BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Assessing the Human Ability to Recognize Synthetic Speech

Daniel Prudký*

Abstract
This work responds to the development of artificial intelligence and its potential misuse in the field of cybersecurity. It aims to test and evaluate the human ability to recognize a subset of synthetic speech, called voice deepfake. This paper describes an experiment in which we communicated with respondents using voice messages. We presented respondents with a cover story about testing the user-friendliness of voice messages while secretly sending them a pre-prepared deepfake recording during the conversation and looked at things like their reactions, their knowledge of deepfakes, or how many respondents correctly identified which message was manipulated. The results of the work showed that none of the respondents reacted in any way to the fraudulent deepfake message and only one retrospectively admitted to noticing something specific. On the other hand, a voicemail message that contained a deepfake was correctly identified by 96.8% of respondents after the experiment. Thus, the results show that although the deepfake recording was clearly identifiable among others, no one reacted to it. And so the whole thesis says that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they are not expecting them.

*xprudk08@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Motivation

Artificial intelligence is evolving at a tremendous speed, bringing us a huge number of possibilities and things that can make our lives easier. It is used in many fields including healthcare, aviation and security. But it is also associated with threats of all types, and deepfakes are one of them.

Deepfakes are media created by artificial intelligence, specifically by using deep neural networks through deep learning methods. In their production, artificial intelligence merges, combines, replaces, or overlays elements of media to create new false representations of things that never happened.

Nowadays, these fake media are reaching a stage where they are not even recognizable by machines, let alone humans who may not even be aware of the existence of such threats in today's digital world. Moreover, within audio, it is no longer just about English models. A lot of multi-language tools for creating voice deepfakes are being developed, and they can appear in almost any language.

There are many attack scenarios in which deepfakes have been used. For example, they could be attacks targeting specific individuals or institutions in the form of vishing or widespread disinformation to spread propaganda, and so on. People should be able to defend themselves against this spread of fraudulent information and media, they should know how to verify such things and how to deal with them. But we don't know if we will ever be able to do that.

Therefore, the goal of this work is to test and then evaluate the human ability to recognize synthetic speech. There have been several attempts to assess whether people can distinguish a deepfake from a real one. However, these experiments first introduced participants to the deepfake problem before presenting them with a medium and asking them to identify it. Their results are quite variable and vary mainly depending on the methodology. Voice deepfakes have been tested, for example, in a survey by Müller et al. [1] who report that the accuracy of identifying a deepfake and a genuine recording is 80%.

2. Experiment

We were inspired by the *Authorizing Card Payments with PINs* [2] experiment, in which the authors replicated an actual attack hidden behind the cover story and we transferred this idea to the voice deepfake field.

Created an experiment in which respondents were confronted with voice deepfakes without being told about them and observe whether they recognize it as a deepfake or at least notice something strange.

We hid the whole experiment behind a cover story and under this curtain, we asked the participants simple questions, in the form of facts for the game Two Truths One Lie, using voice messages in a classic chat. Similarly, respondents answered us with voice messages, this was to support the cover story presented. In between the real voice messages, we also sent one pre-prepared deepfake recording of my voice and then watched to see if the participants recognized the deepfake or at least noticed anything odd. At the end of the experiment, we sent them a questionnaire asking about their knowledge of and attitude towards deepfake, revealed the cover story, and asked if they could identify the deepfake message in the chat. A diagram of the experiment flow is shown in poster Figure 1, using [3].

3. Results

The results of the experiment are quite clear. During the conversation, none of the 31 respondents reacted in any way to the fraudulent deepfake message. When evaluating the questionnaire to ask if they noticed anything, we agreed that only one person noticed and described something specific to deepfakes, as shown in poster Figure 2. On the other hand, when identifying a deepfake message, 96.8% of respondents correctly identified it, shown in poster Figure 3. Thus, the results show that although the deepfake recording was clearly identifiable among others, no one reacted in any way to it.

The experiment also found that 83.9% of respondents had at least heard of deepfakes, primarily from social media, educational videos or simply stumbling across them online, this is shown in poster Figure 4. The survey also asked respondents how confident they were that they would detect a deepfake. We asked them this question twice, at the beginning and at the end of the questionnaire, and had them rate their view on a scale of 1 to 5. The average was 2.29 at the beginning and after the experiment was revealed, at the end of the questionnaire, the average increased to 2.94. The value increased mainly with younger respondents.

At the end of the thesis, we describe our own suggestions for training people in recognizing deepfakes and spreading awareness about them. These suggestions are based on the results of our and related work, or for example, the recommendations of the FBI. We propose here a website that would include interactive training demos of different training approaches, suggestions for tools to detect deepfakes, as well as links to support victims of this technology.

4. Conclusions

The main contributions of this work can be described as follows:

- The work shows that the human ability to recognize voice deepfakes is not at such a level that we can trust it, and it is very difficult for people to distinguish between real and fake recordings.
- It suggests a platform for working with deepfakes and training its recognition, along with recommendations for useful tools or expert help.
- Displays that people's awareness of deepfakes is quite high, especially from informative videos, articles and similar.
- Shows that the production and quality of voice deepfakes in non-English languages is not a problem for artificial intelligence. And at the same time, no one needs professional expertise to create one.

Work has shown that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they are not expecting them.

Acknowledgements

I would like to thank my supervisor Ing. Anton Firc for his insightful suggestions for the work, consultation and other help in the development of the experiment and thesis.

References

[1] 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*, pages 85–91, October 2022. arXiv:2107.09667 [cs, eess].

[2] Vashek Matyas, Jan Krhovjak, Marek Kumpost, and Daniel Cvrcek. Authorizing card payments with pins. *Computer*, 41:64 – 68, 03 2008.

[3] Free Vectors, Stock Photos & PSD Downloads.