Odyssey 2026
23-26 June 2026, Lisbon, Portugal

I Hear, Therefore I Trust: A Socio-Technical Investigation of Humans as Synthetic Speech Detectors

Lelia Erscoi[1], *Tomi Kinnunen*[1]

[1]Computational Speech Group, University of Eastern Finland, Joensuu, Finland
lerscoi@uef.fi

Abstract

Automatic deepfake detection has received considerable research attention, yet the socio-technical environment in which humans actually encounter synthetic speech remains poorly understood. We investigate voice deepfake detection as a perceptual and contextual process, presenting a localization task in which 47 participants marked suspected synthetic segments across authentic, fully synthetic, and partially synthetic utterances under three manipulated trust cues: instructional framing, affective priming, and provenance labeling. Participants provided quality ratings on mechanicalness, expressiveness, intelligibility, clarity, calmness, and confidence of evaluation. Utterance class was the primary determinant of detection accuracy and perceptual quality; trust cues produced no main effects but motivated detection behavior. Fully synthetic speech was detected at below-chance levels. Quality ratings tracked utterance type, indicating implicit discrimination where overt detection failed.

Index Terms: synthetic speech, voice deepfakes, trust and artificial intelligence, socio-technical systems

1. Introduction

Generative artificial intelligence (genAI), particularly conversational agents, extends a socio-cultural transition of epistemic authority and information sourcing, from traditional news outlets, to social media platforms and now to personalized chatbots [1, 2]. Within this shifting landscape, trust becomes a currency of the attention economy [3]. Communities such as Reddit's r/isthisAI[1] have millions of weekly visitors who propose content for open debate on authenticity. The ease of creation and distribution allows for an overflow of synthetic media content that dampens the user detection capabilities and creates a liar's dividend, where anything encountered online might plausibly be synthetic [4]. In this study, we aim to draw attention to speech deepfakes as a socio-technical trust issue [5] that remains under-investigated, particularly given that human detection performance is typically studied in isolation from the broader stream of cues listeners encounter in everyday communicative settings. Even so, detection accuracy tends to be limited [6, 7, 8], raising alarms about user vulnerabilities in the face of increasingly believable generative models.

Speech is especially susceptible to misinterpretation when presented in isolation from other cues that could support holistic authenticity assessment. Beyond serving as a carrier for linguistic content, the human voice occupies a unique position in trust judgments, with paralinguistic dimensions (e.g., tone, pitch, intonation) additionally carrying associations with identity, intent, familiarity, or authenticity [9]. Going beyond traditional voice spoofing attacks (or presentation attacks) targeted against automatic speaker verification systems [10], voice deepfakes blend into the broader content stream, where credibility is inferred from arbitrary metrics such as number of likes or even personal attitudes towards the content [11]. In such contexts, automatic detection systems or simple disclosure labels may be insufficient, since warnings compete within the same attention economy [12]. When reproducibility and ease of dissemination are no longer technical constraints, the risk of voice deepfakes lies in how they are received, as well as in how trust is manufactured in everyday communicative settings [13, 14].

The rapid maturation of speech synthesis technology [15] tends to outpace countermeasure development in an arms race fueled by ever-growing data availability. On the automatic detection front, performance evaluation benchmarks such as the ASVspoof challenge series [16, 17] indicate strong synthetic speech detection accuracy, especially under controlled conditions. Contrastingly, synthetic speech detection accuracy in naive, unattentive listeners hovers near chance levels [6, 7]. More alarmingly, listeners tend to rely on outdated acoustic cues such as robotic prosody or abnormal pausing and breathing, elements that modern synthesis systems no longer reliably produce [6, 18]. This leaves judgments vulnerable to the very contextual and social factors that have received little empirical attention. When only segments of an utterance are manipulated, detection rates fall even among attentive listeners [7].

The present study moves beyond purely computational concerns to examine the broader socio-technical phenomena of voice deepfake detection. Building on previous frameworks describing trust and artificial intelligence [19, 20] or deepfakes specifically [5], we focus on the environmental context as the critical variable often overlooked in controlled lab settings. Crucially, no prior study in human deepfake detection has jointly manipulated the acoustic manipulation type, the contextual trust environment, and the listener response modality within a single ecologically valid protocol. This fragmentation means that existing benchmarks are incomplete by design, that they measure detection under conditions that do not reflect how people actually encounter synthetic speech. The result is a systematic overestimate of human detection capability. We argue that neither automatic detection systems nor human listeners alone offer a silver bullet, as both are subject to distinct biases [12, 18, 21, 22, 23]. Consequently, effective mitigation requires understanding the interaction between these agents and the contextual factors that shape their performance.

This paper offers three primary contributions. First, we delineate the environmental context as a distinct dimension of the socio-technical trust framework, identifying specific subsources of trust (such as instructional framing, affective prim-

[1]https://www.reddit.com/r/isthisAI/

10.21437/Odyssey.2026-50

ing, and provenance labels) that mediate human susceptibility to voice deepfakes. Second, we systematically investigate how these contextual factors influence detection accuracy and perceptual quality, testing four hypotheses: if detection accuracy varies by utterance authenticity (**H1**) and whether this relationship is moderated by trust orientation (**H1a**), and if subjective ratings vary by authenticity (**H2**) and trust orientation (**H2a**). Third, to advance ecological validity, we introduce a synthetic speech localization task. Unlike binary classification tasks, this allows participants to highlight specific segments they deem inauthentic, providing fine-grained insight into the decision-making process and revealing exactly where and why trust fractures in real-world listening scenarios.

2. Socio-technical background

The term "deepfake" was coined in the mid-2010s to describe non-consensual impersonating pornographic synthetic media produced with deep learning techniques [24]. Over time, its definition has expanded to encompass generative techniques across audio, text, images, videos, and multimodal media. Beneficial applications such as language dubbing [25] or personalized voice assistants [26] have fueled public interest in generative AI, leading to a diminished perceived gravity and amplified potential for confusion and harm [27]. When deepfakes disseminate disinformation or cause harm, culpability becomes increasingly diffused, with responsibility often shifted onto users who are framed as insufficiently skeptical [4]. Initiatives such as the European Union's Regulation 2024/1689 (Artificial Intelligence Act) [28] and its corresponding AI Office[2] aim to address AI technologies based on their assigned risk, with deepfakes categorized as a "transparency" risk.

2.1. Trust and synthetic speech

The issue of harm by (speech) deepfakes is not a matter of *how*, but *when* [4, 22, 29]. With the normalization of synthetic media, increasingly realistic synthesis models create an expanding security gap at both societal and technical levels, where subtle alterations of reality threaten privacy, safety, and even the integrity of democratic processes [4].

Mounting a spoofing attack has become increasingly straightforward, driven by open-access Text-to-Speech (TTS) models and the broad availability of voice recordings [30], including those that users willingly provide to build personalized voice-based agents [30]. Automatic detection models often overlook cues that human listeners naturally pick up, such as emotional affect [31] or paralinguistic artifacts [32]. For detection tools to function effectively as decision-support mechanisms, however, they must first establish cognitive trust with their users [19, 13]. In the absence of such trust, users are prone to disregard model outputs entirely. How to optimally combine human and machine capabilities, therefore, remains an open question, particularly given evidence that even minor inconsistencies in algorithmic outputs erode the very trust that makes human-AI collaboration effective [19].

2.2. Human perception and the limits of vigilance

It remains an open question how realistic it is to expect individuals to be constantly vigilant. As with other technological security threats, such as malware, computational safeguards must be complemented by awareness and education, particularly since many people may still (mistakenly) believe that AI is "not quite there yet". However, as personal life increasingly merges with the digital domain (see, for example, digital personas [33]), it becomes easier to tailor attacks that can deceive even trained professionals, especially when the impersonation relies solely on voice. In real-world conditions, spoofing attacks always come with intent and meaning. Attackers exploit trust by imitating a family member[3] or by invoking authority to sway decisions or influence elections[4]. Moreover, synthetic voice generation need not arise from explicitly malicious motives to result in deception and harm, as illustrated by cases in which individuals develop romantic attachments to personalized chatbots, sometimes leaving their actual families [5] or even ending their own lives [6].

Partial manipulation (interweaving synthetic bits within otherwise authentic content) requires far less technical overhead while achieving a significant impact through harvesting the quality and authority of the original signal. The challenge of discriminating authentic from synthetic speech may be contaminated by the perception of the whole, in which authentic regions anchor judgment and suppress suspicion even when manipulation is present, or vice versa, where the presence of manipulation discredits the overall credibility. Empirical evidence favors the latter as detection rates fall substantially when authentic and synthetic speech are intermixed [7]. This suggests that partial manipulation is not merely harder to detect but qualitatively different as a perceptual problem.

2.3. Trust cues as experimental manipulations

Critically, trust in what one hears is not formed on acoustic evidence alone. The present study aims to investigate this multi-source account of trust through three manipulations, hereby referred to as *trust cues*, working at the environmental level of the trust framework introduced in Section 1. These cues each target a distinct channel through which contextual information reaches the listener, thereby supplementing their decision process on top of the acoustic properties of the utterance to be evaluated, whether synthetic, authentic, or a mix of the two.

- **Instruction framing** (I+ / I−) manipulates the cognitive set with which participants approach the task. Participants are presented with one of two scenarios: one framing the audio as originating from a groundbreaking communication technology under positive evaluation (I+), the other framing it as material from a nefarious actor targeting platform users (I−). In both conditions, participants assumed the role of a social media platform moderator whose decisions carry real consequences for the platform's users, and received identical task instructions that remained visible throughout the experiment. This manipulation targets the top-down expectations that govern how attentively and skeptically a listener engages with incoming audio [34].
- **Affective valence priming** (V+ / V−) manipulates the emotional context preceding each trial through images depicting positive- or negative-valence content. Drawing on evidence that affect modulates perceptual judgment [35, 36], this manipulation tests whether emotional state at the moment of lis-

[2] https://digital-strategy.ec.europa.eu/en/policies/ai-office

[3] https://people.com/woman-conned-out-of-usd15k-after-ai-cloned-daughters-voice-terrifying-scam-11775622

[4] https://f24.my/A6cV

[5] https://people.com/man-proposed-to-his-ai-chatbot-girlfriend-11757334

[6] https://edition.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis

tening shifts detection sensitivity or evaluative bias.

- **Provenance labeling** (C+ / C−) manipulates credibility in source attribution by presenting a textual label affirming a trusted origin for the audio, with the negative trust option having no such label. This tests whether an explicit source claim overrides or interacts with acoustic evidence in forming authenticity judgments [37]. Provenance labels of this kind are increasingly being adopted in practice, and are set to become a regulatory requirement under the EU AI Act, with a dedicated Code of Practice on marking and labeling of AI-generated content currently being drafted ahead of the transparency obligations coming into force in 2026 [7].

3. Methods

3.1. Dataset

The LlamaPartialSpoof [38] is a recent dataset that contains both fully synthetic and partially spoofed speech. It contains English utterances from 40 LibriTTS [39] speakers, with synthetic counterparts generated using five open-source models (LJ JETS[8], YourTTS[9], XTTS V2[10], GPT-SoVITS[11], CosyVoice[12]) plus one commercial service (ElevenLabs[13]). Partially synthetic samples are created by splicing genuine and fake segments using cross-fading techniques, simulating realistic localized manipulation attacks.

For the task at hand, 20 utterances were sampled in a ratio of 1:1:2 of authentic-synthetic-partially synthetic speech by balancing the generation model and sentence spoken. To maintain a manageable experiment duration and minimize participant listening fatigue [40], utterances were limited to those over 10 seconds, and for partially synthetic speech, only utterances with a maximum of 3 synthetic sections were chosen. Files were then mixed with randomly selected environmental recordings from the International Soundscape Database [41] at 25 dBSNR to introduce ecologically valid acoustic conditions reflecting real-world listening scenarios [42].

For the affective priming manipulation, images were drawn from the Open Affective Standardized Image Set (OASIS) [43], a standardized dataset containing user-validated arousal (calm-intense) and valence (negative-positive) ratings for each image. Images rated in the top 20% for arousal and at either valence extreme were selected to maximize priming effects, with manual verification to exclude explicitly graphic content.

3.2. Synthetic speech detection

Participants were asked to evaluate the 20 utterances presented in a randomized order to prevent carryover effects [44]. For each clip, participants are instructed to listen to the audio in its entirety and to identify any portions that sound artificial or fabricated. Participants could annotate the suspicious segments using two types of markers: point-based "flags" (timestamps) or interval-based "segments" (start-end ranges). Participants were provided unlimited time to attend to each utterance and instead were instructed to perform to the best of their ability.

[7] https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

[8] https://github.com/espnet/espnet/tree/master/egs2/ljspeech/tts1

[9] https://github.com/coqu$I$−ai/TTS/

[10] https://huggingface.co/coqui/XTTS-v2

[11] https://huggingface.co/lj1995/GPT-SoVITS/tree/main

[12] https://github.com/FunAudioLLM/CosyVoice

[13] https://elevenlabs.io

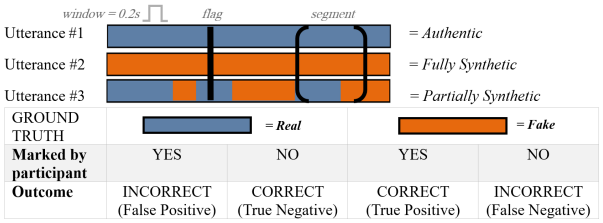

Figure 1: *How user markers translate to evaluation metrics. Each speech excerpt is divided into windows (length* $= 0.2s$*) to extract the ground-truth label and to verify whether the user placed a marker (flag or segment) within that window. A* $\pm 200ms$ *temporal margin is applied to each marker to account for reaction-time delays.*

3.3. Speech evaluation

Participants also completed a set of subjective evaluation statements. For our study, each utterance was treated as a single perceptual entity by participants, regardless of any insertions or edits, ensuring that their judgments reflected the overall impression of the voice rather than isolated segments. The following dimensions were assessed on a five-point Likert scale (*Strongly Disagree* $= -2$, *Disagree* $= -1$, *Unsure* $= 0$, *Agree* $= +1$, *Strongly Agree* $= +2$), with *Unsure* as the default response.

- **Artificiality ("The voice sounds mechanical")**: Listeners are known to be sensitive to prosodic and timing cues that signal non-human origin regardless of prior task experience [6]. This dimension deliberately targets markers of *artificiality* rather than the more ambiguous concept of *naturalness* [45].
- **Expressiveness ("The voice sounds expressive")**: captures emotional variation and prosodic richness [46]. Synthetic voices may be perfectly intelligible yet lack the affective modulation characteristic of human communication [9].
- **Intelligibility ("The voice is easy to understand")**: reflects the ease with which listeners comprehend the speech signal, affecting cognitive load and user acceptance [47].
- **Clarity ("The audio sounds clean")**: captures technical recording or synthesis quality, including compression artifacts or signal anomalies [48].
- **Calmness ("The voice sounds calm")**: assesses the perceived arousal and emotional valence of the speaker. In deception contexts, voice-based credibility judgments are known to be influenced by perceived speaker affect [49, 50].
- **Confidence ("I am confident in my evaluation")**: captures subjective certainty about a judgment. It is well known that confidence does not reliably correlate with detection accuracy, and synthetic speech can remain highly deceptive even to confident listeners [51]. Confidence ratings map to detection behavior: marking no segment implies an authentic classification, meaning high confidence reflects certainty in non-detection rather than abstention.

The evaluation questions were presented in randomized order in each trial to avoid automaticity [52].

3.4. Task design

Detection performance was evaluated using window-based metrics derived from the temporal overlap between participant annotations and ground-truth synthetic regions, with a sliding window of $0.2s$ and a boundary tolerance of $\pm 200ms$. Each win-

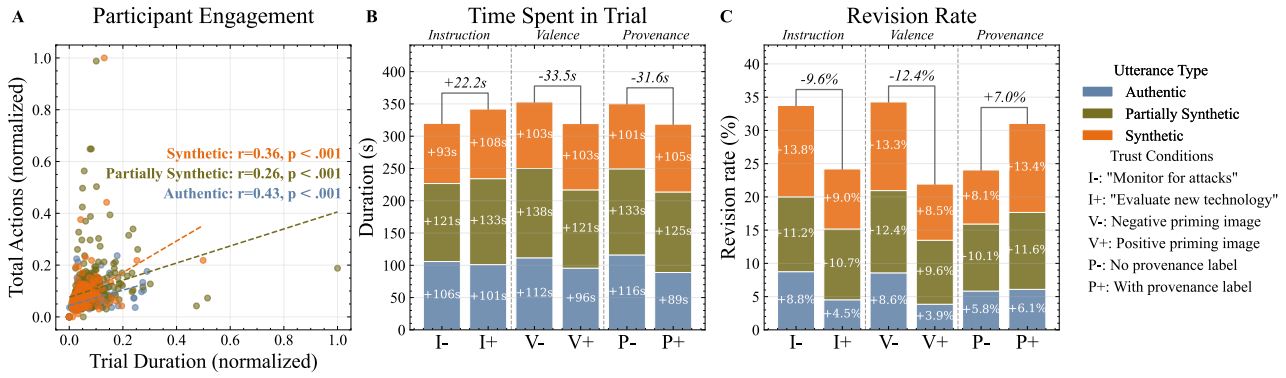

Figure 2: *Decision-making patterns across conditions. (A) Positive correlation between action count and trial duration across all utterance types; synthetic speech trials required significantly more time and annotations. (B) Mean trial duration. The positive instruction framing ($I+$)* increased *duration by $+22$ s, whereas positive valence ($V+$) and provenance labeling ($C+$)* reduced *it by -33 s and -32 s, respectively. (C) Revision rates (proportion of trials where a participant added a marker, then deleted it). Provenance labeling ($C+$) increased revisions ($+7.0\%$), while positive instruction framing ($I+$) and positive valence ($V+$) decreased them (-9.6%, -12.4%, respectively), indicating that contextual cues modulate annotation confidence and the willingness to revise judgments.*

dow is assigned a binary ground-truth label and a binary participant decision, giving four outcomes: a window is a true positive (TP) if synthetic and marked, a true negative (TN) if authentic and unmarked, a false positive (FP) if authentic but marked, and a false negative (FN) if synthetic but unmarked. This allows performance to degrade proportionally to the fraction of the audio incorrectly flagged. Accuracy is the proportion of windows for which the participant's decision matched the ground truth. For utterance types containing only one class of windows (e.g., fully authentic or fully synthetic), accuracy reduces to TNR or TPR, respectively, and the complementary metrics are undefined.

Two forms of attention control were implemented to ensure adequate task engagement. First, participants were instructed to listen to each utterance in full before responding, and a compliance criterion of 90% of all trials was imposed. This threshold was chosen to tolerate occasional accidental clicks or overly rapid responses due to overconfidence, while still discouraging multitasking that could impair performance, given the task's expected complexity [53]. Second, adherence to task instructions was probed with an additional question that appeared on a random one-third of trials, asking participants to indicate which scenario they had been instructed to follow. Participants selected from five options: the two true scenarios ("Monitoring for attacks."/$I-$ and "Evaluating new voice technology."/$I+$), two decoy scenarios ("Creating new synthetic voices." and "Moderating for offensive language."), and a default option ("I did not pay attention"). Because the task scenario remained visible on-screen throughout, submissions were rejected if a participant provided more than one incorrect answer. This defensive design [54] was used to help ensure that the collected data accurately reflected participants' actual abilities within a crowdsourcing framework.

Platform interactions were timestamped throughout the task, with an action log recording marker additions, removals, and response selections. Only responses present at the moment of saving were treated as final answers. Participants were debriefed at the end of their session and shown a visualization of their marker placements against ground truth, alongside their evaluation responses.

3.5. Trust cues

Trust-influencing manipulations were designed to reflect real-world scenarios where users encounter potentially manipulated media under varying degrees of contextual trust.

Each participant was randomly assigned to one of two contextual scenarios, presented at the beginning of the session and confirmed through an initialization button ("I understand"). The positive trust scenario framed the task as evaluating a groundbreaking communication technology designed to benefit users. The negative trust scenario warned participants about a nefarious entity deliberately targeting the platform's users. All participants received identical task instructions: they would act as content moderators for a (fictitious) social media platform, and their decisions would affect the platform's users. These instructions remained visible throughout the experiment to maintain the framing context.

For valence priming, each participant was randomly assigned at the start of the session to either a positive or a negative valence condition. 20 images were sampled for each condition (see Section 3.1), meaning each participant would see a new image each trial, but that all participants assigned to a condition would see the same set of images. Participants viewed images exclusively from their assigned condition, presented in randomized order, before each trial.

Finally, on each individual trial, there was a 50% probability that the speech excerpt would be accompanied by a provenance label. Unlike standard "AI-generated" disclosures which often function as neutral warnings or negative signals [37], our label explicitly affirmed a trusted origin ("This originates from a verified source"). This design choice was motivated by arguments that positive provenance cues may be more effective at building cognitive trust than mere disclosure [4, 55]. The label was randomly assigned and had no correlation with whether the audio was genuine or synthetic.

3.6. Platform

The experiment was conducted using Prolific[14], a web-based platform for recruiting research participants. The decision to

[14] https://www.prolific.com

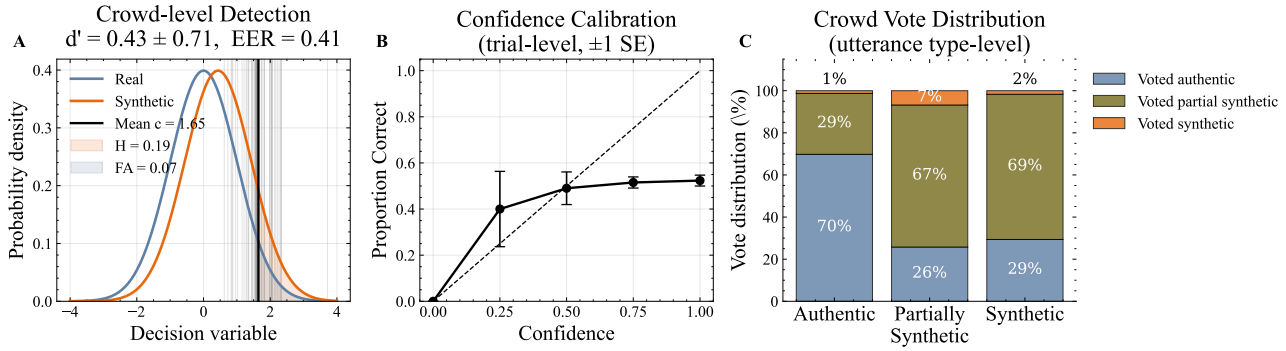

Figure 3: *Crowd-level detection performance. (A) Raw discriminability. Distributions show the internal decision variable for real and fake windows, indicating near-chance discrimination. (B) Confidence calibration. Observed accuracy as a function of self-reported confidence indicates that participants were systematically overconfident. (C) Majority-vote accuracy by utterance type. Authentic and partially synthetic clips were correctly identified in 80% of cases; fully synthetic clips were never correctly classified.*

use a crowdsourcing platform was driven by the need to recruit participants who shared the same linguistic background as the speech dataset, thereby reducing potential effects from non-native speakers [46]. Submissions were monitored and approved on a rolling basis until the participant quota was filled. The experimental task was delivered through a custom Python application implemented in Streamlit[15], a framework for deploying interactive web applications, and was visually designed to evoke the structure of a social media platform (e.g. the priming image presented as the trial thumbnail).

Prior to deployment, an in-house preliminary usability study ($n = 6$) was conducted to assess interface clarity and ease of use. Participants were familiar with spoofed audio content through prior experience. In response to the feedback, the instructions were set to a retrospective annotation paradigm, allowing participants to listen to the audio as many times as needed. An additional pilot study was conducted via Prolific ($n = 5$) to validate end-to-end functionality. Data from this pilot were not included in the analysis and instead used to confirm that the task was feasible within the advertised 40-minute average and to verify data integrity in the web deployment format, with the participants being compensated for their time. Pilot completion times averaged 48 minutes, with the longest submission lasting 76 minutes and 16 seconds. Based on this pilot, additional "sanity checks" were implemented to verify participant engagement, namely the attention-to-task instruction check and confirmation that the full utterance was attended to prior to submitting responses, as described in Section 3.4.

3.7. Participants

The study was conducted online in December 2025 and totaled 152 participants who attempted the study. Submissions were manually validated using pre-established sanity checks (see Section 3.4) to ensure adequate engagement with the task. Inclusion criteria included adults aged 18 or older and normal or corrected hearing. To control for linguistic variability, participants were limited to native English speakers from the United States. No specialized technical or linguistic expertise was required, although participants were informed that the study involved listening to utterances and making authenticity judgments. Of the 50 validated submissions, 3 were excluded from analysis due to missing data, yielding a final sample of n=47 (25 male, 21 female; age range $18-55+, \mu = 41$). Participants with validated submissions were compensated for their time.

[15] `https://streamlit.io`

4. Results

Each utterance was processed by aligning participant marker placements to ground-truth synthetic regions via sliding window scoring (see Figure 1), producing window-level metrics from which an overall verdict per utterance was derived and inter-participant agreement calculated. The results reveal a critical tension in the socio-technical trust framework: while the environmental context was hypothesized to modulate detection sensitivity, utterance type overwhelmingly dictated outcomes.

Utterance Type/ Window Type	Accuracy	Majority Vote	TPR	FPR	TNR	FNR
Authentic	96.4%	4/5	n/a	3.6%	96.4%	n/a
Synthetic	8.3%	0/5	8.3%	n/a	n/a	91.7%
Partially Synth. - Fake	29.3%	n/a	29.3%	n/a	n/a	70.7%
Partially Synth. - Real	88.6%	n/a	n/a	11.4%	88.6%	n/a
Partially Synthetic	56.1%	8/10	29.3%	11.4%	88.6%	70.7%
Overall	**55.8%**	**12/20**	**19.4%**	**7.1%**	**92.9%**	**80.6%**

Table 1: *Window- and trial-level detection performance by utterance type. Fully synthetic speech was never correctly identified at the trial level (0/5).*

Detection performance: Linear mixed-effects models, estimated via maximum likelihood using the Python `statsmodels` package [56], revealed a strong ordinal effect of utterance type on detection accuracy (**H1** supported), with authentic speech yielding highest accuracy and synthetic lowest. The same ordinal pattern held for perceptual quality ratings (**H2** supported): authentic utterances received the highest quality scores, followed by partially synthetic, with fully synthetic rated lowest (see Figure 4).

Trust cue effects: Trust cues exerted no reliable main effects on either detection accuracy or quality evaluation (**H1a** and **H2a** not supported). Influence was confined to interaction terms, suggesting that contextual framing modulates responses only in combination with specific audio or participant characteristics rather than shifting detection performance globally. **Collective decision-making**: aggregating window-level markings revealed low crowd discriminability ($d' = 0.43 \pm 0.71$, EER $= 41\%$, see Figure 3A), with substantial between-person variability. At the trial level, majority voting classified trials as follows:

$$\text{MajorityVote(trial)} = \begin{cases} \text{authentic} & \text{if } c = 0 \\ \text{partially synthetic} & \text{if } 0 < c \leq 70\% \\ \text{fully synthetic} & \text{if } c > 70\% \end{cases} \quad (1)$$

where c denotes coverage (proportion of utterance duration covered by participant-added markers). Trial-level accuracy was 60% (12/20 correct), with average reported confidence of 71.5%. Performance varied substantially by ground-truth utterance type (see Table 1, *Majority Vote*): authentic utterances achieved 80% accuracy (4/5 correct) and partially synthetic 80% (8/10 correct), but fully synthetic speech was *never* correctly classified (0/5 correct). Instead, it was systematically misclassified as partially synthetic in all cases, indicating that participants detected manipulation but consistently underestimated its extent.

Hypothesis	Predictor	Coef	SE	p (adj)	Effect size
H1: Accuracy	Synth. vs Auth.	-0.871	0.020	<.001***	Cohen's d=-4.04, R^2=.67
	Part. vs Auth.	-0.402	0.017	<.001***	Cohen's d=-1.87, R^2=.37
H1a: TPR	Synth. vs Part.	-0.191	0.023	<.001***	Cohen's d=-0.67, R^2=.09
	$I-$	-0.019	0.048	.750	Cohen's d=-0.07
	$V-$	0.040	0.046	.495	Cohen's d=0.14
	$C-$	0.031	0.022	.258	Cohen's d=0.11
H2: Quality Score	Synth. vs Auth.	-1.154	0.070	<.001***	Cohen's d=-1.53, R^2=.23
	Part. vs Auth.	-0.716	0.060	<.001***	Cohen's d=-0.95, R^2=.13
H2a: Quality Score	Synth. vs Auth.	-1.154	0.070	<.001***	Cohen's d=-1.53, R^2=.23
	Part. vs Auth.	-0.717	0.060	<.001***	Cohen's d=-0.95, R^2=.13
	$I-$	-0.137	0.117	.350	Cohen's d=-0.18
	$V-$	0.025	0.112	.824	Cohen's d=0.03
	$C-$	-0.038	0.050	.534	Cohen's d=-0.05

Table 2: *Linear mixed model results, $N = 940$ trials, 47 participants, 20 trials each. All models include a random intercept per participant estimated via maximum likelihood. H1 and H2 show strong effects of utterance type on both detection accuracy and perceived quality. H1a and H2a show no significant effect.*

5. Discussion

Utterance type emerged as the dominant determinant of detection, suggesting that the environmental dimension of trust may fall to a secondary role in real-time detection tasks. This pattern may reflect a mismatch in listener expectations: participants entered the task expecting robotic or background artifacts [6, 18], yet encountered realistic soundscapes and utterances with only subtle word-level manipulations. Such expectation violations could have pulled attention away from contextual cues toward acoustic scrutiny. Evaluative judgments closely tracked audio realism, suggesting that listeners build an internal representational ranking for quality (authentic > partially synthetic > synthetic) that is accessible to conscious evaluation but does not transfer to overt detection decisions. This implicit-explicit dissociation has direct repercussions: attackers may successfully sway into compliance listeners who nonetheless feel something is "off", without that unease surfacing as an actionable response. Participants' engagement metrics offer additional insight into their decision-making patterns (see Figure 2). Participants spending the least time on authentic trials suggests that they can readily recognize real content but are poor at detecting synthetic material. When uncertain, they draw on other information sources (e.g., the trust-related changes in trial duration and revision rates), yet they still perform poorly overall on the task. The combination of below-chance accuracy, overconfidence, and negligible contextual benefits indicates that listeners cannot reliably serve as sole gatekeepers against synthetic speech.

6. Limitations and future directions

Exploring context-driven human decision-making factors is a multidimensional problem. If automated detection studies test the strength of countermeasures [7, 16, 17, 22, 29], with human listener studies exploring themes of detection [6, 7, 18, 46], our study tackled the environment dimension by integrating variables that would otherwise be discarded in protocol development. Containing these factors within a practical platform constrained ecological validity, especially as the explicit task framing primed participants for suspicion and thus likely attenuating the trust cue effects. In naturalistic settings, where synthetic speech is encountered without warning, people's behavior and building of trust would be additionally influenced by their surroundings, disposition, or prior experience with the platform [14, 49]. Research on synthetic speech detection would benefit from in-the-wild designs that study participants who are unaware of its potential presence. More broadly, as acoustic quality continues to improve, research attention should shift from detection performance toward identifying specific risk scenarios that can inform regulatory deterrents [8]. Finally, the stimuli set of 20 utterances limits the generalizability of these findings across the broader landscape of synthesis systems and speaker identities. A beneficial extension of this paradigm would be connecting to ongoing research into speaker trustworthiness and how that is embedded into synthetic speech [57, 58].

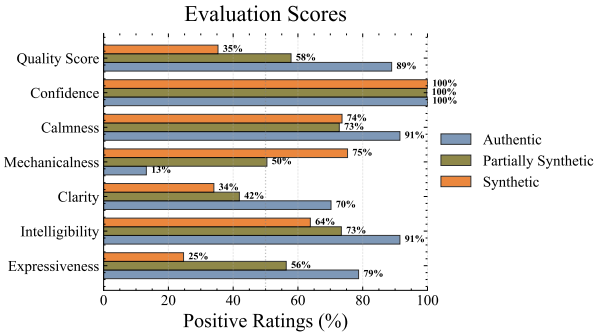

Figure 4: *Percentage of positive ratings ("Agree", "Strongly Agree"). Authentic speech consistently received the highest subjective ratings across all dimensions, while fully synthetic speech received the lowest. Note that mechanicalness is presented as rated (higher = more mechanical), such that lower scores indicate higher perceived quality for this dimension.*

7. Conclusion

This study challenges the "one solution fits all" approach to synthetic speech detection by investigating (partial) voice deepfake detection as a *socio-technical process shaped by trust*. Although participants expressed confidence in their judgments, they were generally unable to detect synthetic speech, with detection performance improving only when authentic speech served as an acoustic anchor for suspicion as in the case of partial synthetic utterances. Contextual trust cues, intended to modulate suspicion, did not significantly alter detection outcomes. As such, it seems that the unreliable, naive listener,

instead of being supported by additional information, may experience cognitive overload rather than improved accuracy. More broadly, if moving towards ecological validity itself drives complexity beyond what naive listeners can manage, the expectation of routine human detection in an increasingly synthetic media landscape warrants serious reconsideration.

If these results generalize, the amount of synthetic speech that passes the perceptual threshold is likely far greater than reported, and more so as people engage with it as they would with authentic speech. If this entanglement of real and synthetic continues, the assignment of truth will become increasingly fuzzy. Using automated decision systems as support will then have to be tailored to specific user needs and affordances.

The path forward seems to point toward cooperative architectures that distribute detection across both human judgment and technical countermeasures. Such joint systems should handle uncertainty rather than demand binary verdicts, and account for the overconfidence and response biases, some of which are documented here. Building such systems will require a clearer understanding of the conditions under which people reason, doubt, and decide, as well as how the platforms they frequent can be designed to support, rather than undermine, that process.

8. Acknowledgments

This work was carried out as part of the VoCS (Voice in Communication Sciences) doctoral network, funded by the European Union's Horizon Europe Framework programme under Grant Agreement No 101168998 and partially supported by the Academy of Finland (Decision No. 349605, project "SPEECHFAKES"). This study was submitted to and received no objection from the ethics committee of the University of Eastern Finland. We thank the reviewers for their constructive feedback, and the participants and volunteers whose involvement made this work possible.

9. References

[1] M. B. Johansen, "Public intellectuals on new platforms: constructing critical authority in a digital media culture," in *Rethinking cultural criticism: New voices in the digital age*. Springer, 2020, pp. 17–42.

[2] J. Nehring, A. Gabryszak, P. Jürgens, A. Burchardt, S. Schaffer, M. Spielkamp, and B. Stark, "Large language models are echo chambers," in *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, 2024, pp. 10 117–10 123.

[3] J. Myers, "Attention economy theory," in *Media Ecology for the 21st Century: Theories of Culture, Communications, and Consciousness*. Springer, 2025, pp. 101–109.

[4] K. R. Harris, "Synthetic media detection, the wheel, and the burden of proof," *Philosophy & Technology*, vol. 37, no. 4, p. 131, 2024.

[5] O. Laas, "Deepfakes and trust in technology," *Synthese*, vol. 202, no. 5, p. 132, 2023.

[6] K. Warren, T. Tucker, A. Crowder, D. Olszewski, A. Lu, C. Fedele, M. Pasternak, S. Layton, K. Butler, C. Gates *et al.*, "" 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*, 2024, pp. 2696–2710.

[7] A. Alali and G. Theodorakopoulos, "Partial fake speech attacks in the real world using deepfake audio," *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, p. 6, 2025.

[8] Z. H. Bhatti, B. Ahtisham, S. Tausif, N. George, M. Javed *et al.*, "Can you tell it's ai? human perception of synthetic voices in vishing scenarios," *arXiv preprint arXiv:2602.20061*, 2026.

[9] G. Fan and D. Liu, "When machines speak with feeling: Investigating emotional prosody, authenticity, and trust in ai vs. human voices," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 47, 2025.

[10] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[11] T. Buchanan, "Why do people spread false information online? the effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation," *Plos one*, vol. 15, no. 10, p. e0239666, 2020.

[12] K. R. Harris, "Ai or your lying eyes: Some shortcomings of artificially intelligent deepfake detectors," *Philosophy & Technology*, vol. 37, no. 1, p. 7, 2024.

[13] E. K. Chiou, N. L. Schroeder, and S. D. Craig, "How we trust, perceive, and learn from virtual humans: The influence of voice quality," *Computers & Education*, vol. 146, p. 103756, 2020.

[14] K. M. Scott, S. Ashby, and J. Hanna, "" human, all too human": Noaa weather radio and the emotional impact of synthetic voices," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–9.

[15] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[16] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[17] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[18] S. Amirkhani, G. Stevens, M. Shajalal, and A. Boden, "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), 2025.

[19] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of management annals*, vol. 14, no. 2, pp. 627–660, 2020.

[20] S. Lockey, N. Gillespie, D. Holm, and I. A. Someh, "A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions," 2021.

[21] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, p. 113368, 2023.

[22] X. Li, P.-Y. Chen, and W. Wei, "Where are we in audio deepfake detection? a systematic analysis over generative and detection models," *ACM Transactions on Internet Technology*, 2025.

[23] K. Somoray, D. J. Miller, and M. Holmes, "Human performance in deepfake detection: a systematic review," *Human Behavior and Emerging Technologies*, vol. 2025, no. 1, p. 1833228, 2025.

[24] K. M. Bezio, "Ctrl-alt-del: Gamergate as a precursor to the rise of the alt-right," *Leadership*, vol. 14, no. 5, pp. 556–566, 2018.

[25] R. Baños, "Key challenges in using automatic dubbing to translate educational youtube videos," *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, vol. 22, 2023.

[26] S. Pan, L. Fortunati, and A. Edwards, "Grooming an ideal chatbot by training the algorithm: Exploring the exploitation of replika users' immaterial labor," *New Media & Society*, vol. 27, no. 10, pp. 5489–5507, 2025.

[27] M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky, "Sycophantic ai decreases prosocial intentions and promotes dependence," *arXiv preprint arXiv:2510.01395*, 2025.

[28] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[29] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 513–566, 2023.

[30] L. Krikheli, S. El-Wahsh, and R. Cave, "It's not just "press record": a viewpoint for providing ethical voice banking," *Evidence-Based Communication Assessment and Intervention*, pp. 1–10, 2026.

[31] A. Mahapatra, I. R. Ulgen, A. R. Naini, C. Busso, and B. Sisman, "Can emotion fool anti-spoofing?" *arXiv preprint arXiv:2505.23962*, 2025.

[32] S. Layton, T. De Andrade, D. Olszewski, K. Warren, C. Gates, K. Butler, and P. Traynor, "Every breath you don't take: Deepfake speech detection using breath," *Digital Threats: Research and Practice*, vol. 6, no. 3, pp. 1–18, 2025.

[33] D. De Kerckhove and C. M. De Almeida, "What is a digital persona?" *Technoetic Arts: A Journal of Speculative Research*, vol. 11, no. 3, pp. 277–287, 2013.

[34] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," in *Speech Prosody 2014*, 2014.

[35] G. Loewenstein, J. S. Lerner *et al.*, "The role of affect in decision making," *Handbook of affective science*, vol. 619, no. 642, p. 3, 2003.

[36] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2823–2832.

[37] D. Gamage, D. Sewwandi, M. Zhang, and A. K. Bandara, "Labeling synthetic content: User perceptions of label designs for ai-generated content on social media," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–29.

[38] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "Llama-partialspoof: An llm-driven fake speech dataset simulating disinformation generation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[39] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[40] Y. Zhang, X. Ding, and N. Gu, "Understanding fatigue and its impact in crowdsourcing," in *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 2018, pp. 57–62.

[41] A. Mitchell, T. Oberman, F. Aletta, M. Erfanian, M. Kachlicka, M. Lionello, and J. Kang, "The international soundscape database: An integrated multimedia database of urban soundscape surveys–questionnaires with acoustical and contextual information," *Online. URL: https://www. zenodo. org/record/5914715# YnwwoGDP00Q*, 2021.

[42] A. Bockstael, L. Samyn, P. Corthals, and D. Botteldooren, "Presenting and processing information in background noise: A combined speaker–listener perspective," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 210–218, 2018.

[43] B. Kurdi, S. Lozano, and M. R. Banaji, "Introducing the open affective standardized image set (oasis)," *Behavior research methods*, vol. 49, no. 2, pp. 457–470, 2017.

[44] J. J. Wehrman, J. Wearden, and P. Sowman, "Decisional carry-over effects in interval timing: Evidence of a generalized response bias," *Attention, Perception, & Psychophysics*, vol. 82, no. 4, pp. 2147–2164, 2020.

[45] E. Cooper, S. L. Maguer, E. Klabbers, and J. Yamagishi, "Good practices for evaluation of synthesized speech," *arXiv preprint arXiv:2503.03250*, 2025.

[46] E. San Segundo, A. López-Jareño, X. Wang, and J. Yamagishi, "Human perception of audio deepfakes: the role of language and speaking style," *Available at SSRN 5954496*, 2025.

[47] Y. Ma and Y. Tang, "The intelligibility benefits of modern computer-synthesized speech for normal-hearing and hearing-impaired listeners in non-ideal listening conditions," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 5, no. 1, p. 5, 2024.

[48] O. Sokol, H. Naumenko, V. Derkach, V. Kuznetsov, D. Progonov, and V. Husiev, "Automatic speaker verification on compressed audio," in *2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE, 2022, pp. 1–7.

[49] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience," 04 2018.

[50] K. Scott, S. Ashby, and J. Hanna, ""human, all too human": Noaa weather radio and the emotional impact of synthetic voices," 04 2020, pp. 1–9.

[51] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, "Unmasking illusions: Understanding human perception of audiovisual deepfakes," *arXiv preprint arXiv:2405.04097*, 2024.

[52] J. Toner, B. G. Montero, and A. Moran, "The perils of automaticity," *Review of General Psychology*, vol. 19, no. 4, pp. 431–442, 2015.

[53] R. F. Adler and R. Benbunan-Fich, "The effects of task difficulty and multitasking on performance," *Interacting with Computers*, vol. 27, no. 4, pp. 430–439, 2015.

[54] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013.

[55] S. A. Fisher, "Something ai should tell you–the case for labelling synthetic content," *Journal of Applied Philosophy*, vol. 42, no. 1, pp. 272–286, 2025.

[56] S. Seabold, J. Perktold *et al.*, "Statsmodels: econometric and statistical modeling with python." *scipy*, vol. 7, no. 1, pp. 92–96, 2010.

[57] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience," in *Proceedings of the technology, mind, and society*, 2018, pp. 1–6.

[58] C. Maltezou-Papastylianou, R. Scherer, and S. Paulmann, "How do voice acoustics affect the perceived trustworthiness of a speaker? a systematic review," *Frontiers in Psychology*, vol. 16, p. 1495456, 2025.