Malinka *et al.*
EURASIP Journal on Image and Video Processing (2024) 2024:24
https://doi.org/10.1186/s13640-024-00641-4

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.

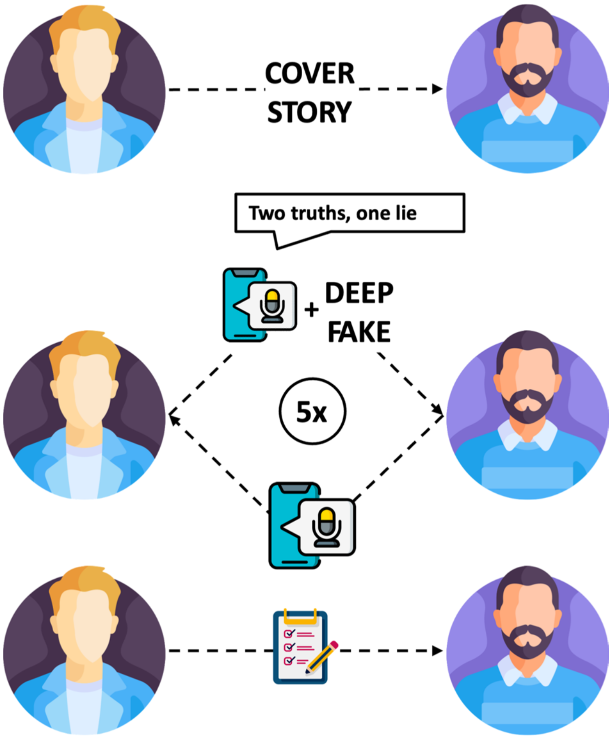


Fig. 1 Flowchart describing the course of the experiment

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.[7][8] 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

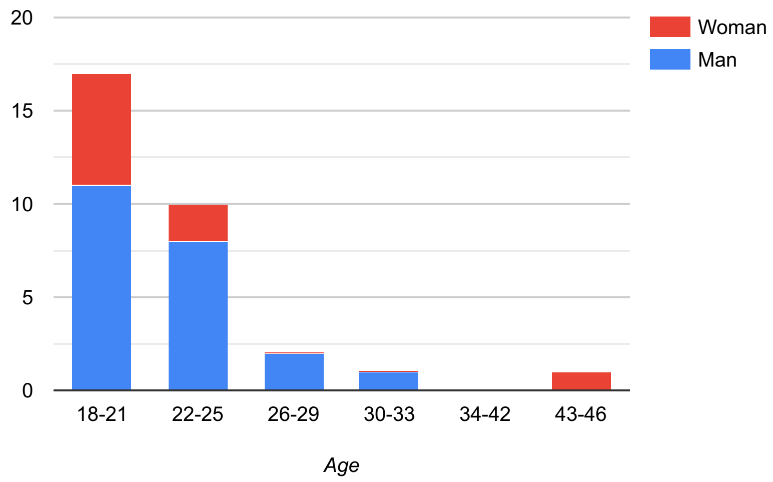


Fig. 2 Age of respondents with a look at the gender ratio in five age groups

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.

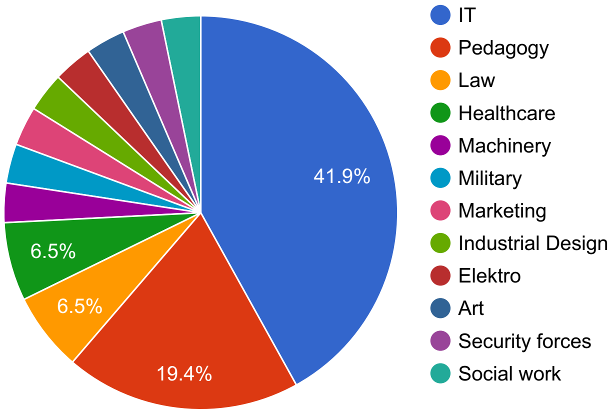


Fig. 3 Proportions of fields in which respondents work

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.

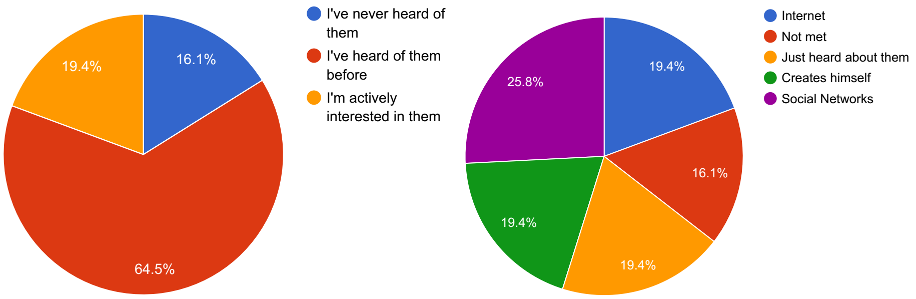
4.1 Experiment one: influence of the prior information

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,

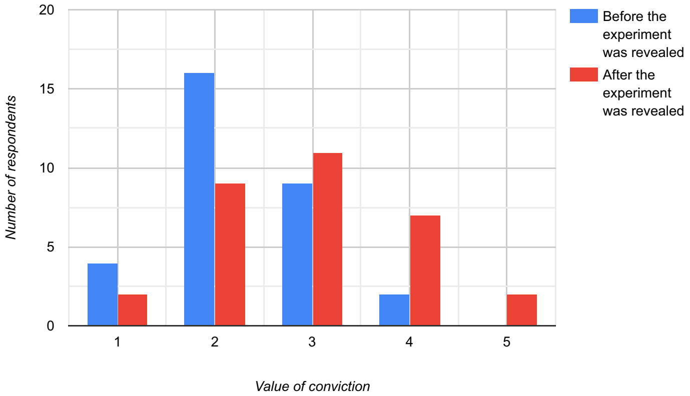


Fig. 5 Responses to the question of how confident respondents are in detecting a deepfake, quantified by a number of respondents

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

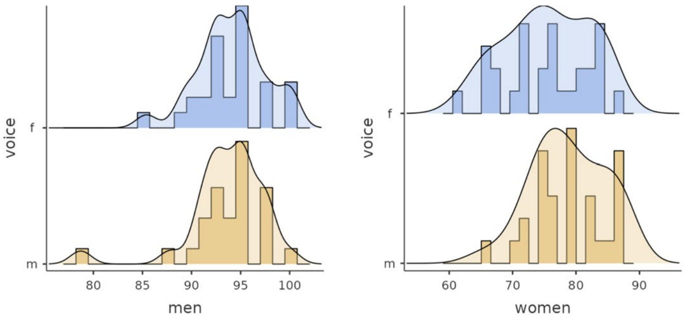


Fig. 6 Plots depicting the accuracy of deepfake detection by gender: Men's accuracy is shown on the left, and women's on the right. The X-axis indicates the percentage of correctly identified deepfakes, while dual Y-axis show the volume of accurately labelled recordings. The graphs employ orange (m) 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.

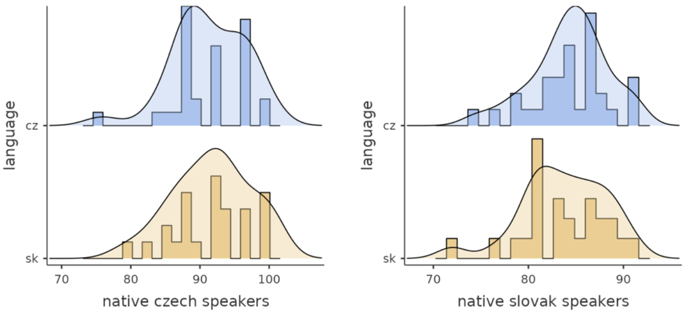


Fig. 7 Plots illustrating the proficiency of native Czech and Slovak speakers in identifying deepfakes, with Czech speakers' results on the left and Slovak speakers' on the right. The X-axis quantifies the percentage of recordings correctly identified. Two distinct colours, blue for Czech (cz) and orange 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