AI Generated Fake Audio as a New Threat to Information Security: Legal and Forensic Aspects

Elena I. Galyashina[a] and Vladimir D. Nikishin[b]

Department of Forensic Expertise, Kutafin Moscow State Law University, Sadovaya-Kudrunskaya Str., 9, Moscow, Russia

Keywords: Information Security, Deep Fake Audio Technology, Cyberattack, Voice Clone, Falsification of Evidence, Biometric Voice Data.

Abstract: A speech pattern synthesis is the classic task for AI which goal is to automate the process of text sounding with natural voice characteristic features. The corresponding sets of techniques are called voice cloning that could be relatively easy recognized by listeners as a machine speech. But in the past two years the new developments in the area of AI and deep generative networks have significantly improved the quality of speech synthesis and brought to life the technology of deep fake audio that appear rather authentic. Cyberattacks using convincing sounding models successfully emulating people's voices have escalated in number, frequency, and impact, drawing increased attention to the vulnerabilities of cyber systems and the need to increase their security. False audio content generated with neural networks has become difficult to be detected that causes a widespread distrust of audio (voice) evidence and brings a serious societal harm. In the face of this threat, there is a significant concern and interest among forensic speech researchers and the juristic public about the malicious implications of AI and risks of fake audio attacks for cybersecurity. This article describes some legal and forensic aspect of exposing fake voice audios generated with neural networks.

1 INTRODUCTION

In 2016, Adobe announced the VoCo tool which was able to simulate the voice of any person. In order to do this, it needed a 20-minute voice sample. A year later, the Canadian start-up Lyrebird launched a beta version of the service with which a neural network can be trained to simulate any voice by a one-minute record. In February 2018, scientists from the Chinese company Baidu published an article describing an approach to voice cloning that works on samples that are only 3.7 seconds long. Last year there was an implementation of speech synthesis with voice cloning Real-Time-Voice-Cloning.

The quality of the result was rather good with some signs of machine voice features but the development continues. With the advent of technology, such voice fakes were used for entertainment, but now it is a dangerous tool exploited by scammers. For example, in March 2019, the director of a British energy company was robbed of 220 thousand euros with his own hands. He sent the entire amount to the Hungarian company on behalf of the head who personally confirmed the task via fake audio (Maras and Alexandrou, 2019).

AI synthesised spech has become a powerful tool of influence and has gained a new strength in the digital age. It is extremely difficult to protect yourself from false information without innovative technologies that will help a person identify fake voices. The main threat of AI generated voice fakes is a devaluation of facts and evidence. If now you can hope for the reliability of these categories, then all sounding messages should be questioned as they could have been generated by a neural network in the future. Of course, it will be the most popular tool for influencing society. All you need is to program the neural network to create the necessary scenario with any person on the planet. Social networks quickly distribute fake materials. Is there any counteraction to audio deepfakes? Neural networks are already being

[a] https://orcid.org/0000-0001-8989-1003
[b] https://orcid.org/0000-0003-2819-8517

Galyashina, E. and Nikishin, V.
AI Generated Fake Audio as a New Threat to Information Security: Legal and Forensic Aspects.
DOI: 10.5220/0010616700003170
In *Proceedings of the International Scientific and Practical Conference on Computer and Information Security (INFSEC 2021)*, pages 17-21
ISBN: 978-989-758-531-9; ISSN: 2184-9862

trained to determine the veracity of the generated video content but cannot expose AI generated speech with natural voice characteristic features. In the face of this threat, there is a significant concern and interest among forensic speech researchers and the juristic public about the malicious implications of AI and risks of fake audio attacks for cybersecurity. In this work, we describe some legal and forensic aspect of exposing fake voice audios generated with neural networks.

2 METHODOLOGY OF RESEARCH

This study relies on the scholarly literature on neural network speech synthesis based on Russian language speech and on the system analysis of the AI generated audio fakes as a new challenge to the security sphere. Our search for news about Russian language-based audio deep fakes via Yandex, Rambler, Google search machines using keywords ‘Russian’, ‘voice cloning’, ‘audio fake’, ‘fake voice’ and the corresponding plural forms was unsuccessful. We did not find any case discussed in mass media with the topic of the fraud conducted with the AI generated Russian speaking fake voice. However, it does not mean that this issue does not exist, as a number of other languages-based spoofs had been reviewed (Chesney, Citron, 2019). Our research dataset includes voice cloning monographs, articles discussing the advantages and disadvantages of the described speech synthesised technologies, the benefits and threats of fake voice usage, and ways of combating them. The authors used a set of methods, including a retrospective analysis of scientific literature assessing the processes and phenomena in the area under study, comparative legal and logical analysis, extrapolation methods, case and scenario analysis to simulate information security-threatening situations.

Based on the conducted research, the main types of ideological and social threats to information (worldview) security caused by deep audio fakes were systematized and classified. Recommendations for updating existing Russian Federation laws described below were formulated to uphold principles of privacy in the processing of voice samples as personal data.

2.1 What is Voice Cloning with AI Technology

Voice cloning is a deep-learning algorithm that takes in voice recordings of an individual and is able to synthesize such a voice into one that is very similar to the original voice. There are numerous apps similar to deepfakes, such as LyreBird, iSpeech, and CereVoice Me which give the public access to such technology.

The algorithm simply needs at most a couple of minutes of audio recordings in order to produce a voice that is similar and it will also take in any text and will read it out loud. AI generated synthetic voices are audibly indistinguishable from the original audio samples (Kinnunen, etc., 2017).

You need to have a number of matched audio recordings and texts to train the system. Short samples of spoken speech would be enough to create sounds which are similar to human voice.

Although this application is still in the developmental stage, it is rapidly developing as big technology corporations, such as Google and Amazon, are investing huge amounts of money in their development.

2.2 Voice Cloning Practical Applications

There is a number of beneficial applications of naturally sounding audio samples of neural networks generated synthetic speech; they include voice greetings, audiobooks, cloning parents' voices for reading fairy tales, audio and video training courses, promotional videos and advertising audio ads, voice bots, personalized voice assistants, etc. Law enforcement officers often have to impersonate another person in order not to be recognized.

It is obvious that AI technologies can also be used for criminal purposes: fraud, prank calls, falsification.

Therefore, it is important to develop tools to prevent illegal use of voice cloning methods and identify gaps in the system of existing Russian Federation laws institutions to prevent risks of the emergence and spread of criminal and fake audios in media sphere.

2.3 Deep Fake Audios as a Treat to Information Security

The modern information space is a breeding ground for cyber-attackers who spread malicious, destructive, false and criminal information via synthesised speech that poses a threat to the personal security of Internet users.

Many websites and communities in social networks can use deep fake audios to glorify fascism and nationalism, xenophobia, religious extremism, promote ideological terrorism, popularize ‘suicide clubs’; promote drugs, underground culture, the cult of violence and cruelty; incite hatred and enmity, humiliate people on the grounds of their social affiliation; harassment and bullying, slander and insults, spread fakes, child pornography and other prohibited information (Galyashina and Nikishin, 2020).

To expose voice fakes, prosecute the attacker, and protect innocent people whose voices are illegally used to commit speech offenses, it is necessary to equip law enforcement agencies with an objective criteria that allow them to detect, identify and prevent the illegal use of other people's voices to commit crimes in the digital media environment (Galyashina, 2021).

3 DISCUSSION OF RESEARCH RESULTS

Our research showed that a significant part of the disseminated criminal information is expressed verbally. Fake audios are designed mainly for the average user’s perception. The difference between an AI generated voice clone and the real speaker sample can be detected by professionally trained expert-listener. The main distinctive features are reflected in the prosodic structure of speech as well as in discursive and intellectual speech skills that are difficult to be imitated (Ladd, 1996). Thus, the first theoretical conclusion is that special phonetic-linguistic knowledge and integrated approach to speaker identification is needed in order to expose faked audio and detect the illegally used cloned voices (Galyashina, 2015).

The second conclusion is connected with the determination of security risks of AI generated speech as these issues have not received sufficient attention from law enforcements. It became obvious that the processes taking place in the information space lead to the emergence of new speech communication phenomena that are represented by unpermitted use of personal voice and speech ideotype that is also not reflected in the legislation. We propose to supplement the current legislation of Article 151.1. of the Civil Code of the Russian Federation with the norm on the protection of not only the image but also the voice of a citizen (Article 151.1). The publication and further use of recordings of a citizen's voice and speech are allowed only with the consent of this citizen. After the death of a citizen, their voice samples can only be used with the consent of the children and a spouse or with the consent of parents. Such consent is not required in cases where: 1) the use of the voice sample is carried out in the state, public or other public interests; 2) the voice sample of a citizen is obtained when audio recording is carried out in places open to the public, or at public events (meetings, congresses, conferences, concerts, performances, sports competitions and similar events), except in cases where such a voice sample is the main object of use; 3) a citizen’s voice sample is recorded for a fee.

The audio recordings of the citizen’s voice, obtained or used in violation of the above-mentioned conditions, shall be removed on the basis of a court decision.

If a recording with a citizen’s voice sample obtained or used in violation of the above-mentioned conditions is distributed on the Internet, the citizen has the right to demand the removal of this voice recording as well as the prohibition of its further distribution.

It is worth noting that until now, there has been no holistic approach to the analysis of fake audio threats arising in the digital media environment, the development of measures to prevent and counteract the spread of destructive information produced by synthetic voice audibly undistinguished from the voice of the real person.

All of this leads to the problem of qualitative changes in the voice during the implementation of measures for the comprehensive protection of speech and voice as biometric data.

This is quite a difficult task, since each person's voice is individual and recognizable (Yarmey, 2001). Moreover, the trained auditory perception helps to identify the most subtle shades of the speech signal. The average human hearing is not accurate in detecting the signs of artificiality or naturalness of AI generated speech. Therefore, in order to solve the problem of fake voice detecting with the preservation the individual features of a natural sound according to a given voice sample, it is necessary to dwell in more detail on the concept of AI generated speech and its main features, to systemize factors that determine audial differentiation of real and faked voices.

The main factor is associated with the acoustic-phonetic structure, that is, with the prosodic similarity of the sound of the AI synthesized speech signal with natural speech. Speech signals can be considered as a physical implementation of a complex hierarchically organized system of linguistic rules, by which the properties of the speech signal are limited by the acoustics of the vocal tract. The acoustic-phonetic structure of natural speech reflects these physical limitations. Synthesized signals are simplified signals, the sound of which is determined only by a limited subset of the set of acoustic parameters that are used to transmit phonetic information in natural speech. In addition, the acoustic parameters used to represent text in synthesized speech are significantly stylized and cannot convey the phonetic context in comparison with natural speech (Piosoni, 1982). Thus, errors in the text-to-speech system mainly occur when calculating and reproducing the suprasegment structure of spontaneous speech. It is possible to adequately identify the cloned voice based on the analysis of the prosodic characteristics of speech analyzed in the aggregate with speech skills of the speaker, his speech competence (Nolan, 1983).

It was proved that it is necessary to check the sound evidence for its possible forgery using artificial intelligence in the course of diagnosing the products of criminogenic speech actions. This might be carried out by assigning and conducting a comprehensive computer-technical, voice and speech comparative examination with voice samples of a person whose voice could potentially be cloned. Thus, a person can be protected from illegal and unjustified prosecution for actions that pose a public danger to information security. The effective forensic support of law enforcement activities in the fight against information threats and the prevention of speech offenses can be carried out.

4 CONCLUSIONS

Cyberattacks using AI generated sounding texts with imitated personal voices have increased in number, frequency, and impact, drawing increased attention to the vulnerabilities of cyber systems and the need to increase their security. False audio content generated with neural networks has become difficult to be audibly detected. It causes a widespread distrust of voice evidence and brings a serious societal harm.

In the face of this threat, there is a significant concern and interest among forensic speech researchers and the juristic public about the malicious implications of AI and risks of fake audio attacks for cybersecurity especially for the Russian language internet sector.

This research enabled us to articulate theoretical problems associated with security of personal voice data and reveal some legal and forensic issues of exposing fake voice audios generated with neural networks.

It has become obvious that it is necessary to check the sound evidence for its possible forgery using AI techniques in the course of forensic diagnosing the products of criminogenic speech actions. It was determined that a success of adequate identification of the cloned voice depends on effectiveness of integrated analysis of prosodic characteristics of speech analyzed together with speech skills of the speaker and his speech competence.

In order to expose voice fakes and prosecute the attacker to protect innocent people whose voices are illegally used to commit speech offenses, it is necessary to equip law enforcement agencies with an legal remedies and objective criteria that allow them to detect, identify and prevent the illegal use of other people's voices to commit crimes in the digital media sphere.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR, project number 20-011-00190.

REFERENCES

Chesney, B., Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107: 1753-1820.

Galyashina, E., Nikishin, V. (2020). Media security of megascience projects: legal experts training. In *J. Phys.: Conf.* Ser.1685 012004 doi:10.1088/1742-6596/1685/1/012004.

Galyashina, E.I. (2021). Law and linguistics in the aspect of worldview internet security. In *The European Proceedings of Social & Behavioural Sciences*. Vol. 102. NININS 2020. doi: 10.15405/epsbs.2021.02.02.37

Galyashina, E.I., (2015). Linguistic analysis in the speaker identification systems: Integrated complex examination approach based on forensic science technology. In *International Conference on Computational Linguistics and Intellectual Technologies*, Dialogue 2015, 27-30 May 2015. 1(14):156-168.

Kinnunen, T., Sahidullah, Md, Delgado H, Todisco, M., Evans, N., Yamagishi, J., and Lee, K. (2017). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In PROC. *Interspeech, Annual Conf. of the Int. Speech Comm. Assoc*., 2–6.

Ladd, D.R. (1996). *International Phonology*. Cambridge: Cambridge University Press.

Maras, M. H., Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *International Journal of Evidence & Proof*. 23(3): 255–262.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambrige: CambrigeUniversity Press.

Piosoni, D. B., 1982. Perception of speech: The human Listener as a cognitive interface. *Speech Tecnol*. 1: 10-23.

Yarmey, A. D. (2001). Earwitness descriptions and speaker identification. *Forensic linguistics*, 8(1): 113-122.