Detecting the Undetectable: Human Judgments and the Challenge of Synthetic Voices

SIMA AMIRKHANI, H-BRS, Germany
GUNNAR STEVENS, H-BRS, Germany
MD SHAJALAL, Siegen University, Germany
ALEXANDER BODEN, H-BRS, Germany

Synthetic voices generated using artificial intelligence (AI) are becoming increasingly indistinguishable from human voices, raising important concerns about trust, deception, and detection in digital communication. This preliminary work synthesizes the current landscape of research on human perception in detecting synthetic voices. We reviewed 13 papers from databases including ACM, IEEE, Springer, and MDPI, and identified five main types of perceptual cues that users rely on to detect voice synthesis: Intuition/Gut Feeling, Liveliness, Emotions, Linguistic Features, and Acoustic and Environmental Features. Our findings highlight the need for further empirical user studies to better understand how individuals perceive and assess the risks posed by synthetic voices. Such research can inform both educational and regulatory strategies aimed at increasing awareness and mitigating the potential harms of synthetic voice technologies.

CCS Concepts: • **Security and privacy** → Human and societal aspects of security and privacy.

Additional Key Words and Phrases: Synthetic voices, voice deepfakes, human perception, AI-generated speech, detection cues, deception, user studies

1 LITERATURE REVIEW: HUMAN DETECTION OF SYNTHETIC VOICES

The rapid advancement of synthetic voice technologies, particularly deepfake and text-to-speech (TTS) systems, has outpaced society's ability to detect and regulate their misuse. As synthetic voices grow increasingly indistinguishable from authentic speech, understanding human perceptual limits has become a pressing area of research. A growing body of user studies (See table 1) has investigated how well people can detect synthetic voices and which perceptual features they rely on in making such judgments.

1.1 Detection Accuracy: Human Struggles with Synthetic Voice

Most studies converge on a consistent finding: humans are generally poor at reliably distinguishing synthetic voices from real ones. Detection accuracy tends to hover near or below chance level, especially when participants are exposed to high-quality deepfake audio. For instance, Barrington et al. (2025) [3] found that even with over 600 participants, human detection hovered around 60%, and participants frequently mistook synthetic voices for real ones. Similarly,

Authors' addresses: Sima Amirkhani, H-BRS, Germany, sima.amirkhani@h-brs.de; Gunnar Stevens, H-BRS, Germany, gunnar.stevens@h-brs.de; MD Shajalal, Siegen University, Germany, Md.Shajalal@uni-siegen.de; Alexander Boden, H-BRS, Germany, alexander.boden@h-brs.de.

Sima Amirkhani, Gunnar Stevens, MD Shajalal, 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), July 21–23, Siegen, Germany. EUSSET. https://doi.org/10.48340/ct2025-1030

Barnekow et al. (2021) [2] reported only 37% accuracy, highlighting the challenge of discerning synthetic speech, especially when familiarity with the speaker is low.

Table 1. Sorted summary of user studies on human detection of synthetic voices.

#	User Study	Year	Sample Size	Population	Result	Pub.
1	Barrington [3]	2025	604	Participants from Prolific platform	Human perception is no longer a consistently reliable method for detecting AI-generated voice clones.	Others
2	Alali and Theodor-akopoulos [1]	2025	148	Native English speakers from UK and US via Prolific	Very low detection rates; Partial manipulation much harder to detect.	MDPI
3	Frank et al. [5]	2024	3002	Cross-national study with diverse participants	Audio was hardest to detect across all media types.	IEEE
4	Groh et al. [6]	2024	2215	Participants in 5 RCTs	Audio-only modality had lowest accuracy. Familiarity and exposure mattered.	Springer
5	Warren [14]	2024	150	General public	Humans outperformed AI but relied on flawed heuristics (e.g., background noise = real).	ACM
6	Li et al. [8]	2024	68	Bilingual participants (English and Chinese)	Semantic content unrecoverable, yet fake detection still feasible.	ACM
7	Tsoi [12]	2024	32	Swedish-speaking individuals	Weak and statistically insignificant correlation between ability to distinguish real vs. fake voices.	Others
8	Sharevski [11]	2024	16	Blind/low-vision participants	Blind users focused on intensity/imperfections; low-vision on pitch/emotion.	ACM
9	Han [7]	2024	12	Blind/low-vision participants	Sighted users relied on semantic coherence; blind users on acoustic anomalies. Trait differences linked to accuracy.	ACM
10	Mai [9]	2023	529	Participants; English & Mandarin samples	73% accuracy. Training marginally helped. Mandarin samples were harder to detect.	Others

#	User Study	Year	Sample Size	Population	Result	Pub.
11	Bhalli et al. [4]	2023	264	Students	Training improved deepfake detection skill measurably.	Others
12	Müller et al. [10]	2022	410	Gamified real/fake detection task	Native speakers had better accuracy. Older users struggled.	ACM
13	Barnekow et al. [2]	2021	100	Students	Only 37% accuracy. Unfamiliarity with voice reduced performance.	Springer

Several studies note that detection performance is contingent on multiple factors, including language, training, and demographic variables. Mai et al. (2023) showed that Mandarin samples were harder to detect than English ones, even after training. While Bhalli et al. (2024) and Müller et al. (2022) [4, 10] found that training interventions slightly improved detection capabilities, the gains were modest and often plateaued quickly. This suggests that while education can help, it cannot fully compensate for the sophistication of current voice synthesis systems.

Interestingly, detection performance also varies based on user characteristics. Tsoi et al. (2024) [12] observed that familiarity with a speaker increased accuracy, while Han et al. (2024) [7] revealed that blind users, despite lacking visual cues, utilized distinct auditory heuristics to detect synthetic speech—often focusing on pitch irregularities or emotional inconsistencies. Sighted users, by contrast, leaned more heavily on semantic coherence and background context.

Across studies, synthetic audio was consistently rated as the most difficult media modality to evaluate when compared to text, video, or image-based deepfakes [5, 6]. Taken together, the literature paints a sobering picture: human intuition alone is not a reliable defense against synthetic audio deception.

1.2 Perceptual Features: Auditory Cues Used by Humans

Beyond accuracy, several qualitative studies have explored the perceptual features that listeners rely on when evaluating whether a voice is real or artificial. While terminology across studies varies, we identified recurring cues and synthesized them into five overarching categories (See Figure 1): *Intuition, Liveliness, Emotions, Linguistic Features, and Acoustic/Environmental Characteristics* (see Table 2).

Human ability to detect synthetic voices is limited and highly inconsistent. While listeners do rely on a rich array of perceptual cues—from intuitive impressions to specific acoustic anomalies—these are often insufficient against sophisticated deepfakes. Furthermore, heuristics like "no background noise means fake" can lead to incorrect conclusions. Training offers some benefits, but as the realism of synthetic audio continues to improve, reliance on human perception alone may be untenable.

Human ability to detect synthetic voices is limited and highly inconsistent. While listeners do rely on a rich array of perceptual cues—from intuitive impressions to specific acoustic anomalies—these are often insufficient against sophisticated deepfakes. Furthermore, heuristics like "no background noise means fake" can lead to incorrect conclusions. Training offers some benefits, but as the realism of synthetic audio continues to improve, reliance on human perception alone may be untenable.

Human ability to detect synthetic voices is limited and highly inconsistent. While listeners do rely on a rich array of perceptual cues—from intuitive impressions to specific acoustic anomalies—these are often insufficient against sophisticated deepfakes. Furthermore, heuristics like "no background noise means fake" can lead to incorrect conclusions.

Table 2. Heuristic Cues Used by Listeners to Detect Synthetic Voices

Heuristic Cue	Description
Intuition / Gut Feeling	Listeners often form impressions based on vague or hard-to-articulate cues. Terms like "robotic," "natural," or "human-sounding" recur across studies [7, 10, 11]. Judgments are rooted in familiarity and expectation rather than detailed analysis.
Liveliness	Breathing, pausing, inflection, and rhythm were commonly cited as indicators of real speech [12, 14]. Their awkward presence or absence suggested artificiality. Signs of "liveness" like inconsistent pacing or vocal tremors were seen as hallmarks of authenticity.
Emotions	Emotional expressiveness was the most frequently mentioned cue. Synthetic voices were perceived as flat or exaggerated. Real voices conveyed subtle emotions—like sadness or frustration—more convincingly, which contributed to perceived credibility [6, 7].
Linguistic Features	Odd phrasing, missing fillers, unusual accents, and overly fluent delivery signaled synthetic origin [7, 12]. Mispronunciations and unnatural sentence flow were especially suspicious in multilingual contexts.
Acoustic and Environmental Features	Listeners noted background noise and sound quality. Clean or static-laden audio was seen as potentially fake, although attackers may now embed such artifacts to simulate realism [14].

Training offers some benefits, but as the realism of synthetic audio continues to improve, reliance on human perception alone may be untenable.

2 WORK-IN-PROGRESS AND FUTURE WORK

Our ongoing work aims to bridge the identified gaps by synthesizing existing findings and laying the groundwork for future empirical studies on human-centered audio deepfake detection. The current stage of our project involves the development of a typology that categorizes the perceptual cues and contextual factors humans use when assessing the authenticity of synthetic voices (See Figure 1). This typology will guide future experiments and contribute to a more systematic understanding of human perception in audio forgery contexts.

As an initial empirical step, we have begun conducting a user study based on semi-structured interviews. 17 Participants are asked to listen to five voice samples—both real and synthetic—generated using state-of-the-art text-to-speech systems. During the interviews, we explore the perceptual cues they use to detect inauthenticity, their perceptions of risks associated with synthetic voices (e.g., misinformation, impersonation), and the personal or technological strategies they believe could mitigate these risks. This qualitative phase complements our literature review by grounding the typology in real-world detection strategies and broadening our understanding of lay users' trust judgments and vulnerabilities.

We are also working on designing a set of controlled user studies to investigate how individual and contextual variables—such as age, gender, familiarity with the speaker, and emotional content—influence the detection of manipulated speech. These studies will explore both intuitive and deliberate decision-making strategies, aiming to uncover the cognitive mechanisms that underlie human judgments of authenticity in synthetic voices. As a future work, we aim to develop design implications for creating user-centered detection interfaces that enhance rather than replace human

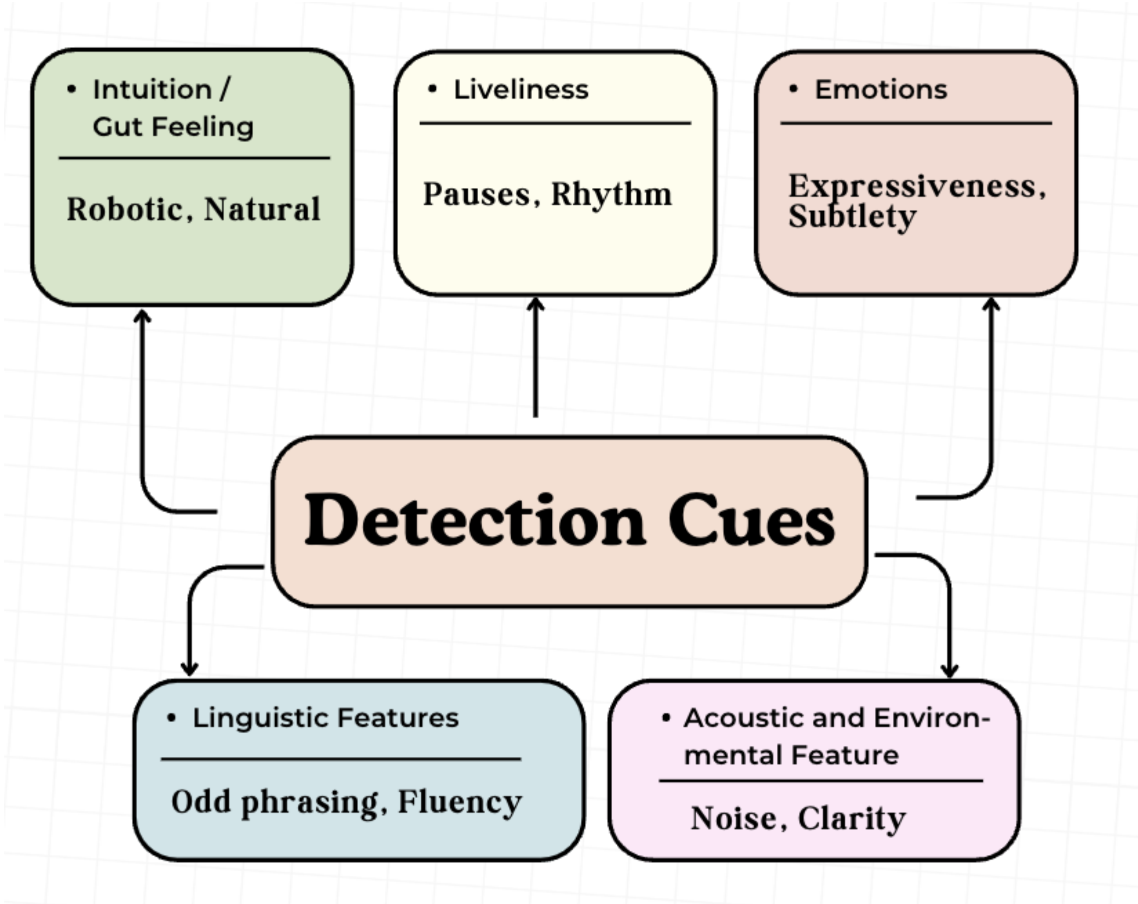

Fig. 1. Heuristic cues commonly used by listeners to assess synthetic voice authenticity.

judgment. These interfaces might incorporate explainable AI (XAI) feedback, multi-modal verification cues, or risk communication strategies that align with human perceptual strengths.

Ultimately, our goal is to contribute to a more holistic approach to synthetic media detection—one that combines technical robustness with human-centered design principles. Future work will emphasize cross-cultural studies, longitudinal assessments of trust, and the development of educational interventions that can improve media literacy and increase public resilience to audio-based manipulation.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the AntiScam project [13] (Grant NO. 16KIS2214).

REFERENCES

[1] 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.

[2] Vanessa Barnekow, Dominik Binder, Niclas Kromrey, Pascal Munaretto, Andreas Schaad, and Felix Schmieder. 2021. Creation and detection of German voice deepfakes. In *International Symposium on Foundations and Practice of Security*. Springer, 355–364.

[3] Farid H. Barrington, S. 2024. People are poorly equipped to detect AI-powered voice clones. *arXiv preprint arXiv:2410.03791* (2024).

[4] Noshaba N 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. *arXiv preprint arXiv:2411.14586* (2024).

[5] 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 *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 55–73.

[6] 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 (2024), 7629.

[7] Mitra P. Billah S. M. Han, C. 2024. Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), 1–14.

[8] Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. 2024. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 3585–3599.

[9] Bray S. Davies T. Griffin L. D. Mai, K. T. 2023. Warning: Humans cannot reliably detect speech deepfakes. *Plos One* 18, 8 (2023), e0285333.

[10] Nicolas M Müller, Karla Pizzi, and Jennifer Williams. 2022. Human perception of audio deepfakes. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*. 85–91.

[11] Zeidieh A. Loop J. V. Jachim P. Sharevski, F. 2024. Blind and Low-Vision Individuals' Detection of Audio Deepfakes. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. 4867–4881.

[12] Gobl I. Tsoi, A. 2024. Listening Between The Lines: Investigating The Correlation Between Voice Recognition and Audio Deepfake Differentiation. (2024).

[13] Hochschule Bonn-Rhein-Sieg Universität Siegen, open.INC GmbH. 2024. AntiScam: Abwehr von Kommunikationsbetrug zum Schutz der digitalen Identität. https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/antiscam. Accessed: 2025-02-28.

[14] Tucker T. Crowder A. Olszewski D. Lu A. Fedele C. ... Traynor P. Warren, K. 2024. "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. 2696–2710.