ai/cloned/	Online tool for text-to-speech synthesis.
Amazon Polly	https://aws.amazon.com/polly/?p=ft&c=ml&z=3	Advanced text-to-speech technology, preparedvoices.
Google TTS	https://cloud.google.com/text-to-speech	Accessible through API, creating custom voice.
Watson	https://www.ibm.com/cloud/watson-text-to-speech	API service, variety of languages and voices.
ESPnet	https://github.com/espnet/espnet	End-to-end speech processing toolkit managed by community.
CoquiTTS	https://github.com/coqui-ai/TTS	Community curated library for advanced TTS.
TensorFlowTTS	https://github.com/TensorSpeech/TensorflowTTS	Collection of real-time state-of-the-art speech synthesis architectures.
TransformerTTS	https://github.com/as-ideas/TransformerTTS	Robust, fast and controllable non-auto-aggressive synthesis transformer model.
Flowtron [270]	https://github.com/NVIDIA/flowtron	Control of speech variation, interpolation and style transfer between speakers seen and unseen during training.
Emotional TTS [122]	https://github.com/Emotional-Text-to-Speech/dl-for-emo-tts	Transition of emotion onto synthesized speech.
YourTTS [36]	https://github.com/edresson/yourtts	Multilingual approach to zero-shot multi- speaker TTS.
Vall-E [279]	https://github.com/enhuiz/vall-e	High-quality personalized speech synthesis with only a 3-s enrollment recording.

Moreover, combined solutions allowing TTS and VC are being published. Lei et al. [160] propose a Glow- WaveGAN 2 architecture. The GAN backbone first learns to extract the latent distribution of speech and reconstruct the waveform from it. Then a flow-based acoustic model only needs to learn the same latent space from texts, which naturally avoids the mismatch between the acoustic model and the vocoder, resulting in high-quality generated speech without model fine-tuning.

Ultimately, Rojc and Mlakar [234] examine the acoustic inventories used for concatenative TTS. As these inventories often grow exponentially with provided data, it is necessary to optimize them. The authors propose using Long Short-Term Memory networks to represent this space, which ultimately reduces the storage requirements by 90% and lookup time by 70%, making the synthesis run faster with fewer resources.

4.1.2. Tools

A vast amount of TTS tools exist. There are commercial and open-source tools. The commercial tools allow easy and quality synthesis with minimal effort. Some allow the creation of custom voices, and some only synthesize the provided pre-trained voices. This easy accessibility broadens the user base and, unfortunately, allows more people to misuse synthetic speech. The open-source tools require deeper knowledge; however, they reward the user with extended usability. The recent advancements even allow the training of the models in an arbitrary language without any special knowledge and effort. This makes the TTS tools very powerful. An overview of available tools is provided in Table 11.

4.1.3. Datasets

The number of datasets has proliferated in the past year, as shown in Table 12. There are various datasets containing synthesized speech using the TTS synthesis. In addition, it is possible to construct a dataset using demo recordings from newly developed tools or from the Blizzard Challenge [325] that aims to bring novel TTS systems. Similarly, the ASVspoof challenge [302,303] provides both TTS and VC synthesized speech. It is possible to identify several issues in the available datasets.

- No paired identities (genuine - deepfake pairs for one speaker). It is then impossible to address the quality of deepfakes. Moreover, in most real-world scenarios, deepfakes are used for impersonation, so generic deepfake speech is insufficient to model real-world use cases.
- Obsolete datasets - with rapid advancements in deepfake creation tools, it is mandatory to keep up to date with this trend. There are significant differences between synthetic speech created a year ago and now. Moreover, deepfake detectors seem to struggle with unseen types of deepfakes (e.g., from a new speech synthesis tool). Using such datasets thus limits the abilities of novel detectors.
- Short utterances - deepfake datasets often contain only standalone sentences. While this might be enough to train deepfake detection methods, real-world usages of deepfakes often require speaking more than one sentence. These sentences also have to follow up with each other, to only by content but also by being persistent in speech quality.
- Language - available datasets primarily contain English speech. There is currently no evidence on whether deepfake detection is language-dependent. To evaluate this behavior, multilingual deepfake datasets are needed. Such datasets are also useable for assessing the human ability to deepfake recognition.

Table 12
Text-to-speech synthesis datasets. Each line represents a different dataset. The first column denotes the dataset name or author, and the second column denotes the dataset size.

Name	Size (syn. speech)
ASVspoof 2019 [302]	NA
ASVspoof 2021 [303]	NA
ADD [307]	NA
TIMIT-TTS [238]	80,000+ utterances
SV2TTS[a] [119]	66 utterances
WaveFake [75]	117,985 utterances
FoR [231]	87,000 utterances
SYNSPEECHDDB [319]	127,890 utterances
FMFCC-A [320]	40,000 utterances
F&M [73]	1600 utterances
FAD [182]	115,800 utterances
FakeAVCeleb [138]	500 utterances

[a] https://google.github.io/tacotron/publications/speaker_adaptation/.

The listed issues lead to the development of methods that struggle with generalization. Moreover, existing datasets are not well suited for exploring human capability in deepfake detection. We thus strongly encourage the development of novel datasets that resolve the stated issues. Inspiration might be found in the facial domain, face morphing exactly, where the community is currently resolving similar problems with generalization.

4.1.4. Attack vector

Bateman [18] discusses the usage of synthetic voice for identity theft. A phone call made in a victim's synthesized voice could trick the victim's executive assistant or financial advisor into initiating a fraudulent wire transfer. This attack is a form of phishing but using a voice, thus called vishing. An incident of this kind already took place in 2019 when a con artist misusing synthetic speech was able to initiate a fraudulent wire transfer of nearly $250 k [208]. As reported, a CEO of an energy company thought he was speaking via phone to his boss. The caller asked him to transfer funds to a Hungarian supplier in an urgent request. The victim, deceived into thinking that the voice was that of his boss, made the transfer. A similar attack has been reported in UAE [31]. Moreover, a recent report by Iacono et al. [117] shows a significant increase in vishing attacks. Vishing attacks were reported by 69% of companies in 2021, which has risen from the 54% experienced in 2020. According to Quarterly Threat Trends - Intelligence Report from Agari and PhishLabs[6] an extreme uptick in the use of vishing in response-based scams between Q1 2021 to Q1 2022 of almost 550% has been spotted.

The synthetic voice might also be used to create bank accounts under false identities. In addition to the bank scenarios, a bad actor might use the synthesized speech of a victim to access his account secured by a voice biometrics system [73,233,244]. A similar attack on speaker recognition might also be used in the corporate environment [32]. Imagine a company using speaker recognition for employee identity verification during phone calls to IT or human resources departments. An attacker collects essential information about a specific employee along with her or his voice samples to synthesize the speech of the selected individual. The attacker then uses the synthetic voice to call the IT department and claims a forgotten password and a need to reset it. The IT department provides the attacker with a temporary password to access the corporate systems.

Finally, synthetic speech might also threaten smart home assistants or smartphone assistants, as they are often reasonably powerful and controlled only using the owner's speech. An attacker misusing synthetic speech might be able to control such devices. We thus evaluate the threat posed to voice biometrics systems by synthesized speech as high.

4.1.5. Detection

Currently, no special methods for text-to-speech synthesis detection exist; thus, we discuss synthetic speech detection as a whole in Section 4.4.

4.2. Voice conversion

Voice conversion is a technique used for modifying a given speech from a source speaker to match the vocal qualities of a target speaker [183,220]. In contrast to TTS, this process is independent of the spoken content and thus does not require transcriptions. Some of the most advanced voice conversion frameworks can separately transfer components of speech such as timbre, pitch, or rhythm [219].

The most widespread usage of voice conversion can be found in online games, voice parodies, and remixed songs. The voice conversion tools can change any of the voice characteristics (age, sex, …) of the original voice to conceal the real identity of an individual [11].

[6] https://info.phishlabs.com/quarterly-threat-trends-and-intelligence-may-2022

4.2.1. Technologies

There are two major approaches to voice conversion: parallel and non-parallel. Their dependency on text transcriptions might make a further distinction [192]: a text-dependent approach requires a word or phonetic transcription for the recordings, a text-independent approach does not use transcription, which forces the system to find speech segments with similar content before building a conversion function. Finally, VC approaches might be divided by the languages that the source and target speaker speak [192]: a language-independent approach allows for different source, and target speaker languages, language-dependent approach restrains the source and target speaker to speak the same language.

Parallel VC is trained using a parallel dataset. The source and target utterances contain the same speech. Training is usually done for each pair of source and target speakers [33]. One of the first approaches for parallel VC was using statistical techniques. Lee et al. [157] propose Gaussian Mixture Model-based (GMM-based) approach, and Ye et al. [306] utilize the maximum likelihood-based approach. In the meantime, more GMM-based [13,116,259,331] and Hidden Markov Models (HMM) based [285,312] methods were proposed. Another approach utilizes linear and non-linear algebra techniques [81,99] such as Bilinear Models, Linear Regression, or kernel transformations. Popa et al. [217] propose the usage of local linear transformations, and Song et al. [253] propose the usage of Support Vector Regression. Signal processing techniques also find a place in VC approaches. There are approaches based on the Vocal Tract Length Normalization [176], frequency warping [265], or PSOLA technique [132]. Ultimately, cognitive techniques such as neural networks or classification and regression trees are used [15,41,191,288].

Non-parallel VC does not require a parallel dataset for training. This technique is newer and offers a more practical use. Non-parallel models must learn the mapping from source to target speaker without aligning frames with the same content. Neural networks are almost exclusively used for this purpose, mainly GANs, Auto-Encoders (AEs), or Variational Auto-Encoders (VAEs). Even further, the non-parallel VC systems might be split depending on the level of the generalization: One-to-One systems where a unique model has to be trained for each pair of speakers and Many-to-Many systems where one universal model is used for a variety of speakers from the training set. A one-shot voice conversion was developed to be independent of the dataset as the model does not require to be trained for a specific speaker; only a short embedding utterance is needed. It thus allows converting speech from source to target even if none of these speakers is contained in the training set [33].

The base technology behind non-parallel voice conversion remains the same throughout different works. The differences appear in modifications and tweaks to these technologies [126,128,211].

Regarding GANs, Nguyen and Cardinaux [201] propose an end-to-end architecture that does not require a vocoder. Instead, VC is performed directly on the raw audio waveform, improving the final speech quality significantly. Kaneko et al. [129] propose masking the input Mel-spectrogram with a temporal mask and encouraging the converter to fill in the missing frames. Kameoka et al. [123] propose using an encoder-decoder architecture with GAN, Wang et al. [281] propose a cycle for restricting the disentanglement, instead of the previous work for reducing speech size to get content, or Li et al. [171] who further improve the StarGAN [127] architecture.

Regarding AEs, Cassanova et al. [36] modify the VITS model proposed by Kim et al. [142] by using raw text instead of phonemes (end-to-end architecture) and modify the encoder, decoder architecture by changing the layer configurations. Chen et al. [48] propose Activation Guidance and Adaptive Instance Normalization to disentangle content from style instead of reducing dimensionality or quantizing content embedding, which results in a better trade-off between the synthesis quality and the speaker similarity.

A portion of the AE-based voice conversion research focuses on disentangle-based learning techniques to separate the timbre and the linguistic content information from a speech signal. Qian et al. [219] propose using three separate encoders, each responsible for a different speech component. Tang et al. [261] use a vector quantization approach in combination with a special bottleneck proposed by Qian et al. [220] to separate content and timbre information from speech more effectively.

Regarding VAEs [49,113,124,146], Lian et al. [171] propose a novel approach where disentangled speech representations are learned using self-supervised learning. The disentanglement is obtained by balancing the information flow between global speaker representation and time-varying content representation. Lian et al. [170] also address content and speaking style disentanglement.

Apart from neural networks, there are still other approaches such as i-vector PLDA [145], restricted Boltzmann machines [159] or transformer models [173]. Niekerk et al. [272] compare using discrete and soft speech units as input features. Based on their findings, they propose soft speech units learned by predicting a distribution over discrete units. The soft units capture more content information, improving the intelligibility and naturalness of converted speech.

Moreover, Huang et al. [108] examine how utterances degraded by noise or reverberation affect voice conversion. Based on their findings, speech enhancement concatenation and denoising training is proposed to enhance the robustness. Ultimately, this approach significantly improves the quality of converted speech when audio is not of studio quality.

Current trends in VC are similar to the TTS area (Section 4.1.1) – one-shot/zero-shot approaches. Xiao et al. [289] study the impact of speaker embeddings on zero-shot voice conversion performance. As a result, the authors propose a novel speaker representation method that provides superior results to D-vector and GST-based speaker embedding systems.

4.2.2. Tools

Most of the available tools for VC provide a non-parallel approach and utilize deep neural networks. In contrast to TTS (Section 4.1), there are currently no commercial tools available. The required user knowledge is thus higher for this category. However, the availability and usability of the open-source implementations are at the same level as for TTS. For further information thus, see Section 4.1.2. An overview of publicly available tools is shown in Table 13.

Table 13
Voice conversion tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding implementation, and the last column special feature of the tool.

Name	Link	Feature
SpeechSplit [219]	https://github.com/auspicious3000/SpeechSplit	Separate transfer on timbre, pitch and rhythm without text labels.
AutoVC [220]	https://github.com/auspicious3000/autovc	Many-to-many zero-shot voice conversion.
FragmentVC [173]	https://github.com/yistLin/FragmentVC	Any-to-any voice conversion.
AdaptiveVC [49]	https://github.com/jjery2243542/adaptive_voice_conversion	Only one source and target utterance required.
Sprocket [147]	https://github.com/k2kobayashi/sprocket	Vocoder-free voice conversion.
StarGAN [123]	https://github.com/hujinsen/StarGAN-Voice-Conversion	Needs only several minutes of training audio.
CycleGAN [125]	https://github.com/leimao/Voice-Converter-CycleGAN	Mapping sequential and hierarchical struc-tures while preserving linguistic information.
CycleGAN-VC2 [126]	https://github.com/jackaduma/CycleGAN-VC2	Improved objective, discriminator and generator.
CycleGAN-VC3 [128]	https://github.com/jackaduma/CycleGAN-VC3	Time-frequency adaptive normalization.
MaskCycleGAN-VC [129]	https://github.com/GANtastic3/MaskCycleGAN-VC	State-of-the-art results of naturalness and speaker similarity.
YourTTS [36]	https://github.com/edresson/yourtts	Zero-shot voice conversion for low-resource languages.
SoftVC [272]	https://github.com/bshall/soft-vc	Improved intelligibility and naturalness of speech.
ASSEM-VC [144]	https://github.com/mindslab-ai/assem-vc	Any-to-many non-parallel voice conversion.
AGAIN-VC [48]	https://github.com/KimythAnly/AGAIN-VC	Improved quality and speaker similarity.
ContolVC [42]	https://github.com/MelissaChen15/control-vc	Time-varying controls on pitch and speed.
FreeVC [163]	https://github.com/olawod/freevc	Improved content extraction, no text annotation needed.

Table 14
Voice conversion datasets. Each line represents a different dataset. The first column denotes the dataset name or author, and the second column denotes the dataset size.

Name	Size
ASVspoof 2019 [302]	NA
ASVspoof 2021 [303]	NA
ADD [307]	NA
SYNSPEECHDDB [319]	127,890 utterances
FMFCC-A [320]	40,000 utterances

4.2.3. Datasets

The situation with datasets in the VC area is worse than in the TTS area. Again, there are no standardized datasets containing converted speech. The converted speech might be collected from the results of the Voice Conversion Challenge [308], demo utterances for VC systems, or the ASVspoof challenge [302,303] database that combines both the TTS and VC synthesized speech. The challenges discussed and possible future development are the same as those discussed for the TTS datasets in Section 4.1. The overview of VC datasets is shown in Table 14.

4.2.4. Attack vector

The attack vector of voice conversion is almost identical to the one proposed for TTS in Section 4.1. An attacker can impersonate an identity of a selected individual and use the converted voice to access his account secured by a voice biometrics system. Additionally, we see voice conversion as a more threatening tool due to the possibility of the conversion happening in real time. This way, the attacker does not need to prepare the media in advance and can adequately respond to a broader spectrum of situations. However, voice conversion might not be the final choice over a TTS tool for the attacker because of the possibility of revealing the attacker's identity. There are beliefs that an original speech might be retrievable from the deepfake utterance – a sort of deconversion. While this is only a theoretical claim, we might see efforts in this area. In the end, it is a similar procedure to demorphing as mentioned before within morphing detection solutions in Section 3.2.5.

Eventually, we evaluate the threat posed by voice conversion to voice biometrics systems as high.

4.2.5. Detection

Currently, no special methods for voice conversion detection exist; thus, we discuss synthetic speech detection as a whole in Section 4.4.

4.3. Speech morphing

The term speech morphing refers to a technique of smooth transformation from one signal to another. This combination creates a new signal with an intermediate timbre [35]. The signals should be sufficiently similar to become reasonably aligned and interpolated into the new signal [215].

Table 15
Voice morphing tools. Each line represents a different tool. The first column denotes the name and publication, the second column link to the corresponding implementation, and the last column special feature of the tool.

Name	Link	Feature
Figaro	https://github.com/symphonly/figaro	Real-time open-source voice modification.
VoiceMorphing	https://github.com/nestyme/voice-morphing	Gender classifier and age regressor for morphing.
PyVoiceChanger	https://github.com/juancarlospaco/pyvoicechanger	Real Time Microphone Voice Changer App.

As the literature review shows, speech morphing, or voice morphing, is often confused with voice conversion [11,209]. For the scope of this work, we understand voice conversion as modifying a speech from a source speaker to match the voice of a target speaker (see Section 4.2) and speech morphing as a technique of combining two voices to create an intermediate one [183].

4.3.1. Technologies

Most morphing approaches are based on interpolating sound parametrizations obtained from analysis or synthesis techniques, such as Short-time Fourier Transform, Linear Predictive Coding, or Sinusoidal Model Synthesis [109].

One of the first attempts in the area of speech morphing was administrated by Abe [5] in 1996. Abe proposed modifying the fundamental frequency and DFT spectrum that outputs high-quality speech. A similar approach was proposed by Kawahara et al. [137]. A different approach proposed by Pfitzinger et al. [215] consists of LPC-based source-filter decomposition, separate interpolation, and composition of the morphed speech signal. Pfitzinger mainly focuses on the alignment and interpolation problems, where he proposes using dynamic programming to find the alignments of speech signals. Chappell et al. [38] propose separately interpolating the residual signal and Line Spectrum Pairs representation.

4.3.2. Tools

Currently, no tools are providing the speech morphing functionality solely. The research in this area fell silent almost a decade ago. The lack of any beneficial usage of this technology causes this. However, some tools offer similar functionality (see Table 15). These tools might provide a solid base stone for building a speech morphing tool with some tweaks and modifications.

4.3.3. Datasets

As previously mentioned, speech morphing has not seen any ongoing research in the last decade. This fact reflects even in the current situation with speech-morphing datasets. To the best of our knowledge, no datasets contain any morphed speech the way we define it. This area of speech deepfakes thus opens many possibilities for further research.

4.3.4. Attack vector

The attack vector for speech morphing is hypothetical, as no particular use case currently allows for an attack of this type. However, the idea behind the attack remains the same as for face morphing (Section 3.2). An attacker creates a morphed speech of two individuals. The attacker then uses this speech to enroll in a voice biometrics system. The created voice profile allows both individuals to be matched against this profile.

4.3.5. Detection

As previously mentioned, there are no recently published works in speech morphing. This fact reflects even in detecting morphed speech, as there are currently no methods designed to detect morphed speech. In addition, it is also questionable how well the current deepfake speech detection systems will perform against this type of synthetic speech. An overview of synthetic speech detection methods is provided in Section 4.4.

4.4. Detecting deepfake speech

The literature review reveals that no methods detect specifically only one type of synthetic speech (TTS, VC, speech morphing). The proposed methods approach the detection problem in general, including all means of deepfake speech creation [46]. Because of these reasons, we merge the deepfake speech detection methods into this section.

One of the possibilities is to extract custom features from speech. Muhammad et al. [12] built a GMM-based detection system to extract differences in the spectral power between live humans and replayed voices. Xue et al. [299].

Use the fundamental frequency (F0) originally used to improve the quality of synthetic speech. This frequency is too average for synthetic speech, which differs significantly from real speech. Martín-Doñas and Álvarez [187], in contrast, propose to extract the features using the wav2vec2 feature extractor. These features are then fed into a downstream classifier.

Data augmentation seems to be a popular solution to enhance the generalization ability of the detectors. Chen et al. [47] propose utilizing data augmentation, a special modification of Neural Network architecture, and one-class learning with channel-robust training strategies. The data augmentation was also utilized by Das [59]. Das evaluated various data augmentation approaches and combined the best ones with the ASVspoof 2021 baseline systems into a detection system. The augmented data is also fed into various types of classifiers in solutions proposed by Refs. [44,46,130]. In addition, Cáceres et al. [55] not only propose the usage of data augmentation but a new loss function called Focal Loss with a linear fusion of classifiers with different input features. Also, Chen et al.

[45] experiments with a custom loss function. They propose a large-margin cosine loss function (LMCL). The target of the LMLC function is to maximize the variance between genuine and deepfake samples while minimizing the intra-class variance.

Moreover, the fusion of classifiers was also proposed by Kang et al. [131]. Instead of different inputs, Kang et al. examined the influence of different activation functions on deepfake speech detection. Finally, they created a detection system consisting of classifiers operated by different activation functions. On a similar note, Zhang et al. [318] propose a score-level fusion from different classifiers.

Tak et al. [258] propose using Graph Attention Networks (GATs) to model the spectral and temporal relationship between neighboring sub-bands or segments. Tak et al. [257] further improve the GAT-based system by using the GAT on directly learned representation from the raw waveform. Another example of an end-to-end deepfake detection system was proposed by Ge et al. [79]. The authors propose to learn the detector architecture using a neural network, instead of using hand-crafted architectures.

Acoustic features might also be used to reveal synthetic speech. Conti et al. [53] exploit the fact that synthesized speech often lacks emotional behavior. The detector thus uses high-level features obtained from an emotional recognition system. Chaiwongyen et al. [37] use timbre and shimmer to discriminate between genuine and deepfake speech.

An analysis of the ASVspoof 2019 challenge revealed that most deepfake detection systems are based on deep neural networks. The best performance is obtained by combining score levels of single systems that vary by training data, feature extraction procedures, architectures, or training policy. For further development of detection methods, it seems beneficial to use mix-up techniques to prevent the model overfitting and FIR filters for coded magnitude response emulation [268].

In contrast to the detection methods utilizing the speech signal or some of its forms, Khochare et al. [139] propose detecting deepfake speech using an image-based approach. This approach relies on converting the input speech to a Mel-spectrogram and feeding it to Deep Neural Classifier. A similar approach has been taken by Fathan et al. [68].

Among the less traditional approaches, we find Wang et al. [283], who propose a framework for a neural network-based speaker recognition system that monitors the activations of neurons in different network layers. Yan et al. [304] propose a Res-Net architecture boosted by neural stitching to improve the generalization ability, or Blue et al. [25] use vocal tract reconstruction, as deepfakes often model impossible or highly-unlikely anatomical arrangements.

Finally, to further improve the quality of detection, an examination of the method's behavior might be beneficial.

Ge et al. [80] propose a tool for revealing the unexpected behavior of classifiers.

5. The future of deepfakes

The previous sections discussed facial and speech deepfakes along with detection mechanisms. We consider it essential to add a few topics that are not exactly related to biometrics systems but are important to understand the full scope of deepfake misuse and possible future developments in affected areas.

5.1. Deepfake creation

The literature review has revealed that a lot of focus has been given to zero-shot (one-shot) methods recently. This means that, very soon, training the deepfake creation tools will not be required to obtain quality results. This will allow even more people to access deepfake creation.

With the constant advancements in the quality of deepfake creation tools, we are approaching a situation where anyone can create deepfakes with excellent quality. Additionally, with the first attempts at migrating the deepfake.

Creation networks to end devices, such as smartphones, there will soon be no limit on who or where can create deepfakes.

These developments, thus, expand the power of attackers and create more room for attacks.

5.2. Deepfake attacks

A deepfake attack might be decomposed into three stages: collecting data, creating a deepfake persona, and using this persona for various malicious intents. These steps are visualized in Fig. 8.

Suitable media for deepfake creation might be retrieved from the internet (social networks, streaming platforms, etc.) or locally (capturing a person on the street). With the growth of social networks, collecting large amounts of personal images, videos or recordings has become easier than ever. This simple access to media makes almost anyone a suitable target for deepfake attacks.

Real-time deepfake creation is only a matter of time. Such an ability will significantly boost the power of an attacker. Moreover, combinations of facial and speech deepfakes are starting to be used maliciously. A recent example that could have had a massive impact is a spoofed video of Ukrainian President Zelenskyy [103]. If the video had been better quality or published earlier, the situation could have had a very different closure. A combination of facial and speech deepfakes is compelling in deceiving humans.

We expect to experience more and more attacks misusing the combinations of deepfake speech and video. One potential use-case is intercepting a corporate video meeting (e.g., Teams call), where the attacker joins looking and sounding like one of the employees. The quality of the deepfake and the ability to respond in real-time allow the attacker to join the conversation and retrieve confidential information.

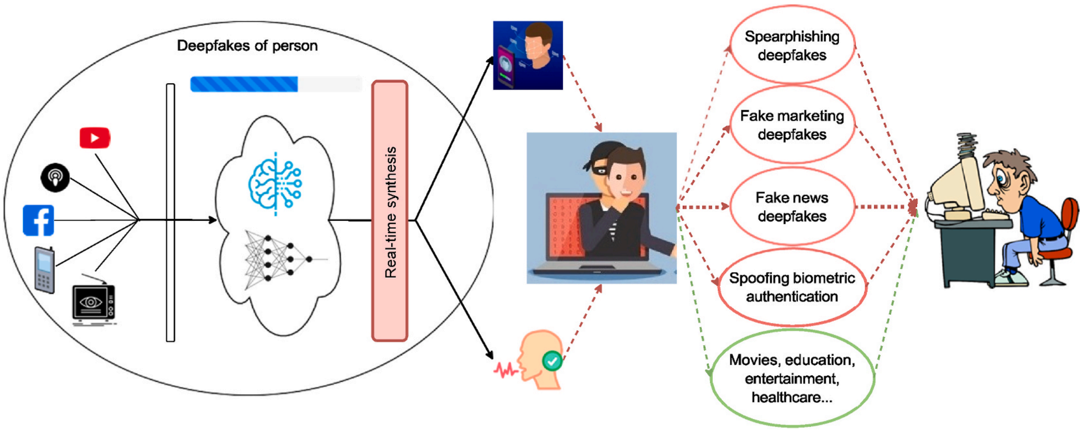


Fig. 8. Deepfake lifecycle.

5.3. Deepfake detection

Deepfake detection is a very timely topic. As the deepfake quality rapidly increased in the past three years, we moved from a situation where no special threats were posed to a situation where deepfakes present a significant problem to computer security. This section provides insight into possible future developments of deepfake detection.

5.3.1. Challenges of deepfake detection

Deepfake detection, being a relatively young area, still faces many challenges that need to be overcome to provide reliable results [185,227]. One of the most acute problems is a generalization. Deepfake detectors perform reasonably well on the known datasets, but as soon as the dataset or operation conditions changes, the reliability and accuracy significantly suffer. This might be decomposed into smaller problems: the lack of datasets, mostly unlabeled data, and unknown types of attacks. While deepfake datasets are still being released, their quality, quantity, and internal structure often do not represent the real world. An ideal dataset should be as diverse as possible, contain thousands of entries, and finally, deepfake and genuine entries should be paired by the same person (ideally with the same content or setting).

To overcome some of these challenges, Raja et al. [227] suggest the following.

- Cross-dataset evaluation – The performance of deepfake detectors should be evaluated using more datasets than just the one used for training.
- Sequestered dataset – An evaluation dataset should be hidden from the creators of deepfake detectors. This way, no solution can be manipulated to perform better during the evaluation, and the unknown type of attack challenge is also being addressed.
- Independent evaluation – Evaluation should be carried out by an unattended third party to simulate the behavior in operational deployment.
- Evaluation platform – To simplify the independent evaluation process, an evaluation platform can be used. This ensures simple access to retrieving the performance of a new detector and the comparability of these results with other detectors.

Another challenge that needs to be addressed is interpretability. While we see the enormous effort in developing new detection techniques, little attention is paid to the underlying and crucial question: How and why does the deepfake detection work? Moreover, to reliably use deepfake detectors, for example, in legal enforcement, it is a must to understand how the used detector works and decides [266].

Ultimately, the vulnerabilities of published detectors should be carefully examined. Deep learning, dominantly used for deepfake detection, allows for the execution of adversarial attacks. Moreover, with regard to the generalization problem, several perturbations and manipulations might be used to fool the detector. It is thus important to understand these vulnerabilities before relying on the detectors [266].

While some efforts occur in the facial recognition domain, speaker recognition hangs behind for a few years.

5.3.2. Obstruction of deepfake creation

While most people image deepfake detectors as a means to ensure our safety from synthetic media, there might be other ways to mitigate the threats posed by deepfakes. One of the approaches is to inhibit the creation of malicious deepfake media proactively. Li et al. [168] propose a method named Landmark Breaker that disrupts facial landmark extraction, which obstructs the creation of deepfake videos. Approaches similar to this mitigate the threats posed by deepfakes before they even can occur, which seems to be more beneficial than the actual detection. Similar solutions might include noise in speech or videos with cryptography properties that prevent the deepfake creation tools from extracting identity-specific information. This area is exciting and promising, so we hope to see more attempts in the future.

5.3.3. Natural defenses against deepfakes

As previously mentioned, detectors might be incorporated to mitigate the threats posed by deepfakes, or deepfake creation might be obstructed. An additional solution might come in the form of naturally spoofing resilient biometrics. The naturally spoofing resilient biometrics are built to reject any spoofing trial without using detectors of any kind.

Recently, Anand et al. [14] proposed a new method for speaker verification using vibrations. This method proved to distinguish between genuine and deepfake attempts reliably. The deepfake attempts utilized synthesized speech. The results indicate that one suitable way to prevent deepfakes is to develop biometrics systems on principles naturally resilient to deepfakes or other spoofing attacks.

Discovering similar methods might again help mitigate the threats posed by deepfakes to biometric authentication.

Finally, the ultimate combination of all the mentioned defenses might be enough to secure today's systems.

5.3.4. Ready-to-use deepfake detectors

While most of the work done in the area of deepfake detection remains in the form of academic publications and their experimental implementations, there are some tools prepared for immediate use, such as Deepware[7] or Sensity[8]. Most of these solutions are proprietary, and their internals is secret. However, this area seems to develop in a direction where deepfake detection is offered as a service. This would allow for easier and more accessible deepfake detection, even for individuals with no particular background in IT.

5.3.5. Public challenges for deepfake detection

The challenges for deepfake detection aim to bring the best detection methods for selected types of deepfakes.

These challenges are held annually and provide insight into the latest defenses against deepfakes.

The image-oriented challenges might be stated as Open Media Forensics Challenge [9] by NIST, Deepfake Detection Challenge [10] b y Meta or Kaggle deepfake detection challenge [11]. Among the speech-oriented challenges, ASVSpoof [302,303] and ADD [307] challenges exist. ASVSpoof is a biannual event that aims to promote the study of spoofing and the design of countermeasures – detectors to protect Automatic Speaker Verification (ASV) systems. The ADD challenge added real-life and challenging scenarios to fill in the gap in deepfake speech detection.

5.4. Societal impacts and legal regulations

In addition to machine-based deepfake detection, it is crucial to understand the human perception of deepfake media. In some use cases (e.g., bank call centers or border crossing), a human operator interacts with the biometrics system and user. In these scenarios, proper human capability to detect deepfakes might improve overall security. Understanding how to differentiate between real and fake media, thus, might help protect society from the posed threats.

As mentioned, the human ability on deepfake detection is already being put to the test. Published research examines the perception of fake faces [83,86,218] or fake speech [195].

The societal aspect is large. Considering the recent developments, we aim towards a future where media can no longer be trusted. As deepfake media quality is improving, we will soon lose our ability to differentiate between real and fake. It is even possible that society will remain dependent on the help of deepfake detectors to reveal what is real. Ultimately, the proper understanding of the impacts on society should lead to developing legal regulations for deepfake media. However, there is a long way to go before we can regulate deepfakes legally. To this extent, it is also important to clarify the responsibilities for creating and spreading deepfakes. Perhaps, fast-acting countries such as China, who already put a legal deepfake regulation to action [21], will serve as a testing ground for further development of such regulations. An interesting fact to add is that China is currently the number one publisher of deepfake-related research, according to the Web of Science. As Fig. 9 shows, almost one-third of the publications originate in this country. Moreover, according to Google Trends, China takes second place in the list of relative interest in deepfake-related topics. This interest is visualized in Fig. 10.

6. Conclusions

Deepfake creation tools are on the rise, and the development of detection methods tries to hold onto this trend. The most popular keywords for deepfake creation seem to be zero-shot, one-shot, and end-to-end. This demonstrates how.

Fast we progress towards generalizable and easy-to-use solutions. We, thus, are only one step from real-time deepfake creation. As soon as we reach this milestone, cyberspace, as we know it today, may change beyond recognition. The development of such methods will significantly increase the power of attackers. With easy access to various media suitable for deepfake creation, more elaborate deepfake attacks will come.

The biometrics systems seem to be threatened by identity-swapping techniques, as they might result in providing unauthorized

[7] https://scanner.deepware.ai

[8] https://sensity.ai

[9] https://mfc.nist.gov/

[10] https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

[11] https://www.kaggle.com/c/deepfake-detection-challenge

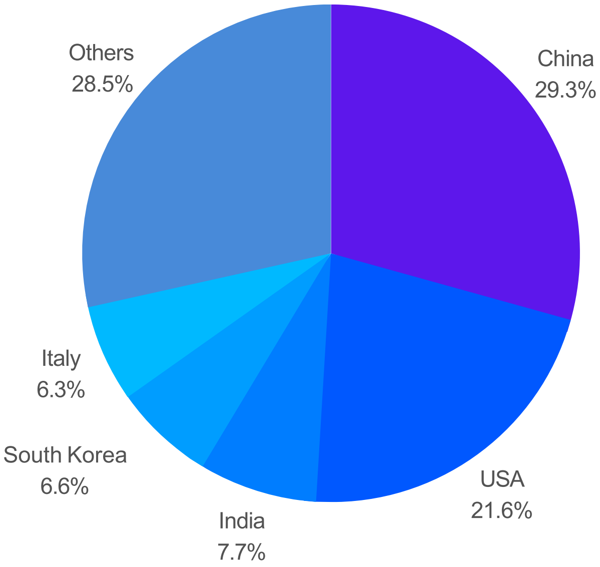


Fig. 9. Publication distribution by world countries for 2021 and 2022 b y keyword "deepfake". The top 5 countries listed according to the WoS, Others represent all remaining world countries. Source: https://www.webofscience.com/wos/woscc/summary/894c369c-e777-4199-89f3-dea9078a35bd-6a94f062/relevance/1.

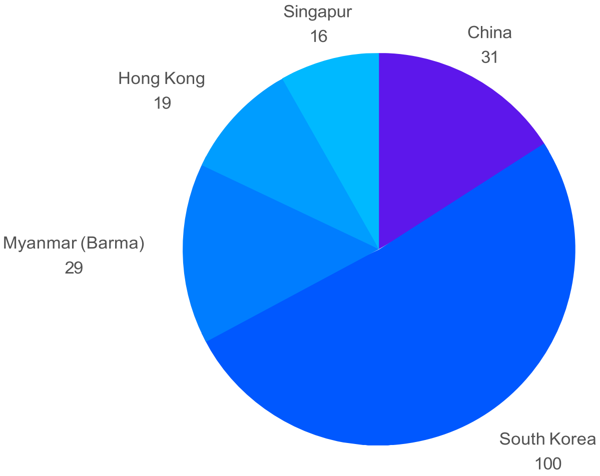


Fig. 10. Popularity of "deepfake" related Google searches. Top 5 countries listed according to Google Trends. The number represents a relative interest compared to other searches in a given country, not the total count of queries. Values are calculated from 0 to 100, where 100 is the location with the most popularity as a fraction of total searches in that location. Source: https://trends.google.com/trends/explore?date=2016-01-01% 202023-01-12&q=deepfake.

access or dodging identification. Face morphing and face swap present primary threats to facial recognition.

Voice biometrics, in contrast, are vulnerable to any deepfake speech. The principle of speech deepfakes causes this – all allow impersonation.

In the deepfake detection domain, generalization and robustness are heavily discussed. Much research focuses on developing novel methods to detect manipulated or noisy media. It is an important milestone we must pass before successfully deploying detection methods to guard cyberspace. Moreover, several attempts to evaluate and improve the human ability to detect deepfakes have been reported. Human-based detection is an important part of deepfake detection, as there is not always the possibility to rely on technology.

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

This work was supported by Fakulta Informačních Technologií, Vysoké Učení Technické v Brně [FIT-S-23-8151].

Data availability statement

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

References

[1] Create AI Voices that sound real, January 16, 2023, https://www.resemble.ai, 2021.

[2] Ultra-realistic voice cloning with Overdub, January 16, 2023, https://www.descript.com/overdub, 2021.

[3] Online Deepfake Maker, January 9, 2023, https://deepfakesweb.com, 2022.

[4] Reface: Be Anyone and Reface Anything, January 9, 2023, https://hey.reface.ai, 2022.

[5] M. Abe, Speech morphing by gradually changing spectrum parameter and fundamental frequency, in: Proceeding of Fourth International Conference on Spoken Language Processing, 4, ICSLP '96, Philadelphia, PA, USA, 1996, pp. 2235–2238, https://doi.org/10.1109/ICSLP.1996.607250. IEEE.

[6] Darius Afchar, Nozick Vincent, Junichi Yamagishi, Isao Echizen, MesoNet: a Compact Facial Video Forgery Detection Network.2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7, https://doi.org/10.1109/wifs.2018.8630761. Dec 2018.

[7] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V. Jawahar, Audio-visual face reenactment, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 5178–5187.

[8] Shruti Agarwal, Hany Farid, Ohad Fried, Maneesh Agrawala, Detecting deep-fake videos from phoneme-viseme mismatches, In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2814–2822, https://doi.org/10.1109/CVPRW50498.2020.00338.

[9] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, Hao Li, Protecting world leaders against deep fakes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[10] Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy Dawson, M. Nasser, Nasrabadi, Morph detection enhanced by structured group sparsity, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV) Workshops, 2022, pp. 311–320.

[11] Ijaz Ahmed, Ayesha Sadiq, Muhammad Atif, Mudasser Naseer, Adnan Muhammad, Voice Morphing: an Illusion or Reality. In2018 International Conference on Advancements in Computational Sciences (ICACS). 1–6, 2018, https://doi.org/10.1109/ICACS.2018.8333282.

[12] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, Hyoungshick Kim, Void: A Fast and Light Voice Liveness Detection System, USENIX Association, USA, 2020.

[13] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, Ariki Yasuo, GMM-based emotional voice conversion using spectrum and prosody features, Am. J. Signal Process. 2 (12 2012) (2012) 134–138, https://doi.org/10.5923/j.ajsp.20120205.06.

[14] S. Abhishek Anand, Jian Liu, Chen Wang, Maliheh Shirvanian, Nitesh Saxena, Yingying Chen, EchoVib: exploring voice authentication via unique non-linear vibrations of short replayed speech, in: Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, Association for Computing Machinery, New York, NY, USA, 2021, pp. 67–81, https://doi.org/10.1145/3433210.3437518. Virtual Event, Hong Kong) (ASIA CCS '21.

[15] Azarov Elias, Maxim Vashkevich, Denis Likhachov, Petrovsky Alexander, Real-time voice conversion using artificial neural networks with rectified linear units, Proc. Interspeech (2013) 1032–1036, https://doi.org/10.21437/Interspeech.2013-113.

[16] Jawadul H. Bappy, Amit K. Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, B.S. Manjunath, Exploiting spatial structure for localizing manipulated image regions, in: 2017 IEEE Int. Conf. Comput. Vis.(ICCV), 2017, pp. 4980–4989, https://doi.org/10.1109/ICCV.2017.532.

[17] Matthias Bastian, Top gun: maverick - ai helps with Val Kilmer's return. https://the-decoder.com/top-gun-maverick-ai-enables-val-kilmers-return/, 2022.

[18] Jon Bateman, Deepfakes and synthetic media in the financial system: assessing threat scenarios, Carnegie. Endow. Int. Peace (2020) i–ii. Technical Report, http://www.jstor.org/stable/resrep25783.1.

[19] David Beniaguev, Synthetic Faces High Quality (SFHQ) Dataset, 2022, https://doi.org/10.34740/kaggle/dsv/4737549.

[20] Nicolas Beuve, Wassim Hamidouche, Deforges Olivier, DmyT: Dummy Triplet Loss for Deepfake Detection (ADGD '21), vols. 17–24, Association for Computing Machinery, New York, NY, USA, 2021, https://doi.org/10.1145/3476099.3484316.

[21] Ananya Bhattacharya, China Goes a Step Further in Regulating Deepfakes, 2023. https://qz.com/china-new-rules-deepfakes-consent-disclosure-1849964709.

[22] Sandika Biswas, Sanjana Sinha, Dipanjan Das, Brojeshwar Bhowmick, Realistic talking face animation with speech-induced head motion, in: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing (Jodhpur, India) (ICVGIP '21), Association for Computing Machinery, New York, NY, USA, 2021, https://doi.org/10.1145/3490035.3490305. Article 46, 9 pages.

[23] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, K. Shree, Nayar, Face swapping: automatically replacing faces in photographs, in: ACM SIGGRAPH 2008 Papers (Los Angeles, California) (SIGGRAPH '08), Association for Computing Machinery, New York, NY, USA, 2008, https://doi.org/10.1145/1399504.1360638. Article 39.

[24] Volker Blanz, Kristina Scherbaum, Thomas Vetter, Hans-Peter Seidel, Exchanging faces in images, Comput. Graph. Forum 23 (09) (2004) 669–676, https://doi.org/10.1111/j.1467-8659.2004.00799.x.

[25] Blue Logan, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, Patrick Traynor, Who are you (I really wanna know)? Detecting audio DeepFakes through vocal tract reconstruction, in: 31st USENIX Security Symposium (USENIX Security 22), USENIX Association, Boston, MA, 2022, pp. 2691–2708. https://www.usenix.org/conference/usenixsecurity.

[26] Giulia Boato, Cecilia Pasquini, Antonio Luigi Stefani, Sebastiano Verde, Daniele Miorandi, TrueFace: a Dataset for the Detection of Synthetic Face Images from Social Networks, 2022, https://doi.org/10.5281/zenodo.7065064.

[27] Piotr Bojanowski, Joulin Armand, David Lopez-Paz, Arthur Szlam, Optimizing the Latent Space of Generative Networks, 2019 arXiv:1707.05776 [stat.ML].

[28] Masha Borak, Tax Scammers Hack Government-Run Facial Recognition System, 2021. https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax.

[29] Stella Bounareli, Vasileios Argyriou, Georgios Tzimiropoulos, Finding directions in GAN's latent space for neural face reenactment, in: 33rd British Machine Vision Conference 2022, BMVA Press, 2022, pp. 21–24. BMVC 2022, London, UK, November, https://bmvc2022.mpi-inf.mpg.de/0383.pdf.

[30] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, Georgios Tzimiropoulos, StyleMask: Disentangling the Style Space of StyleGAN2 for Neural Face Reenactment, 2022, https://doi.org/10.48550/ARXIV.2209.13375.

[31] Thomas Brewster, Fraudsters Cloned Company Director's Voice in $35 Million Bank Heist, Police Find, 2021. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

[32] Tina Brooks, G. Princess, Jesse Heatley, J. Jeremy, Samantha M. Scott Kim, Sara Parks, Maureen Reardon, Harley Rohrbacher, Burak Sahin, S. Shani, S. James, T. Oliver, V. Richard, Increasing Threat of Deepfake Identities, U.S. Department of Homeland Security, 2021. Technical Report.

[33] Brukner Jan, Non-Parallel Voice Conversion, in: Diplomová Práce. Vysoké Učení Technické V Brně, Fakulta Informačních Technologií, 2020. https://www.fit.vut.cz/study/thesis/19207/.

[34] Chris Burt, Morphing attack detection for face biometric spoofs needs more generalization, datasets: biometric Update. https://www.biometricupdate.com/202207/morphing-attack-detection-for-face-biometric-spoofs-needs-more-generalization-datasets, 2022.

[35] Pedro Cano, Alex Loscos, Jordi Bonada, Maarten de Boer, Xavier Serra, Voice morphing system for impersonating in karaoke applications, in: ICMC, 2000.

[36] Edresson Casanova, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, Moacir A. Ponti, YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, in: Proceedings of the 39th International Conference on Machine Learning (Proceedings Machine Learning Research, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, vol. 162, 2022, pp. 2709–2720. PMLR, https://proceedings.mlr.press/v162/casanova22a.html.

[37] Anuwat Chaiwongyen, Norranat Songsriboonsit, Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, Masashi Unoki, Contribution of timbre and shimmer features to deepfake speech detection, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, 2022, pp. 97–103, https://doi.org/10.23919/APSIPAASC55919.2022.9980281.

[38] David T. Chappell, John H.L. Hansen, A comparison of spectral smoothing methods for segment concatenation based speech synthesis, Speech Commun. 36 (2002) 3, https://doi.org/10.1016/S0167-6393(01)00008-5.

[39] Beijing Chen, Weijin Tan, Yiting Wand, Guoying Zhao, Distinguishing between natural and GAN-generated face images by combining global and local features, Chin. J. Electron. 31 (1) (2022) 59–67, https://doi.org/10.1049/cje.2020.00.372.

[40] Lele Chen, K Maddox Ross, Zhiyao Duan, Chenliang Xu, Hierarchical cross-modal talking face generation with dynamic pixel-wise loss, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7832–7841.

[41] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, Li-Rong Dai, Voice conversion using deep neural networks with layer- wise generative training, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12) (2014) 1859–1872, https://doi.org/10.1109/TASLP.2014.2353991.

[42] Meiying Chen, Zhiyao Duan, ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Speed, 2022, https://doi.org/10.48550/ARXIV.2209.11866.

[43] Renwang Chen, Xuanhong Chen, Bingbing Ni, Yanhao Ge, SimSwap: an efficient framework for high fidelity face swapping, in: MM '20: the 28th ACM International Conference on Multimedia, ACM, 2020, pp. 2003–2011, https://doi.org/10.1145/3394171.3413630.

[44] Tianxiang Chen, Elie Khoury, Kedar Phatak, Ganesh Sivaraman, Pindrop labs' submission to the ASVspoof 2021 challenge, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 89–93, https://doi.org/10.21437/ASVSPOOF.2021-14.

[45] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, Elie Khoury, Generalization of audio deepfake detection, in: Proc. Odyssey 2020 the Speaker and Language Recognition Workshop, 2020, pp. 132–137.

[46] Xinhui Chen, You Zhang, Ge Zhu, Zhiyao Duan, UR channel-robust synthetic speech detection system for ASVspoof 2021, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 75–82, https://doi.org/10.21437/ASVSPOOF.2021-12.

[47] Xinhui Chen, You Zhang, Ge Zhu, Zhiyao Duan, UR channel-robust synthetic speech detection system for ASVspoof 2021, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 75–82, https://doi.org/10.21437/ASVSPOOF.2021-12.

[48] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, Hung-yi Lee, Again-VC: a one-shot voice conversion using activation guidance and adaptive instance normalization, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2021, pp. 5954–5958, https://doi.org/10.1109/ICASSP39728.2021.9414257.

[49] Ju chieh Chou, Hung-Yi Lee, One-shot voice conversion by separating speaker and content representations with instance normalization, Proc. Interspeech (2019) 664–668, https://doi.org/10.21437/Interspeech.2019-2663, 2019.

[50] Hyunjae Cho, Wonbin Jung, Junhyeok Lee, Sang Hoon Woo, SANE-TTS: Stable and Natural End-To-End Multilingual Text-To-Speech, ISCA, 2022, https://doi.org/10.21437/interspeech.2022-46. Interspeech 2022.

[51] Byoung Jin Choi, Myeonghun Jeong, Minchan Kim, Hwan Mun Sung, Nam Soo Kim, Adversarial speaker-consistency learning using untranscribed speech data for zero-shot multi-speaker text-to-speech, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC), 2022, pp. 1708–1712, https://doi.org/10.23919/APSIPAASC55919.2022.9979900.

[52] J.S. Chung, A. Jamaludin, A. Zisserman, You said that?, in: British Machine Vision Conference, 2017.

[53] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C. Stamm, Stefano Tubaro, Deepfake speech detection through emotion recognition: a semantic approach, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8962–8966, https://doi.org/10.1109/ICASSP43922.2022.9747186.

[54] Jemine Corentin, in: Real-time Voice Cloning, Master Thesis, Université de Liège, Liège, Belgique, 2019.

[55] Joaquín Cáceres, Roberto Font, Teresa Grau, Javier Molina, The biometric vox system for the ASVspoof 2021 challenge, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 68–74, https://doi.org/10.21437/ASVSPOOF.2021-11.

[56] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, Fadi Boutros, Privacy-Friendly Synthetic Data for the Development of Face Morphing Attack Detectors, 2022, pp. 1606–1617. June 2022).

[57] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, Anil K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5781–5790.

[58] Rashmiranjan Das, Gaurav Negi, Alan F. Smeaton, Detecting deepfake videos using euler video magnification, Electron. Imag. 4 (Jan 2021) (2021) 272, https://doi.org/10.2352/issn.2470-1173.2021.4.mwsf-272.

[59] Rohan Kumar Das, Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 29–36, https://doi.org/10.21437/ASVSPOOF.2021-5.

[60] Aabir Datta, Om Krishna Yadav, Yukti Singh, S. Sountharrajan, M. Karthiga, E. Suganya, Real-time face swapping system using OpenCV, in: In 2021 Third Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), 2021, pp. 1081–1086, https://doi.org/10.1109/ICIRCA51532.2021.9545010.

[61] Tiago José de Carvalho, Christian Riess, Elli Angelopoulou, Hélio Pedrini, de Rezende Rocha Anderson, Exposing digital image forgeries by illumination color classification, IEEE Trans. Inf. Forensics Secur. 8 (7) (2013) 1182–1194, https://doi.org/10.1109/TIFS.2013.2265677.

[62] Debayan Deb, Jianbang Zhang, K. Anil, Jain, AdvFaces: Adversarial Face Synthesis, 2019 arXiv:1908.05008 [cs.CV].

[63] Ilke Demir, Umur Aybars Ciftci, Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking, Association for Computing Machinery, New York, NY, USA, 2021, https://doi.org/10.1145/3448017.3457387.

[64] Brian Dolhansky, Joanna Bitton, Pflaum Ben, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer, The DeepFake Detection Challenge Dīataset, 2020 arXiv:2006.07397 [cs.CV].

[65] Ted Dunstone, New Face Morphing Dataset (For Vulnerability Research), 2018. https://www.linkedin.com/pulse/new-face-morphing-dataset-vulnerability-research-ted-dunstone/.

[66] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, R.J. Skerry-Ryan, Yonghui Wu, Parallel tacotron 2: a non- autoregressive neural TTS model with differentiable duration modeling, in: Proc. Interspeech 2021, 2021, pp. 141–145, https://doi.org/10.21437/Interspeech.2021-1461.

[67] Yuki Endo, User-controllable latent transformer for StyleGAN image layout editing, Comput. Graph. Forum 41 (7) (2022) 395–406, https://doi.org/10.1111/cgf.14686.

[68] Abderrahim Fathan, Jahangir Alam, Woo Hyun Kang, Mel-spectrogram image-based end-to-end audio deepfake detection under channel-mismatched conditions, in: 2022 IEEE Int. Conf. Multimed. Expo (ICME), 2022, pp. 1–6, https://doi.org/10.1109/ICME52920.2022.9859621.

[69] Matteo Ferrara, Annalisa Franco, Davide Maltoni, The magic passport, in: IEEE International Joint Conference on Biometrics. 1–7, 2014, https://doi.org/10.1109/BTAS.2014.6996240.

[70] Matteo Ferrara, Annalisa Franco, Davide Maltoni, On the Effects of Image Alterations on Face Recognition Accuracy, Springer International Publishing, Cham, 2016, pp. 195–222, https://doi.org/10.1007/978-3-319-28501-6_9.

[71] Anselmo Ferreira, Ehsan Nowroozi, Mauro Barni, VIPPrint: validating synthetic image detection and source linking methods on a large scale dataset of printed documents, J. Imag. 7 (2021) 3, https://doi.org/10.3390/jimaging7030050.

[72] Anton Firc, Applicability of Deepfakes in the Field of Cyber Security, in: Brno University of Technology, Faculty of Information Technology, Brno. Supervisor Mgr. (Kamil Malinka, Ph.D), 2021. Master's Thesis.

[73] Anton Firc, Kamil Malinka, The Dawn of a Text-dependent Society: Deepfakes as a Threat to Speech Verification Systems, in: InProceedings of the 37th ACM/ SIGAPP Symposium on Applied Computing (Virtual Event) (SAC '22), Association for Computing Machinery, New York, NY, USA, 2022, pp. 1646–1655, https://doi.org/10.1145/3477314.3507013.

[74] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, VideoForensicsHQ: detecting high-quality manipulated face videos, in: IEEE International Conference on Multimedia and Expo (ICME 2021), IEEE, Shenzhen,China, 2021, https://doi.org/10.1109/ICME51207.2021.9428101 (Virtual).

[75] Joel Frank, Lea Schönherr, WaveFake: A Data Set to Facilitate Audio Deepfake Detection, 2021 arXiv:2111.02813 [cs.LG].

[76] Tao Fu, Ming Xia, Gaobo Yang, Detecting GAN-generated face images via hybrid texture and sensor noise based features, Multimed. Tool. Appl. 81 (18) (2022) 26345–26359, https://doi.org/10.1007/s11042-022-12661-1. (Accessed 1 July 2022).

[77] Xiaomeng Fu, Xi Wang, Jin Liu, Wantao Liu, Jiao Dai, Jizhong Han, MakeItSmile: detail-enhanced smiling face reenactment, in: 2022 International Joint Conference on Neural Networks (IJCNN). 1–8, 2022, https://doi.org/10.1109/IJCNN55064.2022.9892359.

[78] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, Ran He, Information bottleneck disentanglement for identity swapping, in: In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3403–3412, https://doi.org/10.1109/CVPR46.437.2021.00341.

[79] Wanying Ge, Patino Jose, Massimiliano Todisco, Nicholas Evans, Raw differentiable architecture search for speech deepfake and spoofing detection, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, vols. 22–28, 2021, https://doi.org/10.21437/ASVSPOOF.2021-4.

[80] Wanying Ge, Patino Jose, Massimiliano Todisco, Nicholas Evans, Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 6387–6391, 2022, https://doi.org/10.1109/ICASSP43922.2022.9747476.

[81] Mostafa Ghorbandoost, Abolghasem Sayadiyan, Mohsen Ahangar, Sheikhzadeh Hamid, Shahrebabaki Abdoreza Sabzi, Jamal Amini, Voice conversion based on feature combination with limited training data, Speech Commun. 67 (2015) 113–128, https://doi.org/10.1016/j.specom.2014.12.004.

[82] Oliver Giudice, Luca Guarnera, Sebastiano Battiato, Fighting deepfakes by detecting GAN dct anomalies, Journal of Imaging 7 (2021) 128, https://doi.org/10.3390/jimaging7080128. J. Imag.

[83] Sankini Rancha Godage, Froy Lovåsdaly, Sushma Venkatesh, Kiran Raja, Raghavendra Ramachandra, Christoph Busch, Analyzing human observer ability in morphing attack detection -where do we stand? IEEE Trans. Technol. Soc. (2022) 1, https://doi.org/10.1109/TTS.2022.3231450.

[84] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc, 2014, in: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[85] Graphika Team, Fake Cluster Boosts Huawei, 2022. https://graphika.com/reports/fake-cluster-boosts-huawei.

[86] Matthew Groh, Ziv Epstein, Chaz Firestone, Rosalind Picard, Deepfake detection by human crowds, machines, and machine- informed crowds, Proc. Natl. Acad. Sci. USA 119 (1) (2022), e2110013119, https://doi.org/10.1073/pnas.2.110013119.

[87] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, Denis Dimitrov, GHOST—a new face swap approach for image and video domains, IEEE Access 10 (2022) (2022) 83452–83462, https://doi.org/10.1109/ACCESS.2022.3196668.

[88] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, Yuan Lu, Mask-guided portrait editing with conditional GANs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[89] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy Yates, Andrew Delgado, Daniel Zhou, Timothée Kheyrkhah, Jeff Smith, Jonathan, MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation, in: IEEE Winter Conference on Applications of Computer Vision (WACV 2019), Waikola, HI, 2019. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=927035.

[90] Weinan Guan, Wei Wang, Jing Dong, Bo Peng, Tieniu Tan, Robust face-swap detection based on 3D facial shape information, in: Ruiping Wang (Ed.), Artificial Intelligence, Lu Fang, Daniel Povey, Guangtao Zhai, Tao Mei, Springer Nature Switzerland, Cham, 2022, pp. 404–415.

[91] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, Siwei Lyu, Eyes tell all: irregular pupil shapes reveal GAN-generated faces, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 2904–2908, https://doi.org/10.1109/ICASSP43922.2022.9746597.

[92] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, Siwei Lyu, Robust attentive deep neural network for detecting GAN-generated faces, IEEE Access 10 (2022) (2022) 32574–32583, https://doi.org/10.1109/ACCESS.2022.3157297.

[93] Zhiqing Guo, Gaobo Yang, Jiyou Chen, Xingming Sun, Fake face detection via adaptive manipulation traces extraction network, Comput. Vis. Image Understand. 204 (2021) (2021), 103170, https://doi.org/10.1016/j.cviu.2021.103170.

[94] Muhammad Hamza, Samabia Tehsin, Hanen Karamti, Norah Saleh Alghamdi, Generation and detection of face morphing attacks, IEEE Access 10 (2022) (2022) 72557–72576, https://doi.org/10.1109/ACCESS.2022.3188668.

[95] Chol-Jin Han, Un-Chol Ri, Song-Il Mun, Kang-Song Jang, Song-Yun Han, An end-to-end TTS model with pronunciation predictor, Int. J. Speech Technol. 25 (4) (2022) 1013–1024, https://doi.org/10.1007/s10772-022-10008-7. (Accessed 1 December 2022).

[96] Hassani Ali, Hafiz Malik, Efficient face-swap-verification using PRNU, in: 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), 2022, pp. 42–48, https://doi.org/10.1109/CDMA54072.2022.00012.

[97] Hassani Ali, Hafiz Malik, Jon Diedrich, Efficiently mitigating face-swap-attacks: compressed-PRNU verification with sub-zones, Technologies 10 (2022) 2, https://doi.org/10.3390/technologies10020046.

[98] Zhenliang He, Meina Kan, Shiguang Shan, EigenGAN: Layer-Wise Eigen-Learning for GANs, 2021 arXiv:2104.12476 [cs.CV].

[99] Elina Helander, Silén Hanna, Tuomas Virtanen, Moncef Gabbouj, Voice conversion using dynamic kernel partial least squares regression. Audio, speech, and language processing, IEEE Transactions on 20 (04 2012) (2012) 806–817, https://doi.org/10.1109/TA.SL.2011.2165944.

[100] Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, Jun Wan, GuidedStyle: Attribute knowledge guided style manipulation for semantic face editing, Neural Network. 145 (2022) (2022) 209–220, https://doi.org/10.1016/j.neunet.2021.10.017.

[101] Gee-Sern Hsu, Chun-Hung Tsai, Hung-Yi Wu, Dual-generator face reenactment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 642–650.

[102] Gee-Sern Jison Hsu, Hung-Yi Wu, Pose-guided and style-transferred face reenactment, in: 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 2458–2462, https://doi.org/10.1109/ICIP42928.2021.9506315.

[103] Jeremy Hsu, Deepfake Detector Spots Fake Videos of Ukraine's President Zelenskyy, 2022. https://www.newscientist.com/article/2350644-deepfake-detector-spots-fake-videos-of-ukraines-president-zelenskyy/.

[104] Chen Hu, Xianghua Xie, One-shot decoupled face reenactment with vision transformer, in: Nicole Vincent (Ed.), Pattern Recognition and Artificial Intelligence, Mounîm El Yacoubi, Eric Granger, Pong Chi Yuen, Umapada Pal, Springer International Publishing, Cham, 2022, pp. 246–257.

[105] Chen Hu, Xianghua Xie, Lin Wu, Face reenactment via generative landmark guidance, Image Vis Comput. 130 (2023) (2023), 104611, https://doi.org/10.1016/j.imavis.2022.104611.

[106] Jinyu Hu, Yuchen Ren, Yuan Yuan, Li Yin, Lei Chen, PathosisGAN: sick face image synthesis with generative adversarial network, in: 2021 2nd International Conference on Artificial Intelligence and Information Systems (Chongqing, China) (ICAIIS 2021), Association for Computing Machinery, New York, NY, USA, 2021, https://doi.org/10.1145/3469213.3470691. Article 258, 6 pages.

[107] Shu Hu, Yuezun Li, Siwei Lyu, Exposing GAN-generated faces using inconsistent corneal specular highlights, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2500–2504, 2021.

[108] Chien-Yu Huang, Kai-Wei Chang, Hung-Yi Lee, Toward degradation-robust voice conversion, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6777–6781, https://doi.org/10.1109/ICASSP4392.2.2022.9746636.

[109] Dong-Yan Huang, Susanto Rahardja, Ee Ping Ong, High level emotional speech morphing using straight, in: Seventh ISCA Workshop on Speech Synthesis, 2010.

[110] Jiajun Huang, Xueyu Wang, Bo Du, Pei Du, Chang Xu, DeepFake MNIST+: a DeepFake facial animation dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 1973–1982.

[111] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, Zhao Zhou, GenerSpeech: towards style transfer for generalizable out- of-domain text-to-speech, in: Advances in Neural Information Processing Systems, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, 2022. https://openreview.net/forum?id=dmCyoqxEwHf.

[112] Sung-Feng Huang, Chyi-Jiunn Lin, Da-Rong Liu, Yi-Chen Chen, Hung-yi Lee, Meta-TTS: meta-learning for few-shot speaker adaptive text-to-speech, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) (2022) 1558–1571, https://doi.org/10.1109/TASLP.2022.3167258.

[113] Wen-Chin Huang, Hsin-Te Hwang, Yu-Huai Peng, Tsao Yu, Hsin-Min Wang, Voice Conversion Based on Cross-Domain Features Using Variational Auto Encoders, in: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP) (Nov 2018), 2018, https://doi.org/10.1109/iscslp.2018.8706604.

[114] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran Raja, Raghavendra Ramachandra, Naser Damer, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira, Jaime S. Cardoso, João Tremoço, Miguel Lourenço, Sergio Serra, Eduardo Cermeño, Marija Ivanovska, Borut Batagelj, Andrej Kronovšek, Peter Peer, Vitomir Štruc, SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-Aware Synthetic Training Data, 2022, https://doi.org/10.48550/ARXIV.2208.07337.

[115] Arezu Rezgar Hussein, Rasber Dhahir Rashid, KurdFace morph dataset creation using OpenCV, Sci. J. Univ. Zakho 10 (4) (2022) 258–267, https://doi.org/10.25271/sjuoz.2022.10.4.943. Dec. 2022.

[116] Hsin-Te Hwang, Tsao Yu, Hsin-min Wang, Yih-Ru Wang, Sin-Horng Chen, Incorporating Global Variance in the Training Phase of GMM-Based Voice Conversion, in: 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA, 2013, pp. 1–6, https://doi.org/10.1109/APSIPA.2013.6694179.

[117] Laurie Iacono, Josh Hickman, Caitlin Muniz, The Rise of Vishing and Smishing Attacks – the Monitor, issue 21, https://www.kroll.com/en/insights/publications/cyber/monitor/vishing-smishing-attacks, 2022.

[118] Marija Ivanovska, Andrej Kronovšek, Peter Peer, Vitomir Štruc, Borut Batagelj, Face Morphing Attack Detection Using Privacy-Aware Training Data, 2022, https://doi.org/10.48550/ARXIV.2207.00899.

[119] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc, 2018, in: https://proceedings.neurips.cc/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf.

[120] Jun Jiang, Bo Wang, Bing Li, Weiming Hu, Practical face swapping detection based on identity spatial constraints, in: 2021 IEEE Int. Joint Conf. Biom. (IJCB), 2021, pp. 1–8, https://doi.org/10.1109/IJCB52358.2021.9484396.

[121] Liming Jiang, Li Ren, Wayne Wu, Chen Qian, Chen Change Loy, DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection, 2020 arXiv:2001.03024 [cs.CV].

[122] Brihi Joshi, Aditya Chetan, Pulkit Madaan, Pranav Jain, Srija Anand, Eshita, Shruti Singh, An Exploration into Deep Learning Methods for Emotional Text-To-Speech, 2020, https://doi.org/10.5281/zenodo.3876081.

[123] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo, StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks, in: 2018 IEEE Spoken Lang. Technol. Workshop (SLT), 2018, pp. 266–273, https://doi.org/10.1109/SLT.2018.8639535.

[124] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo, ACVAE-VC, Non-Parallel Voice Conversion With Auxiliary Classifier Variational Autoencoder 27 (9) (2019) 1432–1443, https://doi.org/10.1109/TASLP.2019.2917232, sep. 2019.

[125] Takuhiro Kaneko, Hirokazu Kameoka, Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks, 2017 arXiv:1711.11293 [stat.ML].

[126] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo, CycleGAN-VC2: improved CycleGAN-based non-parallel voice conversion, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[127] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo, StarGAN-VC2: rethinking conditional methods for StarGAN-based voice conversion, Proc. Interspeech (2019) 679–683, https://doi.org/10.21437/Interspeech.2019-2236.

[128] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo, CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020.

[129] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Nobukatsu Hojo, MaskCycleGAN-VC: learning non-parallel voice conversion with filling in frames, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2021.

[130] Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan, CRIM's system description for the ASVSpoof2021 challenge, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 100–106, https://doi.org/10.21437/ASVSPOOF.2021-16.

[131] Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan, Investigation on activation functions for robust end-to-end spoofing attack detection system, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 83–88, https://doi.org/10.21437/ASVSPOOF.2021-13.

[132] Srinivasan Kannan, Pooja R. Raju, R. Sai Surya Madhav, Shikha Tripathi, Voice conversion using spectral mapping and TD-PSOLA, in: Advances in Computing and Network Communications, Springer Singapore, Singapore, 2021, pp. 193–205.

[133] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018. https://openreview.net/forum?id=Hk99zCeAb.

[134] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, Timo Aila, Alias-free generative adversarial networks, in: Proc. NeurIPS, 2021.

[135] Tero Karras, Samuli Laine, Timo Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[136] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila, Analyzing and improving the image quality of StyleGAN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[137] H. Kawahara, H. Matsui, Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation, in: 2003 IEEE Int. Conf. Acoust. Speech Signal Process., Proceedings. (ICASSP '03), 1, 2003, https://doi.org/10.1109/ICASSP.2003.1198766.

[138] Hasam Khalid, Shahroz Tariq, Minha Kim, Simon S. Woo, FakeAVCeleb: a novel audio-video multimodal deepfake dataset, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. https://openreview.net/forum?id=TAXFsg6ZaOl.

[139] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, Faruk Kazi, A deep learning framework for audio deepfake detection, Arabian J. Sci. Eng. (2021) 1–12.

[140] Kietzmann Jan, Linda W. Lee, Ian P. McCarthy, Tim C. Kietzmann, Deepfakes: trick or treat? Bus. Horiz. 63 (2) (2020) 135–146, https://doi.org/10.1016/j.bushor.2019.11.006. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING).

[141] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, Youngjung Uh, Exploiting spatial dimensions of latent in GAN for real-time image editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 852–861.

[142] Jaehyeon Kim, Jungil Kong, Juhee Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to- speech, in: Marina Meila, Tong Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, PMLR, vol. 139, 2021, pp. 5530–5540, in: https://proceedings.mlr.press/v139/kim21f.html.

[143] Jiseob Kim, Jihoon Lee, Byoung-Tak Zhang, Smooth-swap: a simple enhancement for face-swapping with smoothness, in: In Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recogn.IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10779–10788.

[144] Kang-Wook Kim, Seung-Won Park, Junhyeok Lee, Myun-Chul Joe, ASSEM-VC: realistic voice conversion by assembling modern speech synthesis techniques, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2022, https://doi.org/10.1109/ICASSP43922.2022.9746139, 6997–7001.

[145] Tomi Kinnunen, Lauri Juvela, Paavo Alku, Junichi Yamagishi, Non-parallel voice conversion using i-vector PLDA: towards unifying speaker verification and transformation, in: 2017 IEEE Int. Conf. Acoust. Speech Signal Process., 2017, pp. 5535–5539, https://doi.org/10.1109/ICASSP.2017.7953215.

[146] Kazuhiro Kobayashi, Wen-Chin Huang, Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Tomoki Toda, Crank: an open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2021, https://doi.org/10.1109/ICASSP3972.8.2021.9413959, 5934–5938.

[147] Kazuhiro Kobayashi, Tomoki Toda, EasyChair. Sprocket: Open-Source Voice Conversion Software, 2018, https://doi.org/10.29007/s4t1. EasyChair Preprint no. 64.

[148] Xianwei Kong, Shengwu Xiong, Self-supervised flow field decoupling for Controllable face reenactment, J. Phys. Conf. 2253 (1) (2022), 012034, https://doi.org/10.1088/1742-6596/2253/1/012034 apr 2022.

[149] Pavel Korshunov, Sebastien Marcel, DeepFakes: a New Threat to Face Recognition? Assessment and Detection, 2018 arXiv:1812.08685 [cs.CV].

[150] Iryna Korshunova, Wenzhe Shi, Joni Dambre, Lucas Theis, Fast Face-Swap Using Convolutional Neural Networks, 2017 arXiv:1611.09577 [cs.CV].

[151] Marek Kowalski, Jacek Naruniec, Tomasz Trzcinski, Deep alignment network: a convolutional neural network for robust face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.

[152] S. Robin, Kramer S, Michael O. Mireku, Tessa R. Flack, Kay L. Ritchie, Face Morphing Attacks: Investigating Detection with Humans and Computers, 2019, https://doi.org/10.1186/s41235-019-0181-4.

[153] Prabhat Kumar, Mayank Vatsa, Richa Singh, Detecting Face2Face Facial Reenactment in Videos, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2578–2586, https://doi.org/10.1109/WACV45572.2020.9093628.

[154] Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David Han, Hanseok Ko, Injecting 3D perception of controllable NeRF-GAN into StyleGAN for editable portrait image synthesis, in: European Conference on Computer Vision, Springer, 2022, pp. 236–253.

[155] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, Gyeongsu Chae, KoDF: a large-scale Korean DeepFake detection dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV), 2021, pp. 10744–10753.

[156] Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux, Giulia Boato, More real than real: a study on human visual perception of synthetic faces [applications corner], IEEE Signal Process. Mag. 39 (1) (jan 2022) 109–116, https://doi.org/10.1109/msp.2021.3120982.

[157] Chung-Han Lee, Chung-Hsien Wu, MAP-Based Adaptation for Speech Conversion Using Adaptation Data Selection and Non-parallel Training, vol. 5, 2006.

[158] Dami Lee, Deepfake Salvador Dalí Takes Selfies with Museum Visitors, 2019. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.

[159] Ki-Seung Lee, Restricted Boltzmann machine-based voice conversion for nonparallel corpus, IEEE Signal Process. Lett. 24 (8) (2017) 1103–1107, https://doi.org/10.1109/LSP.2017.2713412.

[160] Yi Lei, Shan Yang, Jian Cong, Lei Xie, Dan Su, Glow-WaveGAN 2: high-quality zero-shot text-to-speech synthesis and any- to-any voice conversion, Proc. Interspeech (2022) 2563–2567, https://doi.org/10.21437/Interspeech.2022-684.

[161] Ang Li, Jian Hu, Chilin Fu, Xiaolu Zhang, Jun Zhou, Attribute-conditioned face swapping network for low-resolution images, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 2305–2309, https://doi.org/10.1109/ICASSP43922.2022.9747816.

[162] Changlin Li, Zhangjin Huang, An improved face synthesis model for two-pathway generative adversarial network, in: Proceedings of the 2019 11th International Conference on Machine Learning and Computing (Zhuhai, China) (ICMLC '19), Association for Computing Machinery, New York, NY, USA, 2019, https://doi.org/10.1145/3318299.3318346.

[163] Jingyi Li, Weiping Tu, Li Xiao, FreeVC: towards High-Quality Text-free One-Shot Voice Conversion, 2022, https://doi.org/10.48550/ARXIV.2210.15418.

[164] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, Faceshifter: towards High Fidelity and Occlusion Aware Face Swapping, 2019, 13457 arXiv preprint arXiv:1912.

[165] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, Baining Guo, Face X-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[166] Li Qi, Weining Wang, Chengzhong Xu, Zhenan Sun, Learning Disentangled Representation for One-Shot Progressive Face Swapping, 2022, https://doi.org/10.48550/ARXIV.2203.12985.

[167] Yuezun Li, Ming-Ching Chang, Siwei Lyu, in: Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7, https://doi.org/10.1109/WIFS.2018.8630787.

[168] Yuezun Li, Pu Sun, Honggang Qi, Siwei Lyu, Toward the Creation and Obstruction of DeepFakes, Springer International Publishing, Cham, 2022, pp. 71–96, https://doi.org/10.1007/978-3-030-87664-7_4.

[169] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, Siwei Lyu, Celeb-DF: a large-scale challenging dataset for DeepFake Forensics, in: IEEE Conference on Computer Vision and Patten Recognition (CVPR), 2020.

[170] Jiachen Lian, Chunlei Zhang, Gopala Krishna Anumanchipalli, Yu Dong, Towards improved zero-shot voice conversion with conditional DSVAE, in: Proc. Interspeech, 2022, pp. 2598–2602, https://doi.org/10.21437/Interspeech.2022-11225.

[171] Jiachen Lian, Chunlei Zhang, Yu Dong, Robust disentangled variational speech representation learning for zero-shot voice conversion, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6572–6576, https://doi.org/10.1109/ICASSP43922.2022.9747272.

[172] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang, Expressive talking head generation with granular audio-visual control, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3387–3396.

[173] Yist Y. Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, Lin-shan Lee, Fragmentvc: any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5939–5943, https://doi.org/10.1109/ICASSP39728.2021.9413699.

[174] Feng Liu, Minchul Kim, Anil Jain, Xiaoming Liu, Controllable guided face synthesis for unconstrained face recognition, in: In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII (Tel Aviv, Israel), Springer-Verlag, Berlin, Heidelberg, 2022, pp. 701–719, https://doi.org/10.1007/978-3-031-19775-8_41.

[175] Kai Liu, Bicheng Li, Jiale Li, Deep face-swap model combining attention mechanism and CycleGAN, J. Phys. Conf. 2278 (1) (2022), 012037, https://doi.org/10.1088/1742-6596/2278/1/012037 may 2022.

[176] Kun Liu, Jianping Zhang, Yonghong Yan, High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin, in: Fourth Int. Conf. Fuzzy Syst. Knowl. Discov. (FSKD 2007), 4, 2007, pp. 410–414, https://doi.org/10.1109/FSKD.2007.347.

[177] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, S.Y. Kung, 3D-FM GAN: towards 3D-controllable face manipulation, in: Computer Vision – ECCV 2022, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Springer Nature Switzerland, Cham, 2022, pp. 107–125.

[178] Zhengzhe Liu, Xiaojuan Qi, H. Philip, S. Torr, Global texture enhancement for fake face detection in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[179] Min Long, Xuan Zhao, Le-Bing Zhang, Fei Peng, Detection of face morphing attacks based on patch-level features and lightweight networks, Secur. Commun. Network. 2022 (11 Mar 2022) (2022), 7460330, https://doi.org/10.1155/2022/7460330.

[180] Yuanxun Lu, Jinxiang Chai, Xun Cao, Live speech portraits: real-time photorealistic talking-head animation, ACM Trans. Graph. 40 (6) (2021) 17, https://doi.org/10.1145/3478513.3480484. Article 220 (dec 2021).

[181] Sneha Lukose, Savitha S. Upadhya, Text to speech synthesizer-formant synthesis, in: 2017 Int. Conf. Nascent Technol. Eng. (ICNTE), 2017, pp. 1–4, https://doi.org/10.1109/ICNTE.2017.7947945.

[182] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, Le Xu, Ruibo Fu, FAD: A Chinese Dataset for Fake Audio Detection, 2022, https://doi.org/10.48550/ARXIV.2207.12308.

[183] Anderson F. Machado, Marcelo Queiroz, Voice Conversion: A Critical Survey, 2010, https://doi.org/10.5281/ZENODO.849853, 2010.

[184] Andrey Makrushin, Neubert Tom, Dittmann Jana, Automatic generation and detection of visually faultless facial morphs, in: InProceedings of the 12th Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory Appl., 6, 2017, pp. 39–50, https://doi.org/10.5220/0006131100390050. VISAPP, (VISIGRAPP 2017). INSTICC, SciTePress.

[185] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, Neyaz Khan Ahmad, DeepFake detection for human face images and videos: a survey, IEEE Access 10 (2022) (2022) 18757–18775, https://doi.org/10.1109/ACCESS.2022.3151186.

[186] Sophia Martin, Everything You Need to Know about Face Changing App and How Much Will it Cost You?, 2022.

[187] M. Martín-Doñas Juan, Aitor Álvarez, The vicomtech audio deepfake detection system based on Wav2vec2 for the 2022 ADD challenge, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 9241–9245, https://doi.org/10.1109/ICASSP43922.2022.9747768.

[188] McCloskey Scott, Michael Albright, Detecting GAN-generated imagery using saturation cues, in: 2019 IEEE Int. Conf. Image Process. (ICIP), 2019, pp. 4584–4588, https://doi.org/10.1109/ICIP.2019.8803661.

[189] Yisroel Mirsky, Wenke Lee, The creation and detection of deepfakes: a survey, ACM Comput. Surv. 54 (1) (Jan. 2021) 41, https://doi.org/10.1145/3425780. Article 7.

[190] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha, Emotions don't lie: an audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM International Conference on Multimedia (Seattle, WA, USA) (MM '20), Association for Computing Machinery, New York, NY, USA, 2020, pp. 2823–2832, https://doi.org/10.1145/3394171.3413570.

[191] Seyed Hamidreza Mohammadi, Kain Alexander, Voice conversion using deep neural networks with speaker-independent pre- training, in: IEEE Spoken Lang. Technol. Workshop (SLT), 2014, pp. 19–23, https://doi.org/10.1109/SLT.2014.7078543.

[192] Seyed Hamidreza Mohammadi, Kain Alexander, An overview of voice conversion systems, Speech Commun. 88 (2017) 65–82, https://doi.org/10.1016/j.specom.2017.01.008.

[193] Jesus Monge Alvarez, Holly Francois, Hosang Sung, Seungdo Choi, Jonghoon Jeong, Kihyun Choo, Kyoungbo Min, Sangjun Park, CAMNet: a controllable acoustic model for efficient, expressive, high-quality text-to-speech, Appl. Acoust. 186 (2022), 108439, https://doi.org/10.1016/j.apacoust.2021.108439.

[194] Lucio Moser, Jason Selfe, Darren Hendler, Doug Roble, Dynamic Neural Face Morphing for Visual Effects, in: SIGGRAPH Asia 2021 Technical Communications (Tokyo, Japan) (SA '21 Technical Communications), Association for Computing Machinery, New York,NY, USA, 2021, pp. 2–4, https://doi.org/10.1145/3478512.3488596.

[195] Nicolas M. Müller, Pizzi Karla, Jennifer Williams, Human Perception of Audio Deepfakes (DDAM '22), Association for Computing Machinery, New York, NY, USA, 2022, pp. 85–91, https://doi.org/10.1145/3552466.3556531.

[196] Aakash Varma Nadimpalli, Ajita Rattani, GBDF: Gender Balanced DeepFake Dataset towards Fair DeepFake Detection, 2022, https://doi.org/10.48550/ARXIV.2207.10246.

[197] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B.S. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, Amit K. Roy-Chowdhury, Detecting GAN generated fake images using Co-occurrence matrices, Electron. Imag. 31 (5) (Jan. 2019), https://doi.org/10.2352/issn.2470-1173.2019.5.mwsf-532, 532–1–532–7.

[198] National Academies of Sciences, Engineering, and Medicine, in: Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop, The National Academies Press, Washington, DC, 2019, https://doi.org/10.17226/25488.

[199] Pedro C. Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F. Sequeira, Jaime S. Cardoso, OrthoMAD: morphing attack detection through orthogonal identity disentanglement, in: International Conference of the Biometrics Special Interest Group (BIOSIG), 2022, pp. 1–5, https://doi.org/10.1109/BIOSIG55365.2022.9897057, 2022.

[200] João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez, in: GANprintR: Improved Fakes and Evaluation of the State-Of-The-Art in Face Manipulation Detection, 2019 arXiv:arXiv:1911.05351.

[201] Bac Nguyen, Fabien Cardinaux, NVC-net: end-to-end adversarial voice conversion, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7012–7016, https://doi.org/10.1109/ICASSP43922.2022.9747020.

[202] Dat Viet Thanh Nguyen, Phong Tran The, Tan M. Dinh, Cuong Pham, Anh Tuan Tran, QC-StyleGAN - quality controllable image generation and manipulation, in: Alice H. Oh, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho (Eds.), Advances in Neural Information Processing Systems, 2022. https://openreview.net/forum?id=AWeZdGJ89lC.

[203] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, Cuong M. Nguyen, Deep learning for deepfakes creation and detection: a survey, Comput. Vis. Image Understand. 223 (2022), 103525, https://doi.org/10.1016/j.cviu.2022.103525.

[204] Robert Nichols, Christian Rathgeb, Pawel Drozdowski, Christoph Busch, Psychophysical evaluation of human performance in detecting digital face image manipulations, IEEE Access 10 (2022) 31359–31376, https://doi.org/10.1109/ACCESS.2022.3160.596.

[205] Yuval Nirkin, Yosi Keller, Tal Hassner, FSGAN: subject agnostic face swapping and reenactment, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7184–7193.

[206] Yuval Nirkin, Yosi Keller, Tal Hassner, improved subject agnostic face swapping and reenactment, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2023) 560–575, https://doi.org/10.1109/TPAMI.2022.3155571. FSGANv2.

[207] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, Gerard Medioni, On face segmentation, face swapping, and face perception, in: IEEE International Conference on Automatic Face Gesture Recognition, 2018, pp. 98–105, https://doi.org/10.1109/FG.2018.00024.

[208] Lindsey O'Donnell, CEO 'Deep Fake' Swindles Company Out of $243K. https://threatpost.com/deep-fake-of-ceos-voi ce-swindles-company-out-of-243k/147982/, 2019.

[209] Christina Orphanidou, I.M. Moroz, Stephen Roberts, Wavelet-based voice morphing, WSEAS J. Syst. 10 (2004) 3297–3302, 01 2004.

[210] Geon-Woo Park, Eun-Ju Park, Simon S. Woo, Zoom-DF: a dataset for video conferencing deepfake, in: Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes (Nagasaki, Japan) (WDC '22), Association for Computing Machinery, NewYork, NY, USA, 2022, pp. 7–11, https://doi.org/10.1145/3494109.3527195.

[211] Maitreya Patel, Mirali Purohit, Mihir Parmar, Nirmesh J. Shah, Hemant A. Patil, Ada{GAN}: Adaptive {GAN} for Many-to-Many Non-parallel Voice Conversion, 2020. https://openreview.net/forum?id=HJlk-eHFwH.

[212] Bo Peng, Hongxing Fan, Wei Wang, Jing Dong, Yuezun Li, Siwei Lyu, Li Qi, Zhenan Sun, Chen Han, Baoying Chen, Yanjie Hu, Shenghai Luo, Junrui Huang, Yutong Yao, Boyuan Liu, Hefei Ling, Guosheng Zhang, Zhiliang Xu, Changtao Miao, Changlei Lu, He Shan, Xiaoyan Wu, Wanyi Zhuang, DFGC 2021: A DeepFake Game Competition, 2021 arXiv:2106.01217 [cs.CV].

[213] Fei Peng, Le Qin, Min Long, Face morphing attack detection and attacker identification based on a watchlist, Signal Process. Image Commun. 107 (2022), 116748, https://doi.org/10.1016/j.image.2022.116748.

[214] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl shift facenheim, in: Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, Weiming Zhang (Eds.), DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework, 2021 arXiv:2005.05535 [cs.CV].

[215] Hartmut R. Pfitzinger, Unsupervised speech morphing between utterances of any speakers, in: Proceedings of the 10th Australian Int. Conf. Speech Sci. Technol., 2004, pp. 545–550.

[216] Justin N.M. Pinkney, Chuan Li, clip2latent: Text Driven Sampling of a Pre-trained StyleGAN Using Denoising Diffusion and CLIP, 2022, https://doi.org/10.48550/ARXIV.2210.02347.

[217] Victor Popa, Silen Hanna, Jani Nurminen, Moncef Gabbouj, Local linear transformation for voice conversion, in: 2012 IEEE Int. Conf. Acoust. Speech Signal Process., 2012, pp. 4517–4520, https://doi.org/10.1109/ICASSP.2012.6288922.

[218] Ethan Preu, Mark Jackson, Nazim Choudhury, Perception vs. Reality: understanding and evaluating the impact of synthetic image deepfakes over college students, in: 2022 IEEE 13th IEEE 13th Annu. Ubiquitous Comput. Electron. Mobile Commun. Conf. (UEMCON), 2022, pp. 547–553, https://doi.org/10.1109/UEMCON54665.2022.9965697.

[219] Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, Mark Hasegawa-Johnson, Unsupervised speech decomposition via triple information bottleneck, in: Proceedings of the 37th International Conference on Machine Learning (ICML'20), 2020, p. 11. JMLR.org, Article 726.

[220] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson, AutoVC: zero-shot voice style transfer with only autoencoder loss, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th Int. Conf. Mach. Learn. (Proc.Mach. Learn. Res., 97, 2019, pp. 5210–5219. PMLR.

[221] Le Qin, Fei Peng, Min Long, Face morphing attack detection and localization based on feature-wise supervision, IEEE Trans. Inf. Forensics Secur. 17 (2022) (2022) 3649–3662, https://doi.org/10.1109/TIFS.2022.3212276.

[222] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning transferable visual models from natural language supervision, in: Marina Meila, Tong Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763. PMLR, https://proceedings.mlr.press/v139/radford21a.html.

[223] R. Raghavendra, KiranB. Raja, Sushma Venkatesh, Christoph Busch, Face morphing versus face averaging: vulnerability and detection, in: 2017 IEEE Int. Joint Conf. Biom. (IJCB), 2017, pp. 555–563, https://doi.org/10.1109/BTAS.2017.82.72742.

[224] R. Raghavendra, Kiran B. Raja, C. Busch, Detecting morphed face images, in: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016, https://doi.org/10.1109/BTAS.2016.7791169.

[225] R. Raghavendra, Sushma Venkatesh, Kiran Raja, Christoph Busch, Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features, 2018.

[226] Kiran Raja, Matteo Ferrara, Annalisa Franco, Luuk Spreeuwers, Ilias Batskos, Florens de Wit, Marta Gomez-Barrero, Scherhag Ulrich, Daniel Fischer, Sushma Krupa Venkatesh, Jag Mohan Singh, Guoqiang Li, Loïc Bergeron, Sergey Isadskiy, Raghavendra Ramachandra, Christian Rathgeb, Dinusha Frings, Uwe Seidel, Fons Knopjes, Raymond Veldhuis, Davide Maltoni, Christoph Busch, Morphing attack detection – database, evaluation platform, and benchmarking, IEEE Trans. Inf. Forensics Secur. 16 (2021) 4336–4351, https://doi.org/10.1109/TIFS.2020.3035252.

[227] Kiran Raja, Matteo Ferrara, Annalisa Franco, Luuk Spreeuwers, Ilias Batskos, Florens de Wit, Marta Gomez-Barrero, Scherhag Ulrich, Daniel Fischer, Sushma Krupa Venkatesh, Jag Mohan Singh, Guoqiang Li, Loïc Bergeron, Sergey Isadskiy, Raghavendra Ramachandra, Christian Rathgeb, Dinusha Frings, Uwe Seidel, Fons Knopjes, Raymond Veldhuis, Davide Maltoni, Christoph Busch, Morphing attack detection-database, evaluation platform, and benchmarking, IEEE Trans. Inf. Forensics Secur. 16 (2021) 4336–4351, https://doi.org/10.1109/TIFS.2020.3035252.

[228] Kiran Raja, Gourav Gupta, Sushma Venkatesh, Raghavendra Ramachandra, Christoph Busch, Towards generalized morphing attack detection by learning residuals, Image Vis Comput. 126 (2022), 104535, https://doi.org/10.1016/j.imavis.202.2.104535.

[229] Raghavendra Ramachandra, Guoqiang Li, Residual colour scale-space gradients for reference-based face morphing attack detection, in: 2022 25th International Conference on Information Fusion (FUSION), 2022, pp. 1–8, https://doi.org/10.23919/FUSION49751.2022.9841318.

[230] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, H. Andrew, Sung, Deepfake detection: a systematic literature review, IEEE Access 10 (2022) 25494–25513, https://doi.org/10.1109/ACCESS.2022.3154404.

[231] Ricardo Reimao, Vassilios Tzerpos, FoR: a dataset for synthetic speech detection, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, https://doi.org/10.1109/SPED.2019.8906599.

[232] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhao Zhou, Tie-Yan Liu, FastSpeech 2: Fast and High-Quality End-To-End Text to Speech, 2021 arXiv:2006.04558 [eess.AS].

[233] Tim Ring, Europol: the AI hacker threat to biometrics, Biom. Technol. Today (2) (2021) 9–11, https://doi.org/10.1016/S0969-4765(21)00023-0.

[234] Matej Rojc, Izidor Mlakar, An LSTM-based model for the compression of acoustic inventories for corpus-based text-to-speech synthesis systems, Comput. Electr. Eng. 100 (2022) (2022), 107942, https://doi.org/10.1016/j.compeleceng.2022.1.07942.

[235] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M.s Nießner, FaceForensics++: learning to detect manipulated facial images, in: International Conference on Computer Vision (ICCV), 2019.

[236] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, FaceForensics: A Large- Scale Video Dataset for Forgery Detection in Human Faces, 2018 arXiv:1803.09179 [cs.CV].

[237] Tim Sainburg, Marvin Thielk, Brad Theilman, Benjamin Migliori, Timothy Gentner, Generative Adversarial Interpolative Autoencoding: Adversarial Training on Latent Space Interpolations Encourage Convex Latent Distributions, 2019 arXiv:1807.06650 [cs.LG].

[238] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C. Stamm, Stefano Tubaro, TIMIT-TTS: a Text-To-Speech Dataset for Synthetic Speech Detection, 2022, https://doi.org/10.5281/zenodo.6560159.

[239] Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Gabryś Adam, Jaime Lorenzo-Trueba, Cross-speaker style transfer for text-to-speech using data augmentation, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6797–6801, https://doi.org/10.1109/ICASSP43922.2022.9746179.

[240] Sarkar Eklavya, Korshunov Pavel, Colbois Laurent, Marcel Sébastien, Are GAN-based morphs threatening face recognition?, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 2959–2963, https://doi.org/10.1109/ICASSP43922.2022.9746477. Accepted for Publication in ICASSP2022.

[241] Scherhag Ulrich, Christian Rathgeb, Christoph Busch, Morph deterction from single face image: a multi-algorithm fusion approach, in: Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications (Amsterdam, Netherlands) (ICBEA '18), Association for Computing Machinery, New York, NY, USA, 2018, pp. 6–12, https://doi.org/10.1145/3230820.3230822.

[242] Clemens Seibold, Wojciech Samek, Hilsmann Anna, Peter Eisert, Detection of face morphing attacks by deep learning, in: InDigital Forensics and Watermarking, Springer International Publishing, Cham, 2017, pp. 107–120.

[243] Clemens Seibold, Wojciech Samek, Hilsmann Anna, Peter Eisert, Accurate and robust neural networks for face morphing attack detection, J. Inf. Secur. Appl. 53 (2020) (2020), 102526, https://doi.org/10.1016/j.jisa.2020.102526.

[244] John Seymour, Azeem Aqil, Your Voice Is My Passport, 2018.

[245] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael Reiter, Adversarial Generative Nets: Neural Network Attacks on State-Of-The-Art Face Recognition, 2017, 12 2017.

[246] Yujun Shen, Jinjin Gu, Xiaoou Tang, Bolei Zhou, Interpreting the Latent Space of GANs for Semantic Face Editing, CVPR, 2020.

[247] Yujun Shen, Ceyuan Yang, Xiaoou Tang, Bolei Zhou, InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs, TPAMI, 2020.

[248] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang, Few-shot head swapping in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10789–10798.

[249] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe, First order motion model for image animation, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc, 2019, in: https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf.

[250] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, Sergey Tulyakov, Motion Representations for Articulated Animation, CVPR, 2021.

[251] Jag Mohan Singh, Raghavendra Ramachandra, Fusion of deep features for differential face morphing attack detection at automatic border control gates, in: In 2022 10th European Workshop Vis. Inf. Process. (EUVIP), 2022, pp. 1–5, https://doi.org/10.1109/EUVIP53989.2022.9922773.

[252] Berrak Sisman, Junichi Yamagishi, Simon King, Haizhou Li, An overview of voice conversion and its challenges: from statistical modeling to deep learning, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 132–157, https://doi.org/10.1109/TASLP.2020.3038524.

[253] Peng Song, Y.Q. Bao, L. Zhao, C.R. Zou, Voice conversion using support vector regression, Electron. Lett. 47 (09) (2011) 1045–1046, https://doi.org/10.1049/el.2011.1851.

[254] Luuk Spreeuwers, Maikel Schils, Raymond Veldhuis, Una Kelly, Practical evaluation of face morphing attack detection methods, in: Handbook of Digital Face Manipulation and Detection, Springer, Cham, 2022, pp. 351–365.

[255] Yannis Stylianou, Voice transformation: a survey, in: 2009 IEEE Int. Conf. Acoust. Speech Signal Process., 2009, pp. 3585–3588, https://doi.org/10.1109/ICASSP.2009.4960401.

[256] Youcef Tabet, Boughazi Mohamed, Speech synthesis techniques. A survey, in: Int. Workshop Syst. Signal Process. Appl, WOSSPA, 2011, pp. 67–70, https://doi.org/10.1109/WOSSPA.2011.5931414.

[257] Hemlata Tak, Jee weon Jung, Patino Jose, Madhu Kamble, Massimiliano Todisco, Nicholas Evans, End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 1–8, https://doi.org/10.21437/ASVSPOOF.2021-1.

[258] Hemlata Tak, Jee weon Jung, Patino Jose, Massimiliano Todisco, Nicholas Evans, Graph attention networks for anti-spoofing, Proc. Interspeech (2021) 2356–2360, https://doi.org/10.21437/Interspeech.2021-993.

[259] Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura, Modulation spectrum-based post-filter for GMM-based voice conversion, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, 2014, pp. 1–4, https://doi.org/10.1109/APSIPA.2014.7041540.

[260] Anni Tang, Xue Han, Jun Ling, Rong Xie, Li Sang, Dense 3D coordinate code prior guidance for high-fidelity face swapping and face reenactment, in: 2021 16th IEEE 16th IEEE Int. Conf. Automat. Face and Gesture Recogn. (FG 2021), 2021, pp. 1–8, https://doi.org/10.1109/FG52635.2021.9667065.

[261] Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao, Avqvc: one-shot voice conversion by vector quantization with applying contrastive learning, in: ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2022, pp. 4613–4617, https://doi.org/10.1109/ICASSP43922.2022.9746369.

[262] Paul Taylor, Text-to-Speech Synthesis, Cambridge University Press, 2009, https://doi.org/10.1017/CBO9780511816338.

[263] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, Matthias Nießner, Neural Voice Puppetry: Audio-Driven Facial Reenactment, 2020 arXiv:1912.05566 [cs.CV].

[264] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, Matthias Niessner, Face2Face: real-time face capture and reenactment of RGB videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ((CVPR)), 2016.

[265] Xiaohai Tian, Wa Lee Siu, Zhizheng Wu, Eng Siong Chng, Haizhou Li, An exemplar-based approach to frequency warping for voice conversion, IEEE/ACM Trans. Audio Speech Lang. Process. 25 (10) (2017) 1863–1876, https://doi.org/10.1109/TASLP.2017.2723721.

[266] Ruben Tolosana, Christian Rathgeb, Ruben Vera-Rodriguez, Christoph Busch, Luisa Verdoliva, Siwei Lyu, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, Peter Rot, Klemen Grm, Vitomir Štruc, Antitza Dantcheva, Zahid Akhtar, Sergio Romero-Tapiador, Julian Fierrez, Aythami Morales, Javier Ortega-Garcia, Els Kindt, Catherine Jasserand, Tarmo Kalvet, Marek Tiits, Future Trends in Digital Face Manipulation and Detection, Springer International Publishing, Cham, 2022, pp. 463–482, https://doi.org/10.1007/978-3-030-87664-7_21.

[267] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, Javier Ortega-Garcia, Deepfakes and beyond: a Survey of face manipulation and fake detection, Inf. Fusion 64 (2020) (2020) 131–148, https://doi.org/10.1016/j.inffus.2020.06.014.

[268] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Kondratev Alexander, Galina Lavrentyeva, STC antispoofing systems for the ASVspoof2021 challenge, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 61–67, https://doi.org/10.21437/ASVSPOOF.2021-10.

[269] Soumya Tripathy, Juho Kannala, Esa Rahtu, Single source one shot reenactment using weighted motion from paired feature points, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2715–2724.

[270] Rafael Valle, Kevin Shih, Prenger Ryan, Catanzaro Bryan, Flowtron: an Autoregressive Flow-Based Generative Network for Text-To-Speech Synthesis, 2020 arXiv:2005.05957 [cs.SD].

[271] Mariëtte van Huijstee, Pieter van Boheemen, Djurre Das, Linda Nierling, Jutta Jahnel, Murat Karaboga, Fatun Martin, Linda Kool, Joost Gerritsen, Tackling Deepfakes in European Policy, 2021, https://doi.org/10.2861/325063.

[272] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, Herman Kamper, A comparison of discrete and soft speech units for improved voice conversion, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 6562–6566, https://doi.org/10.1109/ICASSP43922.2022.9746484. ICASSP.

[273] Sushma Venkatesh, Kiran Raja, Raghavendra Ramachandra, Christoph Busch, On the influence of ageing on face morph attacks: vulnerability and detection, in: 2020 IEEE Int. Joint Conf. Biom. (IJCB), 2020, pp. 1–10, https://doi.org/10.1109/IJ.CB48548.2020.9304856.

[274] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Christoph Busch, Face morphing attack generation and detection: a comprehensive survey, IEEE Trans. Technol. Soc. 2 (3) (2021) 128–145, https://doi.org/10.1109/TTS.2021.3066254.

[275] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, Christoph Busch, Can GAN generated morphs threaten face recognition systems equally as landmark based morphs? - vulnerability and detection, in: 2020 8th 8th Int. Workshop Biom. Forensic (IWBF), 2020, pp. 1–6, https://doi.org/10.1109/IWBF49977.2020.9107970.

[276] Sushma Venktatesh, Multilevel fusion of deep features for face morphing attack detection, in: International Conference on Electrical, Computer, Communications and Mechatronics Engineering, 2022, pp. 1–7, https://doi.org/10.1109/ICECCME55909.2022.9987842. ICECCME.

[277] Luisa Verdoliva, Media Forensics and DeepFakes: an overview, IEEE J. Sel. Top. Signal Process. 14 (5) (2020) 910–932, https://doi.org/10.1109/JSTSP.2020.3002101.

[278] Konstantinos Vougioukas, Stavros Petridis, Maja Pantic, End-to-End Speech-Driven Facial Animation with Temporal GANs, 2018 arXiv:1805.09313 [eess.AS].

[279] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, He Lei, Sheng Zhao, Furu Wei, Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers, 2023, https://doi.org/10.48550/ARXIV.2301.02111.

[280] Jinwei Wang, Kehui Zeng, Bin Ma, Xiangyang Luo, Qilin Yin, Guangjie Liu, Sunil Kr, Jha, GAN-generated fake face detection via two-stream CNN with PRNU in the wild, Multimed. Tool. Appl. 81 (29) (2022) 42527–42545, https://doi.org/10.1007/s11042-021-11592-7. (Accessed 1 December 2022).

[281] Qiqi Wang, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao, DRVC: a framework of any-to-any voice conversion with self-supervised learning, in: ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2022, pp. 3184–3188, https://doi.org/10.1109/ICASSP43922.2022.9747434.

[282] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, Yang Liu, DeepSonar: towards effective and robust detection of AI-synthesized fake voices, in: Proceedings of the 28th ACM International Conference on Multimedia (Seattle, WA, USA) (MM '20), Association for Computing Machinery, New York, NY, USA, 2020, pp. 1207–1216, https://doi.org/10.1145/3394171.3413716.

[283] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, Yang Liu, DeepSonar: towards effective and robust detection of AI-synthesized fake voices, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1207–1216, https://doi.org/10.1145/3394171.3413716 (Seattle, WA, USA)(MM '20).

[284] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, Jamie Shotton, Fake it till you make it: face analysis in the wild using synthetic data alone, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 3681–3691.

[285] Chung-Hsien Wu, Chi-Chun Hsia, Te-Hsien Liu, Jhing-Fa Wang, Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1109–1116, https://doi.org/10.1109/TASL.2006.876112.

[286] Haotian Wu, Peipei Wang, Xin Wang, Ji Xiang, Rui Gong, GGViT:Multistream vision transformer network in Face2Face facial reenactment detection, in: 2022 26th International Conference on 26th Int. Conf. Pattern Recogn. (ICPR), 2022, pp. 2335–2341.

[287] Yihan Wu, Xu Tan, Bohan Li, He Lei, Sheng Zhao, Ruihua Song, Tao Qin, Tie-Yan Liu, AdaSpeech 4: adaptive text to speech in zero-shot scenarios, in: Proc. Interspeech, 2022, pp. 2568–2572, https://doi.org/10.21437/Interspeech.2022-901.

[288] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, A Study of Speaker Adaptation for DNN-Based Speech Synthesis, 2015, https://doi.org/10.21437/Interspeech.2015-270.

[289] Ruitong Xiao, Haitong Zhang, Lin Yue, DGC-vector: a new speaker embedding for zero-shot voice conversion, in: ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process., 2022, pp. 6547–6551.

[290] Taihong Xiao, Jiapeng Hong, Jinwen Ma, ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[291] Tianyi Xie, Liucheng Liao, Bi Cheng, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, Zejun Ma, Towards Realistic Visual Dubbing with Heterogeneous Sources, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1739–1747, https://doi.org/10.1145/3474085.3475318.

[292] Guodong Xu, Yuenan Hou, Ziwei Liu, Chen Change Loy, Mind the gap in distilling StyleGANs, in: Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, Tal Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 423–439.

[293] Pengxiang Xu, Mei Xue, Wei Yi, Tiancheng Qian, Robust facial manipulation detection via domain generalization, in: In 2021 7th International Conference on Computing and Artificial Intelligence (Tianjin, China) (ICCAI 2021), Association for Computing Machinery, New York, NY, USA, 2021, pp. 196–201, https://doi.org/10.1145/3467707.3467736.

[294] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Pan Jia, Shengfeng He, High-resolution face swapping via latent semantics disentanglement, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR), 2022, pp. 7642–7651.

[295] Yangyang Xu, Xuemiao Xu, Jianbo Jiao, Keke Li, Cheng Xu, Shengfeng He, Multi-View Face Synthesis via Progressive Face Flow, in: IEEE Transactions on Image Processing, vol. 30, 2021, https://doi.org/10.1109/TIP.2021.3090658.

[296] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, Errui Ding, MobileFaceSwap: A Lightweight Framework for Video Face Swapping, 2022, https://doi.org/10.48550/ARXIV.2201.03808.

[297] Xue Han, Jun Ling, Anni Tang, Li Song, Rong Xie, Wenjun Zhang, High-fidelity face reenactment via identity-matched correspondence learning, in: ACM Trans. Multimed Comput. Commun. Appl (nov 2022), 2022, https://doi.org/10.1145/3571857.

[298] Jinlong Xue, Yayue Deng, Yichen Han, Ya Li, Jianqing Sun, Jiaen Liang, ECAPA-TDNN for Multi-Speaker Text-To-Speech Synthesis, 2022, https://doi.org/10.48550/ARXIV.2203.10473.

[299] Jun Xue, Cunhang Fan, Lv Zhao, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, Shegang Shao, Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (Lisboa, Portugal) (DDAM '22), Association for Computing Machinery, New York, NY, USA, 2022, pp. 19–26, https://doi.org/10.1145/3552466.3556526.

[300] Ziyu Xue, Xiuhua Jiang, Qingtong Liu, Zhaoshan Wei, Global & local facial fusion based GAN generated fake face detection, Sensors 23 (2023) 2, https://doi.org/10.3390/s23020616.

[301] Nand Kumar Yadav, Satish Kumar Singh, Shiv Ram Dubey, CSA-GAN: cyclic synthesized attention guided generative adversarial network for face synthesis, Appl. Intell. 52 (11) (2022) 12704–12723, https://doi.org/10.1007/s10489-021-03064-0, 1 September 2022.

[302] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicolas Evans, Tomi Kinnunen, Aik Lee Kong, Ville Vestman, Andreas Nautsch, ASVspoof 2019, in: 3rd Automat. Speak. Verif. Spoofing Countermeas. Chall. Database, 2019, https://doi.org/10.7488/ds/2555.

[303] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Patino Jose, Andreas Nautsch, Xuechen Liu, Aik Lee Kong, Tomi Kinnunen, Nicholas Evans, Héctor Delgado, ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: Proc. 2021 Ed. Automat. Speak. Verif. Spoofing Countermeas. Chall., 2021, pp. 47–54.

[304] Rui Yan, Wen Cheng, Shuran Zhou, Tingwei Guo, Wei Zou, Xiangang Li, Audio deepfake detection system with neural stitching for ADD, in: ICASSP 2022 - 2022 ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2022, pp. 9226–9230, https://doi.org/10.1109/ICASSP43922.2022.9746820.

[305] Jiachen Yang, Guipeng Lan, Shuai Xiao, Li Yang, Jiabao Wen, Yong Zhu, Enriching facial anti-spoofing datasets via an effective face swapping framework, Sensors 22 (2022) 13, https://doi.org/10.3390/s22134697.

[306] Hui Ye, S. Young, Quality-enhanced voice morphing using maximum likelihood transformations, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006), https://doi.org/10.1109/TSA.2005.860839.

[307] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Bai Ye, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, Haizhou Li, Add 2022: the first audio deep synthesis detection challenge, in: ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2022, pp. 9216–9220, https://doi.org/10.1109/ICASSP43922.2022.9746939.

[308] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, Tomoki Toda, Voice Conversion Challenge 2020 — Intra-lingual Semi-parallel and Cross-Lingual Voice Conversion, 2020. https://www.isca-speech.org/archive/VCC_BC_2020/pdfs/VCC2020_paper_13.pdf.

[309] Takato Yoshikawa, Yuki Endo, Yoshihiro Kanamori, Diversifying detail and appearance in sketch-based face image synthesis, Vis. Comput. 38 (9) (2022) 3121–3133, https://doi.org/10.1007/s00371-022-02538-7, 1 September 2022.

[310] Haiming Yu, Hao Zhu, Xiangju Lu, Junhui Liu, Migrating face swap to mobile devices: a lightweight framework and a supervised training solution, in: 2022 IEEE Int. Conf. Multimed. Expo. (ICME), 2022, pp. 1–6.

[311] Yu Ning, Larry S. Davis, Mario Fritz, Attributing fake images to GANs: learning and analyzing GAN fingerprints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019 (ICCV).

[312] Zhenjun Yue, Xiang Zou, Yongxing Jia, Hao Wang, Voice conversion using HMM combined with GMM, in: 2008 Congr. Image Signal Process., 5, 2008, pp. 366–370, https://doi.org/10.1109/CISP.2008.165.

[313] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, Victor Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[314] Haoyu Zhang, Marcel Grimmer, Raghavendra Ramachandra, Kiran Raja, Christoph Busch, On the applicability of synthetic data for face recognition, in: In 2021 IEEE International Workshop on IEEE Int. Workshop Biom. Forensic (IWBF), 2021, pp. 1–6, https://doi.org/10.1109/IW.BF50991.2021.9465085.

[315] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Naser Damer, Christoph Busch, MIPGAN—generating strong and high quality morphing attacks using identity prior driven GAN, IEEE Transactions on Biometrics, IEEE Trans. Biom. Behav. Ident. Sci. 3 (3) (2021) 365–383, https://doi.org/10.1109/TBIOM.2021.3072349.

[316] Jiangning Zhang, Xianfang Zeng, Chao Xu, Yong Liu, Real-time audio-guided multi-face reenactment, IEEE Signal Process. Lett. 29 (2022) (2022) 1–5, https://doi.org/10.1109/LSP.2021.3116506.

[317] Le-Bing Zhang, Juan Cai, Fei Peng, Min Long, Yuanquan Shi, Noise robust face morphing detection method, in: Pushpendu Kar, Steven Guan (Eds.), International Conference on Internet of Things and Machine Learning (IoTML 2021), vol. 12174, International Society for Optics and Photonics, SPIE, 2022, 1217417, https://doi.org/10.1117/12.2628711.

[318] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang, SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation, 2022, https://doi.org/10.48550/ARXIV.2211.12194.

[319] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, Xianfeng Zhao, SynSpeechDDB: a New Synthetic Speech Detection Database, 2020, https://doi.org/10.21227/ta8z-mx73.

[320] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, Xianfeng Zhao, FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection, in: Digital Forensics and Watermarking, Xianfeng Zhao, Alessandro Piva, and Pedro Comesaña-Alfaro, Springer International Publishing, Cham, 2022, pp. 117–131.

[321] Botao Zhao, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao, nnSpeech: speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech, in: ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2022, pp. 4293–4297, https://doi.org/10.1109/ICASSP43922.2022.9746875.

[322] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, Nenghai Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[323] Peng Zhou, Xintong Han, Vlad I. Morariu, Larry S. Davis, Two-stream neural networks for tampered face detection, in: 2017 IEEE Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW), 2017, pp. 1831–1839, https://doi.org/10.1109/CVPRW.2017.229.

[324] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, Jianbing Shen, Face Forensics in the wild, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR), 2021, pp. 5778–5788.

[325] Xiao Zhou, Zhen-Hua Ling, Simon King, The Blizzard Challenge 2020, online, http://www.festvox.org/blizzard/bc2 020/BC20_zhou_ling_king.pdf, 2020.

[326] Yipin Zhou, Ser-Nam Lim, Joint audio-visual deepfake detection, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 14800–14809.

[327] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, Chen Change Loy, CelebV-HQ: A Large-Scale Video Facial Attributes Dataset, in: ECCV, 2022.

[328] Yuhao Zhu, Li Qi, Jian Wang, Chengzhong Xu, Zhenan Sun, One shot face swapping on megapixels, in: Proceedings of the IEEE Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR), 2021, pp. 4834–4844.

[329] Yiming Zhu, Hongyu Liu, Yibing Song, Ziyang Yuan, Xintong Han, Chun Yuan, Qifeng Chen, Jue Wang, One model to edit them all: free-form text-driven image manipulation with semantic modulations, in: Alice H. Oh, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho (Eds.), Advances in Neural Information Processing Systems, 2022. https://openreview.net/forum?id=kb33f8J83c.

[330] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, Yu-Gang Jiang, WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection, 2021 arXiv: 2101.01456 [cs.CV].

[331] Tudor-Cătălin Zorilă, Daniel Erro, Inma Hernaez, Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations, in: Advances in Speech and Language Technologies for Iberian Languages, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 30–39.

[332] Adrian Łańcucki, Fastpitch: parallel text-to-speech with pitch prediction, in: ICASSP 2021 - 2021 IEEE IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. (ICASSP), 2021, pp. 6588–6592, https://doi.org/10.1109/ICASSP39728.2021.9413889.

The dawn of a text-dependent society: deepfakes as a threat to speech verification systems

Anton Firc
Brno University of Technology
Brno, Czech Republic
ifirc@fit.vutbr.cz

Kamil Malinka
Brno University of Technology
Brno, Czech Republic
malinka@fit.vutbr.cz

ABSTRACT

Recent developments in the field of deepfakes bring new threats that take advantage of the fact that it is increasingly difficult to distinguish between real and artificial media. Nowadays, mostly as fake news or disinformation; however, there are still unexplored areas such as using deepfakes to spoof voice verification. We present a real-world use case for spoofing voice authentication in a customer care call center. Based on this scenario, we evaluate the feasibility of attacking such a system and create an attacker profile. For this purpose, we examine three available speech synthesis tools and discuss their usability. We use these tools and acquired knowledge to generate a dataset including deepfake speech and assess the resilience of voice biometrics systems against deepfakes. We prove that voice biometrics systems are indeed vulnerable to deepfake powered attacks. The most significant outcome is the proposal of text-dependent verification as a novel countermeasure for presented attacks. Text-dependent verification provides higher security than text-independent verification and can be used today as the simplest protection method against deepfakes.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; **Biometrics**;

KEYWORDS

deepfakes, speech verification, voice biometrics, machine learning, cybersecurity

ACM Reference Format:
Anton Firc and Kamil Malinka. 2022. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event,* . ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3477314.3507013

1 INTRODUCTION

The term *deepfake* has no agreed-upon technical definition. It is a combination of words *deep learning* and *fake*. Deepfakes are a subset of synthetic media created using deep neural networks that depict events that never happened to entertain, defame individuals, spread fake news, and many others [2].

The recent advancements in machine learning make the creation of deepfakes easier than ever. Even people without any technical background are able to use commercial tools with intuitive UI. These techniques and tools are being used for both illicit and legitimate purposes. One of the unexplored areas of illicit usage is using deepfakes to spoof speaker recognition systems. Public opinion is mixed, and very little scientific work exists in this area. This is why we decided to examine this area in deeper detail.

A survey performed by a biometric security startup ID R&D in December of 2019 concluded that the worries about deepfake voice fraud are slowing down the adaptation of voice biometrics systems [27]. Around two-thirds of Americans are afraid of their voice being spoofed and used to access their accounts secured by voice biometrics systems [8].

According to the article published by the Guardian, the companies behind the voice biometrics technology say there are more than 100 unique physical and behavioral characteristics of each individual. These characteristics are stated to include the length of the vocal tract, nasal passage, pitch, accent, and many more. These characteristics were claimed to be as unique to an individual as a fingerprint is. Nuance communication claimed that even professional voice imitators could not fool their system, while some of the other companies claimed that their voice verification system is even able to spot a difference between identical twins [13].

These statements were already refuted by a BBC reporter when his non-identical twin managed to get access to his account [13, 30]. As stated before, the bank claimed to use more than one hundred different characteristics of voice to verify identity. The second alarming fact is that the bank allowed the reporter and his twin to get seven attempts of verification wrong before the eight successful attempt [30]. Even though this incident has happened in the first half of 2017, and the speech recognition technology has advanced rapidly since then, this incident proves that an attack on a voice biometrics system using only a person with similar voice characteristics was feasible, thus under the right circumstances, there is still a chance to spoof such a system. This incident also sheds a different light on the reality of the statements claiming that these systems cannot be easily fooled.

The creation of synthetic speech using the voice of another person still becomes easier, even without any IT education or knowledge. Thus the only needed thing is a recording of the target person's voice, which can be obtained nowadays simply by downloading a video from a social network, YouTube, or a podcast. These facts led some people to believe that voice verification should serve

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507013

as an additional method to verify a person's identity, not the main one [13].

From the security viewpoint, voice biometrics systems should include techniques to detect replayed or synthetically generated speech. They are also believed to possess these techniques, as stated in the article published by TechRadar [21]; however, this article also mentions that this ability might be compromised if the deepfake technology continues to improve. ID R&D even shows their liveness detection in action as a promo for their voice biometrics system [9]. One of the most recent articles discussing this topic published in May of 2020 [11] supports the idea of voice biometrics systems being fooled by sufficiently advanced deepfakes. It even proposes that this ability might lead to an artificial intelligence arms race, with institutions upgrading their authentication systems and criminals improving the quality of the deepfakes to overcome the implemented measures.

Up to date, no experiments regarding the deepfake resilience of voice biometrics systems in real-world environment exist. We decided to analyze the current situation and assess the advancements of deepfake technology and the difficulty of executing a deepfake powered attack on a scenario of customer verification in a customer care call center. In this scenario, the communication is made exclusively using the telephone, and the voice biometrics system has to be spoofed as well as the human operator.

Firstly, we evaluate the technical feasibility of using speech synthesis tools to create deepfakes. We use the acquired knowledge to better understand the attacker profile in terms of needed knowledge and resources to succeed in the proposed attack scenario.

Secondly, we examine the resilience of two voice biometrics systems to deepfakes and show that deepfakes do present a serious threat to these systems. For this purpose, we create our own deepfake dataset in the Czech and English languages.

Finally, we create a new dataset for experiments to evaluate differences between text-dependent and text-independent verification. Using this dataset, we show that text-dependent verification is more resilient to deepfakes.

The main contributions of this work may be summarized as follows:

- This work proves that deepfakes present a severe threat to speech verification and that the voice biometrics systems might be easily spoofed, so measures to prevent these kinds of attacks have to be implemented.
- We conclude that no expert domain knowledge in speech synthesis or processing is required to execute a deepfake powered attack on a system secured by voice biometrics.
- We present that text-dependent verification is more robust than text-independent verification when dealing with deepfakes by creating a custom dataset to compare the security provided by both verification types.
- A dataset containing English and Czech deepfake speech was created and published for further use.

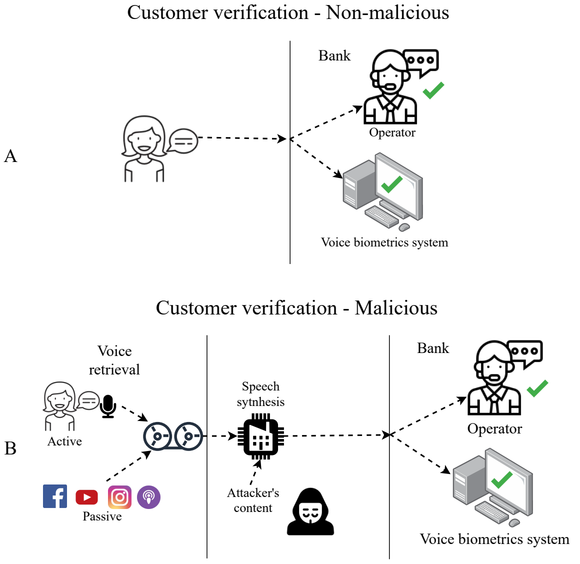


Figure 1: Attack schema. Figure A represents non-malicious (genuine) access to customer care call center, Figure B represents malicious access with target voice retrieval phase and speech synthesis.

The proposed attack scenario and experiments are discussed in Section 2. Section 3 reviews related work, and Section 4 defines the voice biometrics systems and their performance measures as we use them in this work. The experiments carried out for this research are divided into three parts. The first part discussed in Section 5 examines the essentials of deepfake creation. The second part discussed in Section 6 examines the performance of deepfakes in text-independent verification. The third part described in Section 7 examines the difference between text-dependent and text-independent verification. Finally, Section 8 discusses conclusions and further work.

2 PROPOSED ATTACK SCHEMA

While deepfake creation seems to be more accessible than ever, we already know that voice biometrics systems are not foolproof, as mentioned in the previous section. We aim to examine how difficult it is to spoof voice biometrics systems using deepfakes and what measures might be taken to mitigate the threats posed by deepfakes to voice biometrics systems.

A proof-of-concept on spoofing voice verification was presented at the 2018 Black Hat conference by J. Seymour and A. Aqil [28]. They used a text-to-speech system to spoof Apple Siri and Microsoft Speaker recognition API. Now, three years after the publishing of mentioned research, the topic remains more than actual.

Currently, there is only one publicly available incident report involving synthetic speech. Fraudsters impersonated a CEO of an energetic company and manipulated an employee to initiate a fraudulent wire transfer worth 250k USD to their accounts [14].

We decided to focus on the area of customer verification in companies providing customer care call centers (see Figure 1). A non-malicious scenario involves the customer making a telephone call to the call center. While the customer talks to the operator about her's or his request, the voice biometrics system verifies the customer's identity. The operator finally executes the requested action of the customer.

In a deepfake scenario of an attack on a customer care call center, the attacker synthesizes utterances with a speech of his victim in advance. The attacker then makes a phone call to the customer care service and begins to replay the prepared utterances to verify his identity and initiate an unauthorized action of his choice. The success of this scenario relies on the deepfake ability to spoof voice biometrics systems as well as humans. We have already carried out experiments to evaluate the human ability to distinguish between genuine and synthetic speech. The results suggest that the human factor does not play any significant role in a scenario of this kind. These results will be published as a part of our future work. Nevertheless, for the scope of this work, the human factor will be omitted.

2.1 System model

The target system might be a speech recognition system as described in Section 4, or a more complex system incorporating a speech recognition system. Bank or telephone operator call centers use systems of the second category. The voice biometrics system verifies the customer's identity while the operator talks to the client. In this case, not only the voice biometrics system is present as a security measure, but also the human operator might spot the synthetic speech and end the call before any incident happens. Our idea of system model connected with attacker and victim is shown in Figure 1.

2.2 Attacker model

An attacker is a person with the ability to create voice deepfakes, and his goal is to gain access into a system secured by voice authentication, such as a bank call center. The attacker is in possession of all needed personal information about his victim, all needed details about the typical scenario of the voice authentication process, and finally samples of voice belonging to the victim.

The attacker will use all of this information to synthesize utterances reproducing the victim‘s speech, and then in the most believable way possible, try to access the system secured by voice biometrics system and use the granted access to his advantage.

The ability to create voice deepfakes can be understood in two main ways. The attacker is either able to collect and prepare enough data to train his own speech synthesis tool or to get unauthorized access into one of the commercial systems and misuse the stored speech synthesis templates of registered users to synthesize speech. The second type of attacker would be less powerful and probable, as only a tiny portion of people use such commercial systems.

2.3 Victim model

A victim is any person that uses her's or his voice to authenticate into any system. We also expect the victim to be a regular computer user or to possess a telephone. The voice samples of the victim can be retrieved from any of the content posted online, such as videos on social networks, YouTube, or even podcasts if available. Even if the victim does not post on such platforms, the voice shall be obtainable by recording a phone call.

2.4 Research questions

The experiments executed during this research might be divided into three separate parts that follow up each other.

The first part aims to explore the technical feasibility of cloning a voice of a selected individual. Two commercial and one open-source text-to-speech synthesis tools will be tested to evaluate the needed knowledge, data, and quality of the results. This part will answer the following research questions:

- How difficult is it to create a synthetic copy (clone) of an individual's voice?
- How much data is needed to clone an individual's voice in usable quality?

The second part then uses the synthesized speech to verify the resilience of voice biometrics systems to deepfakes. We compare the quality of speech synthesized by each one of the tools with all of the available voice biometrics systems using text-independent and text-dependent verification if available. A dataset containing deepfake and genuine speech for 100 English and 60 Czech speakers will be created. Using the dataset, we will collect information on the performance of deepfakes in text-independent verification. This part will answer the following research questions:

- Are today's voice biometrics systems capable of detecting synthetic speech?
- How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?

Based on observations, we noticed an interesting feature of text-dependent verification we decided to examine more closely. The third and final part thus focuses on the examination of the difference between text-dependent and text-independent verification. As the text-dependent type examines not only the general voice characteristics (independent of the content) but also the way how a specific phrase is spoken, it seems to deliver more security when facing a synthesized speech [15, 16]. To further examine this hypothesis, we will create our own dataset containing phrases for text-dependent verification, synthesize selected phrases, and compare matching scores retrieved from both verification types. If we confirm this hypothesis, we obtain a reliable, easy-to-implement defense mechanism as a result of this part. This part will answer the following research questions:

- Is text-dependent verification harder to spoof using deepfakes than text-independent verification?

3 RELATED WORK

In the previous section, we described the whole attack schema, including the research questions that we focus on. Our work is not the first one to be carried out in the area of voice biometrics systems and synthetic speech. However, none of the currently published researches examines this topic in such depth as we do. The main focus in this area is given towards improvements of speech synthesis, or deepfakes in general, and towards methods for deepfake detection. Then, there remains a gray zone of usability of proposed detection methods and usage in real-world environments.

This section presents the related work in all of the mentioned areas and puts this work into the context of these areas.

Existing datasets. Datasets relevant to the scope of our research might be divided into two categories: *deepfake datasets* that contain genuine and deepfake utterances of each speaker and *datasets for speech processing tasks* that contain only genuine utterances of each speaker. The deepfake datasets are mostly used to train and develop systems for deepfake detection. The datasets for speech processing are used for various tasks ranging from training speaker authentication systems to speech synthesis systems. Unfortunately, up to date, no standardized datasets to measure the performance of voice biometrics systems in terms of robustness to deepfakes exists.

Deepfake datasets might be retrieved from the challenges aimed to develop deepfake detection methods [34, 35] or might be created using data provided for challenges aimed to measure advances of deepfake creation as Voice Conversion challenge [36] or Blizzard challenge [40]. Recently, new datasets designed for deepfake detection tasks were proposed [7, 25, 38, 39].

Speech processing datasets are more easily accessible and also more popular. The most popular ones might be stated as: Common Voice Corpus 6.1 [1] consisting of 60 languages and more than 7k hours of speech, LJ Speech [10] consisting of 24 hours of speech, Vox Celeb [17] consisting of 2k hours of speech, Librispeech [20], LibriTTS [37] providing speech from an extensive library of audiobooks or VoxLingua107 [31] consisting of speech in 107 languages.

Deepfake detection. Up to date, there is a considerable amount of research in deepfake detection methods. The most significant results come from the ASVspoof challenge [34, 35] which is a biannual event focusing on bringing the best deepfake detection methods possible. There are also other important works related to our topic, C. Borrelli et al. [3] proposes a method using short and long-term prediction traces, T. Nguyen et al. [18] provides a survey on the recent usage of deep learning for deepfake detection, R. Wang et al. [33] proposes a framework that monitors the behavior of neurons in a voice biometrics system based on neural networks.

Speech synthesis. A lot of research is being published in the field of speech synthesis. The most notable publications for the scope of this work are a framework for text-to-speech synthesis [12] by Y. Jia and its implementation [4] by J. Corentin that we use in this work. J. Shen et al. in [29] presented an architecture for text-to-speech synthesis that is nowadays very widely used. A. Oord et al. in [32] presented an architecture for a raw audio generation that is also very often used in speech synthesis tools.

Usability. This research area stands somewhere in the middle of the mentioned areas. We use knowledge from all of these areas, while the usage of results does not fall into any of these categories. Currently, the only published research on usability of deepfakes to spoof voice biometrics systems is from J. Seymour and A. Aqil [28]. The authors use TTS synthesis, with models trained for each speaker separately. Unfortunately, no greater detail on the used methodology and data is provided. In contrast, our research examines the aspects of deepfake creation in greater detail, uses tools utilizing transfer-learning, provides deepfake datasets for the reproduction of results, and, most importantly, focuses on the feasibility of such attacks in real-world scenarios.

4 VOICE BIOMETRICS SYSTEMS

We have presented a scenario of using deepfakes to attack customer call center secured by voice biometrics system and experiments that evaluate the feasibility of the presented scenario. This section defines the voice biometrics system as we understand it for the scope of this work and values used to measure the performance of such systems that we use later in experiments to evaluate how credibly do deepfake imitates genuine speech.

The voice biometrics, or speaker recognition, the system provides a biometric-based security process known as speaker authentication [15]. Freely translated, it means that the system authenticates its users based on *what they are* by processing their unique voice characteristics. There are two technologies: speaker verification and speaker identification [15]. Verification is a process of determining whether a person is who she or he claims to be, while identification is a process of determining an identity of a person from a pool of known identities [15, 16]. Speaker verification further divides to text-dependent and text-independent [16]. Text-dependent verification needs the same phrase to be spoken during the enrolment, and in the verification phase, the text-independent verification, on the other hand, has no restriction on the spoken content [16].

Shortly said, the voice biometrics system is a black box that authenticates a person based on their own voice.

4.1 Performance measures of voice biometrics systems

The performance of a biometrics system might be measured and compared using numerous measures. The verification performance evaluation is done by performing many genuine and impostor attempts while the matching scores are saved. A genuine attempt is performed by a user to match his own profile, and an impostor attempt is performed by a user to match someone else's profile [23].

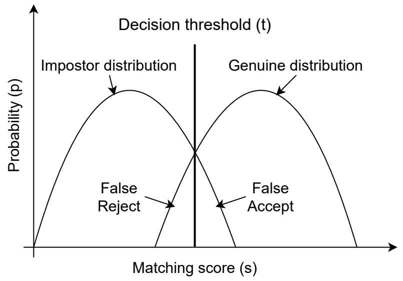


Figure 2: User matching scores distribution by attempt type.

Saved matching scores can be used to plot user matching scores distribution (see Figure 2) and to calculate error rates. The error rates relevant for this paper are false non-match rate (FNMR) and false match rate (FMR). These rates are calculated by applying a varying matching score threshold to the matching score [23]. As shown in Figure 3, FMR and FNMR are then used to calculate the equal error rate (EER), which is a point where the FMR and FNMR equal [23].

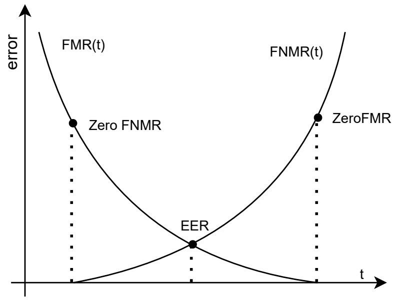


Figure 3: FMR, FNMR and EER curves.

5 EXPERIMENT 1 - TECHNICAL FEASIBILITY OF DEEPFAKE CREATION

Further sections discuss in detail the design, execution, and results of experiments proposed in Section 2. The first experiment aims to evaluate the technical feasibility of deepfake creation and acquire knowledge on the potential attacker’s knowledge and determination. For every tool, a discussion on needed knowledge and final usability is provided. Finally, this section concludes the overall usability of the presented tools and the needed knowledge of an attacker.

5.1 Speech synthesis tools

During this research, three text-to-speech tools were tested. Two commercial tools: *Overdub* [6] and *ResembleAI* [26] and one open-source tool *Real Time Voice Cloning* [4]. The quality of synthesized speech was evaluated simply by listening to synthesized and genuine recordings. We followed the principles described in ITU-T Recommendation P.85 [19].

5.1.1 Descript Overdub. Descript is an audio/video editor providing AI features to edit and enhance recordings [5]. Overdub feature allows the synthesis of speech in users’ own voice from the typed text.

To provide training speech a pre-prepared transcript[1] has to be read. The minimal length of audio provided is 10 minutes. Thirty minutes of recorded audio should provide production-ready results [6].

The usage of this tool is straightforward, and even individuals without any extensive knowledge are able to synthesize speech using the intuitive GUI. The quality of synthesized speech is very high, and it really sounds like the target speaker.

5.1.2 Resemble AI. Resemble AI is an online platform for text-to-speech synthesis. They provide both API and a web interface for speech synthesis [26]. The training consists of reading prepared sentences. A minimum of 50 sentences is required.

Speech synthesis might be modified by adding effects to the written text. Provided effects include pause, emphasis, phoneme, spelling each character, emotion, or substitute.

This tool is simple to use and requires no more than a web browser. The quality of synthesized speech is adequate; however, much more training data than minimum must be provided. The synthesized speech sounds like the targeted individual, even though the pace is sometimes a bit off.

5.1.3 Real Time Voice Cloning. This open-source tool[2] is an implementation of the SV2TTS framework proposed by Y. Jia et al. in [12]. As Figure 4 shows, the tool is composed of three separate parts: *encoder*, *synthesizer* and *vocoder* [4, 12]:

- encoder conditions the synthesis network on a reference speech signal from the desired target speaker
- synthesizer predicts a mel-spectrogram of synthesized speech
- vocoder finally transforms the mel-spectrogram to speech

The tool features a toolbox that provides a simple GUI to synthesize speech using pretrained models[3]. The advanced operation must be done using a console application. The console application provides scripts for data preparation, training, and speech synthesis. When using the pretrained models, the usage simplifies to running provided python scripts and providing needed target audio and transcriptions.

The main feature of this tool is the ability to synthesize speech based on short embedding recording. This makes the tool very versatile, as the pretrained models are independent of the target speaker.

The quality of synthesized speech might be marked as average. The speech is mostly recognizable, and the voice is similar to the target speaker’s voice. Glitches and other imperfections occur mostly as long periods of silence. However, fine-tuning the synthesizer model as proposed in [24] dramatically improves the resulting quality. Thus, in further experiments, we use fine-tuning for synthesizing all speech.

5.2 Experiment conclusions

While the commercial tools provided the most quality synthetic speech, their usage is limited. The training of both tools required to read prepared scripts of moderate lengths, which dramatically increases the difficulty of the voice retrieval phase as shown in Figure 1; moreover, the basic versions do not allow the creation of more than one synthetic voice. Achieving quality of the commercial tools with the open-source one is a demanding task; however, this price is rewarded with absolute freedom regarding all aspects of speech synthesis.

There are several factors why we believe that the attacker would choose the presented open-source tool Real Time Voice Cloning (RTVC):

- Only a very short embedding recording needed for speech synthesis
- No pre-prepared training script needed
- Unlimited access to the tool and its source code
- No usage of third party services for malicious intents

This experiment has shown that synthesizing speech capable of spoofing voice biometrics system might be just a matter of few

[1] https://coda.io/@overdub/overdub-scripts

[2] https://github.com/CorentinJ/Real-Time-Voice-Cloning

[3] https://github.com/CorentinJ/Real-Time-Voice-Cloning/wiki/Pretrained-models

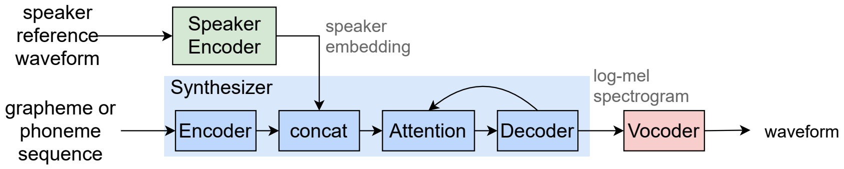


Figure 4: Overview of framework implemented within the RTVC tool [12]

clicks in intuitive GUI. The commercial tools performed the best; however, usage of these tools is limited.

All research questions were answered:

How difficult is it to create a synthetic copy (clone) of an individual's voice?

The difficulty is determined by the requirements on the resulting quality and chosen approach. Synthesizing speech using commercial tools is a very simple process; however, the usability is quite limited. Synthesizing speech using an open-source tool requires more extensive knowledge but provides broad possibilities of usage. In summary, this process is not trivial and requires some knowledge and time to get familiar with it. The demands on knowledge and effort depend on the needed quality and usability of results.

When considering an attacker with no prior experience with deepfake creation or speech synthesis tools, we estimate the time needed to gain all of the needed knowledge to be 2 to 3 weeks.

How much data is needed to clone an individual's voice in usable quality?

The amount of data depends on the chosen approach. The RTVC tool needs as little as a 5-second embedding recording to synthesize target speech. To achieve higher quality results, fine-tuning is needed, which needs at least 0.2 hours of transcribed speech. We estimate this amount of data to be easily obtainable for most of the attackers, making almost every person a suitable target. More data is needed if the attacker plans to create a new model from scratch, around 20 hours of transcribed speech.

6 EXPERIMENT 2 - TEXT-INDEPENDENT VERIFICATION AND DEEPFAKES

The first experiment has shown that speech synthesis does not require expert domain knowledge of speech synthesis or processing. The following experiment examines the usability of synthetic speech to spoof voice biometrics. Firstly, we show that deepfakes present a severe threat to voice biometrics systems. Secondly, we create our own deepfake dataset containing English and Czech speech and examine the aspects of spoofing voice biometrics on a bigger scale. We decided to add the Czech speech as we are curious about the performance of speech synthesis using uncommon models and languages.

6.1 Used voice biometrics systems

During our research, we were able to get hands-on two voice biometrics systems: *Microsoft Speaker Recognition API* [16] and *Phonexia Voice Verify demo* [22]. Unfortunately, no companies developing voice biometrics systems that we reached out to wanted to cooperate.

The Phonexia Voice Verify demo provides only a browser interface and verification through a telephony provider. Also, no numerical value on the result is provided, only the result itself. The Microsoft Speaker Recognition API features REST API and the verification result as a numerical value, making it more suitable for bulk testing and comparison of results. According to these facts, the majority of the experiments were executed using Microsoft Speaker Recognition API.

6.2 Experiment design

At first, we examine how the selected voice biometrics systems behave when authenticating genuine speech. Secondly, we compare speech synthesized using all tools described in Section 5. Finally, we create a deepfake dataset using the RTVC tool, and then we examine the differences between genuine and deepfake speech using Microsoft Speaker Recognition API. We use only the RTVC tool because we believe that a potential attacker would choose this tool and the Microsoft Speaker Recognition API because of the definite result of the verification process that is easy to compare. We examine the differences by collecting matching scores of the genuine, impostor, and deepfake attempts. Then we compare collected matching scores by plotting score distribution plots and FMR, FNMR curves. The matching score calculation depends on the type of the recording and follows this scheme:

Genuine matching scores will be calculated for each of the speakers by creating a voice profile using enrolment recording and then calculating the matching score for each of the speaker's recordings, even the enrolment one against the created voice profile.

Impostor matching scores will be calculated for each of the speakers by creating a voice profile using enrolment recording and then calculating matching scores of recordings of other speakers against the created voice profile.

Deepfake matching scores will be calculated for each speaker by creating a voice profile using enrolment recording and then calculating the matching score of each speaker‘s deepfake recording against the created voice profile.

6.3 Dataset

We created a new deepfake dataset because the current availability of English deepfake datasets is minimal, and no Czech deepfake dataset exists. As we stated before, there is currently only one dataset containing both genuine and deepfake speech, and the others have to be created by combining provided genuine speech and published deepfake speech. Moreover, this way, we can really explore all of the needed knowledge, time, and data needed to synthesize speech intended to be used maliciously and understand the

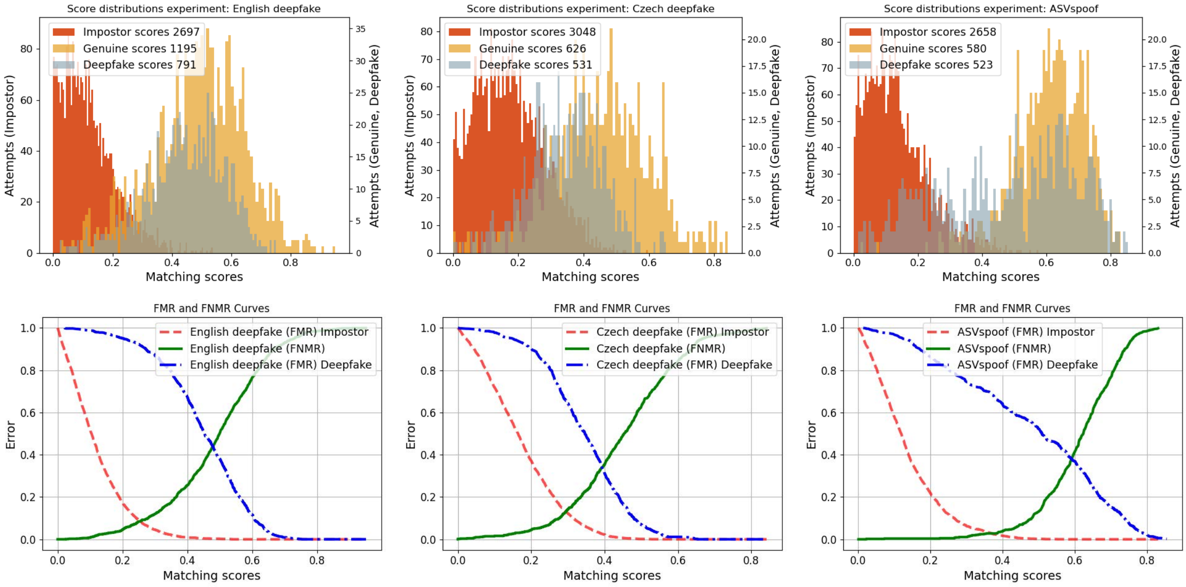


Figure 5: Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom). The left plots represent created English deepfake dataset. The middle plots represent created Czech deepfake dataset. The right plots represent the ASVSpoof 2019 challenge dataset [34].

attacker’s profile even more deeply. The dataset consists of genuine and deepfake speech of 100 English and 60 Czech speakers selected from the Common Voice Corpus [1]. The speakers were selected according to the count of the recordings available; only speakers with the highest count were selected. The longest recording was selected for each speaker as enrollment recording for MS Speaker Recognition API and a template recording for speech synthesis. The remaining recordings were used to fine-tune the synthesizer model. Transcription for synthesized English and Czech utterances can be found in Appendix A. All of the synthesized recordings were then collected into a dataset and published[4].

6.4 Experiment execution

We obtained initial score and behavior baselines by verification of genuine speech. For each verification type and voice biometrics system, we used one person’s speech and collected the results for dozens of attempts. Table 1 shows the average calculated matching scores of Microsoft Speaker Recognition API. The genuine matching scores are from interval $[0.70, 0.95]$, and there seems to be no difference in text-dependent and text-independent matching scores. Phonexia Voice Verify demo accepted all of the genuine attempts even when we tried to change our speech, like speaking slower or in a deeper voice.

Using all of the tested tools, we synthesized speech and verified it against the created profiles. For the Overdub tool, we provided 30 minutes of training audio. For the Resemble AI tool, we provided 150 training sentences. In the case of the RTVC tool, we used the provided pre-trained models. The only modification was fine-tuning the synthesizer model for the target speaker by doing additional 1k iterations using 0.2 hours of speech. As Table 2 shows, the commercial tools were very close to the genuine matching scores, while the RTVC tool fell a bit behind. We compare the best-achieved deepfake matching scores to average genuine matching scores to gain an insight into how an attack under favorable conditions would perform.

Table 1: Achieved average matching scores for each type of verification using Microsoft Speaker Recognition API.

Verification type	Matching score
text-dependent	0.83815
text-independent	0.82174

Table 2: Best achieved matching scores for each TTS tool for each type of verification using Microsoft Speaker Recognition API.

Verification type	Tool	Matching score
text-dependent	RTVC	0.59272
	Overdub	0.64144
	Resemble AI	0.55970
text-independent	RTVC	0.62365
	Overdub	0.79611
	Resemble AI	0.60146

[4] https://drive.google.com/drive/u/2/folders/1vlR-TA7gjKzjYylxzRnA_HzZEyWiLeOk

As the Phonexia Voice Verify demo provides verification only through a telephone, we decided to replay the synthesized speech from a notebook speaker to a phone placed nearby. As shown in Table 3, commercial tools were able to synthesize speech that passed the verification process. The RTVC tool performed worse, with verified and rejected parts of the recording resulting in a rejection.

Table 3: Results of malicious verification in Phonexia Voice Verify demo using the best recording synthesized with each of the used tools.

Tool	Verified
Overdub	yes
Resemble AI	yes
RTVC	no

After validating that the voice biometrics systems tend to accept deepfakes, we proceeded to test the created dataset. As Figure 5 shows, the deepfake dataset performed very well. The deepfake matching score distributions almost identically overlay the genuine matching score distributions. The FMR and FNMR plots show a significant increase in the EER value between genuine and deepfake plots. The overlap of genuine and deepfake matching scores indicates that there was no difference in how the voice biometrics system processed genuine and deepfake speech. The increase in EER value shows that presenting deepfake speech to the voice biometrics system increases the number of mistakes the system makes in terms of false accepts.

For comparison, we collected matching scores for the ASVspoof 2019 challenge dataset [34] as seen in Figure 5. In contrast to our dataset, the deepfake matching scores are distributed almost evenly through the whole matching score range. However, the increase in EER value is very similar to our dataset.

This comparison shows that even an inexperienced attacker is able to reach the quality of deepfakes created by people with extensive domain knowledge. Data obtained from all of the datasets implies that deepfakes indeed have the ability to credibly reproduce genuine speech and to spoof voice biometrics systems. The deepfakes of the highest quality are even able to perform better than the genuine recordings.

6.5 Experiment conclusions

The experiments have shown that deepfakes do have the ability to spoof voice biometrics systems. The created dataset achieved results comparable to other datasets containing deepfake speech used to train detection tools or showcase speech synthesis advancements.

All research questions were answered:

Are today's voice biometrics systems capable of detecting synthetic speech?

This experiment has shown that the tested voice biometrics systems were unable to detect synthetic speech. This indicates that the voice biometrics systems might not be able to detect deepfakes. To further confirm this answer, more robust testing with more voice biometrics systems must be executed.

How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?

As shown in matching score distributions, deepfakes are able to reproduce the genuine utterances very precisely. In the case of our dataset, the deepfake matching scores almost exactly reproduced the genuine ones. This puts deepfakes into a position of a dangerous means to spoof the voice biometrics systems.

7 EXPERIMENT 3 - TEXT-DEPENDENT VS. TEXT-INDEPENDENT VERIFICATION

During the previous experiments we noticed an interesting difference between matching scores calculated using text-dependent and text-independent verification. The text-dependent verification has constantly provided higher genuine matching scores and lower deepfake matching scores than the text-independent variant. We decided to explore this anomaly in deeper detail to find out whether it is just a random event or a feature that can be used as a countermeasure.

7.1 Experiment design

The Microsoft Speaker Recognition API provides text-dependent verification with a set of predefined phrases [16]. Unfortunately, no publicly available dataset contains these phrases, so we will create our own dataset for the scope of this experiment as described in Section 7.2. Using the recorded dataset, for each speaker, the synthesizer model of the RTVC tool will be fine-tuned for additional 1k steps, and then each used phrase will be synthesized ten times. Afterward, matching scores for both genuine and deepfake utterances will be calculated the following way:

A **text-dependent** profile will be created for each speaker, and the genuine and deepfake matching scores for each phrase will be calculated.

A **text-independent** profile will be created, using recordings from the training set. Afterward, for each speaker, the reference utterance and deepfake utterances will be used to calculate the matching scores.

After collecting all the matching scores, we will compare the differences between both of the verification types. We will calculate the average matching scores and the deviations of the scores.

7.2 Dataset

We decided to create a small dataset to create a proof-of-concept. The dataset consists of 5 speakers, four male, and one female. The dataset consists of phrases for text-dependent verification and other utterances to fine-tune the RTVC tool used for speech synthesis. The phrases used for text-dependent verification were selected to include phrases of all lengths:

- my name is unknown to you
- my voice is my passport verify me
- I am going to make him an offer he cannot refuse

Each phrase is recorded five times in total, where the first four recordings are used for text-dependent profile enrolment, and the fifth recording will be used as a reference to calculate the genuine matching score. The training set will consist of 75 sentences randomly selected from the transcripts of the Common Voice Corpus and the Harvard Sentences[5].

[5] https://www.cs.columbia.edu/~hgs/audio/harvard.html

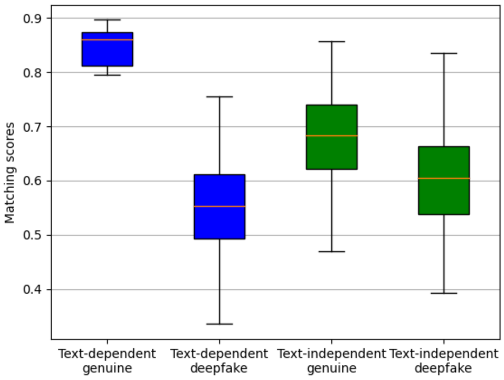


Figure 6: Comparison of scores calculated for text-dependent and text-independent verification using the same genuine and deepfake recordings for both verification types.

7.3 Experiment execution

After collecting all of the matching scores as proposed before in this section, we processed the results. As Figure 6 shows, there is a notable difference between genuine and deepfake matching scores of text-dependent verification, while this difference almost vanishes within text-independent verification matching scores. The drop in genuine matching scores between text-dependent and text-independent verification and the increase in deepfake matching scores between these two is evident. The text-independent genuine and deepfake matching scores are much more similar than the text-dependent ones.

7.4 Experiment conclusions

We created a new dataset to experiment with text-dependent verification provided by MS Speaker Recognition API. Even though the created dataset consists of speech for five speakers, the collected data was enough to show that the difference between text-dependent and text-independent verification when facing deepfakes is not a random event.

All research questions were answered:

Is text-dependent verification harder to spoof using deepfakes than text-independent verification?

As the results show, the deepfake matching scores differ vastly from the genuine ones. This difference almost vanished when using text-independent verification. This implies that it is much easier to reproduce the matching scores of text-independent verification, which puts the text-dependent verification into a position of the more secure one when facing deepfakes. To completely verify this hypothesis, more robust testing must be carried out.

The impact of this finding is crucial for improving the current security of voice biometrics systems when facing deepfakes. Text-dependent verification is a well-known method that is surely implemented in many systems. This way, the resilience of voice verification to deepfakes might be significantly improved without the need for any extensive changes or additions to the voice biometrics system. Even though this finding requires to be examined in deeper detail on more robust experiments, current results look promising.

8 CONCLUSIONS

We have shown that deepfakes do present a serious threat to voice biometrics systems. As shown, the voice biometrics systems that do not explicitly implement any kind of liveness detection can be easily spoofed. The creation of deepfakes with the capability to spoof such systems might be generalized to finding a proper GitHub repository, learning to work with it, collecting voice samples of the victim, transcribing the voice samples, and finally synthesizing speech. Even though most of the models and tools are suited for the English language, it is possible to synthesize speech in different languages in decent quality without any extensive knowledge on speech synthesis or processing.

We show that using text-dependent verification has the ability to mitigate the threats posed by deepfakes to voice biometrics systems. This approach might be used as the currently simplest method of protection against synthetic speech.

To set our results into the context of the proposed attack scenario, we can say with certainty that the scenario is feasible. As shown, spoofing voice biometrics systems is feasible, and as we stated before, adding a human factor into this scenario does not significantly increase security. Regarding these facts, an attack on a customer care call center secured by a voice biometrics system is definitely accomplishable.

To further extend the scope of this research, more robust experiments have to be executed to validate the results and used methods. We plan to test more voice biometrics systems on their resilience to deepfakes, to create a larger dataset for a more extensive comparison of text-dependent and text-independent verification in different languages, or to create a framework and datasets for testing the performance of deepfake detection techniques and finally to extend the research to cover more languages.

A DATASET TRANSCRIPTION

The following sentences were synthesized to create the English deepfake dataset discussed in Section 6:

(1) My voice is my passport verify me.
(2) October arrived, spreading a damp chill over the grounds and into the castle.
(3) Hello. Yes, I would like to inform myself on the topic of spoofed voice and the security implications.
(4) In some cases specific syllables and particular words are consistently represented by specific syllables.
(5) The black cat has reflexes, agility and stamina of an olympic level acrobat.
(6) The grape is still popular in North Africa, Algeria, Morocco and Tunisia.
(7) Eight minutes later she went to general quarters and enemy bodies were reported.
(8) Elisabeth attended university of Chicago Laboratory schools.
(9) Hale docking occurs in one of two ways.
(10) Debate may also end if no senator wishes to make any further remarks.

The following sentences were synthesized to create the Czech deepfake dataset discussed in Section 6:

(1) Tato pláž byla oceněna modrou vlajkou.
(2) Domníval jsem se, že to bude představovat tisícové náklady.
(3) Každoročně se schází vrcholný výkonný orgán tvořený hlavami států nebo předsedy vlád členských států.
(4) Tento typ strukturální podpory je příležitostí, jak dosáhnout našich cílů.
(5) Kvůli nedostatku přesnosti se moderní porodnictví termínu vyhýbá.
(6) Z mezinárodního hlediska se používá vždy mezera.
(7) Musíme i nadále podporovat naše zemědělce při modernizaci jejich podniků.
(8) Menšími úpravami prošla také karoserie.
(9) Kanada patří mezi nejdůležitější partnery, které Evropská unie má.
(10) Hráč si může udělat i své vlastní návštěvníky.

ACKNOWLEDGMENTS

This work was supported by the internal project of Brno University of Technology (FIT-S-20-6427). Computational resources were supplied by the project "e-Infrastr-uktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development, and Innovations Infrastructures.

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4218–4222. https://www.aclweb.org/anthology/2020.lrec-1.520

[2] Jon Bateman. 2020. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Technical Report. Carnegie Endowment for International Peace. i–ii pages. http://www.jstor.org/stable/resrep25783.1

[3] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. 2021. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security* 2021, 1 (2021), 2. https://doi.org/10.1186/s13635-021-00116-3

[4] Jemine Corentin. 2019. *Real-time Voice Cloning*. Master thesis. Université de Liège, Liège, Belgique. https://matheo.uliege.be/handle/2268.2/6801?locale=en

[5] Descript. 2020. Descript webpage. online. https://www.descript.com

[6] Descript. 2021. Overdub. online. https://www.descript.com/overdub

[7] Joel Frank and Lea Schönherr. 2021. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. arXiv:cs.LG/2111.02813

[8] Sudipto Ghosh. 2019. Are You Confident About Distinguishing Between a Computer-Generated Voice and Human Voice? online. https://aithority.com/ait-featured-posts/are-you-confident-about-distinguishing-between-a-computer-generated-voice-and-human-voice/

[9] ID R&D. 2021. Combat Voice Spoofing Attacks. online. https://www.idrnd.ai/voice-anti-spoofing/

[10] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. online. https://keithito.com/LJ-Speech-Dataset/

[11] Ed Jefferson. 2020. Are voice biometrics the new passwords? online. https://www.raconteur.net/technology/cybersecurity/voice-biometrics/

[12] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2019. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. arXiv:cs.CL/1806.04558

[13] Rupert Jones. 2018. Voice recognition: is it really as secure as it sounds? online. https://www.theguardian.com/money/2018/sep/22/voice-recognition-is-it-really-as-secure-as-it-sounds

[14] Valencia A. Jones. 2020. *Artificial Intelligence Enabled - Deepfake technology The Emerge of a New Threat.* Master thesis. Utica College.

[15] Judith A. Markowitz. 2004. *Designing for Speaker.* Springer Netherlands, Dordrecht, 123–139. https://doi.org/10.1007/978-1-4020-2676-8_7

[16] Microsoft. 2020. *About the Speech SDK.* https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview#speaker-verification.

[17] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027. https://doi.org/10.1016/j.csl.2019.101027

[18] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2021. Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv:cs.CV/1909.11573

[19] Recommendation P.85. 1994. Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[21] Jon Petersen. 2019. Combating deepfakes with voice biometric technology. online. https://www.techradar.com/news/combating-deepfakes-with-voice-biometric-technology

[22] Phonexia. 2021. *Phonexia Voice Verify.* https://www.phonexia.com/en/product/voice-verify/.

[23] Precise Biometrics AB. 2014. *Understanding biometric performance evaluation.* https://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation-QR.pdf.

[24] Real-Time-Voice-Cloning 2020. [Online]. Single speaker fine-tuning process and results. Real-Time-Voice-Cloning GitHub. https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/437

[25] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD).* 1–10. https://doi.org/10.1109/SPED.2019.8906599

[26] Resmble AI. 2020. Resemble AI webpage. online. https://www.resemble.ai

[27] Eric Hal Schwartz. 2019. Deepfake Security Concerns Are Limiting Voice ID Adoption: Survey. online. https://voicebot.ai/2019/12/19/deepfake-security-concerns-are-limiting-voice-id-adoption-survey/

[28] John Seymour and Azeem Aqil. 2018. Your Voice is My Passport. https://www.blackhat.com/us-18/briefings/schedule/#your-voice-is-my-passport-11395

[29] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:cs.CL/1712.05884

[30] Dan Simmonss. 2017. BBC fools HSBC voice recognition security system. online. https://www.bbc.com/news/technology-39965545

[31] Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a Dataset for Spoken Language Recognition. In *Proc. IEEE SLT Workshop.*

[32] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:cs.SD/1609.03499

[33] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. arXiv:eess.AS/2005.13770

[34] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicolas Evans, Tomi Kinnunen, Kong Aik Lee, Ville Vestman, and Andreas. Nautsch. 2019. ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database. https://doi.org/10.7488/ds/2555

[35] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv:eess.AS/2109.00537

[36] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. 2020. Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion. https://www.isca-speech.org/archive/VCC_BC_2020/pdfs/VCC2020_paper_13.pdf

[37] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *CoRR* abs/1904.02882 (2019). arXiv:1904.02882 http://arxiv.org/abs/1904.02882

[38] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. 2020. SynSpeechDDB: a new synthetic speech detection database. https://doi.org/10.21227/ta8z-mx73

[39] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. 2021. FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection. arXiv:cs.SD/2110.09441

[40] Xiao Zhou, Zhen-Hua Ling, and Simon King. 2020. The Blizzard Challenge 2020. online. http://www.festvox.org/blizzard/bc2020/BC20_zhou_ling_king.pdf

Resilience of Voice Assistants to Synthetic Speech

Kamil Malinka, Anton Firc( ), Petr Kaška, Tomáš Lapšanský, Oskar Šandor, and Ivan Homoliak

Brno University of Technology, Božetěchova 2, 612 00 Brno, Czech Republic
{malinka,ifirc,ilapsansky,ihomoliak}@fit.vut.cz,
{xkaska01,xsando02}@stud.fit.vut.cz

Abstract. With the increasing integration of voice assistants in smart home systems, concerns regarding their security, especially regarding personal information access and physical entry control, have escalated. This is further amplified by the rapid development of generative AI methods, which bring new types of attacks. Therefore, we focus on modern voice assistants and their resilience against deepfake spoofing attacks. We rigorously assess the resistance of smart devices to sophisticated audio impersonation techniques. In detail, we evaluate voice assistants on four devices (Google Assistant, Siri, Bixby, and Alexa) with 72 test subjects. Subsequently, we conduct a comprehensive security analysis to determine the extent of potential impacts stemming from identified vulnerabilities. Our findings contribute to the enhancement of voice assistant security, ensuring safer and more reliable utilization in domestic environments.

Keywords: Deepfake · Voice Assistant · Security Analysis · Spoofing Attacks · Voice Biometrics

1 Introduction

Digital Assistants (a.k.a. Virtual Assistants, Intelligent Personal Assistants, or Artificial Intelligence Assistants) are becoming increasingly popular due to their growing sophistication and capabilities. These assistants are integrated into devices such as smart speakers, smartphones, or web services and use advanced AI approaches to perform individual tasks, answer questions, maintain conversations with users, and retain information for issuing reminders and warnings based on environmental constraints like time and location [27].

Around 3.25 billion Voice Assistant (VA) devices were purchased globally until 2019. Estimates indicate that by the end of 2024, the number of VA devices will surge to approximately 8.4 billion units, a figure equivalent to the world's population [14]. Many VAs employ speaker recognition to offer individual users personalised responses or authorise access to private data such as calendars or notes [28]. However, the use of VAs in home automation poses a plethora of security risks to users. If the authentication in VAs were easily evaded, it would

J. Garcia-Alfaro et al. (Eds.): ESORICS 2024, LNCS 14982, pp. 66–84, 2024.
https://doi.org/10.1007/978-3-031-70879-4_4

have severe consequences on home automation or leakage of the user's personal information. Moreover, smart home automation may grant physical access, which means that attacks can extend the cyber layer to reach the physical world.

For example, in 2020, an online streamer's residential address was inadvertently disclosed by activating a voice assistant [25]. While conducting a live stream, where the streamer was asleep, a viewer donated $25 and attached a voice command as a message, "Alexa, what is my current location?" This command, read aloud by the stream's text-to-speech system for donations, unintentionally triggered the streamer's voice assistant device. The device responded by audibly revealing the streamer's location, effectively resulting in an unintentional doxxing[1] incident.

On the contrary, developers of voice assistants usually allow only less sensitive operations to be carried out by voice commands. Nevertheless, third-party developers or end users may still use these assistants to unlock doors or authorise payments [1–3]. In the event of misuse of an assistant in this setting, a potential attacker can gain private information about the users, gain physical access to the home, or cause financial harm.

In addition to the traditional threats to the VAs, researchers need to focus on the new challenges and threats brought by the rapid development of Artificial Intelligence (AI), which motivated our research. Therefore, in this paper, we focus on deepfakes attacks on VAs.

Deepfakes are a subset of synthetic media (images, video, speech) automatically generated by AI [7]. There are already several attacks on voice biometric systems which utilise voice deepfakes [16]. Hence, we conjecture that voice assistants are also prone to deepfake spoofing regarding impersonating the victims. Using deepfake technology, the attacker can produce the desired commands in the victim's voice and play them to the assistant [22], which accepts them as legitimate and executes requested unauthorised action (e.g., opening the doors).

Due to the rising concerns about the resistance of these assistants against deepfake spoofing attacks, we conduct an empirical study that assesses the resilience of the most widespread voice assistants nowadays. We examine four separate assistants and their resistance to replay and deepfake spoofing attacks. The replay attacks provide an attack baseline, as they are one of the simplest means of spoofing voice biometrics systems.

In the case of deepfakes attacks, we first select a few publicly available tools to synthesise deepfake speech. Subsequently, minimal voice samples from 72 involved participants are collected to create a synthetic voice of users for each tool. Each user registers at all devices and the created synthesised output is then replayed from another device to attack the VA with the appropriate commands for each user. For better comparison, we also perform a simple replay attack. We utilise a realistic attack vector and exploit that some VAs do not distinguish the source of the sound [6,20,35].

[1] The act of publicly providing personally identifiable information about an individual or organisation.

Contributions. The main contributions of this paper can be summarised as follows:

1. We experimentally demonstrate the vulnerability of four voice assistants to attack based on voice deepfakes and replay attacks.
2. As part of the experiment, we also evaluate the suitability of the selected speech synthesis tools for this type of attack.
3. We analysed the proposed scenarios to evaluate the security impacts of demonstrated attacks.

2 Voice Assistants

Voice assistants belong to the voice-user interface (VUI) category. They are software applications that run in the background of voice command devices and are activated on the signal of particular phrases such as “Hey Siri ...”, “Alexa ...” or “Hey Bixby ...”. The user interacts with the voice assistant using a voice command. The voice assistant has a “keyword spotting” technology that recognises its wake-up command from ordinary speech [24]. Some of the assistants offer automatic speaker recognition (ASV). ASV means that the assistant can identify a person by their voice. Such a system first parses the user’s voice and then creates a unique acoustic model or voiceprint of the user’s voice [19]. Voice assistants such as Siri, Alexa or Bixby are equipped with ASV.

Voice assistants are often used with a smart speaker combination like Apple Homepod or Google Nest. A Smart Home is created when these speakers are connected to other home appliances. Such a home allows the user to use a phone or other input device to remotely control home appliances through the Internet connection. Thus, the user can control, for example, the temperature in the house, the lights or the security access to the house.

The vulnerabilities of voice assistants depend on the policies set by the user. For example, Amazon Alexa offers only limited features, but in combination with ASV it opens up a new set of policies, such as letting Alexa address you by name, entering personal events in the calendar, playing music, creating personal notifications, and letting Alexa say all the notifications or shopping online. All of these functionalities can be limited in the settings. Still, as the limitations increase, the assistant becomes more secure but ceases to be useful due to the functionalities’ limitations, which is counterproductive. One of the most attractive things to an attacker is personal information such as calendar data, contact names or devices linked to the assistant.

3 Related Work

The related work might be split into three logical and follow-up parts: speech synthesis, spoofing attacks on biometric systems, and spoofing voice assistants.

3.1 Deepfake Speech Synthesis

Deepfake speech is currently created (synthesised) using specialised tools that rely on deep learning methods [17]. Generative Adversarial Networks (GANs) or Variational AutoEncoders (VAEs) are often employed. We distinguish two techniques for creating deepfake speech: text-to-speech synthesis (TTS) or voice conversion (VC) [17]. TTS consumes written text and an embedding utterance and produces deepfake speech that sounds like the speaker on the embedding utterance. VC, in contrast, consumes a pair of utterances. A source utterance with the desired phrase and a target utterance and outputs the source phrase spoken in the speaker's voice on the target utterance.

The state-of-the-art speech synthesis tools work in zero or few-shot settings, requiring only a very short embedding (or target) utterance to synthesise the desired target speech. Moreover, the emphasis is given to multilingual models that can synthesise speech in multiple languages, even ones not seen during training. One of the currently best-known open-source tools is CoquiAI[2]. CoquiAI integrates multiple models and provides a user-friendly interface for speech synthesis tasks. The models include VITS [21] model, which combines variational inference, normalising flows, and adversarial training to enhance speech generation. It features a stochastic duration predictor for synthesising speech with diverse rhythms from text, effectively capturing natural speech variations in pitch and rhythm. YourTTS [12] builds on the VITS architecture but adds modifications to allow multi-speaker and multilingual training. These modifications include using raw text as input instead of phonemes, stochastic duration predictor, or the affine coupling layers of the decoder, encoder, and vocoder, which are conditioned on external speaker embeddings. TorToise [8] is an expressive, multi-voice TTS system applying recent advancements in image generation to speech synthesis. The field of image generation has significantly progressed with autoregressive transformers and denoising diffusion probabilistic models (DDPMs), which treat image creation as step-wise probabilistic processes utilising extensive data and computation. Initially developed for images, these techniques are now adapted to enhance speech synthesis.

3.2 Spofing Attacks on Biometrics Systems

Alegre et al. [5] stated that a generic biometric system might become vulnerable to voice synthesis, voice conversion, impersonation, and replay attacks. In the impersonation attack, another person imitates a voice to break biometric authentication. It has been proven that an attacker does not need to be a proficient voice impersonator to fool the ASV technology [30].

Evans et al. [15] tested the robustness of various ASV systems, showing very worrying results. Wu et al. [34] shows that replay attacks of a recording of a male voice were tested with a false acceptance rate (FAR) of 78.36% and a female voice with a FAR of 65.28%, which is enormously high. Replay attacks were

[2] https://github.com/coqui-ai/TTS.

previously seen as a major threat to ASV because of the complexity of creating a synthetic voice, but this is no longer true. As the population's awareness of deepfakes grows, people learn about all the possibilities of what they can do with deepfakes and methods of creating them are becoming more public [17].

Recent studies [16,29] have shown that deepfake spoofing attacks on biometrics systems are possible. Creating high-quality deepfakes is currently just a matter of minutes with paid services that allow fast and reliable voice cloning [17]. As the studies mentioned, the biometrics systems have no default ways to prevent such attacks.

3.3 Spoofing Voice Assistants

Recent studies have scrutinised the security of Voice Assistants (VAs) used in smart devices, given their integration into daily tasks and control of smart home devices. Focusing on the two prevalent VAs, Google Assistant and Siri, Bilika et al. [9] have investigated the robustness of their protection mechanisms, which are designed to limit sensitive operations to device owners. The study involved participants training these VAs to recognise their voices, followed by attempts to breach the systems using deepfake commands from participant-provided voice samples. The findings revealed that over 30% of the synthetic voice attacks successfully triggered the VAs to execute potentially hazardous tasks. Notably, the effectiveness of attacks varied significantly between the two vendors and displayed a gender bias in one instance.

Nacimiento-García et al. [26] explored the potential of spoofing attacks on Amazon Alexa. The approach centred on deploying YourTTS, a text-to-speech synthesis system, through a Telegram bot to generate cloned voice samples. These artificially synthesised voices were then employed to attempt impersonation attacks against Alexa to circumvent the voice profile-based identification mechanisms. The experiments aimed to verify the feasibility of conducting unauthorised activities by deceiving the voice recognition capabilities of these systems.

Finally, a proof-of-concept study [32] with 12 participants examined the potential exploitation of VAs through voice deepfakes. This research aimed to demonstrate the ease with which malicious entities could access privacy-sensitive data via Google Assistant, Alexa, and Siri. The study's experiments involved training a voice deepfake model with samples from participants and testing the model's effectiveness against the three digital assistants. The findings confirmed the viability of voice deepfakes to successfully extract sensitive information, such as birth dates, addresses, and personal contacts.

Our study markedly advances the field by substantially expanding the respondent pool to 72 individuals, which exceeds previous research efforts and aligns with the guidelines for qualitative studies of this nature [10]. Furthermore, we examine a broader range of voice assistants, incorporating tests on the four most popularly used models [11,31]. Crucially, our work includes a comprehensive threat analysis, meticulously evaluating the potential impacts and implications of our identified vulnerabilities.

Table 1. Individual voice assistants, their features and the software used.

Features	Google Assistant	Siri	Alexa	Bixby
Wake Word	"Hey Google"	"Hey Siri"	"Alexa"	"Hi Bixby"
Speaker recognition	Yes	Yes	Yes	Yes
NLP	Yes	Yes	Yes	Yes
Software	Google Assistant	iOS, WatchOS	Alexa app	Bixby app

4 Experiments

The experimental part examines the resilience of Voice Assistants to deepfake spoofing attacks. More specifically, whether voice assistants using automatic speaker recognition to identify a user can be spoofed using deepfake recordings of the user to reveal private information, cause financial harm, etc. We focus on the voice assistants Siri, Alexa, Bixby and Google Assistant. The parameters of the VAs are displayed in Table 1.

There are many methods how to carry out this attack, e.g., the adversary plays back the so-called sound near the VA, or the sound is reproduced by a smartphone or by inserting a malicious command into the TV or radio, which triggers the VA.

Our experiment only targets user authentication in the English language. All the tested assistants allow voice authentication to an enrolled voiceprint, which consists of repeating predefined phrases.

Every subject involved in the experiment was first enrolled into four voice assistants and then recorded to create the deepfake speech. While preparing speech synthesis models, the participant tested the acceptance rate of the bonafide trials, after which the subjects' cooperation was no longer necessary. Finally, we tested the acceptance rate of replay and deepfake spoofing attacks.

Our preliminary experiments observed that speaker recognition is performed only during the wake word recognition. The commands that follow the wake work after successful authentication may thus be spoken by an arbitrary speaker with no effect on the results [13]. Thus, we use this fact to simplify the executed experiments by testing only the wake word, not the whole content of the requests.

4.1 Used Speech Synthesisers

To create deepfake speech, we used four state-of-the-art speech synthesisers in a text-to-speech (TTS) setting. We selected two commercial (paid) and two open-source tools to cover the whole range of available tools, as shown in Table 2.
CoquiAI[3] is a paid service offering text-to-speech services. The synthesis models allowed uploading a short embedding recording and then synthesising speech with the speaker's voice from this recording. The minimal length was set to three seconds.

[3] Discontinued in 12/2023.

Table 2. Overview of employed speech synthesisers.

Name	Type	Min. Enrollment sample	Used Enrollment sample
CoquiAI	paid	3 s	20 s
ResembleAI	paid	25 sentences	25 sentences
TorToiSE	open-source	6 × 10 s	8 × 20 s
XTTS	open-source	3 s	20 s

ResembleAI[4] is a paid service offering text-to-speech and voice conversion. To create a deepfake voice, the user must read and record pre-defined sentences in the application interface. The minimum requirement is 25 sentences.

TorToiSe[5] [8] is an open-source text-to-speech tool. It allows a few-shot speaker adaptation using six ten-second utterances as embeddings.

XTTS[6] is an open-source text-to-speech tool. It allows a few-shot speaker adaptation using one at least three-second utterance as embedding.

4.2 Environment Description

All experiments were conducted in a quiet room with doors and windows closed to simulate the home environment where the voice assistants are being used. The assistants were placed on a table approximately one meter from the respondent, approximately one meter apart. These settings remained uniform for all trials and respondents.

4.3 Details of the Setup

The preliminary part of the experimental part tested whether automatic speaker recognition (ASR) was performed only for wake-word spotting or for the whole voice command. We took six respondents and grouped them into pairs. Person A was registered with the assistants. Person A activated the assistant using the wake word, and person B said an arbitrary voice command. This process was repeated five times for each pair. In every case, the assistant responded to the voice command of unregistered person B. The results thus confirm our hypothesis that the ASR is only performed for wake-word spotting. Thus, the attacker only needs to wake up the assistant using a spoofed voice and then deliver the voice command in his voice. Because of this behaviour, we can simplify further experiments only to test if the wake word is recognised, as the remainder of the voice command does not play a role [13].

[4] https://www.resemble.ai/.

[5] https://github.com/neonbjb/tortoise-tts.

[6] https://github.com/coqui-ai/TTS.

The experiment began with each participant signing a consent to participate in the experiment that collects anonymised data related to voice phrases.[7] Afterwards, the participant is enrolled into all voice assistants using standard procedures as instructed by the enrollment wizard. With all assistants set, each participant performed 30 bonafide authentication trials with each assistant.

The next step was to record the participant's speech and create a deepfake speech. Ninety sentences were recorded in total, where the first eight sentences were the wake words for the assistants (two sentences per assistant) to test the replay attacks. The following 25 sentences were used as enrollment recordings for the Resemble AI tool and the rest for the remaining speech synthesis tools. The sentences were uploaded to the Resemble AI tool to create a deepfake synthesis model. Meanwhile, all the remaining recorded sentences were concatenated into eight recordings, each consisting of approximately 20 s of speech. These recordings were provided to the TorToiSe as embeddings. Finally, the first of the eight concatenated recordings were used as an embedding recording for CoquiAI and XTTS tools. Using each tool, we synthesised one recording containing the wake word for each assistant. We used 16 recordings for each participant (four assistants and four synthesisers). The synthesis was executed in an iterative manner; we subjectively evaluated the naturalness and noise in the deepfake recording every time and repeated the synthesis until the recording contained comprehensible speech without significant noise. On average, we had to repeat the synthesis process one to three times.

After testing the bonafide attempts and recording enrollment samples, the participants' jobs were over, and we continued our experiments as follows. First, we tested the replay attacks by replaying the original sentences with wake words for each assistant. Then, after synthesising all the deepfake speech, we continuously played the wake words synthesised using different tools to all assistants.

The experimental procedure for each participant consisted of several stages, which lasted approximately 1 h and 30 min. The breakdown of this time is as follows: registering with the voice assistants took 10 min; recording the participant's speech was a 15-min process; conducting the genuine trials also took 15 min. The creation of deepfakes varied in time, ranging from 10 min to an hour, largely depending on the server load during ResembleAI's training phase. In addition, replay attack trials were completed in 10 min, while deepfake spoofing attacks took 40 min to test. Consequently, the total time to complete experiments across all respondents was two months.

We opt not to include the bonafide tests in our study, primarily due to their time-consuming nature and lack of variability in results. Our initial testing with 36 respondents yielded consistently high success rates (over 95% accuracy), indicating a plateau in data variability. As a result, we decided to omit this part of the test, thereby reducing the engagement time for each respondent by approximately 15 min. This decision does not affect the validity of our study. It

[7] Note that this experiment was reviewed by our institutional review board who confirmed that no private or personal data are stored while all other collected data are properly anonymised.

Table 3. Devices and their specific versions used in experiments for individual assistants.

Voice Assistant	Device	Software version
Google Assistant	Google Nest Mini 2 gen. 2020	2.57.375114
Siri	iPhone SE	iOS 16.6.1
Bixby	Samsung Galaxy A53 5G	Android 13
Alexa	Echo Dot 4 gen. 2020	9295801732

reflects that voice assistants, as commercial products are optimised for usability, often prioritising usability over security. This optimisation inherently leads to high acceptance rates in bonafide mated trials, as evidenced by our preliminary results, where, at most, only one in thirty trials were unsuccessful.

The success rate was computed to evaluate the efficacy of each verification attempt. The ratio of successful trials to total trials (30) was calculated distinctly for each unique combination of participant, VA and speech synthesiser. This ratio was then converted into a percentage, representing the proportion of successful trials out of the total trials conducted. The *success rate* serves as a critical metric, with an ideal rate approaching 100% for bonafide mated trials[8], indicating high reliability, and conversely, approaching 0% for replay and deepfake spoofing attacks, indicating robust security. The *success rate* for each participant, assistant and synthesiser was calculated using the formula:

$$\textit{success rate (\%)} = \left(\frac{\text{number of successful trials}}{30} \right) \times 100$$

Finally, it was necessary to use the same version of the software throughout the entire measurement period to avoid possible deviations in the measurement that could occur due to fixing various bugs or improving the assistants‘ features. The setup of the assistants can be seen in Table 3.

5 Experimental Evaluation

To test the resilience of voice assistants to deepfake speech, we collected results from 72 respondents. Each respondent created their profile in the tested assistants, and then we evaluated the resilience of these assistants to replay and deepfake spoofing attacks.

The testing group was composed of 72% males and 20% females. 84% of respondents had Czech nationality, 14% were Slovak, and 2% were Ukrainian. The age distribution was as follows: 55 young participants (19–34 years), nine early middle-aged (35–49 years), six late middle-aged adults (50–65 years), and two elderly (66 and more).

[8] Verification attempts where a legitimate user's voice sample is presented to their voice assistant.

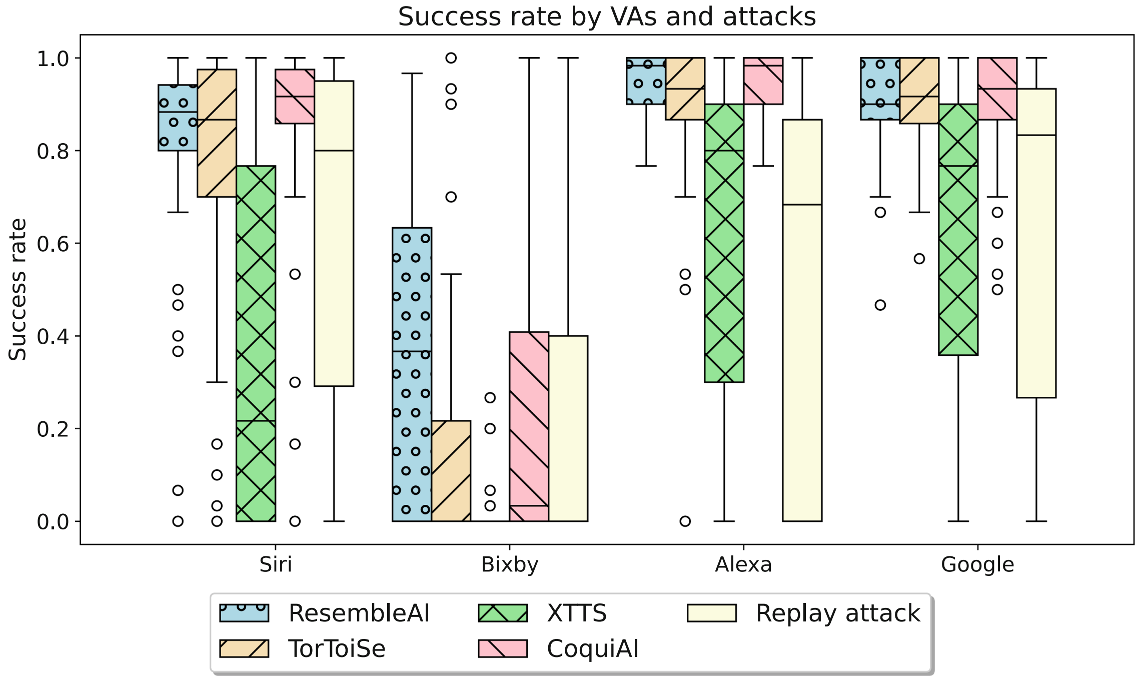


Fig. 1. Success rates of attacks on voice assistants.

The baseline – bonafide attempts were collected for 36 respondents, where the success rate steadily remained over 95%. Due to no changes in the observed success rate, we dropped the bonafide testing to preserve the respondents' time. The assistants are primarily set for usability, documented by the high success observed.

The breakdown of attack success rates is shown in Fig. 1. The replay attacks succeeded approximately every second time, while some of the deepfakes reproduced the bonafide success rates of more than 90%.

The findings reveal that Bixby consistently repelled most of the attempted attacks. However, whether this resilience results from better security or more sensitive ASR is questionable. We noticed that the success rate of attack verification attempts is influenced by the pause length between the words in the wake sentence, as further mentioned in Sect. 7.1. This observed *resilience* may thus only be a result of too-sensitive ASR. In contrast, other assistants accepted most attacks as bonafide attempts. However, due to the proprietary nature of Bixby's internal mechanisms and the lack of published details on the parameters of its deep learning model, it is difficult to determine the specific factors contributing to its enhanced security performance.

The paid synthesisers achieved very high success rates, which shows that such an attack is plausible. Only open-source XTTS deviates from this outstanding spoofing potential, and the success rates are distributed throughout the spectrum. Unfortunately, no pattern is observable in the collected data explaining this distribution. This behaviour may be caused by individual vocal characteristics of individual respondents, where some have a voice similar to one used for training the XTTS tools. Thus, their deepfake achieves higher success rates and vice versa.

As the speech quality of the paid synthesisers is generally better, it is evident that the paid synthesisers performed the best. However, even the open-source TorToiSe was able to approach the paid synthesisers.

However, even the worst-performing tool (XTTS) succeeded at least once for most respondents. Since there is no limit on authentication attempts, the attacker thus only requires more time to try multiple times until one of the attempts succeeds. The lower success rate thus only increases the time complexity of an attack.

Finally, we assessed the impact of demographics on observed attack success rates. The trials are independent, and observed attack success rates do not follow the normal distribution. To evaluate the influence of gender and nationality[9] used the Mann-Whitney U test with a significance level $\alpha = 0.5$ to compare the rates for each pair *assistant – attack*. The only significant difference was found in the case of Bixby with ResembleAI and CoquiAI attacks. This difference was measured for male/female and Czech/Slovak success rates. To assess the impact of age, we used the Kruskal-Wallis H Test with a significance level $\alpha = 0.5$ for all age groups across *assistant – attack* pairs. No significant difference in observed success rates was found.

The demographics, thus, do not influence the success rate of evaluated attacks. There are minor differences only for the Bixby assistant, which further supports the hypothesis of Bixby‘s too-sensitive ASR.

Overall, the success rates of the deepfake spoofing attacks are considerably high. Spoofing voice assistants is thus an undemanding process, raising many security concerns.

6 Threat Analysis

We have shown that voice assistants, specifically the automatic speaker recognition implemented in such assistants, are vulnerable to deepfake spoofing attacks. The next step is to assess the real impact of a potential attack.

For example, let us consider the scenario where the attacker can throw a wireless speaker into a room through a window and gain complete control of a smart home by spoofing a voice assistant. As we demonstrate that such an attack is feasible, it is crucial to understand the security implications of the presented vulnerability. To this extent, we perform a security analysis of voice assistants' standard functions and assess how easily these functions may be misused and the potential damage of such misuse.

While tested assistants primarily share the same functionality, there are some differences. This section, thus, describes and breaks down the functions that could be abused.

The categorisation is based on the following factors:

Difficulty of execution:

[9] Only Czech and Slovak. We excluded Ukrainian since there was only one respondent.

- *Low* – short sentences
- *Medium* – medium-length sentences, including complicated sequences, such as phone numbers
- *High* – long sentences and follow-up questions

Device state:

- *Locked* – the device is locked
- *Unlocked* – the device is unlocked

Attack severity:

- *Low* – an inconvenience for the victim or minimal privacy breach
- *Medium* – exploitable information, low financial loss or defamation
- *High* – unauthorised access to an object, significant financial loss, major privacy breach

Next, we present a detailed description of functions. The functions' and properties' summary is depicted in Table 4.

Table 4. Overview of scenarios and their parameters. VAs column abbreviations: A – Alexa, B – Bixby, G – Google and S – Siri.

Scenario	Difficulty	State	Severity	VAs
Phone call	**High**	Locked	Med-High	All
Sending messages	**High**	Locked	Med-High	All
Reading notifications	Low	Locked	Low-Med	All
Reading text messages	Low	Unlocked	Low-Med	All
Operating camera	Low-Med	Locked	Low-Med	All
Accessing digital wallet	Med-High	Unlocked	**High**	Siri
Subscription management	Low-Med	Locked	Med-High	Alexa
Controlling smart home	Low-Med	Locked	Low-High	ABS
Calendar, schedules access	Low-Med	Locked	Low-Med	All
Online shopping	**High**	Unlocked	**High**	AG
Information retrieval	Low-Med	Both	Low-Med	All

Phone Calls: The ability to make calls from a stranger's device can be abused in several ways, such as making scam calls from a stranger's number or calling premium rate numbers. For smart speakers, exploiting things like calls to emergency services is impossible since very few providers have this functionality enabled. Dialling premium rate numbers are only available on Siri and Google since Alexa does not support dialling such numbers. The difficulty of the attack has been classified as medium to high because it is necessary to pronounce the whole number quickly. A slight pause in pronouncing the phone number will interrupt the action. Possible damages have been classified as medium to high since making phone calls via a paid line and dialling premium numbers is possible, thus causing financial damage. These conditions apply to the use of assistants via mobile phones. Smart speaker devices have this functionality limited to specific locations.

Sending Messages: Sending text or multimedia messages can be misused to transmit scam messages, send advertisements or send dangerous links that can be part of SMS phishing. However, these attacks are challenging as the entire message content must be dictated to the device. This attack can also be completed without unlocking the phone or device. The damage factor has been rated medium to high mainly because messages can be sent to someone in contact. This increases the chance that the recipient will be fooled by a phishing SMS message, considering that they will receive a message from someone they know. Even if the phishing is unsuccessful, SMS is a paid service, so the victim can still be financially ill.

Reading Notifications: The possible misuse of reading notifications can vary widely, as reading all the content in notifications is possible. Primarily, it can be used to read personal conversations. However, it can also be used to read messages containing a verification code to log into a bank account or to read notifications from applications providing two-factor authentication. The difficulty of executing this attack is low simply because the sentence to trigger the action is straightforward. Even though the state is categorised as locked, it depends on the phone settings, which must be set so that the content of the message is displayed in the notification, even on the locked screen. It is also possible to read the notification only once. The potential damage caused by this attack is categorised as low-medium because it depends on the phone's settings, and the stand-alone code the attacker gets cannot be exploited. For a possible exploit, the attacker must trigger the notification via a bank login or use the code to receive a package in a stranger's name.

Reading Text Messages: An attack is working on the same principle as reading notifications, with the difference being that it is only possible to read text messages in the preconfigured application for sending and reading messages. Unlike reading notifications, this function is only available when the phone is unlocked. It is possible to read older messages and read messages more than once. However, getting information from other applications is impossible, as in the case of reading notifications.

Taking Pictures and Recording Videos: Using camera functions can also be exploited by an attacker, as taking photos and videos using only voice commands is possible. With Google Assistant and Siri, this function can be invoked even in locked mode. Alexa also has this function but requires a specific kind of device called Echo Show, which has its display and camera. Taking photos and videos is not considered high-risk, but it could be a dangerous combination with messaging.

Misusing Digital Wallet: Misusing a digital wallet like Apple Pay can be very easy on devices using Siri, as all it takes is a short sentence to send a payment between known accounts. However, it is classified as medium to high in difficulty. This attack also requires payment confirmation by tapping on the smartphone screen, so the attack cannot be carried out by voice alone.

Managing Subscriptions: If the user of a device with the Alexa voice assistant has filled in all the necessary details for payment, it is possible to subscribe to the Amazon Music app using voice only. As this is a pay-as-you-go plan with per-month billing, a significant financial loss is doable if this event goes unnoticed.

Using Smart Home or Internet of Things (IoT) Devices: As more and more devices can be connected to voice-controlled systems, the following devices are under threat of being misused: Smart Televisions, Thermostats, Lights, Locks, and Cameras. Due to the significant variation of different devices with different functionalities, it is impossible to determine the possible amount of damage that such an attack could cause. A case in which an attacker lights a light bulb in a room might not have as many financial or other consequences as in which a perpetrator unlocks the front door of a house or sets the thermostat to the highest possible temperature.

Managing Calendars, Schedules, To-Do Lists, Timers, and Routines: All assistants can store and manage large amounts of information that may be of little to no value to an attacker. These functionalities would probably only inconvenience the victim, but some could be considered vulnerable. For example, information about a person's schedule could provide his whereabouts, which the criminal could exploit.

Making Online Purchases: Making online purchases is a broad term, and for each voice assistant, it can mean something different. In some countries, Google Assistant allows users to authorise payments and make in-app purchases through Google Play. Alexa enables users to manage their shopping cart and purchase through Amazon shop. In the case of Siri, Apple has decided not to provide purchases through the voice assistant due to privacy concerns and the unreliability of authentication.

Retrieving Information: Different systems store the information provided to the assistant differently. Google Assistant can remember specific information such as the front door code or package shipments. The process of storing and retrieving data is as follows:

"Hey, Google, remember that my front door code is 1110."
"Hey Google, what's my front door code."

With this request, it is possible to get a response containing the code from the front door. This function is also available when the phone is locked. Alexa stores the same information in its notes; therefore, it cannot be obtained by asking. Siri stores this information like Alexa, so reading the notes is required to retrieve the data. However, this is impossible on iOS devices from a locked state, so the device must be unlocked.

7 Discussion

Our experiments demonstrate how exposed VAs are to deepfake spoofing attacks. Even though tools like XTTS were not particularly successful, and the voice assistant Bixby rejected most attack attempts, it is essential to point out a few key facts.

In our study, we observed that voice assistants, by design, do not restrict the number of access attempts, as they continually listen for activation phrases like "Hey Siri" without distinguishing between the device owner and others. This characteristic implies that even a single successful trial can be deemed effective for an attacker, as unlimited attempts are available. In this context, a lower success rate merely extends the time needed to execute an attack rather than preventing it successfully. For instance, achieving access in just one out of 30 trials is sufficient to consider the attack successful. In our experiments involving 72 subjects, Bixby, the most secure, denied access in all 30 attempts for only seven subjects. In contrast, for other voice assistants, every subject managed to gain access at least once. Therefore, while a higher success rate indicates a more efficient attack, any success rate higher than zero ultimately leads to the same outcome-the attacker gains access.

7.1 Observations

During the experiments, we have obtained four observations that are worth mentioning:
O1: Google assistant sometimes responds to the "Hey Siri" wake word. This behaviour was noticed with bonafide and deepfake attempts.
O2: Bixby has a better success rate if there is a longer pause between "Hey" and "Bixby" words.
O3: Some respondents read the sentences for deepfake creation unnaturally fast, resulting in lowered deepfake quality.
O4: Some respondents mispronounced the wake words, such as "Hey Siiiiiiri"; however, such mispronunciation seems to have no impact.

These observations may impact the final results; however, examining them would require a different experiment setting and is thus out of the scope of our research. These observations may be further explored in future research.

7.2 Mitigation Methods

In the advancement of voice assistant (VA) technologies, the emergence of deepfake spoofing attacks presents a significant challenge, necessitating the development of countermeasures. To mitigate these risks, continuous authentication represents a possible strategy, extending identity verification beyond the initial login to cover the entire user session. This approach will certainly make the attack more difficult to execute by requiring a longer deepfake; however, based on developments to date, attackers can be expected to manage this as well. For example, one can expect to create a deepfake in real time soon using voice conversion methods. Thus, we recommend focusing more on limiting VA activities under weak voice authentication or further strengthening the authentication process. This could involve utilising multifactor authentication techniques, such as passphrase verification or the requirement for a recognized device to be near the VA system, enhancing the dynamic security landscape.

Moreover, the integration of liveness [4,18,33] or deepfake [17,23] detection modules into the authentication process is imperative. By continuously analyzing audio inputs, these systems can discern between genuine human voices and synthetic reproductions, thus preventing unauthorized access attempts. Additionally, limiting the scope of actions available via voice commands, especially concerning sensitive data, further secures VA systems against the potential misuse stemming from successful spoofing attempts.

Physical security measures, such as deactivating the VA device when not in use and safeguarding it against unauthorized physical access, play a supportive role in the overarching security framework. The collective application of continuous authentication, detection technologies, command restrictions, and physical security forms a comprehensive defence strategy. Such an approach is pivotal in addressing the multifaceted threats posed by deepfake technologies, thereby safeguarding the integrity of voice assistant systems in the face of evolving cyber threats.

8 Conclusions

We have shown that the currently and dominantly used voice assistants are not resilient to replay or deepfake spoofing attacks. The attacker can easily synthesise the victim's speech and then replay this deepfake speech to a voice assistant in hold of the victim to reveal personal information or cause financial harm. This shows the importance of choosing the appropriate authentication mechanism for each use case. The rigorous threat analysis reveals the possible privacy breaches and financial harms. At the same time, most voice assistant developers understand the security risks associated with speaker recognition implemented in their voice assistants and do not allow them to operate critical functions through voice commands. However, third-party developers or end users might try to use these speaker recognition functionalities, for example, to control devices such as smart locks or authorise online payments, which brings several severe security concerns. In future work, different devices that the voice assistants operate on, such as smart speakers, smartphones, and smartwatches, should be tested to see if they provide the same level of security.

Acknowledgments. This work was supported by the national project NABOSO: Tools To Combat Voice DeepFakes (with code VB02000060) funded by the Ministry of the Interior of the Czech Republic and the Brno University of Technology internal project FIT-S-23-8151.

References

1. Bixby Developers — bixbydevelopers.com. https://bixbydevelopers.com/dev/docs/bhs-dev-guide. Accessed 29 Nov 2023
2. Google Assistant for Android—Documentation — Android Developers — developer.android.com. https://developer.android.com/guide/app-actions/overview. Accessed 29 Nov 2023
3. SiriKit — Apple Developer Documentation — developer.apple.com. https://developer.apple.com/documentation/sirikit/. Accessed 29 Nov 2023
4. Ahmed, M.E., Kwak, I.Y., Huh, J.H., Kim, I., Oh, T., Kim, H.: Void: a fast and light voice liveness detection system. In: 29th USENIX Security Symposium (USENIX Security 2020), pp. 2685–2702. USENIX Association, August 2020. https://www.usenix.org/conference/usenixsecurity20/presentation/ahmed-muhammad
5. Alegre, F., Janicki, A., Evans, N.: Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In: Proceedings of the Conference Name. EURECOM and Warsaw University of Technology, Sophia Antipolis, France and Warsaw, Poland (2023)
6. Alepis, E., Patsakis, C.: Monkey says, monkey does: security and privacy on voice assistants. IEEE Access **5**, 17841–17851 (2017). https://doi.org/10.1109/ACCESS.2017.2730220
7. Bateman, J.: Deepfakes and synthetic media in the financial system: assessing threat scenarios. Technical report, Carnegie Endowment for International Peace (2020). http://www.jstor.org/stable/resrep25783.1
8. Betker, J.: Better speech synthesis through scaling (2023)
9. Bilika, D., Michopoulou, N., Alepis, E., Patsakis, C.: Hello me, meet the real me: voice synthesis attacks on voice assistants. Comput. Secur. **137**, 103617 (2024). https://doi.org/10.1016/j.cose.2023.103617. https://www.sciencedirect.com/science/article/pii/S0167404823005278
10. Boddy, C.R.: Sample size for qualitative research. Qual. Market Res. Int. J. **19**(4), 426–432 (2016). https://doi.org/10.1108/qmr-06-2016-0053. http://dx.doi.org/10.1108/QMR-06-2016-0053
11. BotPenguin: which are the 7 best voice assistants of 2023? November 2023. https://botpenguin.com/blogs/which-are-the-7-best-voice-assistants-of-2023
12. Casanova, E., Weber, J., Shulby, C., Junior, A.C., Gölge, E., Ponti, M.A.: YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone (2023)
13. Combs, M., Hazelwood, C., Joyce, R.: Are you listening? – an observational wake word privacy study. Organ. Cybersecur. J. Pract. Process People **2**(2), 113–123 (2022). https://doi.org/10.1108/ocj-12-2021-0036. http://dx.doi.org/10.1108/OCJ-12-2021-0036
14. Daniel Ruby: 65 Voice Search Statistics for 2023 (Updated Data) (2023). https://www.demandsage.com/voice-search-statistics/
15. Evans, N., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification. In: Proceedings of INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 2013. https://doi.org/10.21437/Interspeech.2013-288
16. Firc, A., Malinka, K.: The dawn of a text-dependent society: deepfakes as a threat to speech verification systems, pp. 1646–1655 (2022). https://doi.org/10.1145/3477314.3507013, cited by: 2

17. Firc, A., Malinka, K., Hanáček, P.: Deepfakes as a threat to a speaker and facial recognition: an overview of tools and attack vectors. Heliyon **9**(4), e15090 (2023). https://doi.org/10.1016/j.heliyon.2023.e15090
18. Gupta, P., Gupta, S., Patil, H.: Voice liveness detection using bump wavelet with CNN. In: 9th International Conference on Pattern Recognition and Machine Intelligence, Kolkata, India, December 2021. https://hal.science/hal-03690065
19. Hoy, M.B.: Alexa, siri, cortana, and more: an introduction to voice assistants. Med. Ref. Serv. Q. **37**(1), 81–88 (2018). https://doi.org/10.1080/02763869.2018.1404391
20. Wakefield, J.: Burger King advert sabotaged on Wikipedia (2017). https://www.bbc.com/news/technology-39589013
21. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech (2021)
22. Lien, J., Al Momin, M.A., Yuan, X.: Attacks on Voice Assistant Systems, pp. 61–77. IGI Global (2022). https://doi.org/10.4018/978-1-7998-7323-5.ch004. http://dx.doi.org/10.4018/978-1-7998-7323-5.ch004
23. Liu, X., et al.: Asvspoof 2021: towards spoofed and deepfake speech detection in the wild. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 2507–2522 (2023). https://doi.org/10.1109/TASLP.2023.3285283
24. Lopez-Espejo, I., Tan, Z.H., Hansen, J.H.L., Jensen, J.: Deep spoken keyword spotting: an overview. IEEE Access **10**, 4169–4199 (2022). https://doi.org/10.1109/ACCESS.2021.3139508
25. Memey-McMemeFace: Alexa what is my current location (2020). https://www.reddit.com/r/WatchPeopleDieInside/comments/iky0qd/alexa_what_is_my_current_location. Accessed 14 Dec 2023
26. Nacimiento-García, E., Caballero-Gil, C., Nacimiento-García, A., González-González, C.: Alexa, do what i want to. Implementing a voice spoofing attack tool for virtual voice assistants. In: Bravo, J., Ochoa, S., Favela, J. (eds.) UCAm I 2022. LNNS, vol. 594, pp. 413–418. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-21333-5_41
27. Poushneh, A.: Humanizing voice assistant: the impact of voice assistant personality on consumers' attitudes and behaviors. J. Retail. Consum. Serv. **58**, 102283 (2021). https://doi.org/10.1016/j.jretconser.2020.102283. https://www.sciencedirect.com/science/article/pii/S0969698920312911
28. Qualcomm: Getting personal with on-device AI (2023). https://www.qualcomm.com/news/onq/2023/10/getting-personal-with-on-device-ai
29. Seymour, J., Aqil, A.: Your voice is my passport (2018). https://www.blackhat.com/us-18/briefings/schedule/#your-voice-is-my-passport-11395
30. Simmons, D.: BBC news, May 2017. https://www.bbc.com/news/technology-39965545
31. Staff, R.: The best voice assistant, September 2021. https://www.zdnet.com/home-and-office/smart-home/the-best-voice-assistant/
32. Ubert, J.: Fake it: attacking privacy through exploiting digital assistants using voice deepfakes. Ph.D. thesis (2023). https://www.proquest.com/dissertations-theses/fake-attacking-privacy-through-exploiting-digital/docview/2811176534/se-2. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-05-18
33. Wang, Y., Cai, W., Gu, T., Shao, W., Li, Y., Yu, Y.: Secure your voice: an oral airflow-based continuous liveness detection for voice assistants. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(4) (2020). https://doi.org/10.1145/3369811

34. Wu, Z., Gao, S., Chng, E.S., Li, H.: A study on replay attack and anti-spoofing for text-dependent speaker verification. In: Proceedings of the Conference Name. Centre for Speech Technology Research, University of Edinburgh, United Kingdom and Human Language Technology Department, Institute for Infocomm Research, Singapore and School of Computer Engineering, Nanyang Technological University, Singapore (2021)
35. Zhang, R., Chen, X., Lu, J., Wen, S., Nepal, S., Xiang, Y.: Using AI to hack IA: a new stealthy spyware against voice assistance functions in smart phones. arXiv preprint arXiv:1805.06187 (2018)