engineering proceedings

MDPI

Proceeding Paper

AI-Driven Vishing Attacks: A Practical Approach [†]

Fabricio Toapanta *,‡, Belén Rivadeneira ‡, Christian Tipantuña ‡ and Danny Guamán ‡

Facultad de Ingeniería Eléctrica y Electrónica, Escuela Politécnica Nacional, Quito 170525, Ecuador; belen.rivadeneira@epn.edu.ec (B.R.); christian.tipantuna@epn.edu.ec (C.T.); danny.guaman@epn.edu.ec (D.G.)

* Correspondence: fabricio.toapanta@epn.edu.ec; Tel.: +593-987084772

[†] Presented at the XXXII Conference on Electrical and Electronic Engineering, Quito, Ecuador, 12–15 November 2024.

[‡] These authors contributed equally to this work.

Abstract: Today, there are many security problems at the technological level, especially in telecommunications. Cybercriminals invade and steal data from any system using vector attacks such as phishing through scam mail, fake websites and phone calls. This latter form of phishing is called vishing (phishing using voice). Through vishing and using social engineering techniques, attackers can impersonate family members or friends of potential victims and obtain information or money or a specific target objective. Traditionally, to carry out vishing attacks, attackers imitated the vocabulary, voice and tone of a person known to the victim. However, with current artificial intelligence (AI) tools, obtaining synthetic voices similar or identical to the person to be impersonated is more straightforward and precise. In this regard, this paper, using ChatGPT and three AI-enabled applications for voice synthesis presents a practical approach for deploying vishing attacks in an academic environment to identify the limitations, implications and possible countermeasures to mitigate the effects on Internet users. Results demonstrate the effectiveness of vishing attacks, and the maturity level of the employed AI tools.

Keywords: cybersecurity; ChatGPT; social engineering; artificial intelligence; LLM; vishing

check for updates

Citation: Toapanta, F.; Rivadeneira, B.; Tipantuña, C.; Guamán, D. AI-Driven Vishing Attacks: A Practical Approach. *Eng. Proc.* **2024**, *77*, 15. https://doi.org/10.3390/engproc2024077015

Published: 18 November 2024

1. Introduction

Phishing attacks attempt to obtain personal information such as usernames, passwords and credit card details by pretending to be a trustworthy entity in an electronic communication. Communication purporting to be from popular social websites, auction sites, online payment processes or IT administrators is commonly used to lure the unsuspecting public. Furthermore, phishing emails may contain links to websites that are infected with malware [1]. Vishing attacks, a particular case of phishing, are also used to obtain private information using phone calls [2]. Even today, artificial intelligence (AI) tools are used to clone voices, so the victim hears a familiar voice, increasing the probability that the launched attack succeeds. Moreover, with suitable feedback (e.g., by obtaining personal data), AI solutions can impersonate a person and imitate their vocabulary, jargon or expressions. In this regard, this paper uses AI tools to present a practical approach for deploying vishing attacks. In this paper, ChatGPT was used to generate the text that will later be used to create audio dialogs (scripts) to send to victims. Texts from the victims, which could be obtained through social engineering or espionage techniques in real-world environments, have previously fed ChatGPT. Three AI-enabled tools have been used to generate the synthetic voices. The tests were conducted in an academic environment, and the targets were relatives or friends of university students who were aware of the nature of the experiments and their implications. Results demonstrated the effectiveness of the vishing attacks, and all collected information could be used not only to know the maturity level of the AI tools to perform this vector attack but also the possible countermeasures and actions to avoid or mitigate the effects on the users. At the end of the experiments, all individuals involved were informed about the study, the tools employed and the results obtained.

2. Literature Review

In this work, it was possible to identify how phishing attacks, especially those that escape detection by security tools, can be generated using ChatGPT. Phishing websites, after being hosted on a web domain, are widely shared through several online communication resources, such as emails, SMS, social media, etc., and indexed by search engines. Since these attacks can cause a lot of harm within a short duration [3], threat intelligence actors, such as anti-phishing bots, utilize various automated, rule-based [4] and Machine Learning ML-based approaches [5] to identify phishing websites.

2.1. Threat Model

An attacker provides multiple prompts to ChatGPT with the intention of generating the source code for a phishing website that: (a) closely imitates the design of a popular organization's website, (b) employs various regular and evasive tactics to deceive users into sharing their sensitive information and (c) incorporates mechanisms for transmitting the obtained credentials back to the attacker [5]. These prompts are specifically designed to bypass ChatGPT's content filtering measures, making it difficult to detect the malicious intent. Once the attacker receives the generated source code, they proceed to host it on a domain, creating a live phishing website that poses a significant risk to unsuspecting users. Generating phishing attacks using ChatGPT provides the attacker with the following advantages:

Rapid Deployment: Attackers take advantage of the low cost and ease of use of ChatGPT to quickly iterate on their phishing attacks, making it more difficult for security vendors to identify and counter them at scale [6].

Technical Expertise Variability: The ease of use of ChatGPT allows attackers with varying levels of technical expertise to generate phishing attacks which employ various evasive techniques that can avoid anti-phishing detection, such as text encoding, browser fingerprinting or clickjacking [6].

Hosting and Accessibility: Attackers can utilize free hosting platforms to deploy their phishing websites, further lowering the barriers to entry and making large-scale attacks more feasible [6].

2.2. Ethical Considerations

This paper acts as a proof-of-concept to demonstrate the feasibility of generating evasive vishing attacks using ChatGPT. Since the authors are in the process of releasing their set of prompts used to design these attacks to OpenAI, they also describe their methodology for generating these attacks to encourage the development of more generalized approaches to detect and mitigate the possibility of such attacks in ChatGPT and other large language models (LLMs).

3. Background Technologies

This section will review some technologies used to carry out this work.

3.1. Large Linguistic Model (LLM)

Large Language Models (LLMs) represent an evolution from language models. Initially, language models were statistical in nature and laid the groundwork for computational linguistics. The advent of transformers has significantly increased their scale. This expansion, along with the use of extensive training corpora and advanced pre-training techniques is pivotal in areas such as AI for science, logical reasoning and embodied AI. These models undergo extensive training on vast datasets to comprehend and produce text that closely mimics human language. Typically, LLMs are endowed with hundreds of billions, or even more, parameters, honed through the processing of massive textual data. They have spearheaded substantial advancements in the realm of Natural Language Processing (NLP) [7] and find applications in a multitude of fields (e.g., risk assessment [8], programming [8], vulnerability detection [8], medical text analysis [9] and search engine optimization [7]).

Large language models (LLMs) can be used to analyze cyber threat intelligence (CTI) data from cybercrime forums, which contain extensive information and key discussions about emerging cyber threats. However, to date, the level of accuracy and efficiency of LLMs for such critical tasks has yet to be thoroughly evaluated. One study, hence, assessed the accuracy of an LLM system built on the OpenAI GPT-3.5-turbo model [9] to extract information. Various ways to enhance the model were uncovered, such as the need to help the LLM distinguish between stories and past events, as well as the need to be careful with verb tenses in prompts. Nevertheless, the results of this study highlight the efficiency and relevance of using LLMs for cyber threat intelligence [9].

3.2. How LLMs Work?

At a basic level, LLMs are based on ML, a subset of AI that refers to feeding a program a large amount of data to train it to identify functions from that data without human intervention [8]. LLMs use a type of ML called deep learning. Deep learning models can be trained to recognize distinctions without human intervention, although some human adjustment is usually necessary [8]. Deep learning uses probability to "learn". For example, in the sentence "The quick brown fox jumped over the lazy dog", the letters "e" and "o" are the most common, appearing several times each. From this, a deep learning model might conclude (correctly) that these characters are among those most likely to occur in Spanish text. A deep learning model cannot conclude anything based on a single sentence. However, after analyzing billions of sentences, it could learn enough to predict how to logically finish an incomplete sentence or even generate its sentences [8].

3.3. Social Engineering (SE)

This term has a long history, but it can be summarized by saying that social engineering is the name given to the various manipulation techniques cybercriminals use to obtain sensitive information from users. First, it was taken from computer science, specifically hackers, and second, it was taken from personal. For hackers or people engaged in computer security in general, this method is the process of using cognitive strategies and social skills to engage a specific target [9].

SE jeopardizes the security of networks, regardless of the type of perimeter security in place, from firewalls, cryptographic methods, virus detection system, etc., as it exploits the possibility that humans trust other humans more than computers or technology, thus making humans the weakest link in the security chain. Malicious activities carried out through human interactions psychologically influence a person to divulge confidential information or violate security procedures. Because of these human interactions, SE attacks are the most powerful, as they threaten all systems and networks and cannot be prevented by software or hardware solutions if people are not trained to prevent them. Cybercriminals choose these attacks when there is no way to hack a system without technical vulnerabilities [8].

Social Engineering Attacks

Given that SE is a technique that uses emotions, feelings, vulnerabilities and other mental processes and human behavior, it becomes a very dangerous tool to affect the potential victim; for example, by impersonating family members, technical support people, co-workers or trusted individuals in order to appropriate personal data and passwords or to impersonate the identity of the deceived person using memories, moments or personal information [7].

Although SE attacks differ from each other, they have a common pattern with similar phases. The pattern involves four stages [9,10]:

- Gather information about the target;
- Develop a relationship with the target;
- Exploit the available information and execute the attack; and
- Exit without leaving a trace.

Figure 1 illustrates the different stages of an SE attack.

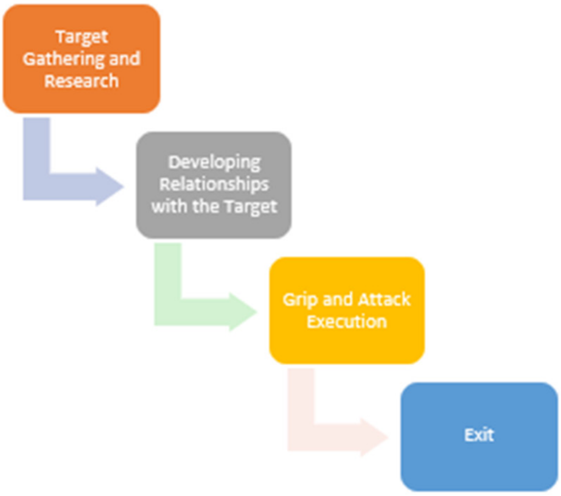
Target Gathering and Research
Developing Relationships with the Target
Grip and Attack Execution
Exit

Figure 1. Stages of SE attack, sources [11].

4. Vishing

Voice phishing is a form of phone criminal attack carried out using social engineering with the use of telephone system to look at private personal and financial information; for the use of financial work it is also referred as "vishing" [12].

Vishing is short for "voice phishing", which involves defrauding people over the phone and enticing them to divulge sensitive information. In this definition of vishing, the attacker attempts to obtain the victim's data and use it for their own benefit—typically, for their own financial advantage [12].

Voice Cloning

Voice cloning is the process in which a computer is used to generate the speech of a real person, generating a replica of their specific and unique voice through AI. This technique has become popular thanks to advances in AI and ML, which allow the replication of a person's voice with surprising accuracy; this has allowed the creation of increasingly realistic synthetic voices, and it can have both positive and negative implications [13]. For this process in general, the programs are based on artificial neural networks, which are ML systems that can identify complex patterns in training data. To clone a voice, the program first needs to be trained with a large amount of voice data from the person to be cloned, the more data provided, the more accurate the cloning will be [14].

Once the program has been trained, it can generate new audio clips that sound like the person in question. This is achieved by manipulating voice parameters such as fundamental frequency, syllable duration and intonation [15].

The expressive speech cloning framework is a multi-speaker Text-to-Speech (TTS) model conditioned by speaker encodings and style aspects of speech. Style conditioning in expressive TTS models is usually done by learning a dictionary of latent style vectors called Global Style Tokens (GSTs). Although GSTs can learn meaningful latent codes when trained on a dataset with a large variation of utterances, they have been empirically found to provide limited style control when trained on a large multi-speaker dataset with mostly neutral prosody [16].

A high-level schematic of the expressive speech cloning framework with the main components being trained separately is shown in Figure 2 [17].

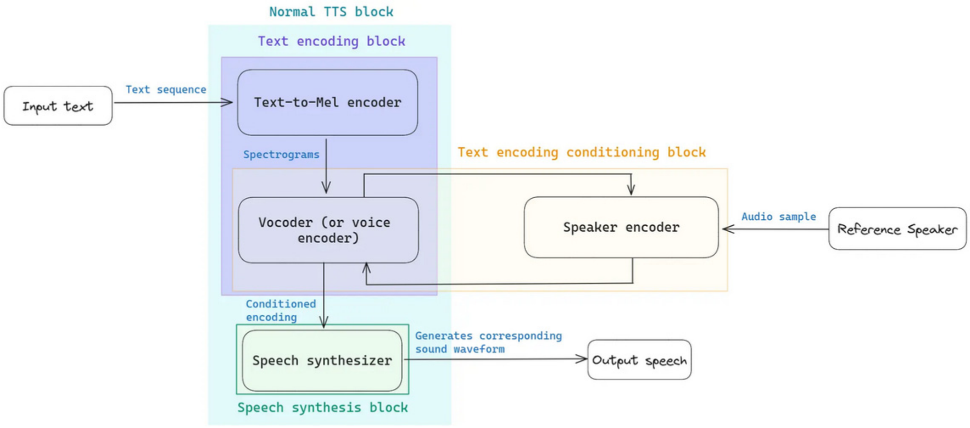

Figure 2. Expressive voice cloning model: Generic TTS model, sources [17].

5. Attack Methodology

To carry out the present project, first, the appropriate methodology is established to execute a vishing attack in a controlled environment. Also, the research participants are identified, where the target population of this study is composed of university students and other individuals.

Since this is a research project, and, to ensure the safety of the participants, informed consent was signed. In this document, participants give their consent for the use of their voice, as well as the information requested, including telephone numbers and the personal data of the persons on whom the analysis will be carried out.

The research is classified as qualitative, as it seeks to explore people's behavior through the simulation of three distinct scenarios. The first scenario involves pretending to be an agent of a financial or state entity to request financial and/or personal information. The second scenario involves simulating being a family member or friend in need of financial assistance. Finally, the third scenario consists of posing as a sales agent offering promotions, prizes or the buying and selling of goods or services.

Similarly, for this project, a conceptual model was used that includes the main aspects of SE, such as psychological factors, technical factors and sensitive information like bank data, personal information, names of family members, workplace, etc., all of which meet the established objectives as shown in Figure 3.

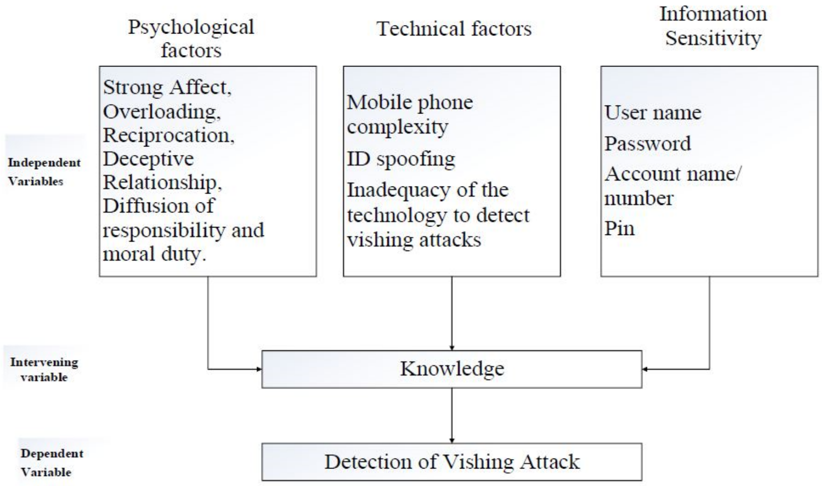

Figure 3. Conceptual model of social engineering factors for a vishing attack, based on [16].

5.1. Research Desing

For this study, a participatory action research (PAR) design was used, as its purpose is practical; to enable participants to respond to a problem using their resources (knowledge and reflection, intervention, action and resolution) [18]. A study of this nature is proposed with at least fifty participants for qualitative research, as it helps the researcher build and maintain a close relationship and thus improve the "open" relationship. A frank exchange of information [18] can help to mitigate some bias and validity.

For this purpose, an objective table is established in which six phases are established by which the type of attack and the possible results or actions to be obtained are classified and prioritized. These are the same ones that are described in Table 1, and they set the limits of the investigation, bearing in mind that the information to be received is of absolute confidentiality.

Table 1. Vishing Attack Classification and Action Phase.

Convention Phase		**Action Phase**
Specification	**Data**	**Operation**
Buy—Sell/Services	• Banks • Personals • Reference Person	• Bank transfer
Courier calls	• Family • Relationship	• Transfer (PayPal packages PayPal customs)
Banks operations	• Credit card (Photography Text)	
Home calls	• Access credentials • Application for purchase • Preparing special events	
Laboral calls	• Access credential (Physic Servers PC)	• Users
Personals calls	• Warranty • References • Contacts • Photo blackmail Extortion tactics	• Discover personal data • Conversations • Intimate photos • Threats, intimidation to obtain money or other benefit

5.2. Test Environment

For the testing environment, several regulatory guidelines were established to carry out the attacks; these are:

- The attacks will be carried out within a specified timeframe and with a limited scope regarding the dialogues used for the proposed scenarios.
- The use of phrases, jargon and words that do not discriminate against or insult the victim.
- The conversation should not exceed topics that are intimate, sexual, or of any other nature that might incite the victim to actions that could harm their physical appearance or psychological state.
- The attack will only focus on the scenarios presented, with the prior express authorization of the participants.
- The attacks will only take place on weekdays and at times that do not disrupt the victims' activities or at night.
- For the attacks, only the information provided by the participants will be used; no additional information may be used without prior authorization.
- The attacks will only be carried out on the phone numbers provided by the participants.

Additionally, it is important to highlight that a participant cannot be targeted in other scenarios, as a measure to prevent the victim from realizing that they are the object of a vishing attack.

5.3. Screenplay Composition

At this stage and as mentioned above, ChatGPT will be used; this AI tool will create the dialogues to be used in the calls and/or audio messages, using the cloned and synthesized voices of the participants so that at the time of the attack they are as realistic as possible.

Three main scenarios will be developed for this project based on three scenarios:

- Request for confidential information: generate content so that the victim accedes to giving sensitive information: mainly bank records, credit cards, etc.
- Request for financial assistance: generate content using SE to encourage the victim to agree to deliver financial assistance, using a known voice and with information obtained through social networks and from participants so that the victims offer no resistance.
- Request for information of a buy–sell service: generate content to obtain information about some good or service that the victim may possess, with the aim of trying to get more accurate information.

5.4. Procedure for Attack

Figure 4 shows the flowchart of the vishing attack procedure, which consists of three stages. In the first stage, content is generated using words that the victim commonly uses, entering them into ChatGPT to create the dialog. In the second stage, selected AI applications are used to clone and synthesize the voices, generating the voice message.

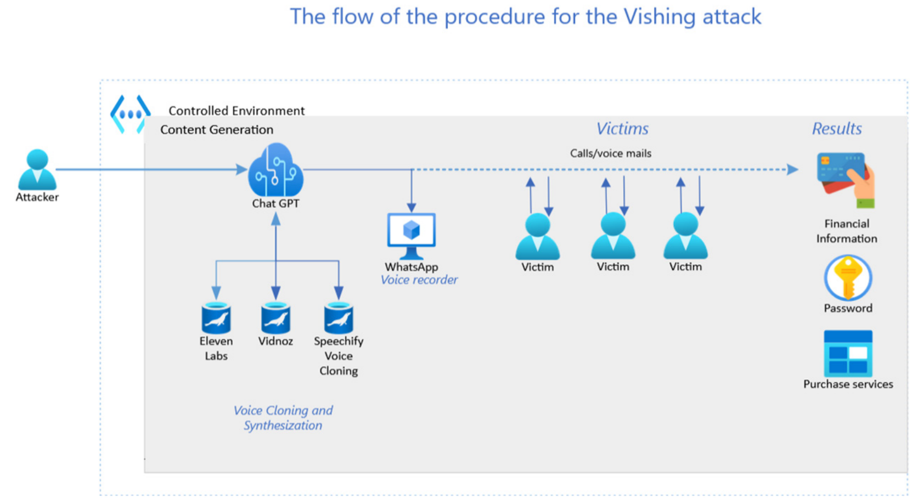

Figure 4. Flowchart of the procedure to be followed for vishing attacks.

This message is sent via WhatsApp or played during a call to the victim. Finally, in the third stage, the results are expected according to the scenarios proposed, i.e., if the victim agrees to provide the requested information or help.

6. Results

In the analysis of the attacks carried out, various types of responses were collected. The most common issue is the lack of response from the victims to audio messages and calls. Another observed reaction is, upon not recognizing the numbers from which the attacks were made, the victims chose to block those contacts.

Finally, to a lesser extent, some people responded to the audio messages and calls and even agreed to fulfill the requests made.

As can be seen in Table 2, the effectiveness percentage is calculated, and the total number of calls or audio messages sent using voices that were cloned from the different applications is counted, as are the number of blocked numbers, unanswered messages or successful attacks.

Table 2. Result of the vishing attack on the three applications.

Action/Application	Eleven Labs	Speechify Voice Cloning	Vidnoz
Attacks Performed	50	50	50
Successful	35	26	29
Locked	2	14	7
No Answer	13	10	14
% Success	70%	52%	58%

Based on the collected data, Figure 5 graphically shows the results of the attacks and the applications with which they were performed.

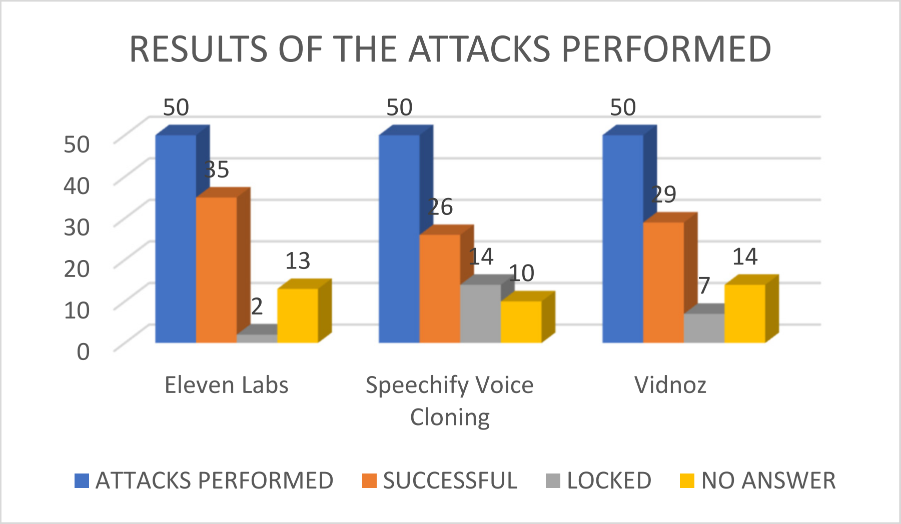

Figure 5. Overall results of vishing attacks for each application.

7. Conclusions and Lessons Learned

In conclusion, this document highlights that vishing represents a significant current global issue. The study shows that ChatGPT can generate both common vishing attacks and more specialized ones. Although it focuses on fifty individuals, this type of attack has proven effective and this methodology can easily be extrapolated to other categories of evasive attacks. Currently, the supervision of malicious content generated by large-scale language models, such as ChatGPT, is in its early stages, allowing attackers to exploit this resource to quickly implement new attacks.

Additionally, the process of voice generation and cloning can be significantly accelerated using the ChatGPT API. As for the AI applications for voice cloning and synthesis, all three applications had phonetic and word pronunciation errors. This resulted in improper phonetics, producing sounds in a spelled-out way or in a robotic manner. This problem occurred in some attacks when using unfamiliar terms for the speech synthesis applications.

This study makes three contributions:

First, an analysis of the phenomenon of social engineering is offered from theoretical and practical approaches. Through artificial intelligence, it is possible to generate content and replicate the voices of family members who are in emergency situations, thereby moving and appealing to people’s solidarity.

Second, it delves into the use of artificial intelligence applications that not only replicate and synthesize voices but also create dialogues that incorporate common words and jargon, enhancing the authenticity of the interaction.

Third, it is noteworthy that artificial intelligence has various capabilities, among which is the ability to transform texts into audible dialogues. This technology can replicate any tone of voice in such a way that it becomes difficult to distinguish between the generated voice and a real voice.

The first lesson learned is that there is a limit to the number of words that can be cloned; when exceeding 1500 words, applications tend to translate inaccurately. The second lesson relates to the number of times voices can be cloned and synthesized. Since this is a paid tool, the number of available clones is limited; in the case of the applications used, the limit is 500 clones per month. Additionally, it has been observed that some uncommon words are pronounced incorrectly by the application when converted to audio, even sounding robotic at times, as they sometimes just spell them out. Although this is an emerging field and there is a lack of information about it, voice cloning applications are still under development. However, their future looks promising, with encouraging growth prospects.

Author Contributions: Conceptualization, F.T. and C.T.; validation, D.G. and B.R.; writing—original draft preparation, F.T.; writing—review and editing, C.T., B.R. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rainie, L.; Anderson, J.; Connolly, J. Cyber-attacks likely to increase. *Digit. Life* **2019**, *2025*, 33–35.
2. Fischer, E.A. *Cybersecurity Issues and Challenges: In Brief*; Congressional Research Service: Washington, DC, USA, 2019; pp. 3–6.
3. Ali, B.; Awad, A. Cyber and physical security vulnerability assessment for IoT-based smart homes. *Sens. J.* **2018**, *18*, 817. [CrossRef] [PubMed]
4. Jang-Jaccard, J.; Nepal, A. A survey of emerging threats in cybersecurity. *J. Comput. Syst. Sci.* **2014**, *80*, 973–993. [CrossRef]
5. Ahlam, F.; Aymen, A. An Effective Blockchain-Based Defense Model for Organizations against Vishing Attacks. *Appl. Sci.* **2022**, *12*, 13020. [CrossRef]
6. Chen, Z.; Liu, B. Synthesis Lectures on Artificial Intelligence and Machine. In *Lifelong Machine Learning*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1–145.
7. Batta, M. Machine Learning Algorithms—A Review. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 381–390.
8. Aroyo, A.M.; Rea, F.; Sandini, G.; Sciutti, A. Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble? *IEEE Robot. Autom Lett.* **2018**, *3*, 3701–3708. [CrossRef]
9. Friedman, A.A.; West, D.M. *Privacy and Security in Cloud Computing*; Issues in Technology Innovation; The Center for Technology Innovation: Washington, DC, USA, 2018; pp. 1–13.
10. Goodall, J.; Lutters, W.; Komlodi, A. Developing Expertise for Network Intrusion Detection. *Inf. Technol. People* **2009**, *22*, 92–108. [CrossRef]
11. Mukkamala, S.; Sung, A.; Abraham, A. Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools. *Enhancing Comput. Secur. Smart Technol.* **2020**, 125–163.
12. Sun, N.; Zhang, J.; Rimba, P.; Gao, S.; Zhang, L.Y.; Xiang, Y. Data-driven cybersecurity incident prediction: A survey. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 1744–1772. [CrossRef]
13. Sigler, K. Crypto-jacking: How cyber-criminals are exploiting the crypto-currency boom. *Comput. Fraud Secur.* **2018**, *2018*, 12–14. [CrossRef]
14. Kalniņš, R.; Puriņš, J.; Alksnis, G. Security evaluation of wireless network access points. *Appl. Comput. Syst.* **2017**, *21*, 38–45. [CrossRef]
15. Pokrovskaia, N. Social engineering and digital technologies for the security of the social capital' development. In Proceedings of the International Conference of Quality Management, Transport and Information Security, St. Petersburg, Russia, 24–30 September 2017; pp. 16–19.

16. Janczewski, L.J.; Lingyan, R.F. Social Engineering-Based Attacks: Model and New Zealand Perspective. In Proceedings of the International Multiconference on Computer Science and Information Technology, Wisla, Poland, 18–20 October 2010; pp. 847–853.
17. Salahdine, F.; Kaabouch, N. Social Engineering Attacks: A Survey Future. *Future Internet* **2019**, *11*, 89. [CrossRef]
18. Rodríguez, M.E. La Investigación Acción Participativa Compleja como Transmétodo Rizomático Transcomplejo en la Transmodernidad. *Rev. Int. Form. Profesores* **2020**, *5*, e020026.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.