Does Cognitive Load Affect Human Accuracy in Detecting Voice-Based Deepfakes?

Marcel Gohsen
marcel.gohsen@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Nicola Libera
nicola.lea.libera@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Johannes Kiesel
johannes.kiesel@gesis.org
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany

Jan Ehlers
jan.ehlers@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Benno Stein
benno.stein@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Abstract

Deepfake technologies are powerful tools that can be misused for malicious purposes such as spreading disinformation on social media. The effectiveness of such malicious applications depends on the ability of deepfakes to deceive their audience. Therefore, researchers have investigated human abilities to detect deepfakes in various studies. However, most of these studies were conducted with participants who focused exclusively on the detection task; hence the studies may not provide a complete picture of human abilities to detect deepfakes under realistic conditions: Social media users are exposed to cognitive load on the platform, which can impair their detection abilities. In this paper, we investigate the influence of cognitive load on human detection abilities of voice-based deepfakes in an empirical study with 30 participants. Our results suggest that low cognitive load does not generally impair detection abilities, and that the simultaneous exposure to a secondary stimulus can actually benefit people in the detection task.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; *Sound-based input / output*.

Keywords

empirical study, voice-based deepfake detection, deepfake detection under cognitive load

ACM Reference Format:
Marcel Gohsen, Nicola Libera, Johannes Kiesel, Jan Ehlers, and Benno Stein. 2026. Does Cognitive Load Affect Human Accuracy in Detecting Voice-Based Deepfakes?. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3786304.3787868

CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/2026/03
https://doi.org/10.1145/3786304.3787868

1 Introduction

Deepfakes refer to synthetic media content that is the result of manipulating source media in order to depict situations that did not occur in reality. Among the most popular applications of deepfake technology is to "fake" people by replicating their appearance or their voice [47]. While there are beneficial applications, such as depicting historical figures [62] or objects [43, 64] in a museum, deepfakes are predominantly used unethically and without consent of the depicted individuals [18]. Examples of malicious deepfake attacks include fraud [11, 17], non-consensual depiction in pornographic contexts [46, 58], and the spread of disinformation [2].

One way to reduce these negative consequences of the malicious use of deepfakes is to educate individuals on how to recognize manipulated media [56]. However, with the advancement of deep learning technologies, these deepfakes are getting hard to spot. Various studies on the human ability in detecting voice [9, 29, 45, 48], video [26, 34, 51, 57], text [15], and multimodal deepfakes [16, 21, 27, 30] come to similar conclusions: human detection performance of deepfakes is often not much better than a coin toss.

Social media platforms are especially popular channels for spreading disinformation and fake news (e.g., through the use of deepfakes) [40, 65]. The consumption of social media causes cognitive load for users [50] in the form of divided attention across multiple posts and advertisements [31] or even information overload [24]. We believe that this cognitive load has the potential to interfere with the human detection performance of deepfakes.

To our knowledge, this paper represents the first study on the impact of cognitive load on the human detection abilities of voice-based deepfakes. We specifically investigate voice modality as it can be effectively used to influence people [23], was among the most difficult modalities to detect in previous studies [16, 27], and requires only a small amount of source material to create realistic-sounding clones of individuals. Although video is the primary modality in social media, voice clones pose a real threat in terms of spreading disinformation, as attackers can imitate news videos by adding cloned voices of newsreaders to widespread B-roll ("symbolic") footage, which provides no visual clues to spot a manipulation.

This paper reports on a study involving 30 participants in which their ability to distinguish cloned voices of newsreaders from their real counterparts is examined in two experiments. In Experiment 1, an audio stimulus is presented to a participant, who has to decide whether this stimulus is real or fake under varying cognitive loads. We emulate this cognitive load with a 1-back task that has to be solved while listening to the stimulus. In Experiment 2, we create a different kind of distraction by showing participants related B-roll footage in parallel to the audio stimulus.[1] We find that light cognitive load, as caused by a 1-back task, does not in general impact human accuracy in detecting voice-based deepfakes. Furthermore, we observe that the simultaneous consumption of a secondary stimulus, such as the B-roll videos in Experiment 2, causes participants to get significantly better at the detection task.

In this paper, we contribute an extensive analysis of the body of literature that conducted empirical studies to examine human voice-clone detection abilities. Further, we outline a realistic attack that makes use of voice clones to spread disinformation on social media. Then, we describe a pipeline to obtain realistic-sounding voices-clones of real newsreaders and present our study. Finally, we report and discuss our findings and possible explanations.

2 Related Work

In the literature, we found 12 papers that report about studies that investigate and quantify the human ability in distinguishing between real (also known as *bona fide*) or fake (also known as *spoofed*) audio stimuli. A recent overview about those studies was provided by Amirkhani et al. [5]. In the following, we discuss and compare these studies with respect to experimental design, stimuli design, variables, findings, and identified indicators. The indicators are organized into a novel taxonomy.

Experimental Approaches. An established procedure is to ask participants to listen to a single audio stimulus and assign one of two labels to the audio clip: bona fide or spoofed [9, 27, 29, 45, 48, 55, 60, 61]. For their study, Alali and Theodorakopoulos [4] mixes bona fide and spoofed audio into a single clip and hence require a third assignable class ("partially fake"). Some studies take into account the confidence of participants in their decision-making. [13] enable participants to choose the option "unsure" instead of making a decision. Barnekow et al. [7] allow participants to assign one of five labels, namely "real", "rather real", "rather fake", "fake" and "no idea". Similarly, Frank et al. [21] ask their participants to rate deepfaked media (including speech) on a 5-point Likert scale ranging from "definitely non-human" to "definitely human." In addition to the standard binary classification, Barrington et al. [9] presents participants with two audio stimuli and ask them to decide if they originated from the same identity. These pairs of audio stimuli originate from either two different human speakers, two identical human speakers, or one human speaker and their cloned voice. A similar pairwise task is included in the study procedure of Mai et al. [45], in which participants are shown a pair of bona fide and spoofed audio stimuli and asked to identify which is which.

A notable distinction between these studies is the format in which they are conducted. Eight of these studies are conducted as an online survey (e.g., crowdsourced via Prolific[2]) [4, 7, 9, 21, 27, 45, 48, 60]. Although this format allows researchers to acquire numerous participants—between 102 [7] and 3,002 participants [21]—the acoustic properties of the environment (e.g., speaker vs. headphones, background noise) can not be sufficiently controlled. Two studies collect their data via web conferencing services (e.g., Zoom) [29, 55]. While Han et al. let participants play the audio stimuli on their local computer through the browser, Sharevski et al. share the audio through the web conferencing software, which could introduce compression artifacts that might interfere with the experiment. Lastly, two studies are conducted in a controlled laboratory environment [13, 61].

Experimental Stimuli. When conducting a study on the human ability to detect auditory deepfakes, it is common practice to use a bona fide and spoofed stimuli from the same speaker. Previous studies have employed one of three methodologies to obtain such stimuli. The first method uses existing datasets that contain such stimuli. The second method involves using an existing dataset of bona fide stimuli, the speakers of which are spoofed using a state-of-the-art voice clone model. The third method creates a corpus of bona fide and spoofed stimuli, which may entail recording or crawling custom audio data.

Popular datasets for these studies, which contain both bona fide and spoofed stimuli, originate from the Automatic Speaker Verification Challenge[3] (ASVspoof). Specifically, the datasets of the ASVspoof challenge of the years 2017 [37], 2019 [59], and 2021 [42] have been used as target stimuli in the studies consulted [13, 29, 48, 55, 60]. These datasets represent a substantial database of speech recordings from the VCTK dataset [63] and spoofed audios, generated with a broad range of state-of-the-art TTS, voice conversion, and deepfake models. Alali and Theodorakopoulos [4] and Barrington et al. [9] resort to their own previously created datasets from 2024, namely the RFP [3] and the Deepspeak dataset [8], respectively. The RFP dataset is a collection of bona fide audios obtained from the VCTK, the UK and Ireland English speech [19], the DiPCo [54], and the YouTube-8M [1] datasets which were used as target speakers for TTS, voice conversion, and partial deepfake approaches to produce spoofed audios. The Deepspeak dataset was created by asking crowdworkers to record a video of themselves speaking utterances, and then used audio and video manipulation such as voice cloning and lip-sync to generate spoofed media. Bhalli et al. [13] use the "Fake or Real" (FoR) dataset [52] from 2019 in their study, which comprises bona fide stimuli from the CMU ARCTIC [39], LJ Speech [33], and Voxforge [44] datasets and synthetic speech from commercial TTS systems. Since Groh et al. [27] mostly focus on video deepfakes in the political domain, they use the Presidential Deepfakes dataset [53] from 2021 to obtain stimuli for their study. Next to the ASVspoof 2021 dataset, Warren et al. [60] also use the Wavefake [22] and FakeAVCeleb [35] datasets (both created in 2021) in their study.

Factors Influencing the Recognition Performance. The aforementioned studies examined the impact of different variables on the accuracy with which humans recognize voice clones. The studies assume that four main factors could influence human accuracy: the characteristics of the participant, the speaker, the spoken content, and the synthesis.

In terms of participant characteristics, the impact of demographics such as gender [9, 13], age [9, 21, 48], educational background

[1] Stimuli are available at: webis.de/data/webis-newsreader-voice-clones-26.html
[2] www.prolific.com
[3] https://www.asvspoof.org/

[21, 61], and country of residence [21] have been analyzed. Additionally, the effect of language proficiency characteristics such as native language [13, 48] and level of fluency in a second language [13] have been investigated. Furthermore, the levels of computing experience of a participant has been considered in three studies [13, 48, 61]. Two studies focus on the human detection abilities of blind and visually impaired individuals [29, 55]. Other studies investigate the impact of familiarity with the speaker's voice [4, 45] and of the level of training [13, 45, 48].

The speaker's characteristics are for the most part underexplored in the aforementioned studies. So far, only the effects of a speaker's gender and their spoken dialect have been examined [4].

An investigation of the impact of spoken content characteristics has been conducted by Watson et al. [61]. In their study, they analyze whether the level of complexity (e.g., tongue twisters versus simple sentences) of voice clones has an impact on human detection abilities. Furthermore, they test whether mentioning a political candidate in the speech affects human accuracies. Finally, two studies examine the effects of spoken language on the human detection accuracy, in order to ascertain whether the language of the spoken content has an effect on this. While Mai et al. [45] draw parallels between English and Mandarin, Frank et al. [21] compare English, German, and Mandarin.

The last category of variables that are varied in studies of human voice clone detection abilities are characteristics of the synthesis. These characteristics include the choice of voice synthesis system [4, 55, 60], the type of the synthesis (e.g., TTS versus deepfake) [29, 48], and duration of the synthesized audio [9, 45, 61]. Of the analyzed studies, Frank et al. [21] and Groh et al. [27] also consider different kind of modalities of the synthesized deepfakes such as video and text.

Study Findings. As the human performance in detecting voice clones depends on a variety of factors, the measured human accuracy varies dramatically from study to study. A large proportion of the studies find that human accuracy in detecting voice clones range between 60% and 80% [9, 27, 45, 48, 60]. Below that range, Frank et al. [21] find accuracies slightly above chance of between 50% and 60% depending on a participant's nationality and the language of the audio stimuli. Accuracy is measured in a comparable range by Han et al. [29] and Sharevski et al. [55]—both studies that investigate the detection performance of blind or low vision individuals. However, there are outliers reporting on accuracies that are considerably below chance. The study by Watson et al. [61] reports on an average accuracy of 42% across university students. Similarly, Barnekow et al. [7] clone the voice of a university professor familiar to their participants and find that they could only recognize the cloned voice correctly in 37% of the cases. The lowest found accuracy is 16% for detecting partial fake speech (i.e., single media items that contain both bona fide and spoofed voices) [4].

In addition to detection accuracies, studies investigated which factors significantly impact these values. One of the most detrimental factor is what approach was used to create the spoofed stimuli [60] which can cause a difference of up to 20% in average accuracy. Furthermore, Müller et al. [48] find that native speakers have an advantage in detecting voice clones speaking in their native language. According to Watson et al. [61], the spoken content has a major impact; complex spoken content is easier to recognize than simple content in spoofed audios. Finally, Bhalli et al. [13] find that training sessions can help to increase the human detection performance.

(1) Artifacts
- Background noises [7, 9]
- Frequency profile [7]
- Recording quality [9, 29]
- Reverb and echo [7, 29]

(2) Inflection
- Emotions [9, 29]
- Intonation [45]
- Monotonousness [9]
- Roboticness [45]

(3) Pronunciation
- Accents [9, 29]
- Filler words [29]
- Mispronunciations [9, 29]
- Stutters [9]

(4) Rhythm
- Breathing [9, 29]
- Pace [9, 45]
- Pauses [9, 29, 45]

Figure 1: Our taxonomy of indicative characteristics of stimuli that participants in various studies reported using to distinguish between bona fide and spoofed voices. The categories are detailed in the last part of Section 2

On the contrary, some studies also identified factors that do not affect the human detection performance. Barrington et al. [9] do not find any effect of the speaker's gender. On the participant side, minor effects are associated with demographics [21] except the performance decline with increased age [48]. Particularly, an educational background in computing does not affect the human detection performance [13, 48, 61]. Lastly, the listening behavior of a participant does not play a significant role such as the number of stimulus repetitions or the time spend on the detection task [45].

Indicative Characteristics of Stimuli. In the different studies that we analyze, participants report on different indicators that helped them to distinguish between bona fide and spoofed stimuli. In Figure 1, we organize these properties into a taxonomy to provide an overview of these indicators.

The artifacts group (1) comprises acoustic artifacts that are unwanted byproducts of speech synthesis algorithms, which include background noises, unnatural frequency profiles, varying levels of perceived audio quality, and inconsistent reverb and echo behavior. Most of these lead participants to decide that a stimulus is spoofed.

The inflection group (2) is concerned with the "melody" and stress of the speech. The existence or lack thereof of perceived emotions can be an indicator of either bona fide or spoofed audio stimuli, respectively. Intonation can be a tool to express such emotions by varying the pitch of pronounced syllables in an utterance. If these intonations are underexpressed or unnaturally frequent, participants perceived stimuli as monotone or robotic.

The pronunciation group (3) describes how the words in an utterance are pronounced. Accents (e.g., New York accents) influence this pronunciation and can impact whether a participant perceive an audio as bona fide or spoofed. Added filler words and stutters are usually associated with bona fide stimuli, while odd mispronunciations can be indicators for spoofed stimuli as well.

Finally, the rhythm group (4) describes temporal properties of a spoken utterance. Breathing at anticipated times, although reproducible with modern voice synthesis algorithms, are typically indicators for bona fide stimuli. The same is true for pauses in the

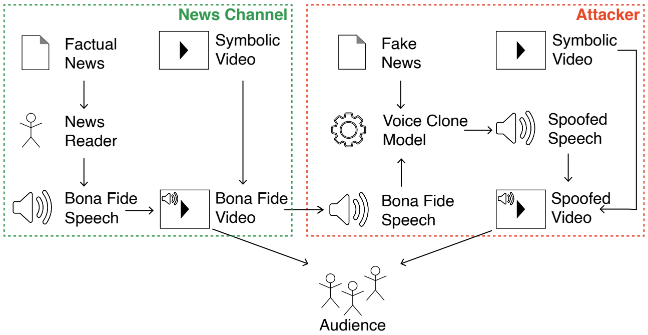

Figure 2: Real-world attack model for spreading fake news videos with voice-based deepfakes. The audience is tricked into believing to hear some well-known newsreader—and thus that the news is spread by the reader's news outlet. Our study is designed to mirror corresponding attacks.

speech. The overall pace of a spoken utterance can also give away if a stimulus is spoofed or bona fide.

3 Fake News Attack Model

To demonstrate how voice-based deepfakes can be instrumentalized to spread disinformation, we outline a real-world attack model. This attack model is based on the premise that newsreaders, journalists and social media influencers publish numerous high-quality videos in which they report on news on social media platforms or news portals. These videos are easily accessible to potential attackers, who can use voice cloning to imitate the voices of these speakers once they have collected enough high-quality material. For example, Elevenlabs[4] suggests a minimum amount of source material of 30 minutes. The attackers could use the cloned newsreader voices to report fake news that fits their agenda, which would appear believable due to the credibility of the imitated newsreader. When news is presented in a voice-over style, combined with symbolic B-roll video footage, with the newsreader not visible, all the visual cues indicating that the media is fake are eliminated.

Figure 2 outlines the attack model differentiating the roles of a news channel, an attacker, and an audience. On the side of the news channel, a newsreader reads out factual news in voice-over style, which is combined with symbolic video material to create the bona fide news video. The attacker collects audio material through, for example, the extraction of audio channels from news videos and other available sources and then uses it to train or condition a voice clone model. This model generates a voice-over narration for attacker-provided fake news in the newsreader's voice, which is paired with original news footage or a symbolic video to produce the spoofed news video ready for publication on social media.

This attack model is not just a thought experiment, but is currently being used in real attacks. In the beginning of 2025, journalist Georgina Findlay became victim of such an attack, in which far right social media channels cloned her voice to spread disinformation following fascist ideologies [20]. In 2023, German news program "Tagesschau" was under attack, in which the voice's of their newsreaders were cloned to spread disinformation about the Ukrainian war and the Covid-19 pandemic [10].

4 Stimulus Design

To obtain bona fide and spoofed stimuli for our experiments, we follow a methodology that is plausible for potential attackers in the described attack model in Section 3. We suspect that a potential attacker is likely to acquire source media from social media platforms. Therefore, we apply the same strategy, manually collecting videos from YouTube. The news videos that we deem usable are clips from well-known news channels (e.g., NBC, BBC) that are of high audiovisual quality and have low levels of background noise. As the effectiveness of the attack depends on the acoustic properties of a voice, we include different speakers in our experiments. Specifically, we collect videos of four renowned newsreaders (two female and two male). The following four newsreaders are selected based on their vocal diversity and the amount of obtainable high-quality source media on YouTube:

- Andrea Mitchell. *American journalist at NBC News.*
- Carl Nasman. *American journalist at BBC News.*
- Lester Holt. *American journalist at Dateline NBC.*
- Sophie Raworth. *English journalist at BBC News.*

The source videos undergo an extensive manual filtering and editing process. Duplicate video segments are removed, and noisy parts are cut out. This collection, filtering, and editing process is repeated until at least one hour's worth of training data has been collected for each of the four voices. During this process, suitable candidate segments for bona fide stimuli in the study are selected and set aside, i.e., not used for training the voice clone model.

We apply additional automatic preprocessing steps to the training portion of the videos. Music and background noises are removed by employing an MDX-Net model [36] for music demixing and a model called "Kim Vocal 1" from the Ultimate Vocal Remover v55[5] application for the isolation of the voice. The result of this step is a clean vocal track of the video. To make the training process more efficient, these audio clips are split into semantically coherent segments of less than 10 seconds using an automatic audio splitter that uses WhisperX [6], a long form audio transcription approach, to find appropriate segmentations.[6] Following this step, we obtained 2,872 audio clips, each averaging about six seconds in lengths.

In initial experiments with various open-source Text-to-Speech systems (TTS) that feature voice cloning, such as F5-TTS [14], StyleTTS 2 [41], TorToise [12], and VoiceCraft [49], TorToise produced the most authentic reproduction of the four selected voices. TorToise is an autoregressive transformer combined with a diffusion-based denoising algorithm and a MEL-based vocoder. Another advantage of TorToise is that its GUI-based training process[7] is sufficiently simple, so that potential attackers do not require expert skills in AI applications to clone a voice.

We train a model for each of the four selected speakers, with the help of the aforementioned GUI. Each model is trained with a batch size of 80 for 400 epochs with model checkpointing for every 50

[4] https://elevenlabs.io

[5] https://ultimatevocalremover.com/

[6] https://github.com/JarodMica/audiosplitter_whisper

[7] https://github.com/JarodMica/ai-voice-cloning

epochs. We then use each of the resulting models to generate example outputs and to manually determine the ideal number of training epochs per voice. For the voices of Lester Holt and Sophie Raworth, 300 epochs yielded the best results, whereas for Andrea Mitchell and Carl Nasman, 350 training epochs were preferable. In terms of training hyperparameters, the default values as suggested by the training GUI are mostly used, such as a learning rate of 10^{-5}. However, some optimizations of the training hyperparameters were performed in the initial experiments, resulting in MEL and text learning rate weights of 1.0 and 0.6, respectively.

In order to eliminate the influence of spoken content and limit human detection characteristics to acoustic features, the bona fide stimuli are transcribed, and the transcriptions are used to reproduce the same content with the trained voice clone models. The generated spoofed stimuli then undergo further manual post-processing. All artifacts and noise are filtered out, and the frequency response curves are brought closer to the original audio signal by applying equalizer curves. Finally, the volume of the bona fide and spoofed stimuli is normalized to ensure identical loudness levels.

In total, we prepare 12 audio stimuli for each voice, six of which are bona fide and six of which are spoofed (48 in total). The audio clips are between 11 and 15 seconds long and correspond to two to three sentences of spoken content. All audio clips are stored as lossless PCM waveforms with a sampling rate of 44.1 kHz.

For Experiment 2, in which videos are shown to participants while listening, the pool of audio stimuli is extended with additional stimuli following the methodology described above. For each of the four selected newsreaders, we select two news videos that are publicly available. To avoid providing any visual cues whether a stimulus contains bona fide or spoofed audio, we collect video material in which the speaker is not visible. The collected videos feature symbolic visualizations to engage the viewer (sometimes referred to as B-roll footage [32]), while the news are provided as a voice-over. In total, we create a pool of 16 video clips, 8 of which have unchanged audio tracks and 8 of which involve spoofed audio. The clips are between 12 and 15 seconds long.

5 Empirical Study

In order to investigate the extent to which humans are capable of recognizing the attacks mentioned above, we conducted two experiments. Experiment 1 is a laboratory-controlled study in which the cognitive load is systematically varied using a dual-task scenario. The aim is to examine how this affects the ability to distinguish between bona fide and spoofed voices. However, recent research shows that humans are less accurate at detecting deepfakes in audio-only scenarios compared to audiovisual media [16, 27]. We therefore opted for a simple secondary task (similar to the established 1-back assignment [25]) to ensure the primary discrimination task (bona fide versus spoofed voice) could still be carried out. Experiment 2 constitutes an application-oriented extension of the research design, with the objective of validating the results using real-life video sequences. The methodological approach employed in both experiments is outlined below.

5.1 Experiment 1: Auditory Sequences

Experiment 1 uses a dual-task technique to control and vary the utilization of working memory capacities (cognitive load). The aim is to investigate how this factor influences the reliability with which bona fide and spoofed voices can be distinguished.

5.1.1 Design and Procedure. Experiment 1 takes place in a small laboratory room. Participants are greeted and requested to sign a form of consent. The experimental procedure is depicted in Figure 3 and described below. Participants are asked to complete a demographic survey through which they could specify their age and gender, and rate their experience with artificial voices and their confidence in detecting them. The latter two were collected on a 7-point Likert scale ranging from no expertise (1) to expert (7), and from not at all confident (1) to very confident (7), respectively.

Following this, participants are asked to familiarize themselves with the bona fide voices of the four newsreaders. For each speaker, we provide an approximately 15-second audio clip. Participants are asked to listen to all four clips from beginning to end at least once. The respective clips are not included in the pool for the single- and dual-task exercises. Once participants become used to the voices, the experiment starts.

To vary the working memory load, participants perform a single and a dual task. The order of the assigned condition is counterbalanced across participants. In the single-task condition, participants are asked to listen to 24 audio clips that are played in random order via speakers. After listening to each clip, they are asked to indicate whether the stimulus is bona fide or spoofed and to describe what leads them to their decision. In addition, they are asked to rate their confidence on a 7-point Likert scale and to evaluate the perceived quality of the audio clips. At the end of each task, participants are asked to rate the perceived difficulty and report heuristics that helped them distinguish between bona fide and spoofed stimuli. In the dual-task condition, a secondary task has to be completed along with the primary task. Specifically, participants are required to observe a series of digits displayed on a monitor and have to press a button when a "3" follows a "1." The numbers on the screen are updated at a rate of 0.8 seconds. The proportion of numbers that are target stimuli range from 13% to 33%.

The 48 audio stimuli (obtained as described in Section 4) form a combined pool from which stimuli are randomly drawn for each participant in the single-task and dual-task conditions. However, we ensure that bona fide and spoofed audios of identical utterances never become stimuli for the same task. Participants are not informed about the distribution of bona fide and spoofed stimuli.

5.2 Experiment 2: Video Sequences

Experiment 2 is conducted with the same participants immediately after the completion of Experiment 1. However, while Experiment 1 uses an established approach to control and vary cognitive load, Experiment 2 incorporates videos as an application-oriented component, which constitutes the most common type of content on social media. Both experiments focus on the manipulation of auditory information. As in Experiment 1's dual-task condition, decision-making in Experiment 2 (bona fide vs. spoofed voices) is made more difficult by incorporating visual information, although no controlled variation of the cognitive load is performed here.

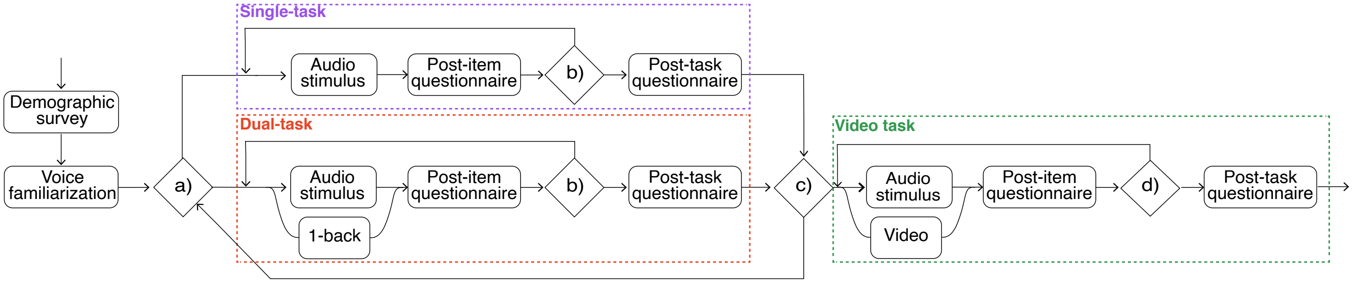

Figure 3: Flow chart of the conducted study consisting of single-task and dual-task conditions of Experiment 1 and the video condition in Experiment 2. The decisions a) and c) are mechanisms to randomize the order of conditions and ensuring that both conditions have been performed once. The decision b) and d) repeat the listening task for 24 and 8 different stimuli in Experiment 1 and 2, respectively.

5.2.1 Design and Procedure. Experiment 2 follows a similar study procedure as Experiment 1. Participants look at a sequence of 8 videos, four of which containing spoofed audio, and have to indicate after being exposed to a stimulus whether the stimulus contains bona fide or spoofed audio. The post-item and post-task questionnaires in Experiment 2 are identical to questionnaires in Experiment 1. From the array of 16 video stimuli, 8 are selected at random, bearing in mind that videos reciting the same utterance (with bona fide and spoofed audio) are excluded from appearing within the same task for the same participant.

5.3 Apparatus

The study application is implemented as a Flask server[8] responsible for data logging and assignment of stimuli and tasks. The front end is implemented using native HTML and JavaScript to provide survey forms and to collect data as well as to implement the stimulus-response architecture (Experiment 1) and media playback (Experiment 1 and 2). As the popularity of consuming social media content on mobile devices in speaker mode increases, participants listen to the stimuli through the internal dual speaker setup of an HP OMEN 17 laptop in both experiments.

5.4 Participants

30 volunteers (22 males, none of whom are non-binary, with an age between 20 and 36 years) participated in both experiments. Due to the composition of the sample, age-related limitations in hearing ability should be ruled out [48]. Furthermore, none of our test subjects reported any issues with their ability to perceive sound. Written informed consent was obtained prior to the data collection.

6 Results

In the following, we discuss the results for the two experiments—Experiment 1 with auditory sequences and Experiment 2 with video sequences—with respect to the accuracy of the participants in detecting bona fide and spoofed voices.

6.1 Results: Experiment 1

Under single-task and dual-task conditions, participants achieve an average accuracy of 0.67. The detection accuracy is slightly lower under the dual-task ($\mu = 0.66$, $\sigma = 0.13$) than under the single-task condition ($\mu = 0.68$, $\sigma = 0.12$). According to a paired t-test, the mean accuracies achieved under these two conditions do no differ significantly ($p = 0.6$). Independent of the task conditions, the mean accuracy in detecting bona fide stimuli is higher ($\mu = 0.7$, $\sigma = 0.14$) than for spoofed stimuli ($\mu = 0.63$, $\sigma = 0.18$), accompanied by a substantial increase in standard deviation. In Figure 4, we compare the detection accuracy of bona fide and spoofed stimuli under individual experiment conditions. While the median accuracies are seemingly unaffected, we can see that the detection accuracies of spoofed stimuli are much less consistent between participants under the dual-task condition, which causes an increase of standard deviation from 0.17 to 0.24. On the contrary, participants are more consistent to detect bona fide stimuli under the dual-task condition, causing a drop in standard deviation from 0.19 to 0.14.

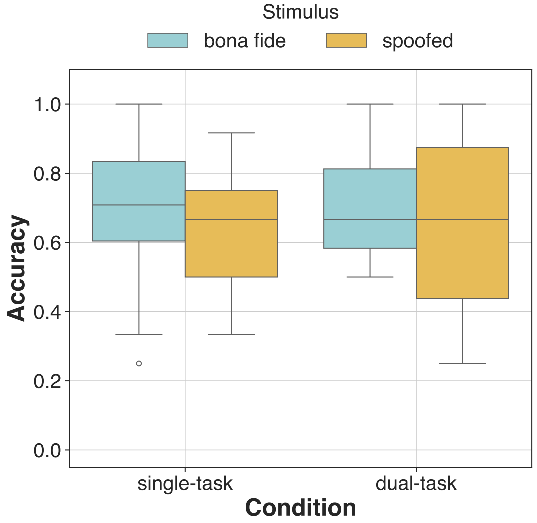

Figure 4: Accuracy in detecting voice clones in the single- and dual-task conditions for spoofed and bona fide stimuli averaged by participant.

As the mean accuracy seems to be unaffected by applying cognitive load, we examine if there is a trade-off in primary (detection)

[8] https://flask.palletsprojects.com

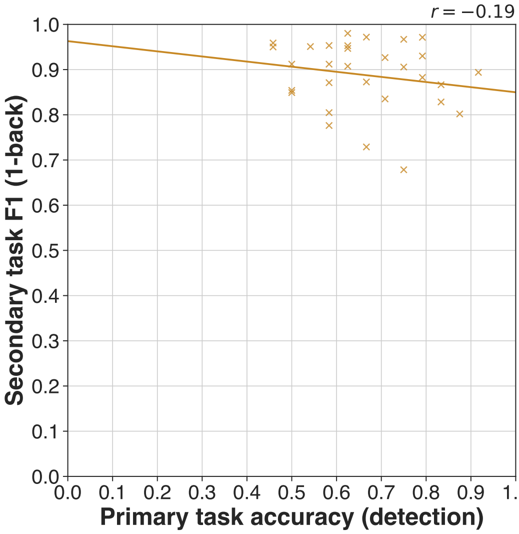

Figure 5: Analysis of correlation between primary (voice-clone detection) task accuracy and secondary task (1-back task) F_1 of the participants. There is a weak negative correlation (Pearson $r = -0.19$) indicating that a trade-off of cognitive capacities exists.

Table 1: Effects of various variables in predicting whether a participant guesses correctly according to a logistic regression model.

Variable	**Coefficient**	***z*-value**	***p*-value**
Decision confidence	0.147	4.050	**<0.001**
Ground truth	0.282	2.444	0.015
Speaker: Mitchell	-0.623	-1.251	0.211
Participant gender	0.143	1.079	0.281
Stimulus	-0.013	-0.757	0.452
Participant age	0.008	0.555	0.572
Condition	0.052	0.454	0.650
Speaker: Nasman	-0.232	-0.410	0.682
Speaker: Raworth	0.185	0.220	0.826
Speaker: Holt	-0.028	-0.041	0.967
Participant experience	-0.004	-0.103	0.918

and secondary task (1-back) performances under the dual-task condition. If such a trade-off exists, participants exhibit a limited cognitive capacity, which is divided across the primary and secondary task. In Figure 5, the participant's accuracy and F_1 in the individual tasks are compared. According to Pearson's correlation coefficient, a weak negative correlation ($r = -0.19$) can be observed, which implies that participants trade off performance in the individual tasks. However, this correlation is not significantly different from zero ($p = 0.32$).

To estimate the effects of individual variables, we fit a logistic regression model that predicts whether a participant made a correct decision. A Likelihood-ratio test on the model reveals that at least one of the variables significantly improves the model fit ($p = 0.0003$).

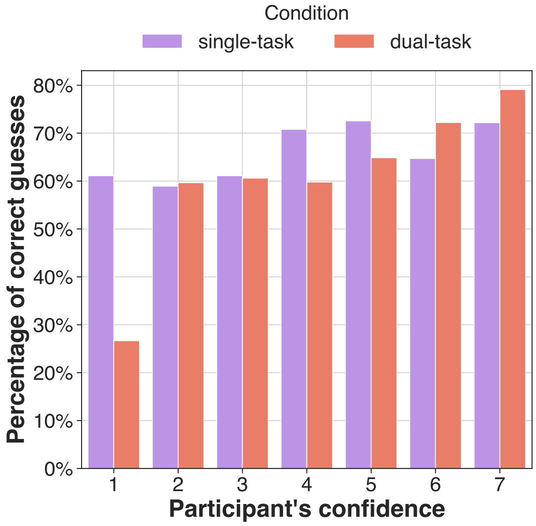

Figure 6: Percentage of correct guesses by participants, broken down by their self-assessed confidence in their guess (higher score = greater confidence in the guess), separated by single-task and double-task conditions.

Table 1 contains the resulting coefficients, z-values, and p-values for each analyzed variable, where the p-values originate from a post hoc Wald-Test. To avoid significant findings due to chance caused by the problem of multiple comparisons, we perform an alpha correction using the Holm-Bonferroni method ($\alpha = 0.005$ after correction). Decision confidence (i.e., how certain a participant is about their decision) has a significant influence, suggesting that participants have reflected well on their decisions. The ground truth (i.e., whether a stimulus is bona fide or spoofed) may or may not have a significant effect, depending on whether one accepts the conservative correction. The dummy variable for bona fide is 1, which means that the positive coefficient shows that if the stimulus is bona fide, there is a higher probability that the participant guesses correctly. The other variables have p-values that are undoubtedly consistent with the null hypothesis and we omit their interpretation.

As the results of the logistic regression identify a participant's confidence in the decision as a good predictor of the participant's success, we dig deeper to see how accurate the self-reflection is for the single- and dual-task conditions. Figure 6 shows the percentage of correct decisions under the condition that a participant assigned a specific confidence score. Under the dual-task condition, the function of the proportion of correct guesses resembles a monotonically increasing function based on a participant's confidence rating. However, a participant's self-reflection abilities under the single-task condition seem to be worse. A point-biserial correlation analysis reveals a correlation of $r = 0.06$ and $r = 0.17$ under the single- and dual-task conditions, respectively, where only the latter is significantly different from zero ($p \approx 7 \times 10^{-6}$). Participants under the single-task condition have a tendency to underestimate their detection abilities. Although they assigned the lowest possible confidence of one, they still guess correctly above chance (60% correct guesses). The highest proportion of correct responses in the single-task condition are collected when participants assigned a confidence score of five.

Decision Indicators. As part of the experiment, we asked participants for a rationale on every decision to find common indicators that humans use to decide between bona fide and spoofed. These free text answers are manually analyzed to obtain an overview.

We find comparable indicators to what prior studies reported (see Figure 1). Participants identified artifacts such as "glitches", drops in volumes, existence of digital effects (e.g., chorus, reverb, echo), static and interference which make participants vote for spoofed. However, also artifacts of human recording are identified such as smacking of lips, breathing or microphone pops that make participants believe that a stimulus is bona fide. Similar aspects are identified with respect to the frequency profile of a stimulus where dull, flat, muffled or distorted profiles are associated with spoofed and clear and scratchy voices with bona fide stimuli. Voices labeled as husk or hoarse are identified as either bona fide or spoofed. In terms of inflections, participants identify absence of emotions, robotic intonation, and monotone utterances as traits of spoofed stimuli. One participant stated that the utterance exhibited a relaxed way of speaking and decided for bona fide while another participant declared a voice as sounding irritated and labeled the stimulus as spoofed. Most reported indicators are about pronunciation aspects of the utterances, which in most cases is an indicator for spoofed stimuli. Participants categorize accents as fake and point out mispronunciations, mumbling, stumbling over words, and badly articulated syllables and word endings. Rhythm aspects are also present, where inconsistent pace, or too fast or too slow pace is associated with spoofed audio. Pauses are more frequently associated with bona fide stimuli.

Some indicators are found in our study that are not part of our taxonomy. One participant established a connection to the spoken content, even though identical utterances are used for bona fide and spoofed stimuli. Participants said that the content sounded like it was generated by AI or that the news content was implausible, which made them choose spoofed. This indicator aligns with an observation made by Watson et al. [61] that if a political candidate appeared in the utterance, more participants decided it was fake. Furthermore, the familiarization seems to have an effect causing participants to justify their choices by saying that the stimulus sounded different (spoofed) or similar (bona fide) to the remembered voice. Finally, some decisions are justified by simply having a "gut feeling" or intuition.

6.2 Results: Experiment 2

Participants in Experiment 2, in which a video is shown in parallel with bona fide and spoofed audio clips, achieved an average accuracy of 0.75. This accuracy is significantly higher than under the single ($\mu = 0.68$, $p = 0.03$) and dual-task ($\mu = 0.66$, $p = 0.005$) conditions in Experiment 1. In Figure 7, the accuracies of detecting bona fide and spoofed stimuli in the video condition are compared. On average, participants detect bona fide stimuli with an accuracy of 0.78 and spoofed stimuli with an accuracy of 0.71. According to a t-test, these differences are not significant ($p = 0.24$). Noteworthy are also the even larger standard deviations of 0.22 and 0.25 for bona fide and spoofed stimuli, respectively.

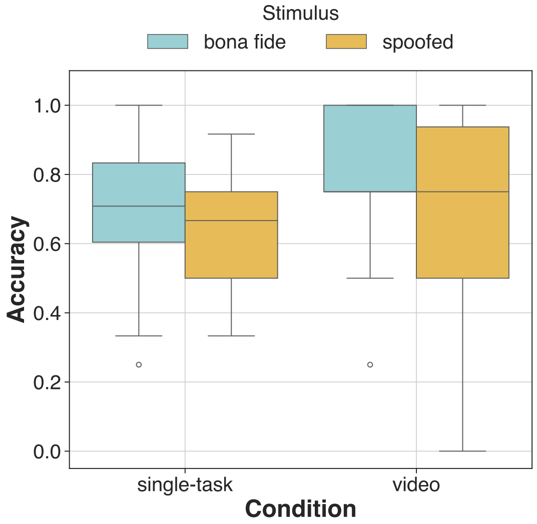

Figure 7: Accuracy in detecting voice clones in the video condition of Experiment 2 in comparison to the single-task condition of Experiment 1 for spoofed and bona fide stimuli averaged by participant.

As Experiment 2 was conducted after Experiment 1, we analyze if learning effects could explain the higher achieved accuracies. Independent of what the condition was (order was randomized), the average accuracy under the first condition is 0.65 while under the second condition is 0.68. Although higher accuracies are achieved in the second condition, the difference is not significant according to a paired t-test ($p = 0.2$).

Finally, we wondered whether the participants who perform well in Experiment 1 also perform well in Experiment 2. Figure 8 shows individual task performances of each participant in single-task, dual-task and video conditions. In the plot, we can observe that in many cases, participants perform similarly good or bad across Experiment 1 and 2 with a few exceptions. Participant 1 got a perfect accuracy in the video condition while only achieving 50% accuracy in the single-task condition. Similarly, participant 3 did well in the dual-task and video conditions, but also only made 50% correct guesses in the single-task condition. A correlation analysis reveals that there is a moderate Pearson correlation between accuracies of single-task versus dual-task ($r = 0.32$), single-task versus video-task ($r = 0.33$), and dual-task versus video-task ($r = 0.49$).

Decision Indicators. Although participants are shown visual stimuli at the same time, the decision indicators on which participants base their decisions do not differ substantially between Experiment 1 and Experiment 2. The only noticeable difference is that at least one participant stated that the videos made the stimulus more credible and prompted him to vote for "bona fide."

7 Discussion

Based on the results of our two conducted experiments, the impact of cognitive load on human detection abilities of voice-based deepfakes remains ambiguous. There is no significant difference in

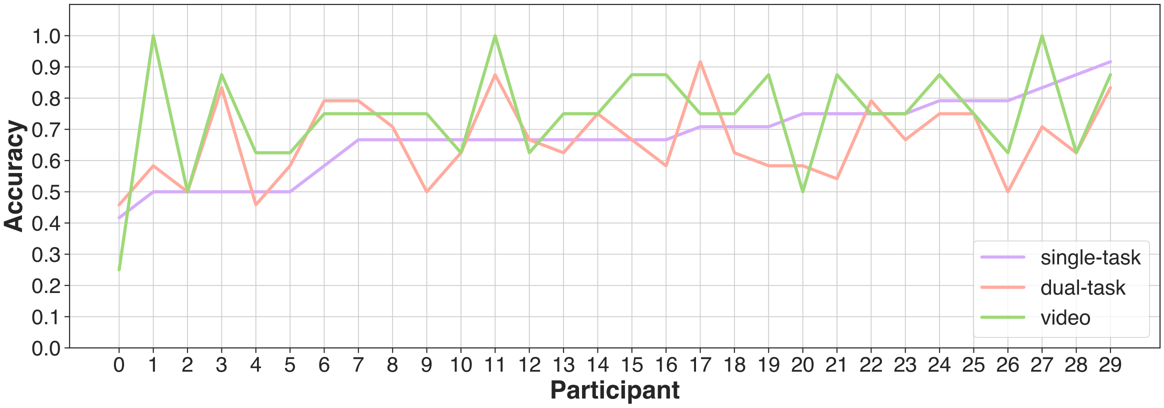

Figure 8: Accuracy in detecting voice clones for each participant in the single-task, dual-task and video conditions. Participants sorted by single-task accuracy.

average accuracy achieved under single-task and dual-task conditions. However, we observe a drastic increase in standard deviation of the accuracy going from the single-task to the dual-task condition for detecting spoofed stimuli. We discuss three plausible theories which could explain these observations.

The first theory entails that a 1-back task does not cause sufficient cognitive load to significantly impact the detection accuracy. This theory is backed up by the difficulty ratings of the participants, of which only 11 found the single-task easier than the dual-task. 13 participant found dual-task to be as easy as the single-task condition, while 6 participants even found the dual-task to be the easier condition. An argument against this theory is the found correlation between the primary and secondary task performance. Although this correlation is weak, it shows that some cognitive capacities have to be divided across both tasks.

Given the fact that some participants found the dual-task condition easier than the single-task condition, our second theory states that a second stimulus or task helps some participants to focus more on the primary stimulus. For example, some studies found that background music can increase focus on sustained attention tasks [38]. Another example is the accessory stimulus effect, in which task-irrelevant “accessory” stimuli can reduce the reaction time of participants in a primary reaction task [28]. This theory would explain the significantly higher accuracy in the video condition in Experiment 2. On the contrary, the theory does not explain the variable accuracy of participants under the dual-task condition.

Assuming that the accessory stimulus effect applies for the video condition of Experiment 2 but not for the dual-task condition of Experiment 1, our third theory tries to explain the found accuracies of the dual-task condition. This theory hypothesizes that there is no interference between cognitive load caused by a 1-back task and voice-based deepfake detection. While the 1-back task makes use of working memory, detecting voice clones is an auditory processing task which might require a different kind of attention. If this theory is true, then the non-significant differences between single-task and dual-task condition can be considered random noise.

The detection abilities of voice clones vary a lot from human to human, which gets evident by the overall high standard deviation across all tasks. The worst participant achieved an accuracy below chance in all three conditions, while the best participant maintained an accuracy above 80% across all tasks. However, participants are somewhat consistent in their performance independent of the task as shown by their moderately correlating task accuracies. This shows how subjective the detection abilities are and that the different accuracies were not achieved due to draws of easy stimuli.

8 Conclusion

People often encounter deepfakes in situations in which they are exposed to cognitive load. Therefore, we examined the effect of cognitive load on the accuracy of detecting voice-based deepfakes with an empirical study. In two experiments, we compared the baseline human detection performance to the accuracy while solving a 1-back task in parallel or while watching symbolic video footage. We found that a light cognitive load as caused by a 1-back task does not systematically affect human detection accuracy. Furthermore, we observed that participants who watched video footage in parallel performed significantly better in the detection task, which could be related to the accessory stimuli effect.

As the 1-back task caused potentially light cognitive load in the participants, we want to repeat the experiment with considerably harder secondary tasks. To determine whether these loads are realistic compared to social media use, we plan to add a social media simulation with scrolling through other posts and advertisements as an additional experiment. In our experiments, only a single stimulus-generating pipeline has been used in, which can be added as an independent variable in a follow-up study. Moreover, the role of the familiarity with the cloned speaker has not been analyzed yet, which could be subject of a future study.

As we have observed, deepfakes can be hard for some people to identify. We encountered participants with below chance accuracies on the task. Supporting those people by educating them or implementing intervention strategies on social media platform is detrimental to protect our society from waves of disinformation. We hope that this and earlier studies are a call to action to mitigate the negative effects of disinformation and make social media platforms a safer place for information access.

Acknowledgments

This work was partially supported by the Federal Ministry of Research, Technology and Space (BMFTR) under grant 16KIS2470 (PADSE: Person-centered, audio- and voice-based deepfake and shallowfake detection).

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. doi:10.48550/arXiv.1609.08675

[2] Samer Al-khazraji, Hassan Saleh, Adil Khalid, and Israa Mishkhal. 2023. Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics* 23 (Oct. 2023), 429–441. doi:10.55549/epstem.1371792

[3] Abdulazeez Alali and George Theodorakopoulos. 2024. An RFP Dataset for Real, Fake, and Partially Fake Audio Detection. doi:10.48550/arxiv.2404.17721

[4] Abdulazeez Alali and George Theodorakopoulos. 2025. Partial Fake Speech Attacks in the Real World Using Deepfake Audio. *Journal of Cybersecurity and Privacy* 5, 1 (2025), 6 pages. doi:10.3390/JCP5010006

[5] Sima Amirkhani, Gunnar Stevens, M. D. Shajalal, and Alexander Boden. 2025. Detecting the Undetectable: Human Judgments and the Challenge of Synthetic Voices. In *Proceedings of the 12th International Conference on Communities & Technologies (C&T 2025)*. European Society for Socially Embedded Technologies (EUSSET), Siegen, Germany, 6 pages. doi:10.48340/ct2025-1030

[6] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023*, Naomi Harte, Julie Carson-Berndsen, and Gareth Jones (Eds.). ISCA, Dublin, Ireland, 4489–4493. doi:10.21437/INTERSPEECH.2023-78

[7] Vanessa Barnekow, Dominik Binder, Niclas Kromrey, Pascal Munaretto, Andreas Schaad, and Felix Schmieder. 2021. Creation and Detection of German Voice Deepfakes. In *Foundations and Practice of Security - 14th International Symposium, FPS 2021 (Lecture Notes in Computer Science, Vol. 13291)*, Esma Aïmeur, Maryline Laurent, Reda Yaich, Benoît Dupont, and Joaquín García-Alfaro (Eds.). Springer, Paris, France, 355–364. doi:10.1007/978-3-031-08147-7_24

[8] Sarah Barrington, Matyas Bohacek, and Hany Farid. 2024. DeepSpeak Dataset v1.0. doi:10.48550/ARXIV.2408.05366

[9] Sarah Barrington, Emily A. Cooper, and Hany Farid. 2025. People Are Poorly Equipped to Detect AI-powered Voice Clones. *Scientific Reports* 15, 1 (March 2025), 11004. doi:10.1038/s41598-025-94170-3

[10] Matthias Bastian. 2023. AI-generated Fake Audio of Germany's Top News Program "Tagesschau" Spreads Disinformation. https://the-decoder.com/ai-generated-fake-audio-of-germanys-top-news-program-tagesschau-spreads-disinformation/.

[11] Jon Bateman. 2020. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Carnegie Endowment for International Peace, Washington D.C., United States.

[12] James Betker. 2023. Better Speech Synthesis through Scaling. doi:10.48550/arxiv.2305.07243

[13] Noshaba Nasir Bhalli, Nehal Naqvi, Chloe Evered, Christine Mallinson, and Vandana P. Janeja. 2024. Listening for Expert Identified Linguistic Features: Assessment of Audio Deepfake Discernment among Undergraduate Students. doi:10.48550/arxiv.2411.14586

[14] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A Fairytaler That Fakes Fluent and Faithful Speech with Flow Matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 6255–6271.

[15] Alicia Tsui Ying Chong, Hui Na Chua, Muhammed Basheer Jasser, and Richard T. K. Wong. 2023. Bot or Human? Detection of DeepFake Text with Semantic, Emoji, Sentiment and Linguistic Features. In *13th IEEE International Conference on System Engineering and Technology, ICSET 2023*. IEEE, Shah Alam, Malaysia, 205–210. doi:10.1109/ICSET59111.2023.10295100

[16] Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. 2024. As Good As A Coin Toss: Human Detection of AI-generated Images, Videos, Audio, and Audiovisual Stimuli. doi:10.48550/arxiv.2403.16760

[17] Audrey de Rancourt-Raymond and Nadia Smaili. 2022. The Unethical Use of Deepfakes. *Journal of Financial Crime* 30, 4 (May 2022), 1066–1077. doi:10.1108/JFC-04-2022-0090

[18] Deeptrace. 2019. *The State of Deepfakes*. Technical Report.

[19] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-Source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 6532–6541.

[20] Georgina Findlay. 2025. 'You're Gonna Find This Creepy': My AI-cloned Voice Was Used by the Far Right. Could I Stop It? https://www.theguardian.com/commentisfree/2025/jan/07/ai-clone-voice-far-right-fake-audio. *The Guardian* (2025).

[21] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. 2024. A Representative Study on Human Detection of Artificially Generated Media Across Countries. In *IEEE Symposium on Security and Privacy, SP 2024*. IEEE, San Francisco, CA, USA, 55–73. doi:10.1109/SP54263.2024.00159

[22] Joel Frank and Lea Schönherr. 2021. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.), Vol. 1. NeurIPS, Virtual Event, 17 pages.

[23] Marcel Gohsen, Johannes Kiesel, Mariam Korashi, Jan Ehlers, and Benno Stein. 2023. Guiding Oral Conversations: How to Nudge Users Towards Asking Questions?. In *8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023)*. ACM, New York, United States, 34–42. doi:10.1145/3576840.3578291

[24] Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Bernhard Schölkopf. 2014. Quantifying Information Overload in Social Media and Its Impact on Social Contagions. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014*, Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice Oh (Eds.). The AAAI Press, Ann Arbor, Michigan, USA, 170–179. doi:10.1609/icwsm.v8i1.14549

[25] Janine Grimmer, Laura Simon, and Jan Ehlers. 2021. The Cognitive Eye: Indexing Oculomotor Functions for Mental Workload Assessment in Cognition-Aware Systems. In *CHI '21: CHI Conference on Human Factors in Computing Systems*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi (Eds.). ACM, Yokohama, Japan, 428:1–428:6. doi:10.1145/3411763.3451662

[26] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds. *Proceedings of the National Academy of Sciences* 119, 1 (Jan. 2022), e2110013119. doi:10.1073/pnas.2110013119

[27] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2024. Human Detection of Political Speech Deepfakes across Transcripts, Audio, and Video. *Nature Communications* 15, 1 (Sept. 2024), 7629. doi:10.1038/s41467-024-51998-z

[28] Steven A. Hackley and Fernando Valle-Inclán. 1999. Accessory Stimulus Effects on Response Selection: Does Arousal Speed Decision Making? *Journal of Cognitive Neuroscience* 11, 3 (May 1999), 321–329. doi:10.1162/089892999563427

[29] Chaeeun Han, Prasenjit Mitra, and Syed Masum Billah. 2024. Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024*, Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski (Eds.). ACM, Honolulu, HI, USA, 949:1–949:14. doi:10.1145/3613904.3642817

[30] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. 2024. Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes. doi:10.48550/arxiv.2405.04097

[31] Nathan Oken Hodas and Kristina Lerman. 2012. How Visibility and Divided Attention Constrain Social Contagion. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012*. IEEE Computer Society, Amsterdam, Netherlands, 249–257. doi:10.1109/SOCIALCOM-PASSAT.2012.129

[32] Bernd Huber, Hijung Valentina Shin, Bryan C. Russell, Oliver Wang, and Gautham J. Mysore. 2019. B-Script: Transcript-based B-roll Video Editing with Recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, Glasgow, Scotland, UK, 81. doi:10.1145/3290605.3300311

[33] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[34] Emilie Josephs, Camilo Fosco, and Aude Oliva. 2023. Artifact Magnification on Deepfake Videos Increases Human Detection and Subjective Confidence. doi:10.48550/arxiv.2304.04733

[35] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.), Vol. 1. NeurIPS, Virtual Event, 14 pages.

[36] Minseok Kim, Woo-Sung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung. 2021. KUIELab-MDX-Net: A Two-Stream Neural Network for Music Demixing. doi:10.48550/arXiv.2111.12203

[37] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee. 2017. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *18th Annual Conference of the International Speech Communication Association, Interspeech 2017*, Francisco Lacerda (Ed.). ISCA, Stockholm, Sweden, 2–6. doi:10.21437/INTERSPEECH.2017-1111

[38] Luca Kiss and Karina J. Linnell. 2021. The Effect of Preferred Background Music on Task-Focus in Sustained Attention. *Psychological Research* 85, 6 (Sept. 2021), 2313–2325. doi:10.1007/s00426-020-01400-6

[39] John Kominek and Alan W. Black. 2003. CMU ARCTIC Databases for Speech Synthesis. http://www.festvox.org/cmu_arctic/.

[40] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. *Science* 359, 6380 (March 2018), 1094–1096. doi:10.1126/science.aao2998

[41] Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.), Vol. 36. Curran Associates, Inc., New Orleans, LA, USA, 19594–19621.

[42] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE ACM Trans. Audio Speech Lang. Process.* 31 (2023), 2507–2522. doi:10.1109/TASLP.2023.3285283

[43] Irene López García, Ephraim Schott, Marcel Gohsen, Volker Bernhard, Benno Stein, and Bernd Fröhlich. 2024. Speaking with Objects: Conversational Agents' Embodiment in Virtual Museums. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2024)*, Ulrich Eck, Misha Sra, Jeanine K. Stefanucci, Maki Sugimoto, Markus Tatzgern, and Ian Williams (Eds.). IEEE, Greater Seattle Area, USA, 279–288. doi:10.1109/ISMAR62088.2024.00042

[44] Ken MacLean. 2018. Voxforge. https://www.voxforge.org/home.

[45] Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans Cannot Reliably Detect Speech Deepfakes. *PLOS ONE* 18, 8 (Aug. 2023), e0285333. doi:10.1371/journal.pone.0285333

[46] Karolina Mania. 2024. Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study. *Trauma, Violence, & Abuse* 25, 1 (Jan. 2024), 117–129. doi:10.1177/15248380221143772

[47] Yisroel Mirsky and Wenke Lee. 2022. The Creation and Detection of Deepfakes: A Survey. *Acm Computing Surveys* 54, 1 (2022), 7:1–7:41. doi:10.1145/3425780

[48] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. 2022. Human Perception of Audio Deepfakes. In *DDAM@MM 2022: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, Jianhua Tao, Haizhou Li, Helen Meng, Dong Yu, Masato Akagi, Jiangyan Yi, Cunhang Fan, Ruibo Fu, Shan Lian, and Pengyuan Zhang (Eds.). ACM, Lisboa, Portugal, 85–91. doi:10.1145/3552466.3556531

[49] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12442–12462. doi:10.18653/V1/2024.ACL-LONG.673

[50] Matthew Pittman and Eric Haley. 2023. Cognitive Load and Social Media Advertising. *Journal of Interactive Advertising* 23, 1 (Jan. 2023), 33–54. doi:10.1080/15252019.2022.2144780

[51] Swaroop Shankar Prasad, Ofer Hadar, Thang Vu, and Ilia Polian. 2022. Human vs. Automatic Detection of Deepfake Videos Over Noisy Channels. In *IEEE International Conference on Multimedia and Expo, ICME 2022*. IEEE, Taipei, Taiwan, 1–6. doi:10.1109/ICME52920.2022.9859954

[52] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019*, Corneliu Burileanu and Horia-Nicolai Teodorescu (Eds.). IEEE, Timisoara, Romania, 1–10. doi:10.1109/SPED.2019.8906599

[53] Aruna Sankaranarayanan, Matthew Groh, Rosalind Picard, and Andrew Lippman. 2021. The Presidential Deepfakes Dataset. In *CEUR Workshop Proceedings*, Vol. 2942. CEUR-WS, Aachen, Germany, 57–72.

[54] Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2020. DiPCo - Dinner Party Corpus. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng (Eds.). ISCA, Shanghai, China, 434–436. doi:10.21437/INTERSPEECH.2020-2800

[55] Filipo Sharevski, Aziz Zeidieh, Jennifer Vander Loop, and Peter Jachim. 2024. Blind and Low-Vision Individuals' Detection of Audio Deepfakes. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024*, Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie (Eds.). ACM, Salt Lake City, UT, USA, 4867–4881. doi:10.1145/3658644.3690305

[56] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H. Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating Disinformation in a Social Media Age. *WIREs Data Mining and Knowledge Discovery* 10, 6 (2020), e1385. doi:10.1002/WIDM.1385

[57] Klaire Somoray and Dan J. Miller. 2023. Providing Detection Strategies to Improve Human Detection of Deepfakes: An Experimental Study. *Computers in Human Behavior* 149 (2023), 107917. doi:10.1016/J.CHB.2023.107917

[58] Soyoung Wang and Seongcheol Kim. 2022. Users' Emotional and Behavioral Responses to Deepfake Videos of K-pop Idols. *Computers in Human Behavior* 134 (2022), 107305. doi:10.1016/J.CHB.2022.107305

[59] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas W. D. Evans, Md. Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, and Zhen-Hua Ling. 2020. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech. *Computer Speech and Language* 64 (2020), 101114. doi:10.1016/J.CSL.2020.101114

[60] Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin R. B. Butler, Carrie Gates, and Patrick Traynor. 2024. "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024*, Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie (Eds.). ACM, Salt Lake City, UT, USA, 2696–2710. doi:10.1145/3658644.3670325

[61] Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja. 2021. Audio Deepfake Perceptions in College Going Populations. doi:arXiv.2112.03351

[62] Nathan Wynn, Kyle Johnsen, and Nick Gonzalez. 2021. Deepfake Portraits in Augmented Reality for Museum Exhibits. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR 2021 Adjunct*. IEEE, Bari, Italy, 513–514. doi:10.1109/ISMAR-ADJUNCT54149.2021.00125

[63] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. doi:10.7488/ds/2645

[64] Matteo Zaramella, Irene Amerini, and Paolo Russo. 2023. Why Don't You Speak?: A Smartphone Application to Engage Museum Visitors Through Deepfakes Creation. In *Proceedings of the 5th Workshop on analySis, Understanding and proMotion of heritAge Contents, SUMAC 2023*, Valérie Gouet-Brunet, Ronak Kosti, and Li Weng (Eds.). ACM, Ottawa, ON, Canada, 29–37. doi:10.1145/3607542.3617359

[65] Xichen Zhang and Ali A. Ghorbani. 2020. An Overview of Online Fake News: Characterization, Detection, and Discussion. *Information Processing & Management* 57, 2 (March 2020), 102025. doi:10.1016/j.ipm.2019.03.004