N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation

Daniel Prudký[1], Anton Firc[2], Kamil Malinka[3]

Abstract: This work assesses the human ability to recognize synthetic speech (deepfake). This paper describes an experiment in which we communicated with respondents using voice messages. We presented the 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. We examined their reactions, knowledge of deepfakes, or how many could correctly identify which message was deepfake. The results show 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 83.9% of respondents after revealing the nature of the experiment. Thus, the results show that although the deepfake recording was clearly identifiable among others, no one reacted to it. In summary, we show 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 do not expect them.

Keywords: deepfake, synthetic speech, artificial intelligence, cybersecurity, deepfake detection

1 Introduction

Mirsky and Lee [ML21] define a deepfake simply as a *"Believable media generated by a deep neural network."* A more extensive definition says it is media created by artificial intelligence (AI), specifically using deep neural networks through deep learning (DL) methods. In their production, artificial intelligence merges combines, replaces, or overlays features of the media to create new fake representations of things that never happened. This media can be practically unnoticeable from authentic ones. Deepfake technology brings many benefits, it can be used for entertainment, but it can also be used for revenge porn, bullying, spreading fake news, political sabotage and more [FM22, FMH23, We19].

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. Many multi-language tools for creating voice deepfakes are being developed, and they can appear in almost any language.

There have been 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

[1] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, xprudk08@stud.fit.vut.cz
[2] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, ifirc@fit.vut.cz
[3] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, malinka@fit.vut.cz

or widespread disinformation to spread propaganda, and so on. People should be able to defend themselves against this spread of fraudulent information and media, and 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.

A recently widespread method is vishing, derived from the two words defining it: voice and phishing. It is a version of phishing in which identity theft is carried out using voice devices such as the telephone, voice assistant, etc. Its use is described by Firc et al. [FMH23]. The authors point out that one such attack happened in 2019 when a fraudster using deepfakes created a transaction of almost $250,000. The CEO of an energy company thought he was talking to his boss on the phone, and when the caller asked him for an urgent transfer of this money, the victim did not hesitate and sent the money, believing he was completing a task from his boss. There are many cases like this today. The same article says that vishing was reported by 69% of companies in 2021, a big increase from 2020, when 54% of companies reported it. Spoofing is also very often associated with this scam, giving the scam much more credibility. For example, the fraudster can call the victim from the real phone number of the person they are playing. Spoofing can also be used with the phone number of a bank, the police, etc. In this way, the attackers try to exert authority on the victim, who is then more likely to disclose the required information in fear.

Therefore, this work aims to assess 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 exposing them to deepfakes. Their results are quite variable and vary mainly depending on the methodology. For example, voice deepfakes have been tested in a survey by Müller et al. [MPW22], who report that the accuracy of identifying a deepfake and a genuine recording is 80%. In contrast, our approach first exposes the respondents to deepfakes and then asks if they noticed anything unusual or if they can identify the deepfake set among other sets.

The whole experiment is hidden behind a cover story of testing the usability of voice messaging. Respondents play the game *Two Truths One Lie*. They receive 5 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. 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 work described in this paper results from a previously completed bachelor's thesis [Pr23].
The main contributions might be stated as follows:

- We assess the human ability to recognize deepfakes in the Czech language.
- We show that people cannot distinguish real and deepfake speech in common conversation.

2 Related work

The research that deals with detecting voice deepfakes by humans is described in the scientific article by Müller et al. [MPW22]. The authors focused on the ratio of the success rate of deepfakes detection by humans and artificial intelligence. The experiment compared human and machine detection capabilities, using a game-based challenge in which the respondent always played a recording and then determined whether it was fake or real. They made the same decision with machine learning models. For the experiment, the authors used the ASVspoof 2019 dataset, created for the ASVspoof 2019 Challenge, which aims to test Automatic Speaker Verification (ASV) systems resistant to spoofing attacks. Through the experiment, the authors found that the human ability to recognize deepfake and real recordings reaches 80%. Further, the experiment found that recordings created using TTS fooled humans much more than voice-conversion or waveform concatenation systems. The authors believe it could be because it used GAN as the waveform generator. Other interesting results are that native speakers handled recognition better than non-native speakers. At the same time, the level of IT experience did not affect performance, and people's ability to recognize deepfakes decreases with age. It is also interesting to note that people learned very quickly, and as the article says, after the first ten rounds, the success rate improved from 67% to 80%, but promptly stabilised at those levels and did not improve.

Other works focus mainly on deepfakes in the form of images and videos. The success rate of respondents in these experiments varies depending on the methodology and dataset used in the experiment itself. For experiments with deepfake images, success rates for better deepfakes and deepfakes with poorer image quality range roughly between 58-70%, while for poorer deepfakes, respondents have been close to the 90% success rate [Gr21, Go23, Ro19]. In terms of videos, success rates again depended on the quality, and for better quality and harder-to-detect deepfakes, the success rate dropped to the 20% mark, while for lower quality fakes, the success rate reached over 80% [KM20, Gr22, Ta21]. A paper by Tahir et al. [Ta21] describes the training of people in identifying deepfakes, and through a more sophisticated analysis of people's behaviour in identifying deepfakes and other parameters, they were able to develop training procedures that increased the success rate of the trained group by 33%. On average, we're talking about a deepfakes detection success rate of roughly 60-65%, depending on multiple factors.

In all former experiments, the participants knew they were exposed to deepfakes and, therefore, might have targeted it. This is where our research differs very fundamentally from others. Another significant difference is the execution in the Czech language.

3 Experiment

The design of the experiment is inspired by Matyáš et al. [Ma08], who propose using a cover story. Moreover, unlike other works, respondents do not know they must reveal deepfakes. 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 was conducted in the Czech Republic; therefore, all communication was in the Czech language. This is also related to producing deepfake voices in the Czech language. While most models and tools are suited for the English language, we show the feasibility of other languages that require individual approaches to training and using the speech synthesis models.

3.1 Research questions

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

RQ1: Are humans able to identify deepfake recording during casual conversation?
We are interested in whether people notice during the interview that they have received a computer-generated recording and how they react to it.

RQ2: Are humans able to detect a deepfake recording among genuine ones?
We want to determine whether people can retrospectively identify which of the messages in a conversation was a deepfake recording.

RQ3: What is people's awareness of deepfake technology?
Given that victim knowledge of deepfakes is critical to detecting these scams, we are interested in how many had heard of the technology or were actively interested in it and what is their experience with deepfakes.

3.2 Experiment execution

To synthesize the deepfake set, we use YourTTS [Ca22] in the voice conversion setting. This decision was motivated mainly by the easy access to the tool via a demo on Google Colab and the fact that we possess a version with a trained model in the Czech language. After synthesis, we improved the quality of this set using post-processing. We removed the noise added during creation and smoothed out the frayed phonemes by cutting out the part of the recording where the phonemes resonated. We also adjusted the pitch of the voice. The test run revealed a significant difference in background noise between real (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 real utterances.

Next, we performed a quality assessment of the synthesized set. We used an evaluation inspired by the *Mean Opinion Score (MOS)* subjective listening test method described by Loizou [Lo11]. We played the recording to 12 experts working with deepfakes regularly. Therefore, we expect their knowledge about deepfake recordings' qualities. Each expert rated the quality on a scale of 1 (poor) to 5 (excellent). The final mean score was 3.0; therefore, the recording qualitatively corresponds to the rating "*Fair*".

As previously mentioned, 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 starts 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 important for us to get them 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 respondents listen to these sets and reply with voice messages as well. This way, we send five sets (voice messages), including one pre-prepared deepfake set. If any respondent raises any suspicion or questions about the deepfake set, we refer them directly to the questionnaire. 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. When creating the questionnaire, it was important to determine the correct sequence of questions so that the questions could not influence those yet to follow.

4 Results

During the experiment, we collected 31 responses. In terms of gender, 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. In focus on the field of work, IT has the highest representation, with 41.9% of respondents. The next common field is education with 19.4%, law and healthcare with 6.5%, and other fields like machinery, marketing, military, art, etc.

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 being revealed the true nature of the experiment. This gives us a deepfake detection success rate of 3.2%. 13 respondents mentioned a lower quality of this recording; however, we cannot consider this as 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, stating that they considered the lower quality to be normal. These results are summarized in Tab. 1.

Reaction during conversation	
Reacted	0%
Described unnatural things from the conversation	
Poorer audio quality	41.9%
Deepfake sign	3.2%

Tab. 1: RQ1 summary.

RQ2: Are humans able to detect a deepfake recording among genuine ones?
After revealing that one of the sets is a deepfake, 83.9% of all respondents correctly identified this set. Respondents who marked the deepfake set, along with it other options, are not counted as successful. Counting these responses as successful would result in 96.8% respondents identifying the deepfake set.

54.8% of respondents justify selecting the deepfake set because it was different to others. The second most-stated reason was the lower quality compared to real recordings, as mentioned by 29% of respondents. Finally, the third most-stated reason is the presence of typical deepfake artefacts, mentioned by 22.6% of respondents. Some respondents gave a combination of stated reasons. These results are summarized in Tab. 2.

Identify deepfake set	
Marked	96.8%
Correctly identify	83.9%
Justification for identification	
Different from the others	54.8%
Lower quality than others	29%
Deepfake sign	22.6%

Tab. 2: RQ2 summary.

RQ3: What is people's awareness of deepfake technology?
Respondents had a choice of three options, 16.1% of respondents answered, "I've never heard of deepfakes", 64.5% answered, "I've heard of deepfakes before", and 19.4% answered, "I'm actively interested in deepfakes". Where they heard about deepfakes is variable but can still be classified into several groups. More than a quarter of people (25.8%) said that they heard about deepfakes on social media, mainly in some informative videos, articles, etc. One respondent said to encounter deepfake videos of politicians on TikTok. Consistently, 19.4% 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. In summary, 83.9% of the participants have at least heard of deepfakes, mainly from social media and informative videos.

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 51.6% of respondents increased this value, while 45.2% did not

Heard of deepfakes	
Heard of them	64.5%
Actively interested	19.4%
Never heard of them	16.1%
Where they heard about them	
Social media	25.8%
Internet	19.4%
Not specify	19.4%
Create them themselves	19.4%
Never heard of them	16.1%

Tab. 3: RQ3 summary.

change it, and only 3.2% decreased it. Younger respondents mainly increased the value of their certainty.

Additionally, after completing the experiment, 74.2% of the respondents said they were surprised by the quality of today's voice deepfake in the Czech language.

4.1 Limitations

The major problem of the experiment was the quality of the recordings because of the artificial noise in the background. And although when we played the recordings back (on iPhone 11), the noise was minimal, and we could understand everything without any problems, many people reached back saying that the quality of the recordings was really bad mainly because of the noise. We suppose it depends on the device on which the respondent listened to the recordings; some devices can reduce the noise, while others can't. Poor quality and noise was also the most common thing that respondents identified as odd about the conversation. In total, 13 respondents mentioned the lowered quality.

4.2 Results discussion

Related work evaluating human ability often reports more than 60% success rate. The success rate of deepfake detection in our scenario is 3.2%, 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 did know they were presented with deepfakes, the success rate of 83.9% is comparable to other research in this field.

These results give an interesting observation. During the conversation, no one responded to deepfakes, but when directly asked to identify the deepfake set, almost every respondent correctly identified it. Many respondents admitted to us that they didn't notice anything on the first listen. Still, when they listened a second time and focused on finding the computer-generated voice, they were immediately sure which one it was. There may

be several reasons for this, but we lean towards something similar to a psychological phenomenon called *The Monkey Business Illusion* [SC10], 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 right answers and therefore ignored the difference in the voice recordings. However, when we told them to focus on quality and find the deepfake, they detected it easily. These results thus demonstrate how crucial role the knowledge of deepfakes plays in their correct identification and that the education of the broad public on this topic is inevitable.

5 Conclusions

This 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. The human ability to detect deepfakes is largely influenced 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 any knowledge of deepfakes cannot reliably identify deepfake recordings in conversation. Combined with the Czech language, we show this problem is general and poses a significant threat to society. 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 real sets. It is thus important to address these imperfections in future and assess what role the audio quality play in the detection process.

Acknowledgements

This work was supported by Fakulta Informačních Technologií, Vysoké Učení Technické v Brně [FIT-S-23-8151].

References

[Ca22] Casanova, Edresson; Weber, Julian; Shulby, Christopher; Junior, Arnaldo Candido; Gölge, Eren; Ponti, Moacir Antonelli: , YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone, February 2022. arXiv:2112.02418 [cs, eess].

[FM22] Firc, Anton; Malinka, Kamil: 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, Association for Computing Machinery, New York, NY, USA, p. 1646–1655, 2022.

[FMH23] Firc, Anton; Malinka, Kamil; Hanáček, Petr: Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. Heliyon, 9(4):e15090, April 2023.

[Go23] Godage, Sankini Rancha; Lovasdaly, Froy; Venkatesh, Sushma; Raja, Kiran; Ramachandra, Raghavendra; Busch, Christoph: Analyzing Human Observer Ability in Morphing Attack Detection -Where Do We Stand? IEEE Transactions on Technology and Society, pp. 1–1, 2023.

[Gr21] Groh, Matthew; Epstein, Ziv; Obradovich, Nick; Cebrian, Manuel; Rahwan, Iyad: Human detection of machine-manipulated media. Communications of the ACM, 64(10):40–47, October 2021.

[Gr22] Groh, Matthew; Epstein, Ziv; Firestone, Chaz; Picard, Rosalind: Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1):e2110013119, 2022.

[KM20] Korshunov, Pavel; Marcel, Sébastien: , Deepfake detection: humans vs. machines, September 2020. arXiv:2009.03155 [cs, eess].

[Lo11] Loizou, Philipos C: Speech quality assessment. Multimedia analysis, processing and communications, pp. 623–654, 2011.

[Ma08] Matyas, Vashek; Krhovjak, Jan; Kumpost, Marek; Cvrcek, Daniel: Authorizing Card Payments with PINs. Computer, 41:64 – 68, 03 2008.

[ML21] Mirsky, Yisroel; Lee, Wenke: The Creation and Detection of Deepfakes. ACM Computing Surveys, 54(1):1–41, January 2021.

[MPW22] Müller, Nicolas M.; Pizzi, Karla; Williams, Jennifer: Human Perception of Audio Deepfakes. In: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. pp. 85–91, October 2022. arXiv:2107.09667 [cs, eess].

[Pr23] Prudký, Daniel: 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.

[Ro19] Rossler, Andreas; Cozzolino, Davide; Verdoliva, Luisa; Riess, Christian; Thies, Justus; Nießner, Matthias: , FaceForensics++: Learning to Detect Manipulated Facial Images, August 2019. arXiv:1901.08971 [cs].

[SC10] Simons, Daniel J; Chabris, Christopher F: The monkey business illusion. Cognition, 119(1):23–32, 2010.

[Ta21] Tahir, Rashid; Batool, Brishna; Jamshed, Hira; Jameel, Mahnoor; Anwar, Mubashir; Ahmed, Faizan; Zaffar, Muhammad Adeel; Zaffar, Muhammad Fareed: Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, May 2021.

[We19] Westerlund, Mika: The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9:40–53, 11/2019 2019.