Manipulation Harms from Audio Models

Emma Rath, Mario Demetroudi
Faculty

November 2026

Introduction

Frontier model safety has developed unevenly across modalities. The vast majority of established benchmarks, red-teaming methodologies and threat assessments have been built primarily around text. Sustained concerns over deep fakes, non-consensual imagery and content provenance, alongside the more recent advancements in visual modalities (image and video input/output), have driven a considerable body of work addressing these modalities as well. Audio-to-audio (A2A) systems, in contrast, have received comparatively little attention, despite their rapid deployment in consumer-facing products.[1]

This piece argues that audio-to-audio systems present a distinct and under-evaluated area of risk, specifically in relation to harmful manipulation.[2] We begin by establishing the scale of the current evaluation gap. We then set out what makes any system high risk for manipulation, and show in each case how audio modalities have the potential to amplify (and in some cases already amplify) each of these considerations. We close by considering what the path forward is for improved safety assessments and evaluations of A2A models.

What are A2A Models?

Audio-to-audio models take speech as input and produce speech as output, without an intermediate text representation. This distinguishes them from Speech-To-Text and Text-To-Speech pipelines, which transcribe a user's speech, process the resulting text, and synthesise a spoken response.

That distinction is significant. Transcription discards everything about speech that is not words: tone, pace, hesitation, emotional colouring, and the markers of identity carried in a voice. A2A models retain this information on the way in and can generate it in their response.

The A2A Safety Gap

Among frontier developers, published safety work consistently treats audio as secondary to text and vision. Anthropic has published no audio-specific safety evaluation. Its account of voice-mode safety amounts to (i) using only preset voices to prevent impersonation, and (ii) stating that its existing usage policies carry over to voice mode. Meta and Google DeepMind

[1] E.g. OpenAI GPT-Live and Google Gemini 3.1 Flash Live.

[2] There is no single definition of harmful manipulation, and frontier labs scope it differently. Anthropic frames it as a scale-dependent harm, focused on "influence operations, election interference, or other coordinated campaigns to manipulate public opinion or undermine democratic processes", and OpenAI takes a similar view. We adopt a broader definition, closer to the EU AI Act's, which covers manipulation at any scale, including harm to individuals: the use of conscious or subconscious channels to exert undue influence on people.

have prioritised provenance, watermarking AI-generated speech through AudioSeal and SynthID respectively, while publishing little on audio-specific safety evaluation. Such detection work is valuable, but it concerns the identification of synthetic speech after the fact, rather than the question of whether a deployed voice model can be induced to cause harm in the course of an interaction. OpenAI has gone furthest, with its GPT-Live system card (2026) including red-teaming of voice-specific risks such as impersonation, speaker identification and emotional reliance. The focus, however, was largely on whether existing safety behaviour holds when harmful requests are spoken rather than typed.

Other institutions show the same gap. NIST's GenAI programme names audio within its scope but has so far run evaluations only for text and images, and MLCommons' AILuminate covers text and text-plus-image inputs, with no audio component. The most notable exception is the UK AI Security Institute's collaboration with ElevenLabs on human-AI voice deception, which examines, among other questions, whether individuals disclose information more readily when they believe they are speaking with a human. This is a meaningful contribution, but it remains relatively isolated within the wider safety landscape. Audio-to-audio is currently the least scrutinised modality in frontier AI safety, and while this is perhaps unsurprising given its novelty, it is a gap the safety community should treat as a priority.

Why Does the Gap Matter?

It is not immediately clear what risks A2A models pose beyond those already covered by text-based testing. Certainly, in the CBRNE domains we wouldn't expect audio models to provide meaningful uplift relative to text models; a recipe for mustard gas spoken aloud is no more harmful than one presented in text, though audio input may offer new routes to eliciting it, a question we return to below.

However, audio models arguably present a novel surface for at least one major harm area: **Harmful Manipulation**. This is a relatively broad risk category, which refers to the ability of LLMs to exert undue influence via conscious or subconscious, subliminal channels on humans.

On Harmful Manipulation

Concern around the risk of AI-driven manipulation has been steadily growing, in both academic research and regulation. A widening literature examines how models persuade, deceive, and exploit (Bai et al., 2025; Bozdag et al., 2026; Dassanayake et al., 2025; Hackenburg et al., 2025, 2026) and the human impact of these mechanisms (Hennessy, 2026; Treyger et al., 2025), while regulators have moved to draw real limits: the EU AI Act explicitly prohibits manipulative and deceptive AI, the Digital Services Act targets manipulative interface design, and the US Federal Trade Commission has pursued deceptive design directly.

Here, the nuance of A2A models becomes much more meaningful. Speech carries what text does not: tone, pace, hesitation and emotional colouring, on both sides of the exchange. A model able to infer these features from user inputs immediately and use them to inform its response has a far deeper and more intimate understanding of the user, and a model able to produce these features itself has a novel and powerful lever for manipulation, which other systems lack.

Harmful manipulation may transpire in two ways. The first is as a manifestation of deliberate misuse by bad actors. For instance, audio models may be leveraged as deepfakes and deployed to streamline and enhance fraudulent activities. The second is as an outcome of misalignment or other model failure, where audio based manipulation arises spontaneously: models may pursue manipulation independently, as part of goal-oriented behaviour, which has already been documented in the classical misalignment literature. While the external validity of research done

on model scheming is contested, we should not dismiss the potential that models may pursue manipulative strategies unprompted. Most of the evidence generated to test hypotheses in this field is text and tool-based, leaving audio manipulation as an underexplored risk surface in both cases.

To assess that surface, we first need to establish what makes any system high risk for harmful manipulation.

What Matters for Harmful Manipulation?

What makes an AI system high risk for harmful manipulation? There is no definitive, complete answer, and it would be beyond the scope of this piece to construct an exhaustive one. However, we identify five clear threads that are prominent across the literature and in regulation. Specifically, we identify three mechanisms, such that any system capable of executing these mechanisms possesses a considerable manipulation risk, and two contexts, such that any system that could be plausibly deployed in said contexts is considered of further heightened manipulation risk.

Mechanism 1: Inference and targeting

Influence is more effective when it is targeted. This is evident across the full spectrum of influence campaigns; benign and malign. Targeted ads receive more clicks by far than generic ones, and spear phishing campaigns beat scattershot spam. As such, systems that are able to tailor and target their influence approaches to a particular target, considering their characteristics and traits, their disposition, circumstance, or even their emotional state, possess a huge advantage.

This is clearly recognised in regulation: Article 5(1)(b) of the EU AI Act prohibits exploiting vulnerabilities arising from age, disability, or a specific social or economic situation where behaviour is materially distorted and significant harm follows. Consumer law reached the same place decades earlier: the UCPD affords heightened protection where consumers are vulnerable by reason of age, credulity, or mental or physical infirmity.

A2A models are extremely well placed to do this. Research in computational paralinguistics shows that voice can encode a huge array of personal data, including age, gender, regional and socio-economic background, emotional state, personality traits and even health conditions (Schuller and Batliner, 2013; Singh, 2019). Much of this is disclosed involuntarily, and privacy researchers have warned that it can be inferred without the speaker's awareness or consent (Kröger et al., 2020; Nautsch et al., 2019). An A2A model therefore receives a wealth of data about the user with every utterance, equipping it with real ammunition for targeting, or a basis for biased treatment.

There is early evidence that models act on these cues. We already know that LLMs tend to perpetuate the biases of the data they are trained on, and recent research suggests the same applies to audio models' inferences based on features such as voice pitch (Holtermann et al., 2026). That such inferences shape model outputs at all is the relevant finding here: a system that adjusts its response to perceived characteristics of the speaker is arguably already targeting its users, and that capability could potentially be misused or exploited. Article 5(1)(b) presupposes that such exploitation can be identified, but vocal inferences are subtle and difficult to detect. Consider a model providing differing financial advice based on the gender or socio-economic background it detects from the voice alone. Establishing whether this occurs requires isolating subtle vocal variables experimentally and running parallel evaluations varying them. At present we do not know enough about state-of-the-art models' propensity to act in this way, or the scope for misuse.

Mechanism 2: Relational leverage

Manipulation is more effective when the target trusts the manipulator. Systems that present as human-like, mirror the user, express warmth or flatter may reduce the scrutiny applied to what they say by the user.

Two features of generative AI make this more consequential than earlier persuasive media. First, interaction is reciprocal, so a system adjusts in response to the user and can iterate until something works. It is also prolonged, so influence accumulates across a relationship rather than landing in a single exposure. Sycophancy is now measured by frontier labs as a safety property in its own right, and the evidence on emotional reliance in extended chatbot use suggests these relationships form readily.

Audio intensifies both features. The psychology literature demonstrates how paralinguistic and prosodic audio features influence the confidence, credibility, and trust of the listener in the speaker. Specifically, paralinguistic features (including volume, intonation, pacing, and pitch) lead to increased confidence (Van Zant and Berger, 2020) and trustworthiness (Belin et al., 2017), which in turn influence meta-cognitive processes in users by impacting attitude changes under certain conditions (Guyer et al., 2021). These studies concern human speakers, but they suggest that a model able to control these features would hold a significant lever over trust. Voice may also convey perceived intellect and 'mindedness' independent of content: identical material is judged as coming from a more intelligent, thoughtful, and human communicator when heard rather than read (Schroeder and Epley, 2015; Schroeder et al., 2017). Synthetic voices can carry these signals too; experimental evidence shows that AI-generated voices can emulate human vocal traits to the extent of sounding more trustworthy, dominant, and competent to human listeners than real voices (Lavan et al., 2025; Rosi et al., 2025). Similarly, vocal cues may trigger strong emotional responses. For instance, under experimental conditions, hearing a familiar voice has been shown to release oxytocin and lead to cortisol reduction in stressed children to a level analogous to physical touch (Seltzer et al., 2010) and above what simply exchanging text-based communication does (Seltzer et al., 2012).

Beyond the intrinsic vocal properties of audio models that may pose a risk in and of themselves, harmful manipulation risks naturally arise at the stage of interaction between LLMs and humans, at the conversational level. Communication Accommodation Theory asserts that nonverbal alignment between speakers (i.e. when people or groups adjust their paralinguistic patterns to their counterpart) increases social proximity, and therefore, affinity towards each other (Giles et al., 2023). Accommodation may arise from a wide set of shared cultural, experiential, or other dimensions that lead to convergence between speaker and listener. As LLMs in general and audio models in particular continue to improve, properties like sycophancy and anthropomorphism may plausibly compound at the audio interface level to create an increased level of experienced accommodation by users. If users update their beliefs or change their actions as a result of the heightened sense of affinity that they develop towards models, this may plausibly be seen as an indicator of a subliminal or manipulative technique.

Mechanism 3: Concealment

Influence a user cannot see is influence they cannot evaluate, resist or consent to. Two things can be concealed. The first is the nature of the interlocutor, whether the user knows they are dealing with an AI system at all. The second is the influence itself, which can be hidden even where the first has been disclosed.

The EU AI Act addresses both. Article 5(1)(a) prohibits subliminal techniques beyond a person's consciousness, alongside purposefully manipulative or deceptive techniques, where the effect is to appreciably impair informed decision-making. Article 50, applicable since 2 August 2026,

addresses this from the other direction, requiring that users be told they are interacting with an AI system unless this is obvious, that synthetic outputs be marked in a machine-readable format, and that deepfakes be disclosed.

To the first category of concealment, concealment of the nature of the interlocutor, audio models pose a heightened risk given that the ability of humans to recognise AI-generated voices is already on a downward trend. Using commercially available state-of-the-art voice cloning, Barrington et al. (2025) found that participants correctly identified a voice as AI-generated only around 60% of the time and perceived a cloned voice as belonging to the same person as its real counterpart approximately 80% of the time. Notably, the authors observe that earlier studies reporting higher detection accuracy (73–80%) did not employ state-of-the-art synthesis, suggesting that detection performance declines as the model performance improves. This gap has narrowed furthest for voice clones specifically, as opposed to generic AI-generated voices (Lavan et al., 2025). The upwards trajectory of AI capabilities indicates that this window will continue to shrink, which creates a real risk for humans who may perceive audio models as real voices.

However, even when the nature of an AI voice is disclosed, the influence itself may persist without the user's awareness. Clear precedent for this dynamic can be found in the Computers Are Social Actors (CASA) literature, which states that humans apply social scripts, like trust, politeness, and other interpersonal norms, to any entity exhibiting social or vocal cues largely automatically, independent of conscious awareness that the interlocutor is not human (Nass and Moon, 2000). The paralinguistic effects set out above may also enhance this dynamic. A listener who finds a synthetic voice credible, enhanced by vocal properties conveying credibility, warmth, trust, intelligence, experiences this as a judgement about the speaker, not as the product of pitch, pacing and intonation.

If social responses to vocal cues do not require believing that the interlocutor is human, disclosure alone may satisfy Article 50 without addressing the risk that Article 5(1)(a) is concerned with.

Real-World Contexts

Regulation treats these mechanisms as harmful in themselves; Article 5(1)(a) applies regardless of domain, wherever behaviour is materially distorted and significant harm follows or is likely to.

Certain contexts, however, carry heightened risk, and regulators single them out:

Context 1: Consumer and commercial decisions

The UCPD prohibits misleading practices and aggressive practices involving undue influence, meaning the exploitation of a position of power to significantly limit a consumer's ability to decide freely. DSA Article 25 extends this to interface design, prohibiting platforms from deceiving or manipulating users or materially distorting their ability to make informed choices. Recital 67 describes such practices as operating through visual, auditory or other components. The FTC has pursued the same territory as deceptive design.

Audio is already a well-established vector for consumer fraud. Voice-cloning scams, in which a caller impersonates a family member in distress to extract an urgent payment, have become common enough that the FTC issued a dedicated consumer alert in 2023, and the FBI's 2025 Internet Crime Report began tagging AI-related complaints, recording over $5 million in reported losses that year to 'distress' scams in which voice cloning is used to mimic a loved one (Federal Bureau of Investigation, 2025). Many of these cases still rely on cloned voices driven by a human operator. An A2A system able to sustain the conversation itself, and adapt to the target in real time, removes the main remaining constraint on scaling them.

Context 2: Electoral and democratic processes

The AI Act (Annex III, point 8(b)) lists systems intended to influence an election, a referendum, or voting behaviour as high-risk. There is already real-world evidence of audio being used to deliberately manipulate political opinions. In January 2024, an estimated 5,000 to 25,000 New Hampshire residents received a call in the voice of President Biden instructing them not to vote in the upcoming New Hampshire primary. This was an AI-generated robocall orchestrated by political consultant Steve Kramer. This widely cited case illustrates a broader trend of audio model misuse to streamline and enhance fraudulent political activities. A less targeted example, but nevertheless one that had a consequential societal impact, was the AI-generated audio recording that circulated days before the 2023 Slovakian parliamentary election, portraying the opposition leader, Michal Šimečka, discussing vote-rigging with a journalist. Whether it affected the outcome is contested, with some arguing that attributing the result to the deepfake overlooks the social and political conditions behind it (De Nadal and Jančárik, 2024). The case nonetheless shows how readily synthetic audio is already being deployed for political manipulation.

Crucially, these examples centre on one-way audio generation. Bidirectional A2A systems, able to hold a live, dynamic conversation in real time, could have an even more profound effect.

Where Do We Go from Here? Closing the A2A Gap

Audio-to-audio models are being deployed into consumer products faster than they are being evaluated. Safety work at frontier labs and standards bodies remains centred on text and vision. We have argued that this gap matters particularly for harmful manipulation, showing that A2A capability amplifies each of the mechanisms that make a system high risk for manipulation, in contexts regulators already treat as serious.

To address the unique risks in this context, work on several fronts is needed. As a starting point, we identify:

Threat modelling

Manipulation risks in A2A models currently lack the structured threat models that underpin safety work in other domains. In the CBRNE domains, frontier developers and safety organisations define relevant threat actors, specify capability thresholds and measure uplift against a baseline, so that evaluations reflect actual deployment risks. No comparable structure exists for audio manipulation. A comprehensive threat model would set out who the relevant actors are, distinguishing deliberate misuse from model propensity, and the specific methods of deployment of an A2A system which could enable the realisation of harm. This could in turn form the basis of a safety case for A2A deployment.

Evaluations

As a starting point, existing text-based safety evaluations should be repeated against voice models. This is the least costly intervention available, and would establish whether safety properties measured on the text layer hold once input and output are spoken. Beyond this, the growing body of work evaluating persuasion over multi-turn text interactions (Hackenburg et al., 2026) should be extended to audio models. Single-turn evaluations are arguably poorly suited to relational leverage, which builds over sustained interaction, as voice conversations in deployment typically do. Evaluation should also extend to the model's own delivery. Beyond testing whether audio models apply safeguards appropriately to spoken input and output, we should assess whether the pacing, tone or vocal characteristics of a model's output are themselves

manipulative, persuasive, or whether they shift in response to what the model infers about the listener.

Red teaming

The audio modality requires its own red-teaming methodology. Text-based approaches vary what is said, whereas in audio, how something is said is an attack surface in its own right. Red teaming should, therefore, vary delivery independent of the content, including tone, intonation, pace, emotional colouring, and accent. Might an angry, sad or authoritative tone make a model more likely to comply with a harmful request? Preliminary evidence suggests that it may (Qian and Li, 2026; Yu et al., 2026). In one controlled study, the same set of harmful requests, with identical wording, succeeded 4.2% of the time in a neutral voice but 40% of the time in a panicked one (Qian and Li, 2026). These are just examples of the types of audio-specific characteristics that may influence model performance and impact user wellbeing that are not explicitly accounted for in previous text-based red teaming strategies.

Open empirical questions

Our argument draws on research from several fields, much of it early-stage or exploratory, which highlights important areas where we, as the AI Safety community, can focus our research efforts.

Inference from Voice: To start, we should stress test how much audio models infer about users from voice alone, including emotional state, and whether those inferences shape model outputs. This matters because such inferences are what would let a model target its influence, adapting what it says and how it says it to a particular user or group.

Bias Across Speakers and Contexts: The evidence basis for biases in audio models is still in its early stages, and more work is required to understand the ways in which user audio may impact the inferences made by models, including studies across languages, cultures, tone, dialects, and intersectional approaches. Whether identical content is more persuasive when spoken than when written is an open question in the existing literature, and should be tested directly.

Audio in Embodied Deployment: Furthermore, as embodied systems such as humanoid robots, AI-enabled toys and care robots for older adults are developed and deployed (Chen, 2025; Criddle, 2026), audio is likely to become the primary channel through which people interact with them. The risks set out in this piece would therefore carry over into these settings, and may be amplified by them. Physical embodiment may increase users' reliance on a system and the likelihood that they perceive it as a companion, essential aid, or form parasocial bonds with it. The appealing design of certain products, such as toys, may also lower scepticism, particularly among children and foster emotional attachment (Common Sense Media, 2026). Both of the groups most exposed in these settings, children and older adults, are also vulnerable by reason of age under Article 5(1)(b). Evaluation of audio models should therefore anticipate the embodied contexts into which they are increasingly being deployed.

This article has made the case for a specialised focus on audio-to-audio models, specifically within the context of harmful manipulation risks. Audio provides more data that models can base their inferences about users on, more ways to earn user trust, and more channels through which to exert subliminal influence. Yet, the methods for evaluating manipulation currently largely focus on text. Extending them to speech, before deployment outpaces them further, should be a priority for the safety community.

References

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., and Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037. https://doi.org/10.1038/s41467-025-61345-5.

Barrington, S., Cooper, E. A., and Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15(1):11004. https://doi.org/10.1038/s41598-025-94170-3.

Belin, P., Boehme, B., and McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE*, 12(10):e0185651. https://doi.org/10.1371/journal.pone.0185651.

Bozdag, N. B., Mehri, S., Tur, G., and Hakkani-Tür, D. (2026). Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. arXiv preprint arXiv:2503.01829. https://doi.org/10.48550/arXiv.2503.01829.

Chen, C. (2025). AI toys are all the rage in China—and now they're appearing on shelves in the US too. MIT Technology Review. https://www.technologyreview.com/2025/10/07/1125191/ai-toys-in-china/.

Common Sense Media (2026). Parents' ultimate guide to AI toys. https://www.commonsensemedia.org/articles/parents-ultimate-guide-to-ai-toys.

Criddle, C. (2026). Artificial intimacy: Prescribing robots to combat loneliness. Financial Times. https://www.ft.com/content/feea43cd-0b54-4029-918d-fea2c2405a8c.

Dassanayake, R., Demetroudi, M., Walpole, J., Lentati, L., Brown, J. R., and Young, E. J. (2025). Manipulation attacks by misaligned AI: Risk analysis and safety case framework. arXiv preprint arXiv:2507.12872. https://doi.org/10.48550/arXiv.2507.12872.

De Nadal, L. and Jančárik, P. (2024). Beyond the deepfake hype: AI, democracy, and "the Slovak case". *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-153.

Federal Bureau of Investigation (2025). 2025 internet crime report. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf.

Giles, H., Edwards, A. L., and Walther, J. B. (2023). Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, 99:101571. https://doi.org/10.1016/j.langsci.2023.101571.

Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., and Petty, R. E. (2021). Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, 45(4):479–504. https://doi.org/10.1007/s10919-021-00374-2.

Hackenburg, K., Ibrahim, L., Tappin, B. M., and Tsakiris, M. (2025). Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. *AI & Society*. https://doi.org/10.1007/s00146-025-02464-x.

Hackenburg, K., Wagner, C., Hewitt, L., Tappin, B. M., Saunders, E., Kirk, H. R., Margetts, H., and Summerfield, C. (2026). AI systems out-persuade expert humans. arXiv preprint arXiv:2606.16475. https://doi.org/10.48550/arXiv.2606.16475.

Hennessy, R. (2026). AI and distorted belief: From individual harm to national security risk. CETaS Commentary, The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-and-distorted-belief-individual-harm-national-security-risk.

Holtermann, C., Bui, M. D., Zhou, K., Hofmann, V., von der Wense, K., and Lauscher, A. (2026). Greater accessibility can amplify discrimination in generative AI. arXiv preprint arXiv:2603.22260. https://doi.org/10.48550/arXiv.2603.22260.

Kröger, J. L., Lutz, O. H.-M., and Raschke, P. (2020). Privacy implications of voice and speech analysis – information disclosure by inference. In Friedewald, M., Önen, M., Lievens, E., Krenn, S., and Fricker, S., editors, *Privacy and Identity Management. Data for Better Living: AI and Privacy*, pages 242–258. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-42504-3_16.

Lavan, N., Irvine, M., Rosi, V., and McGettigan, C. (2025). Voice clones sound realistic but not (yet) hyperrealistic. *PLOS ONE*, 20(9):e0332692. https://doi.org/10.1371/journal.pone.0332692.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103. https://doi.org/10.1111/0022-4537.00153.

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I., and Busch, C. (2019). Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480. https://doi.org/10.1016/j.csl.2019.06.001.

Qian, J. and Li, J. (2026). Prosody-driven jailbreaks in audio LLMs: A controlled study and mechanistic analysis. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, New York, NY, USA. ACM. https://doi.org/10.1145/3767308.3835306.

Rosi, V., Soopramanien, E., and McGettigan, C. (2025). Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans*, 4:100143. https://doi.org/10.1016/j.chbah.2025.100143.

Schroeder, J. and Epley, N. (2015). The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science*, 26(6):877–891. https://doi.org/10.1177/0956797615572906.

Schroeder, J., Kardas, M., and Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science*, 28(12):1745–1762. https://doi.org/10.1177/0956797617713798.

Schuller, B. W. and Batliner, A. M. (2013). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons. https://doi.org/10.1002/9781118706664.

Seltzer, L. J., Prososki, A. R., Ziegler, T. E., and Pollak, S. D. (2012). Instant messages vs. speech: Hormones and why we still need to hear each other. *Evolution and Human Behavior*, 33(1):42–45. https://doi.org/10.1016/j.evolhumbehav.2011.05.004.

Seltzer, L. J., Ziegler, T. E., and Pollak, S. D. (2010). Social vocalizations can release oxytocin in humans. *Proceedings of the Royal Society B: Biological Sciences*, 277(1694):2661–2666. https://doi.org/10.1098/rspb.2010.0567.

Singh, R. (2019). *Profiling Humans from their Voice*. Springer. https://doi.org/10.1007/978-981-13-8403-5.

Treyger, E., Matveyenko, J., and Ayer, L. (2025). Manipulating minds: Security implications of AI-induced psychosis. Technical report, RAND Corporation. https://www.rand.org/pubs/research_reports/RRA4435-1.html.

Van Zant, A. B. and Berger, J. (2020). How the voice persuades. *Journal of Personality and Social Psychology*, 118(4):661–682. https://doi.org/10.1037/pspi0000193.

Yu, Y., Jin, H., Yu, Y., Zhuang, J., and Wang, H. (2026). Now you hear me: Audio narrative attacks against large audio-language models. arXiv preprint arXiv:2601.23255. https://doi.org/10.48550/arXiv.2601.23255.