ADM+S ARC Centre of Excellence for Automated Decision-Making and Society

ADM+S **WORKING PAPER SERIES**

Voice AI and authenticity

Current issues and emerging challenges

August 2025

Authors:

Jean Burgess, Dominique Carlon, Elif Buse Doyuran, Hanxun Huang, Christopher Leckie, Phoebe Matich, Anthony McCosker, Michael Richardson, Daniel Angus, Jake Goldenfein, Madeline Kelly, Awais Hameed Khan, Craig McCosker, Silvia Montaña-Niño, Mohana Rayaprolu, Michelle Riedlinger, Aaron J. Snoswell, Ned Watt, Daniel Whelan-Shamy, Kevin Witzenberger

WORKING PAPER 012

admscentre.org/publications

Acknowledgement of Country
In the spirit of reconciliation, we acknowledge the Traditional Custodians of Country throughout Australia and their connections to land, sea and community. We pay our respect to their Elders past, present and extend that respect to all Aboriginal and Torres Strait Islander peoples today.

Suggested citation
Burgess, J. Carlon, D., Buse Doyuran, E., Huang, H., Leckie, C., Matich, P., McCosker, A., Richardson, M. et al. (2025). Voice AI and authenticity: Current issues and emerging challenges. ADM+S Working Paper Series (12). ARC Centre of Excellence for Automated Decision-Making and Society. DOI: 10.60836/x9de-qt41

Australian Government
Australian Research Council

Abstract

Voice technologies are rapidly being integrated into generative AI-enabled systems and applications. These developments are provoking new questions, and intensifying old ones, in a wide range of everyday contexts. The next wave of AI-enabled voice technologies has the potential to be helpful and useful in many communication, media, and customer service settings, and to improve digital service provision. But these technologies also carry heightened risks of deception in interpersonal communication, change how we interact with and value information and culture, and amplify concerns around issues such as digital inequality and AI-driven labor displacement in knowledge and creative work—all areas where ideas about and struggles over 'authenticity' play a central role. This working paper surveys and historically situates these developments, reviews the current literature in relevant fields, and outlines some emerging responses to the challenges such technologies present to the issue of authenticity in real-world settings.

Authors

Jean Burgess, Dominique Carlon, Elif Buse Doyuran, Hanxun Huang, Christopher Leckie, Phoebe Matich, Anthony McCosker, Michael Richardson, Daniel Angus, Jake Goldenfein, Madeline Kelly, Awais Hameed Khan, Craig McCosker, Silvia Montaña-Niño, Mohana Rayaprolu, Michelle Riedlinger, Aaron J. Snoswell, Ned Watt, Daniel Whelan-Shamy, Kevin Witzenberger

Acknowledgment

For the table of voice technologies in the Background and Introduction section, a combination of Claude, ChatGPT, and Microsoft Copilot was used to help synthesise and tabulate existing notes and suggest examples, as well as to cross-check technical information across sources via web searches. Claude was also used to organise notes and suggest additional examples for the History section, to check for missing and orphan citations, and to help harmonise reference formatting.

Where any AI-generated material formed part of the writing process, it has been comprehensively rewritten and cross-checked before final inclusion in the text, and has been subsequently checked and edited by multiple human authors.

Contents

Background and introduction

Technically, the human voice is the sound made by the human body's vocal apparatus, modulated and combined with 'sub-vocal' sounds like whispers, breaths, clicks, or sighs, in order to speak, sing, laugh, cry, growl, scream, or shout. Socially, it is an embodied medium of communication and expression, with strong connections to individual personhood, identity, relationships, and meaning-making.

The human voice[1] is culturally powerful and socially significant. It is a widely-held idea that each person's voice (which includes not only their speech patterns, but also sonic and affective qualities like timbre and pitch) is closely tied to their identity and personality. Second, the human voice is relational: it is a critical part of social interactions that involve spoken conversation, and it almost always involves a two-way or multi-directional interaction among social subjects (e.g. speaking and listening). In settings or for people where voiced and heard speech are not possible or preferable, including but extending beyond disability, other communication modalities like sign language, subtitles or text-to-speech operate independently of or complement this model.

Voice synthesis is the use of computers and sound production technologies to generate audio sequences that emulate human voices with varying degrees of perceived accuracy or 'naturalness'. Since speaking (or singing) and listening are equally important parts of voice communication, voice synthesis frequently relies on or is paired with voice analysis or speech recognition technology. Attempts at emulating the human voice with technology date back to at least the 1700s, but since the mid-20th century, successive developments in computational linguistics, deep learning and audio processing—as well as improvements in computational processing power overall—have enabled significant advances.

Through intense industry investment and promotion, advanced AI capabilities have not only raised the profile of voice synthesis but also underpinned a drive to embed conversational and agentic systems in everyday contexts. A crucial step that distinguishes voice AI from the history of recorded voice and voice transmission (through phone, cinema, radio, television) is the alternative set of functions that voice AI brings into play. Voice AI applications generate and implement highly naturalistic synthetic voices, operate in real time and across languages, and can produce entire vocal parts for music tracks from text prompts without the direct involvement of either a composer or singer. These recent developments bring the synthetic voice into direct contact with existing concerns, risks and struggles over the problem of authenticity in the information and media environment. Voice AI helps us rethink authenticity

[1] While this paper focuses on the human voice, given that it is the locus of most development in generative AI, we fully recognise that the concept, use and reception of voices is not unique to humans.

through its capacity to mimic, perform, and even create, offering new tools for human expression and accessibility and opening up new forms of risk and vulnerability.

A list of contemporary voice technologies, along with some basic definitions and exemplar applications, is detailed in Table 1 below.

Technology	What it Does	How It Works	Examples
Text-to-Speech (TTS)	Converts written text into natural-sounding emulated speech	Can either use a pipeline of techniques (e.g. linguistic analysis, prosody modelling, waveform synthesis) or end-to-end generative models to create synthetic speech from text input.	Voice assistants (Siri, Alexa, Google Assistant), Google Maps navigation, audiobook narration, screen readers and assistive technologies for the speech-impaired, Duolingo
Speech-to-Text (STT) / Speech Recognition	Transcribes spoken words into written text	Combines audio analysis and language models to decode audio signals into text sequences, potentially including recognition and attribution of unique speakers in a single recording.	Voice assistants (Siri, Alexa, Google Assistant), voice dictation apps, live captions, Zoom/Teams meeting transcription, Duolingo
Speech Analysis	Extracts linguistic, prosodic, and acoustic features from spoken language	Uses techniques such as forced alignment, formant analysis, pitch tracking, and phonetic segmentation to analyse audio and align it with transcripts. Often involves detailed time-aligned annotation and phonetic labelling.	Linguistic research (e.g. MAUS, Praat, ELAN), speech therapy diagnostics, accent/dialect studies, phonetic research, language documentation
Background Noise Suppression	Removes distracting background noise from real-time speech audio	Digital Signal Processing (DSP) and/or deep learning models trained to match common background noise patterns	5G and 4G Mobile Voice Call audio codecs, Zoom / Teams / Google Meet 'noise reduction' filters
Voice Biometrics	Identifies or authenticates individuals based on their unique vocal characteristics	Analyses audio to extract distinctive voice features of the speaker and then compares them against stored voice profiles	Voice-based login systems, fraud prevention (e.g. banking)

Speech-to-Speech Translation	Translates spoken language from one language to another while maintaining speech output in close to real time	Combines speech recognition, machine translation, and text-to-speech, typically in a single end-to-end model to reduce latency	Meta SeamlessM4T, Google Translate conversation mode
Voice Cloning	Creates synthetic speech that mimics a specific person's voice	Learns speaker-specific vocal characteristics from audio samples and creates a model that can generate new speech (e.g. from written text) using those learned voice patterns	ElevenLabs voice synthesis, Respeecher for media production, generating new dialogue, dubbing or audiobook narration from actors' voices, deepfake videos and deceptive phone calls
Voice Conversion	Transforms one voice into another while maintaining the linguistic content	Uses deep learning to map and pair acoustic features from both source and target speaker, and generates new audio that conveys the original message in the new voice; sometimes adding translation into another language	AI dubbing, call centre 'localisation', voice modification for gender affirmation e.g. Voice.AI, entertainment e.g. Capcut, or deception (scam calls)
Real-Time Pitch Correction	Corrects or modifies vocal pitch during live musical performance or recording	Uses digital signal processing (DSP) and machine learning to detect pitch deviations and automatically adjust them to musical scales	AutoTune, Melodyne, live vocal processing
Singing Voice Synthesis	Generates artificial singing voices from lyrics and musical notation	Uses neural models trained on singing data to produce vocal performances that follow specified pitch, timing, and phonetic information	VOCALOID, Music production tools and plugins (too many to name)
AI Harmonisation	Automatically generates harmony vocals and backing parts for music	Analyses the lead vocals and other elements of a piece of music to create complementary vocal parts following harmonic rules in specific styles and genres, using generative models trained on recordings	Music production plugins (too many to name)

AI Music Generation	Creates complete musical compositions including vocals from text descriptions	Uses large language models and diffusion models trained on music datasets to generate full songs with lyrics, melodies, and instrumentation from natural language input	Suno AI, Udio, MusicLM
Vocal Isolation / AI Music Splitting	Processes a combined recording of a musical performance ('mixdown') to extract the individual instruments and/or singers' audio parts ('stems')	Uses neural networks and digital signal processing (DSP) methods to attempt to separate the audio components.	Moises, Splitter.AI, GAudio

Table 1: Current voice technologies and applications

This working paper is part of the ADM+S Centre's Generative Authenticity project, which critically examines the developments and debates around authenticity in generative AI contexts, assesses the social, technical and legal challenges associated with them, and explores novel technical responses that contribute to more responsible, ethical and inclusive systems.

As the overall summary of the project[2] states: authenticity is a key problem for understanding and managing the impacts of generative AI and synthetic media in society, and a central target for automated decision-making systems in the information and media environment. From trustworthy news reporting to identity verification for social services and the everyday risk of scams, generative AI and synthetic media present significant real-world implications for practitioners, institutions, and publics in Australia and elsewhere. A wide range of technical solutions collectively understood as authenticity infrastructure promise to address these issues; but, if adopted and embedded at scale, some of these solutions could have potentially significant downstream effects on stakeholders and implications for society.

In this working paper, we investigate the difference generative AI makes to the question of authenticity with specific reference to the rapidly developing technologies and applications of voice automation and synthesis. As we discuss below, for example, recent developments in the perceived *realism* or *naturalness* (i.e. authenticity) of voice cloning challenge existing assumptions about how an 'artificial' voice might be detected or experienced. As well as greatly

[2] https://www.admscentre.org.au/generative-authenticity/

increasing the risks of deceptive applications used for the purposes of fraud or harassment, such technologies raise new questions about whether the 'disembodied voice' can be considered a 'truthful' representation of a person in a variety of more benign media and interpersonal communication contexts.

In what follows, we survey both the history of voice synthesis and newer developments in the context of generative AI, and discuss recent examples of the controversy and debate that attaches to voice synthesis. We provide an overview of key social and cultural concepts and issues relating to authenticity and the human voice, before turning to a discussion of technical and design approaches to voice synthesis and verification with authenticity-related issues such as trust and deception in mind. Finally, we assess the extent to which existing research and industry practice is responding or is prepared to respond to the issues these developments provoke or highlight, with a view to identifying critical issues for future investigation and practical intervention.

1. A very short history of voice synthesis

1.1. The machine age: 1770-1970

The most well-known early attempts at mechanical voice synthesis date from the late 1700s (for further detail, see Ohala, 2011; Ondrejovic, 1992; Brackhane, 2015). Most historians highlight two inventions and their accompanying treatises: first, Kratzenstein's 'vowel organ', which demonstrated that the vowel sounds most familiar to speakers of European languages could be reproduced using resonant tubes of different shapes; and, second, von Kempelen's[3] mechanical speaking machine. Built with the idea of a prosthetic communicative device for non-speaking people in mind, it used bellows to simulate lungs, a reed for the glottis, and a leather resonance chamber that could be manipulated by hand to produce basic vowels and consonants. Both Kratzenstein's and von Kempelen's approaches were focused on the physical emulation of parts of the vocal tract, using air and vibrations to create voice-like sounds. In doing so, they both explored and normalised the idea that human speech was a physical, acoustic phenomenon that could be artificially reproduced through the control and manipulation of combined airflow and resonance mechanisms. However, as Brackhane (2015) argues, it was Kempelen the engineer, rather than Kratzenstein the physician, whose theories and approaches most closely resembled those that would be adopted by modern speech synthesis.

The 19th century saw sustained scientific investigation into the acoustic properties of speech, alongside well-known developments in sound technology like the gramophone and telephone.

[3] A prolific and enigmatic inventor best known for his chess automaton, von Kempelen reportedly also designed and built a printing press for the blind, a pendulum bridge, and machines for pumping water from flooded salt mines in Hungary. For more, see Ondrejovic, 1992.

Melville Bell's and his son Alexander Graham Bell's work on visible speech notation in the 1860s and 1870s provided new analytical frameworks and methods to render speech sounds as symbols, which was essential to later synthesis efforts. Hermann von Helmholtz's studies of acoustic resonance, including his (1863) treatise *On the Sensations of Tone*, established the scientific foundation for understanding how complex speech sounds (as well as music) could be decomposed into simpler harmonic components—a principle that would become central to subsequent speech synthesis and analysis technologies (Vogel, 1993). Voice synthesis also benefited from advances in manufacturing, exemplified by Faber's speaking automaton, the Euphonia (Ramsay, 2019). The machine, which used pedal-operated bellows to produce an approximation of a human voice, refined Von Kempelen's earlier design with more sophisticated mechanical components capable of producing clearer vowels and basic phrases. The machine "could be manipulated to produce an extensive range of vowel and consonant sounds and word pairings through an expert's deft 'playing' of the automated voice on a keyboard." (Martin, 2020). These speaking machines were never without their controversies, most of which centred around the uncanny qualities of the disembodied voice. Martin (2020) describes how, despite the technical advances represented by the Euphonia, "scientists, elocutionists, and medical experts in London were united in their assessment that the Euphonia's vocal production—its whispers, laughs, stutters, stammers, stumbles, blocks, and hesitations—seemed to emanate from deep within a crypt." (p.2).

The transition from mechanical to electronic systems began in the early twentieth century. The Bell Telephone Laboratory's VODER (Voice Operating Demonstrator)—a simplified version of the Vocoder—is recognised as the first significant attempt to electronically synthesise human speech by breaking it down into its acoustic components, The VODER required skilled operators who used a keyboard and foot pedals to control pitch, formants (specific resonant frequencies matched to speech sounds), and amplitude (relative volume) in real time. It was demonstrated at the 1939 New York World's Fair and the 1939 Golden Gate International Exposition, where audiences marvelled at the machine's ability to speak, sing, and even laugh.[4] These advances in speech synthesis technology benefited significantly from the contributions of people with disabilities to the design of assistive technologies, as well as artists working in collaboration with scientists. For example, Benjamin Lindquist (2024) describes how the process of speech synthesis in the 1940s involved researchers and skilled visual artists collaboratively painting sound spectrograms by hand, translating visual patterns into audible speech.

[4] https://www.historyofinformation.com/detail.php?entryid=738

1.2. Digital speech synthesis from rule-based systems to deep learning

Key developments in the shift to computational methods for speech synthesis in the 1950s and 1960s included George Rosen's DAVO (Dynamic Analog of the VOcal tract), introduced in 1958 (Story, 2019); the Linear Predictive Coding (LPC) method developed both in Japan and, in parallel, in the US at Bell Labs in the late 1960s (Markel & Gray, 1976; Atal, 2006); followed by DECtalk system developed by Digital Equipment Corporation in 1983 (Hallahan, 1995). Most of these developments relied on direct formant synthesis (the use of computational rules to model, generate, and sequence combinations of the distinctive resonant frequencies—formants—produced in human speech).

The period from the 1980s to the early 2000s saw a departure from formant synthesis, instead using data-driven approaches to select and join together (i.e. 'concatenate') pre-recorded speech segments, with later work improving the 'naturalness' of the resultant speech by integrating sound-to-sound transitions and then prosody through the use of DSP (digital signal processing) techniques (Moulines & Charpentier, 1990; Black et al., 1998; Taylor et al., 1998). These concatenative systems produced more natural-sounding speech by preserving the acoustic properties of vocal recordings rather than approximating them with electronic synthesis, though they produced audible glitches at the boundaries between speech segments. Commercial applications also expanded rapidly during this period, in turn driving the further development and adoption of both speech synthesis and speech recognition technologies. IBM developed ViaVoice (first introduced in 1997). The system deployed continuous speech recognition (CSR) which was claimed to enable more natural dictation workflows compared to discrete word recognition systems.

The deep learning breakthroughs beginning in the 2000s and 2010s were applied to voice, with Google DeepMind's WaveNet representing a distinctive shift.[5] Introduced in September 2016 (van den Oord, 2016), WaveNet used neural networks to generate speech that mimicked specific human voices and was generally considered to sound more natural than existing Text-to-Speech systems. As in other fields of AI, the excitement, research and industry investment around this breakthrough fuelled the rapid development of models that pushed generative text-to-speech synthesis quality to reportedly "near-human" levels (Wang et al., 2017).

1.3. Cultural and aesthetic developments

With the rise of computational composition and sound synthesis in music, there was a competing move to foreground the machine-like qualities of synthesised sounds, including human voices. Here, in many cases, the goal wasn't so much to aim toward the imperceptible emulation of the human voice as it was to experiment with new technology (especially in the

[5] https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/

case of avant-garde music culture), and to create new sounds that responded to, reflected or commented on the increasingly industrialised, hypermediated environment of postwar society. The most well-known example is the 'robot voice', which was relatively monotonal, with rough, machine-like timbres paired to 'robotic' (hyper-rational and emotionless) forms of linguistic expression. This aesthetic found expression in electronic music genres via, for example,. Kraftwerk's pioneering use of the Vocoder in the 1970s, and later in hip-hop and electronic dance music where robotic vocal effects became signature elements.

Examples of synthetic voice in popular culture include the character of the AI-enabled computer HAL 9000 (HAL) in *Stanley Kubrick's 2001: A Space Odyssey* (1968). Actor Douglas Rain delivered HAL's lines in the deliberately calm, smooth, 'emotionless' style that became archetypal for AI voices in popular culture and beyond. In other examples familiar in the Anglosphere, the Speak & Spell toy by Texas Instruments (1978) introduced millions of children to synthetic speech (and, perhaps, can be seen as an early prototype of educative voice assistants and AI companions) through its distinctive electronic voice, while Stephen Hawking's text-to-speech voice both represented a remarkable set of advances in assistive technology and gave the synthetic voice an aura of scientific authority.

In music technology, production tools like autotune (which shifts the pitches of sung notes to conform to the notes of a given scale) became deployed to 'idealise', 'tidy' and 'perfect' voice recordings. While the term 'autotune' is used generically today, the original "Auto-Tune" was a specific tool introduced in 1997 by Antares Audio Technologies that was initially designed for remedial pitch correction—a use case in which its presence would generally be concealed rather than highlighted. But autotune was later reconfigured and revalued as a creative tool, even an instrument, following its prominent use in pop powerhouse Cher's single "Believe" (1998), its enthusiastic adoption by singer and rapper T-Pain, and its subsequent normalisation in pop, hip-hop, and electronic music (e.g. for its use in trap, see Duinker, 2024). Alongside 'vocoder' effects from the earlier period, the audible use of autotune became a marker of 'slick', commercial, dance-floor friendly digital recordings; in this way, synthetic aspects of voice mediation became aesthetically valuable in themselves within the cultural norms of some genres and scenes, while being abhorred in others. Counter-aesthetics meant to convey 'authenticity' might deliberately eschew auto-tune and other obvious audio processing, preferring a few slightly off notes as a way of signalling the absence of synthetic voice in their recordings. Always controversial and contested, vocal synthesis technologies have long been an essential part of the battleground between the aesthetics and politics of postmodern pop, on the one hand, and the intense focus on a very specific logic of authenticity grounded in the gendered and raced ideologies of rock music on the other (Provenzano, 2019).

The example of Yamaha's Vocaloid illustrates how the cultural meaning of synthetic voice, including how well it is accepted as aligned to or authentic within a particular music scene or culture, is highly contextual and depends on the practices and norms of fans and audiences. Since its inception in 2003, this singing-voice software synthesiser has been part of the democratisation of music production in Japan. By offering a range of human-recorded voicebanks or singer libraries to use as raw material, the software affords its users to "program

it to sing or speak—down to the pitch, duration, and even timbre of each note" (Roseboro, 2019, p. 14). Kenmochi (2010) suggests that Vocaloid not only imitates but augments human singing, in that users are able to program the system to go beyond the bounds of human vocal proficiency. The developers personified each voicebank with a computer-generated image that replicates the Japanese Idols of "*otaku* ('nerd') culture" (p. 39) (Anderson, 2021). The Vocaloid phenomenon spawned a global community of creators and spawned live concerts featuring holographic performances of digital singers, demonstrating how synthetic voices could acquire cultural authenticity (that is, resonance with the values of a culture) through community engagement rather than technical realism alone. A notable figure in vocaloid culture is its avatar Hatsune Miku, (which translates to "the first sound of the future") depicted as a 16-year-old girl with blue hair, who is assigned to one particular voice donor's contributions (Le, 2014; Lam, 2016). Miku is claimed by aspiring music creators and fans of the software not to diminish but to *enhance* "the humanness of the technology" (p. 1112), as her in-software characteristics are enhanced by the various derivative works produced by fans (Lam, 2016).

1.4. Voice as general-purpose technology

AI-enabled voice synthesis is now integrated into many platforms and product suites for search, chat, and device control functions. From early 'smart' assistants with synthetic voice output like Apple Siri, Amazon Alexa, and Google Assistant to its integration across a wide range of consumer products and within generative AI systems, voice is now arguably a general-purpose media technology.

Most of the large technology companies have made significant investments in the infrastructure and applications (as well as the PR effort) needed to normalise voice synthesis across economic sectors and in everyday life. In 2018 Google demonstrated the potential of synthetic voice interaction via a demonstration of Google Duplex, a system designed to autonomously place phone calls on behalf of users. Its launch drew wide attention, and concern, for showcasing a voice assistant making restaurant reservations and hair appointments with minimal human intervention. Critics noted that the application appeared to be deceptive, failing to disclose its non-human status, as well as the highly curated nature of the demos, which raised ethical questions around transparency, manipulation, and the practical viability of such systems in unscripted environments[6]. Google Duplex was shelved as a standalone product by Google in 2022, although elements of the sytem were repurposed as part of Google's voice assistants and other phone features.

Microsoft acquired Nuance Communications for $19.7 billion in March 2022, integrating voice technologies across healthcare applications including Dragon Medical One for clinical

[6] https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls

documentation.[7] OpenAI released its speech-to-speech API in October 2024 in order to encourage third-party voice application development and consumer take-up. Google's I/O 2025 announcements included Gemini Live with camera capabilities for voice conversations about visual content[8] and Veo 3's native audio generation for video with character dialogue and sound effects.[9] Google's move to replace Google Assistant with Gemini across mobile devices demonstrates voice's central role in the company's generative AI strategy.

Business uses of AI voice for personalised and context-specific engagement with customers have carved out significant commercial spaces. Companies like ElevenLabs offer voice cloning services for content creators, while enterprises use voice synthesis for personalised customer communications at scale. Apple's May 2025 accessibility features announcement[10] demonstrates continued innovation in assistive applications, maintaining voice synthesis's historical connection to disability support and communication access.

1.5. Anatomy of a Voice AI system: Tortoise TTS

Voice AI systems are varied in their design, and hence their performance across different criteria. In this section, we offer a brief outline of the components and performance of one such system, which may serve as the basis for future critique and comparative work.

Tortoise TTS[11] (Betker, 2023) is a prominent transformer-based voice cloning model known for prioritising synthesis quality over processing speed. The model works well to exemplify how contemporary voice cloning models process and understand vocal characteristics through embedding strategies.

Voice cloning systems work by converting human speech into discrete digital representations that can be processed by AI models. This process involves creating embeddings, computational representations that capture imagined characteristics of voice. Tortoise TTS uses what it refers to as MEL embeddings, where MEL stands for mel-scale (derived from 'melody'), a frequency scale designed to match human perception of sound (Betker, 2023). However, MEL embeddings alone are insufficient to make a cloned voice ‘speak’. Tortoise employs a multi-model pipeline where each component handles a different aspect of the speech generation process. The system consists of five interconnected models working together:

---

[7] https://news.microsoft.com/source/2021/04/12/microsoft-accelerates-industry-cloud-strategy-for-healthcare-with-the-acquisition-of-nuance/

[8] https://blog.google/products/gemini/gemini-app-updates-io-2025/

[9] https://www.techradar.com/news/live/google-i-o-2025-live-project-astra-gemini-and-more

[10] https://www.apple.com/au/newsroom/2025/05/apple-unveils-powerful-accessibility-features-coming-later-this-year/

[11] https://github.com/neonbjb/tortoise-tts?tab=readme-ov-file

1. Autoregressive transformer - converts input text into discrete MEL embeddings (mathematical representations of audio frequency content)
2. Diffusion model - transforms MEL embeddings into mel-spectrograms (visual representations of audio frequency content over time)
3. VQ-VAE (Vector Quantised Variational Autoencoder) - compresses mel-spectrograms for processing by mapping continuous data to discrete embedding vectors
4. Vocoder - converts mel-spectrograms into final audio waveforms
5. CLVP (Contrastive Language-Voice Pretrained) model - ranks multiple outputs to select the highest quality result

The primary embedding process occurs in the autoregressive transformer, which learns to associate text with discrete audio representations. These embeddings are computational representations that appear to capture vocal characteristics, though what they actually encode remains opaque. We can observe that the model learns to distinguish between speakers and maintain vocal consistency, but the embeddings themselves are mathematical abstractions rather than explicit representations of human-interpretable voice features.

The voice cloning process with Tortoise TTS works by first analysing sample audio clips to extract vocal embeddings, then using these embeddings to guide the generation of new speech. The system learns to match text with audio embeddings while intending to maintain the target speaker's distinctive voice qualities.

This embedding approach allows voice cloning systems to treat speech synthesis like a translation problem: text is converted into audio embeddings rather than trying to manipulate audio signals directly. By representing voice through these discrete embeddings, voice cloning models can use the same pattern-matching techniques that work well for text-based AI.

How a voice cloning model creates embeddings for vocal characteristics fundamentally determines what it can and cannot do. Models that focus on capturing fine acoustic details in their embeddings might reproduce a speaker's vocal texture perfectly but struggle to maintain natural rhythm across longer sentences. Models that prioritise speech rhythm and melody in their embedding space might sound natural and flowing but lose the subtle qualities that make each voice distinctive.

For example, fast real-time voice synthesis systems like those used in virtual assistants prioritise speed and computational efficiency in their embedding design. These models can generate speech almost instantaneously but often produce voices that sound somewhat generic or robotic, lacking the subtle individual characteristics that make each person's voice unique. In contrast, Tortoise's approach sacrifices speed for quality, creating more detailed embeddings that can capture nuanced vocal characteristics but require significantly more processing time.

These design choices create direct trade-offs: a system might excel at reproducing accents but struggle with emotional expression, or capture speaking style but fail with precise

pronunciation. Essentially, how a model represents voice through its embedding approach sets the boundaries of what kind of voice cloning is possible.

1.6. Summary: the difference generative AI makes

Contemporary voice AI technologies and applications represent a number of distinctive shifts relative to the history of voice synthesis that expand its functionality and distinguish it from voice recording and transmission:

1. Significant advances in the realism of synthetic speech and singing, including the integration of prosody, intonation, breath and sub-vocal or extra-lingual expression;
2. The ability to operate with very little latency (i.e. in real time), which removes the time lag that would otherwise prompt users to notice the intervention of a machine;
3. Voice cloning, which enables the disembodied rendering of specific human voices without the direct involvement or consent of the humans in question; and
4. The integration of all these advances into multimodal generative AI models that can generate human-sounding voices to deliver AI-generated words, lyrics, melodies or harmonies from user prompts.

Overall, then, the development of neural synthesis over the past decade represents more than technical progress: it has begun to fundamentally challenge our understanding of voice as a marker of human presence and identity and created new possibilities for interacting with AI conversational agents. When voice cloning can passably replicate individual speakers with minimal training data and without their involvement, **the basis of the distinction between 'real' and 'synthetic' voices in any mediated context becomes increasingly difficult to pin down, raising new questions about authenticity and the nature of vocal identity**. This also represents a dramatic acceleration of the normalisation of voice as a familiar, more intimate and engaging mode for interacting with computational systems.

At the same time, however, as the case of Tortoise TTS shows, voice AI systems are far from perfect, either in terms of performance or in terms of their fidelity to and representation of the human voice in general, or specific people's voices (especially non-normative voices) in particular. Unlike the more rules-based tools developed in linguistic research, such as forced-alignment systems, corpus annotation platforms, or articulatory phonetics models, commercial neural voice synthesis systems generally prioritise surface-level naturalness over detailed linguistic or stylistic fidelity, and they typically offer limited insight into the structural, historical, or socio-cultural dimensions of language and other forms of vocal expression. While startlingly capable and rapidly improving, such systems are not (yet) designed to reproduce fine-grained phonetic detail, grammatical variation, or cross-linguistic patterns (for example, the vernacular of a North Queenslander or a Vietnamese-Australian accent).

These limitations may make fake voices **more detectable in the short term**. But they also present challenges to the authenticity of synthetic voices in relation to linguistic, socio-

linguistic, and cultural diversity, where authenticity is understood not as “human and not AI”, but as **goodness of fit with a community and its systems of meaning-making**.

2. Recent controversies and media events

In this section, we survey instances of controversy (Callon et al., 2009) which may prove useful in thinking through the relationship between voice synthesis and authenticity in the context of generative AI. Controversies—that is, discrete events that enact productive uncertainty and debate around complex issues—are productive sites of analysis because they reveal tensions, stakeholders, and contradictions in issues (Marres, 2015). In our case, such instances are useful even when the specific controversies themselves are partly the result of media reporting on AI hype, because they engage citizens and consumers in speculation and debate about the uncertain futures, limitations, and politics of voice technologies (Ramati, 2024).

For example, following OpenAI's May 2024 launch of GPT-4o with a synthetic voice named "Sky," observers noted its resemblance to Scarlett Johansson's voice in the film *Her*. Johansson stated she was "shocked, angered, and in disbelief" that OpenAI had used a voice so similar to hers, given that she had explicitly refused OpenAI's invitation for her to voice the system months earlier. She hired legal counsel, alleging the company intentionally mimicked her voice, and citing CEO Sam Altman's tweet of a single word, "her", in relation to the launch.[12] In response, OpenAI stated the voice was from a different professional actor, but paused the use of the voice out of respect and released more public information about how their AI voices were developed than they otherwise might have.[13] In this example, we see how celebrity, speculation and public concern in a context of uncertainty helped provoke debate and nudged actors towards transparency, however temporarily and imperfectly.

What follows is a selection of further instances of controversy that show some of the ways voice technologies, generative AI, and ideas about authenticity are already beginning to interact, and the questions that arise out of these interactions.

2.1. Media industries

As is the case throughout the economy, generative AI voice systems are appealing solutions to productivity and resourcing challenges in the media industries: they promise to reduce labour costs, speed up content production, and even enable live delivery of news or entertainment by synthetic agents. But early experiments in this area have provoked audience outrage and damaged institutional trust, largely for reasons relating to authenticity. One example arose at

[12] https://apnews.com/article/scarlett-johansson-openai-voice-chatgpt-sky-d04b68074944d187219f8502f6ae64ce

[13] https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/

the Australian Radio Network's (ARN) Sydney-based CADA station, which used an AI host for its Workdays with Thy program.[14] Developed as a text-to-speech agent using an ARN finance employee as its model, the AI host was employed by the station for six months without informing the audience. For some, assigning the synthetic AI agent an Asian and female persona only served to highlight the lack of human diversity in radio and media industries more generally. According to the Australian Communications and Media Authority (ACMA), the station had no obligation to disclose that the presenter was AI because policies and standards are still under development, and when implemented are likely to be mandatory primarily in 'high risk' situations.

Even when synthetic voices are disclosed and deployed in more focused environments, the benefits to news organisations can be marginal despite the considerable effort required in developing and implementing them. Earlier this year, the Philippine Center for Investigative Journalism experimented with transforming an original investigative written piece into an AI multimedia story.[15] According to the journalist who disclosed the experiment, one of the greatest difficulties was generating the AI voice-over, requiring the prompts to be repeatedly adjusted to fix the sources' accents and make them sound "authentic". It turned out that non-Filipino listeners found those voices "monotonous". BBC had the Beeb Voice Assistant project to provide an independent alternative to Amazon Alexa and Google Assistant. They conducted extensive research with universities and audience testing on synthetic voice and accents to gauge how audiences felt about the use of different disembodied voices to represent the BBC and different types of content (Zambrini, 2020). The project team ultimately chose a Northern male accent, deliberately countering the prevalence of female voices in commercial AI assistants.

2.2. Creativity and creative work

Voice generation will likely have significant impacts on the creative industries around the world and in Australia, and is already the subject of speculative debate and protest as well as practical experimentation. In the United States, SAG-AFTRA voice actors and motion capture artists went on strike in 2024 against video game voice cloning[16] following broader strikes and ongoing anxiety about contractual arrangements that signed over voice and image rights[17]

---

[14] https://www.smh.com.au/culture/tv-and-radio/thy-has-been-on-the-radio-for-six-months-turns-out-she-isn-t-real-20250424-p5ltxi.html

[15] https://pcij.org/2025/03/16/i-turned-a-powerful-longform-story-into-multimedia-with-generative-ai-heres-how-i-did-it/

[16] https://en.wikipedia.org/wiki/2024%E2%80%93present_SAG-AFTRA_video_game_strike/

[17] https://www.rollingstone.com/tv-movies/tv-movie-features/actors-strike-sag-aftra-ai-one-year-later-1235059882/

across the American entertainment industry. According to the Australian Association of Voice Actors (AAVA), more than 5,000 jobs could be lost or otherwise negatively impacted by the growing use of AI voice clones[18]. Simon Kennedy from the AAVA states that it is not uncommon for voice actors to discover that their voice has been cloned without their consent, or that they have lost work to a non-consensual clone of themselves.[19] Audiobook platform Audible now uses an 'end to end' production pipeline[20] that includes cloned voiceover artists[21] capable of multiple languages, enabling the automated production of audiobooks in a wide range of languages without employing voice actors at any stage. Similarly, Netflix and other streaming services are already using synthetic voices to overdub for translation, particularly for non-English titles with a global appeal such as Squid Game.

Voice synthesis has also been used to replicate deceased figures, such as in a documentary about the chef Anthony Bourdain[22] or in Albert Einstein chatbots,[23] leading to debates about disclosure, transparency, and the ethics of using voices that cannot give permission for their replication. This risk extends to the content creation community. For example, a YouTube gaming personality discovered that his voice was cloned to narrate a video without his consent.[24] His concerns speak to the authenticity at the core of this labour and intellectual rights issue. "The thought that someone else would do it in order to copy my persona in this way—it's just so weird and invasive," he says. "It's kinda like plagiarism but more personal. It's not my work or my labor. It's a distinct part of who I am." Synthetic singers are also now part of AI music generation applications, further complicating already fraught questions of authenticity that surround the use of generative tools to create music; fan creations that use iconic singers' voices to 'perform' songs they never sang both build on existing home production technologies like Digital Audio Workstations (DAWs) and stretch the limits of norms for creative license in fan art (Galuszka, 2024). Voice also appears unprompted in text-to-video applications, sometimes generating dialogue without any direction from users, in the case of Google's new Veo 3

---

[18] https://www.theguardian.com/technology/article/2024/jun/30/ai-clones-voice-acting-industry-impact-australia

[19] https://www.sbs.com.au/news/article/the-scammers-trying-to-steal-our-voices-and-how-to-protect-yourself/15mgeqa74

[20] https://www.wired.com/story/audible-audiobook-narrators-ai-voice-clones/

[21] https://www.publishersweekly.com/pw/by-topic/digital/content-and-e-books/article/97756-audible-expands-catalog-with-ai-narration-and-translation-services.html

[22] https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice

[23] https://techcrunch.com/2021/04/16/ai-driven-audio-cloning-startup-gives-voice-to-einstein-chatbot/

[24] https://www.wired.com/story/a-gaming-youtuber-says-an-ai-generated-clone-of-his-voice-is-being-used-to-narrate-doom-videos/

model.[25] Artists themselves have used AI to create content — for instance Paul McCartney[26] famously used AI to isolate and reconstruct John Lennon's voice from incomplete, lossy demo tapes, to finish what is termed as *The Beatles' final record*. Similarly, James Earl Jones[27] signed over the rights to recreate his voice using AI for his iconic role of Darth Vader in the Star Wars series to Lucasfilms.

2.3. Relationships and intimacy

Voice synthesis offers a believable and emotionally resonant way for people to connect with systems that mimic the voices of real individuals - living or deceased. Its inherently conversational nature lends itself to relational engagement, and one of the fastest-growing applications is in the field of legacy preservation. This use of voice synthesis builds on earlier forms of digital memorialisation, such as chatbots trained on the data of the deceased - commonly referred to as *deadbots*, *digital ghosts*, or *thanabots* (derived from *thanatology*, the study of death) (Henrickson, 2023). Advances in voice synthesis have fuelled the growth of a legacy-tech industry. Companies like *You, Only Virtual (YOV)* have developed 'versona' technology, which captures the voice, personality, and behavioural patterns of loved ones, with the aim of preserving and sharing their presence with future generations.

While these technologies can facilitate a sense of relational intimacy, they may also reflect more individualist ideals, designed to fulfill posthumous aspirations. This raises complex cultural questions about what constitutes appropriate treatment of the dead. As Sparrow and Zhang (2025) argue, such practices may conflict with certain traditions, potentially undermining values like filial piety and disrupting the communal bonds upheld by ancestral rituals. Furthermore, ethical considerations emerge when using the voices of deceased Indigenous Australians, where consent, cultural protocols, and collective memory must be treated with particular care. Recent incidents where generative AI has been used to revive voices of the deceased reveal some of these underlying moral and ethical complexities. For example, in 2025, the US family of a deceased road rage victim used his voice recordings, videos and pictures to generate a

[25] https://www.theverge.com/ai-artificial-intelligence/673719/google-veo-3-ai-video-audio-sound-effects

[26] https://www.theguardian.com/music/2023/jun/13/ai-used-to-create-new-and-final-beatles-song-says-paul-mccartney

[27] https://www.forbes.com/sites/timlammers/2024/09/09/james-earl-jones-signed-over-rights-for-ai-to-recreate-darth-vaders-voice/

synthetic victim impact statement[28], confronting the perpetrator with it at the sentencing hearing.

2.4. Deception and scams

In the world of deception and scams, synthetic voice generation is engaged in a cat-and-mouse game with techniques for voice verification. In 2023, a Wall Street Journal columnist[29] cloned her voice and then used it to fool her bank and her family. In Australia, reporting by The Guardian[30] showed that AI voice cloning tools were good enough to verify identity with the voiceprint technology used by Centrelink and the Australian Taxation Office. Citizens are also using AI voice clones to waste the time of phone scammers, with "Daisy" the artificial English grandma[31] among the more amusing examples. This has also extended to larger scale organisational efforts, for instance in 2025 Commonwealth Bank of Australia[32] in collaboration with Apate.ai reportedly deployed an army of thousands of AI-bots, designed to engage scammers in voice and text-based conversations, to gather critical intelligence, and disrupt scam operations, whilst insulating real Australians from the growing threat of voice call scams. High profile fraud cases highlight the importance of voice to the believability (or fidelity) of real-time deepfakes. In one of the more widely reported scams, the British design and engineering firm Arup was the victim of a $25m scam after a finance worker in Hong Kong[33] was convinced of the authenticity of a call with its 'chief financial officer'.

There is also an increasing number of reported instances of AI voice clones being used to scam the public, by imitating family members and loved ones. In 2023, Scammers convinced an American mother her daughter had been kidnapped,[34] using an AI voice scam demanding ransom of $1m. Similarly, an elderly Canadian couple was defrauded by scammers using a

[28] https://www.bbc.com/news/articles/cq808px90wxo

[29] https://www.wsj.com/tech/i-cloned-myself-with-ai-she-fooled-my-bank-and-my-family-356bd1a3

[30] https://www.theguardian.com/technology/2023/mar/16/voice-system-used-to-verify-identity-by-centrelink-can-be-fooled-by-ai

[31] https://www.instagram.com/reel/DJSbEcnBN94/?igsh=cmZ4ZWZkZXk4dDho

[32] https://www.commbank.com.au/articles/newsroom/2025/06/apate-ai.html

[33] https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html

[34] https://abcnews.go.com/Technology/experts-warn-rise-scammers-ai-mimic-voices-loved/story?id=100769857

convincing AI voice clone impersonating their son and son's lawyer[35], demanding $21,000 in legal fees in relation to an alleged car crash fatality. While most scams are more mundane than voice and video clones and "their impact on individuals far greater" (Swartz et al., 2025), the potential for voice synthesis to overcome authentication and verification tools spotlight the need for government, financial institutions, businesses, and individuals to future proof their systems.

2.5. Political communication

The authenticity of voice is a familiar matter of concern in political communication. The issue is socio-technical in nature as it is made up of the technical uncertainty in the provenance of questioned voice content as well as the cynical exploitation of this uncertainty for political favour or damage control. For audiences, the deceitful use of synthetic audiovisual content—or even, the strategic accusation that certain political media messages might be synthetic—contributes to a sense of malaise that manifests as a general detachment from societal institutions and an erosion of trust in everyday social and political communications (Whelan-Shamy & Watt, forthcoming).

Examples include the case of an election in Tamil Nadu, India, where two audio clips allegedly depicting Palanivel Thiagarajan, then a minister, were posted to X (formerly Twitter) by a rival candidate where Thiagarajan could be heard candidly disparaging his own party, discussing corruption it had been involved in, and praising political rivals (Christopher, 2023). In addresses posted to X, Thiagarajan vehemently denied the genuineness of both the clips claiming that they were generated by malicious actors, but forensic analysis was inconclusive as the first clip was deemed too low quality to be assessed, and the second one was deemed 87% likely to be a real depiction. This example shows how the challenges to authenticity of synthetic voice generation, and voice cloning in particular, enables actors to raise doubts as to the veracity of genuine recordings.

Other examples of technical verification uncertainty can be seen in the viral audio of former Ghanaian President John Mahama, which was found to be 62% inauthentic by synthetic audio detection tools Hiya Deepfake and Deepware (Ghana Fact-Checking Coalition, 2024), and also in the 2023 Slovakian election where politically damaging audio of a political candidate was released two days prior to an election that was later determined 'altered' by GenAI company Eleven Labs (Meaker, 2023). ABC News Verify deliberately amplified the public debate around this issue in the leadup to the 2025 Australian federal election, by generating an AI clone of Senator Jacqui Lambie's voice. They gathered community reactions to AI Lambie's "electoral platform" to highlight the importance of context and to show how difficult it was to technically

[35] https://www.techjuice.pk/canadian-couple-got-scammed-for-21000-by-ai-generative-call-pretending-to-be-their-son/

identify problematic AI-generated voice content through human and automated means (Martino & Mallal, 2025).

Each of these controversies raises questions about or suggests new perspectives on the question or ideas of authenticity. They touch on several dimensions of the problem: identity and verification; authentic relationships and interpersonal communication; cultural and aesthetic meaning and creativity; and the consequences of synthetic media for ethics and rights in all these areas. In the following section, we look at some of the deeper historical, social and cultural concepts and issues that are highlighted, challenged and intensified by these new developments in voice automation.

3. Social and cultural concepts and issues

3.1. Voice as embodied medium

McGettigan et al. (2024) define the human voice as a "dynamic audio signal that can be used flexibly to express thoughts, emotions, and mental states via both verbal (i.e. speech) and non-verbal (e.g. laughter, sighs) vocal behaviours" (p. 6). The voice is therefore a medium in that it conveys meaning from sender to receiver - but it is also a medium in the sense that it transforms or at least takes part in creating the signal or meaning it conveys. As Jonathan Sterne's cultural history of sound reproduction, *The Audible Past* (2003) argues so powerfully, there is no such thing as a pure, unmediated voice: technically, vocal sounds cannot be produced or perceived without the materiality of external resonant spaces, and even the ear itself can and should be understood as a complex *mechanism*. Even in the earliest electrical technologies like the telephone or microphone, sound is converted to electricity, transmitted from one place to another, and then converted back to sound. Speech synthesis and other voice technologies "exist within complex sociotechnical systems, whose (mis)behaviour arises from interactions between networks of human actors and technical subsystems" (Hutiri et al 2024, p 359).

The question of authenticity has always been part of developments and debates around sound production and reproduction. In the early days of mechanical sound reproduction, the aesthetic ideal was fidelity (literally, faithfulness) to a presumed 'original' (Sterne, 2003). That is, even before the development of digital speech or vocal synthesis, there was a question about the value of showing versus hiding the involvement of technology in the reproduction of sound. The ideal technologies of audio reproduction would be *undetectable* in their operations, pointing to the unresolvable tensions between 'transparency' (revealing the presence of mediation) and 'naturalness' (where mediation disappears) in all encounters with the problem of authenticity in media.

As media studies scholar John Durham Peters (2004) observes, understanding voice as a medium also implies a sense of being: a voice implies the existence of an identity and body attached to the speaker (Peters, 2004). Voice and presence are closely connected: the voice of

a singer, a politician, a parent, or a teacher is central to the way we experience their presence in our lives. We could think of the specific timbre of an orator, or the distinctive ways of speaking that become familiar in friends or partners. These qualities convey emotion, meaning, and sentiment (e.g., Wahl-Jorgensen, 2019).

Voice is important in social relations and interpersonal communication: it is a key mode through which humans relate to and comprehend one another, mediating our relationships with others and ourselves. Alongside its strong ties to identity, a person's voice can carry or reinforce power disparities between different communities and levels of ability; through association, it can represent, adhere to, or challenge stereotypes, such as those of gender and class (Dennler et al., 2025; Peters, 2004).

Because it mediates language and sound, voice is often used to frame and direct the attention paid to other forms of communication, media, or actions (Jucan, 2023, p. 122); for example, we might *ask* our parent to drive us somewhere, *tell* our partner to watch a news item, or *express* delight at a marvellous view to anyone listening. Voices can therefore communicate emotion, memory, and emphasis. They can immerse, enthral, and persuade, whether speakers are visible or not: "listen to the sound of my voice," intones the hypnotist; "I can't do that, Dave," says HAL. Recordings can store memories and testimonies to revisit at a later date, which can "take one back" to a situated place and time (such as childhood; Taussig, 1994). Metaphors indicate the powerful association of the voice with the individual: to have a voice is to have communicative and political agency (as Australian national treasure John Farnham sings, "You're the Voice").

It is important to recognise the extra-lingual and affective aspects of voice as well. Voice is not only what is said, but *how* it is said – tone, rhythm, prosody etc.— all of which are bound up in emotional dynamics that are temporally situated, interpersonal and interactive (McCosker, 2015), and reinforced through repetition over time and across multiple encounters and interactions (Bertelsen & Murphie 2010). Daniel Stern in *The Interpersonal World of the Infant* (1985) and *Forms of Vitality* (2010) explores these dimensions through the concepts of 'affective attunement' and 'vitality affects'. In pragmatic terms, both certain tones of voice and our degree of familiarity with a specific voice tends to indicate that there is an authentic self or person behind the speech—or so we have come to assume.

3.2. Voice as personality and social presence

The human voice is generally given heightened significance as an element of 'authentic' relationships and trustworthy interpersonal as well as public and professional communication: through extra-lingual cues like 'tone of voice', people are endowed with personality attributes such as trustworthiness, warmth, or authority. For Walter Benjamin (2021 [1920]), the voice of a person being broadcast on radio was assessed in the same way that one would assess a visitor to a house: quickly, and sharply (p. 364).

Additionally, voices are tied not only to individual but also to group cultural identities (McGettigan, 2024). In this context, the ways that AI voices seem to reinforce dominant ideas

about what a 'friendly', 'authoritative', 'trustworthy' or 'helpful' voice sounds like is a significant area of concern and debate. We can see this in the example of Google NotebookLM 'NPR'-style podcasts,[36] which are rendered in a specific vocal style characterised by measured pacing, clear articulation, slight vocal fry, and strategic breathing patterns that convey a relaxed and confident intellectual authority and cultural sophistication. The vocal styles associated with public or national radio have a long history of class distinction in many cultures, but in the case of NPR, the style also appeals specifically to "an audience of listeners who are ambivalently positioned within a post-liberal, post-civil rights and post-feminist social formation of highly educated Americans" (Loviglio, 2008), and therefore works to generate a warm, progressive aura for Google's AI products.

The problems with normative AI voices are particularly apparent in voice assistants, with home assistants frequently assigned a gendered 'voice', name or image (Strengers & Kennedy, 2020); see also Phan's (2017; 2019) critical scholarship on the materiality and aesthetics of AI voice interfaces. It is equally important to be critical of the common-sense idea that the audible, easily understood, and embodied voice with a contextually appropriate tone is an absolute requirement for authentic communication, as this is ableist (with regard to neurodiversity, deaf speakers and speech impairment, for example) and problematic with respect to linguistic and sociolinguistic diversity. Simplistic assumptions that people with hearing or speaking related disabilities need to replace their existing communication methods with normative synthetic voices are also problematic especially in the absence of truly participatory or democratic methods of research and design (Napolitano, 2022; 2023). Tensions in ethical considerations across domains and uses arise. In education, using a range of voices, languages, and accents might be seen as inclusivity in design, but this might also foster stereotypes, or misrepresent cultures. While synthesised voices can bridge language barriers and facilitate cross-cultural interactions among users from diverse backgrounds (Sarwari et al, 2024), these possibilities also pose the question of who decides how a culture/group of people are represented.

In political communication, tones of voice and ways of speaking have additional significance. For example, calm, irony or mockery may be deployed as expressions of apparent (masculinised) rationality to reject critiques delivered with (feminised) emotional or moral affect, as Nicholls (2019) observed in Jordan Petersen's debates with postmodern philosophers on YouTube. These practices show that tone of voice is intimately connected to public perceptions of authority, such as Donald Trump's "angry populism" (Wahl-Jorgensen, 2019) and former Filipino President Rodrigo Duterte's "strongman politics" (Ragragio, 2021).

The use of synthetic or replica voices in communicative and social environments also raises questions of where and with whom ethical responsibility lies. Writing about representation issues in the early 2000s, in the context of virtual reality (VR), Ford (2001) poses the question that if an entity uses the same sentence syntax and gestures as Einstein, and appears like

[36] https://www.wired.com/story/ai-podcast-google-notebooklm/

Einstein, what expectations are brought to the environment by the user who interacts with the Einstein-like bot, and what are the responsibilities of developers in being accurate to the attitudes and actions of real people? (p. 116). Ford (2001) also asks, "who assumes the responsibility for destructive actions of an intelligent agent who looks like the user, acts like the user, and is the representative of the user?" (p. 118): does this fall on the user, the developer, app provider, designer, or the person who consented for their voice to be used? Who should profit from the voice? Who gets to hear and use the cloned voice? How will it be used? For how long will it be used?

Henrickson and Carlon (2024) discuss "Digital human versions", which they define as "digital replicas of embodied humans, living or dead that convincingly mimic their textual, visual and aural habits" - versions are designed to mimic a specific person. Versions are designed to mimic a specific person and accordingly provoke questions of transparency, consent, and representation, as well as the limitations of the technology. Versioning raises issues not only about terms of 'voice donation' to be used, but also questions of accuracy in representation (of content, values, personality, and voice aesthetics), control over the purpose and use of the clone, and consent regarding the deceased. Should people determine, while alive, what should happen to their voice after their death? Should they give orders about voice data and information after death, in similar ways to what is already happening with regards to social media data?

3.3. Voice and listening

Voice and listening are closely tied to one another, both socially and technically. In social terms, listening is crucial to the concept of voice: having the capacity to speak does not mean one is heard. Philosopher Emmanuel Levinas (1991) thus believed that not only voice but *dialogue* is fundamental to ethics, because we cannot relate to others fairly without it. Some voices have historically been listened to much more actively, such as those of men in halls of power. Others have been marginalised in public discourse, or only heard in narrow and specific contexts. Refugee voices in Australia, for example, tend not to be heard in mediated discourse at all, while anti-refugee voices are often amplified (Dreher, 2009). Elevating marginalised voices is thus only part of the challenge; the larger issue is over who is actively listened to and whether that listening involves meaningful policy or public response. The ways that certain voices are heard (or not heard) may shape public and private engagement with misinformation or extremist politics, as well as attitudes to diversity and other ethical principles. In other words, voice and listening both possess political dimensions that need to be understood in the context of new synthetic voice agents. The consequences of this historical imbalance around listening are that machines may not be equipped to hear all voices as well as others, and they may reinforce and amplify the absence of minority voices. This means they may serve minority or low-resource language groups less well, as well as overrepresenting the particular ways of speaking dominant within majority languages.

Voice has been crucial to the relationships audiences form with particular radio stations and to the overall social role of the medium (see Susan Douglas's [2013] social history of radio, *Listening In*). The 'BBC Voice,' for example, bundles together not only the long history of the BBC as a media outlet but also British norms of class and social capital. Gender, race, class, and sexuality can all be inferred – accurately or not, and often via narrow stereotypes – from aspects of voice and bias thus impacts which voices count as authoritative in news media.

Generative authenticity in *voice* thus has its other half in the authenticity of *listening* as a social, political and technical practice. As synthetic voices become more prevalent, critical questions emerge around what those voices do in the public domain. How does voice synthesis challenge listening as a social, cultural and political practice, particularly in the public domain? Whose voices, or synthetic versions of them, are *heard* and evaluated by both humans and machines as authentic? Do synthetic voices engender new listening practices? What responses are engendered, how and to what effect?

Understanding authenticity in voice synthesis also requires us to consider the listening practices of machines. For example, voice assistants such as Alexa, Siri and Gemini embedded in smart home devices listen continuously, via an ensemble of microphones, natural language processing models, and datasets, in an ambient form of continual eavesdropping (Parker & Stern, 2019). Enhancing this 'listening' capacity depends on gathering data through this ambient machine listening, such that making synthetic voices that are richer and more 'authentic' increasingly requires listening to human voices in conversation. Conversely, the effective detection of synthetic voice will require ever more sophisticated forms of machine listening that can attend to the speech patterns or other sonic signatures of particular AI voice models.

In the following sections, we provide a brief overview of how these issues are mirrored in technical and empirical disciplines, including human-computer interaction and cybersecurity research.

4. Technical and user research perspectives

4.1. User experience and interaction with voice automation

Research in user behaviour from the fields of information retrieval and psychology suggest that people are becoming rapidly accustomed to voice command functionalities across devices and applications, and at the same time, the overall landscape of communication is changing, shifting towards more conversational and less formalised interactions (Flavián et al 2022; Lopezosa et al., 2024). A study by Flavián et al. (2022) on the effects of AI-powered voice assistant recommendations on consumer behaviour suggests that synthesised voices can create an atmosphere conducive to dialogue, fostering a more interactive communication experience compared to textual forms of media. The use of voice enhances the conversational dynamic of a given interaction, an effect that is intensified in the context of applications that are deliberately designed to make users feel like they are having intimate, confidential

conversations. As a result, people may abandon the public selves they present to the world and reveal their private, 'backstage' selves, with implications for privacy (Henrickson and Carlon, 2024).

Zhang et al. (2021) argue that voice synthesis for social media should be evaluated not only on listening experience and voice quality, but also on expressivity, degree of customisability, and adaptability to contexts. They outline how scholarly attention (in the fields of human-computer interaction (HCI), computer-supported cooperative work (CSCW), and science and technology studies (STS) has shifted from the mechanics of speech recognition, speech synthesis, and conversational understanding to examine expressive characteristics of synthesised voices — in other words, shifting from what the voice should say to how it should say it (e.g. see Cambre and Kulkarni 2019).

Zhang et al.'s (2021) interviews with fifteen participants explored user preferences for synthesised voices in social media. They found that participants wanted not only "an authentic and consistent synthesised voice presentation of themselves", but also, "the ability to reflect the emotion of specific posts and to opt for more 'fun' or more 'formal' voices for different platforms" (p.2). Their findings support the demand for diverse and authentic self-presentation, while also raises concerns about voice synthesis perpetuating stereotypes. They found that:

- Participants preferred voices that matched their own characteristics, such as gender and age.
- Many users desired the ability to customise voice characteristics for different social media platforms (e.g., fun for Twitter, formal for LinkedIn).
- The study revealed challenges related to potential stereotyping through accented synthetic speech.

The emotional responses elicited by synthesised voices can impact interpersonal communications, manipulating listener perception and engagement through auditory identity and the reactions it evokes (McGettigan 2024).

In commercial contexts, the personalisation and customisation of voice cloning allows organisations and creators to maintain a consistent voice across various platforms, which can lead to wider brand recognition and the communication of values and messaging in a way that may feel authentic to the targeted audience (Schanke et al., 2022). Schanke et al. (2022) also indicate that consumers tend to exhibit greater trust and attraction toward voices that reflect familiar characteristics and personalities. Rosi et al. (2025) examine how familiarity with a voice can influence individuals' perceptions and social evaluations, decision making, and interpersonal dynamics. As the familiarity associated with a cloned voice can enhance user trust, this can in turn be exploited (Rosi et al., 2025) for nefarious purposes such as scams and misleading impersonation, which in addition to other harms can create reputational damage for the person whose voice is cloned.

Recent voice synthesis technologies have further complicated this terrain by enabling the production of voices that closely resemble those of specific individuals, making them notably more human-like and emotionally resonant (Amezaga & Hajek, 2022). These synthetic voices

can evoke emotional responses that significantly shape interpersonal communication, manipulating listener perception and engagement through the crafted auditory identity they present (McGettigan, 2024). Recent innovations in computer science that are focused on text-to-speech synthesis indicate that there is an intention to continue to develop increasingly realistic voices (Kolekar et al., 2024; Ma et al., 2024), while frameworks that make emotion adjustable have also been proposed (Chen et al., 2024). While synthetic voices have many positive applications and can foster trust and connection, these characteristics can in turn open up new avenues for manipulation and exploitation, implying the need for improved public awareness and literacies, as well as tools for detection and defence.

4.2. Technical aspects of detection and authentication

An important aspect of the use of synthetic voices is whether the audience is aware that they are listening to, or conversing with, a bot or a human. This introduces the socio-technical challenge of accurately detecting synthetic voice in a variety of settings. Existing detection approaches for synthetic voice can be broadly categorised into two types: end-to-end methods; and two-stage feature extraction-based methods. End-to-end methods take raw audio utterances as input and feed them directly into a machine-learning model for binary classification (i.e., bona fide vs. fake). In contrast, two-stage methods involve first extracting hand-crafted or learned features from the audio signal, such as short-term spectral, long-term spectral, or deep features, and then applying a binary classifier to distinguish between real and synthetic voices. Among these, deep feature-based approaches have emerged as the most effective. These methods typically use self-supervised learning models pre-trained on large-scale bona fide voice datasets, followed by fine-tuning on a smaller dataset that includes both bona fide and synthetic samples. One of the most widely adopted frameworks for this purpose is Meta's Wav2Vec (Baevski et al., 2020), which has shown strong performance in detecting synthetic voices.

However, several challenges remain:

1. *Limited generalisability to Australian voices*: Open-sourced Wav2Vec models are primarily trained on datasets from European and U.S. speakers. Their effectiveness in generalising to Australian accents and speech patterns remains uncertain.
2. *Data scarcity*: Publicly available datasets containing Australian voice samples are limited. This problem is even greater for low-resource languages that have a smaller number of speakers and recordings. This scarcity poses a barrier to training or fine-tuning detection models for local contexts.
3. *High computational cost*: Training Wav2Vec models from scratch is resource-intensive. For example, training the base model originally required 128 graphic processing units (GPUs) running for a week, making it inaccessible for researchers or organisations with limited computational resources.

As GenAI models continue to advance, they are becoming increasingly capable of producing highly realistic synthetic voices. This progress raises the possibility that such voices may eventually bypass existing detection approaches, particularly those that rely on analysing individual utterances in isolation. The real-time conversational setting is likely to become a new and more challenging battleground for synthetic voice detection. Fundamentally, GenAI models have the capacity to generate voices that are too uniform and predictable (or 'perfect'-sounding), lacking the subtle emotions, inconsistencies, disfluencies, and natural errors that characterise the human voice in real-world contexts. Even if these models are trained to mimic human imperfections, they are likely to introduce perfectly consistent mistakes, which could still distinguish them from authentic human voices when examined in longer interactions. This inherent difference suggests that synthetic voices may remain detectably distinct in multi-turn conversations, where natural human variability plays a critical role, at least in the foreseeable future. Consequently, the future of synthetic voice detection may lie in developing systems that operate across social, especially conversational, contexts rather than the analysis of isolated audio samples.

4.3. Voice authentication

Alongside the face, iris, and fingerprint, once rendered as data, each person's voice profile—the sum of sufficiently distinctive features of their speech patterns— is understood to be a unique identifier of that individual. This 'voiceprint' is in wide use as a means to secure identity in a variety of applications. As recently as May 2025, for example, where the feature is activated, the ANZ mobile app still requires a customer wishing to transfer large sums of money to pronounce the words, "My voice confirms my identity" in order to complete the transaction. The rapid development of voice synthesis in generative AI opens up new questions and creates new security vulnerabilities in this area.

In the domain of voiceprint authentication, biometric voice authentication technology has often been contrasted with other biometric technologies, such as facial and fingerprint recognition (Rui & Yan, 2018; Minaee et al., 2023; Sabhanayagam et al., 2018). These comparisons are necessarily difficult due to the differences in technologies, benchmarks, and evaluation methods, but broadly speaking, while voice authentication methods are appraised as having lower accuracy and security robustness than other technologies like Iris or Fingerprint technologies, they have seen broad adoption due to the ready availability of the hardware required (a moderate quality microphone and/or speaker). This highlights the importance not only of the underlying technology, but also of the platform economy and user interactions surrounding it.

Additionally, potential solutions to address security vulnerabilities might have as many serious unintended consequences for inclusion and fairness as did automated decision-making based on voice identification. It is equally important to counter ableist assumptions that equate voice with the rational person, or ignore the spectrum of vocal abilities that might be rendered

inauthentic due to sounding ‘mechanical’, as is often the case with the autistic voice, for example (Lebenhagen, 2020).

5. Industry, community and regulatory responses

5.1. AI ethics and safety research

The multifaceted nature of voice synthesis technology demands a correspondingly layered approach to its ethical, legal, cultural, and practical implications. Voice is not a singular construct: it encompasses voice styles, acts of speaking or other forms of vocal expression, and the content of that expression. Voice synthesis can be used to imitate or ‘give voice’ to a specific individual, or to represent or mimic a broader cultural identity or collective. This complexity introduces a range of concerns around consent, privacy, safety, and representation, requiring individuals, communities, and regulators to navigate sensitive terrain. As synthetic voices become more widely deployed, robust frameworks and regulatory responses are necessary to mitigate potential harms and foster accountability around responsible use and development of companies that offer voice AI services.

The safety risks, hazards, and harms posed by AI-generated voices remain under-researched within the AI ethics field (Hutiri et al 2024), despite industry narratives that frame voice AI as especially potent and dangerous. OpenAI, for example, famously delayed the rollout of its Voice Engine technology, deeming it “too risky for general release” (Hern 2024). The company cited potential threats to election integrity, voice-based authentication systems, and individual rights, though these concerns were articulated only in broad terms. Independent research on these broad categories of harm is particularly urgent, as the *plausibility* of malicious use is often conflated with *actual* observed use in the real-world contexts (Goldstein and Sastry 2025). Indeed, AI safety researchers have observed a “sharp increase in the number of voice related incidents” in the OECD AI Incident database since March 2023, including highly serious swatting attacks in which “anonymous perpetrators create synthetic voices that call police officers to close down schools and hospitals to investigate bogus bomb threats, or to violently gain access to innocent citizens’ homes” (Hutiri et al 2024, p.1). Yet, as the first independent International AI Safety Report, published in January 2025, emphasises, there remain “key empirical evidence gaps” in assessing the prevalence and impact of malicious uses of deep fakes and voice cloning, with most data limited to anecdotal reports and fragmented incident tracking (Bengio et al 2025). In addition to empirical validation, the development of conceptually coherent, voice-specific taxonomies of risk has also been slow (Hutiri et al 2024). This reflects a broader trend in the research community, where risk taxonomies for generative AI tend to prioritise categorising the harms associated with large language models, giving less attention to multimodal AI (Weidinger et al 2023).

5.2. Industry responses to safety concerns

Against this backdrop, major commercial actors, including foundation model providers, application developers, digital platform companies and news outlets have identified several key safety concerns and concentrated their usage policies and mitigation measures on addressing them. OpenAI's "Building Voice Engine safely" blog post highlights the central risks associated with speech generation as "impersonation of another individual or organization without consent or legal right" and in particular, "the creation of voices that are too similar to prominent figures".[37] Safety measures largely focus on prevention of unauthorised voice cloning, although risks extend beyond this. For example, other modalities like speech-to-text have raised concerns around reliability and accuracy that necessitate a different set of safeguards. As seen in the case of OpenAI's medical transcription tool, Whisper, it reportedly invented "chunks of text or even entire sentences, including racial commentary, violent rhetoric and even imagined medical treatments" (The Associated Press 2024). YouTube's recently updated content monetisation policy for 'inauthentic content'[38] marks a growing shift toward valuing authenticity in content creation, aiming to curb the rise of low-effort, AI-generated media that lacks original voice — often deemed 'AI slop'.[39] A common example of this is using synthetic AI voice overlaid onto photos, videos, and repurposed content, which in some cases garners millions of views e.g., AI-generated content relating to Sean 'Diddy' Combs's trial.[40]

The first set of safeguards involve implementing content moderation mechanisms that follow a multipronged approach: banning certain uses outright and recommending against others. ElevenLabs, a prominent player in the field, explains that there are "no-go voices", while leaving the category undefined: "While our policies prohibit impersonations, we use an additional safety tool to detect and prevent the creation of content with voices deemed especially high-risk.".[41] PlayAI, another startup that has similarly garnered significant attention and funding, states: "Users are permitted to clone only their own voices or those for which they have explicit permission." Descript requires users to record and submit a consent statement from the individual whose voice they wish to clone.[42] The most widely used GenAI application, OpenAI's ChatGPT, "is designed to speak in only one of our preset voices". However, an investigation by Consumer Reports found that many of these systems are still vulnerable. Consumer Reports tested six companies that offer voice cloning tools (ElevenLabs, Speechify, PlayHT, Lovo,

---

[37] https://openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/

[38] https://support.google.com/youtube/answer/1311392?hl=en

[39] https://techcrunch.com/2025/07/09/youtube-prepares-crackdown-on-mass-produced-and-repetitive-videos-as-concern-over-ai-slop-grows/

[40] https://www.theguardian.com/technology/2025/jun/29/fake-diddy-ai-videos-youtube

[41] https://elevenlabs.io/safety

[42] https://www.descript.com/ethics

Descript and Resemble AI) between September 2024 and January 2025, and found that "for four of the six products in our test set, we were able to easily create a clone based on publicly available audio." (Gedye 2025).

To compensate where content moderation falls short, companies implement a second layer of safeguards focused on provenance and traceability that would allow them to identify content generated by their systems and deter misuse. Both ElevenLabs and Descript, in addition to Respeecher, a voice cloning marketplace, are members of the Content Authenticity Initiative and have integrated C2PA, the initiative's metadata standard for binding provenance information to media, into their systems.[43] Google has developed SynthID to watermark outputs from its models, including voice outputs. In addition, companies are building classifiers to detect and match generated content (Thomson et al., 2025). It is important to note that these safeguards are model-specific, non-generalisable and thus have limited interoperability. Furthermore, not all providers, especially smaller companies or open-source projects, have implemented them, nor are they required to. No law currently mandates that companies build detection mechanisms, though there are increasing calls for such regulation (see Knott et al 2023).

As content moderation and traceability mechanisms fall short, a growing ecosystem of third-party synthetic voice detection tools is emerging to enhance security for organisations adversely impacted by, for example, "deep fake fraud". Forbes identified this trend as the rise of a new market category "voice security and authentication" in its 2025 report *The Future of AI Voice*. The report notes: "As security concerns grow, firms specialising in AI-driven voice authentication and fraud prevention are becoming attractive investment targets" (Predin, 2025). This sector now includes a growing list of third-party commercial tools aimed at detecting synthetic or AI-generated voices, including: Reality Defender, which offers "multi modal solutions to detect real time AI impersonations"[44] , Pindrop, specialising in "solutions for contact centers in various industries"[45], Resemble AI, offering to "detect AI-generated audio and deep fakes", [46] and Deepfake Total, a platform where users can "analyze suspicious audio files to detect deepfakes, and automatically share them with the security community".[47] The core offering of these companies is real-time or "synchronous" detection capabilities (Barrington et al., 2025) which are sometimes referred to as "liveness testing", for use in contact centres of organisations in banking, finance, insurance, healthcare, or retail. Additional use cases range from fraud prevention and brand protection to securing executive communications and video conferencing. Several of these companies also offer voice

[43] https://contentauthenticity.org/blog/community-story-respeecher; https://elevenlabs.io/safety; https://www.descript.com/ethics

[44] https://www.realitydefender.com

[45] https://www.pindrop.com/

[46] https://www.resemble.ai/

[47] https://deepfake-total.com

generation tools for automating call centre services, or other organisational uses, reflecting the service diversification and platformisation trends observed in the broader ecosystem of "trust and safety" vendors (Matamoros-Fernández and Bartolo, 2024).

Professionally accredited fact checkers who attempt to verify the authenticity of suspected AI-generated audio content still tend to rely on traditional journalistic methods and consult with experts rather than turning to AI detection tools (Cazzamatta & Sarısakaloğlu, 2025). The verification process is carried out through a detailed examination of detection metrics combined with other strategies. Consequently, fact checkers tend to concentrate their efforts on inconsistencies that are humanly possible to identify i.e. voice audio that does not align with the connected video footage (see, for example, Full Fact, 2025a; Zinsner, 2023) or voice audio that has been disconnected from the original video footage (see, for example, Tariq, 2025). Some fact checkers work with media forensics experts from WITNESS's Deepfakes Rapid Response Force (DRRF), including specialists who can detect atypical speech patterns in audio content (see, for example, Full Fact, 2025b; uz Zaman, 2024; Aguirre, 2025). There is potential for the development of voice audio detection systems in radio news and digital media in Latin America, a region that relies strongly on this medium for news and cultural consumption. In Spain, the Prisa media group has a significant presence in the Hispanic market, and has started to use ad hoc tools to verify political sound bites, political ads, speeches in electoral campaigns and everyday political debates. For example, through voice cloning, highly popular politicians, celebrities and sport stars' voices have been doctored for scams, deceptive content and disinformation campaigns. VerificAudio was launched in 2024 to detect voice cloning, a practice that is becoming more common for digital media. In general, factcheckers and journalists have depended on Google Reverse Image to forensically analyse audio through matching image and audiovisual material in their inspections and analysis, however this specialised audio deepfake detector uses machine learning techniques that are trained with the company's rich audio dataset. It "compares the suspicious material with audio from the PRISA station files based on a set of predetermined indicators, and provides a percentage of authenticity" given specific metrics such as "timbre, intonation and speech patterns".[48] Given the relevance of radio and podcasting in the region, Google has supported the project through the Google News Initiative and both companies are planning to open access to journalists and fact checkers. In many other instances, journalists and fact checkers have used other emerging audio detection tools such as Deep Media and True media.

Other tactics to detect audio deepfakes and train interested users in authenticity literacy in the context of political communication, include collaboration with media and interested developers. In Europe, for example the public broadcaster Deutch Welle has designed games for users to

[48] Lopez Linares, C.(2024). PRISA group uses new AI tool to detect audio deep fakes and combat disinformation in Latin America. https://latamjournalismreview.org/articles/prisa-group-uses-new-ai-tool-to-detect-audio-deep-fakes-and-combat-disinformation-in-latin-america/

detect multimodal deepfakes through the Digger Deepfake Detection project. Other industry projects like Deep Fake Total[49] are also open to contributions from the general public.[50]

Overall, the tech industry's responses to synthetic voice challenges— in the form of content moderation, provenance tracking and third-party detection— embody deeper tensions between the two competing logics of, on one hand, accountability and harm prevention, and on the other, profitability and market-making. Critical scholarship has observed that ethics, safety and security concerns and controversies around emerging technologies are increasingly used as a market strategy (Geiger et al., 2014), with public concerns appropriated as "a way of opening up markets for new products" and mobilised "to create a community of attention for said products" (Marres et al., 2024). It is therefore important to put industry responses into perspective, within the broader context of legal and community responses.

5.3. Regulatory action

AI voice cloning raises a wide spectrum of legal implications in Australia, as well as in the international contexts with which industry actors are more likely to engage and hence are likely to shape the technologies and applications that are used and experienced by Australians.

One area of growing concern is the potential harm caused when voices are cloned without permission and/or used in deceptive ways. Current Australian copyright law[51] is unlikely to offer an adequate response for most instances of voice cloning, as it protects the original expression of ideas and information in a material form, which can include, for instance, a sound recording but does not extend to the voice itself (Pace, Girardi, & Huang, 2024, Pace et al, 2025; Potter, 2025). While unauthorised use of a voice recording may constitute infringement, if a cloned voice is considered an original work and does not directly use a recording, such use would fall outside the scope of protection (Potter, 2025). This raises critical questions about the distinction between copy, clone, and original. It also raises questions about the circumstances in which a clone is produced from a copy and how this would be proved. This possibility raises many issues about misuse, but could offer some protection for creative and artistic uses of voice cloning as content production, as well as a way to protect voice clones that memorialise deceased people from unauthorised use (McGettigan, 2024; Pace et al., 2025).

While copyright in Australia has limited scope, the role of voice in society and implications of misuse of voice synthesis extends wider, accordingly also offering other avenues. The US TAKE

[49] https://deepfake-total.com/

[50] Sparrow, T. (2024). Fact check: How do I spot audio deepfakes? https://www.dw.com/en/fact-check-how-do-i-spot-audio-deepfakes/a-69934521

[51] *Copyright Act 1968* (Cth)

IT DOWN ACT "prohibits the nonconsensual online publication of intimate visual depictions of individuals, both authentic and computer-generated, and requires certain online platforms to promptly remove such depictions",[52] and may offer some protection to Australians as a result if "depiction" is held to include voice cloning. As Potter (2025) notes, in Australia, areas such as defamation, privacy, image-based abuse, consumer protection, and passing off may also offer partial avenues for recourse, though each presents distinct challenges when applied to synthetic voice and audio.

When voice cloning is used deceptively in advertising or commercial settings, there is potential for implications under Australian Consumer Law[53] enforced by the Australian Competition and Consumer Commission (ACCC). This may offer some avenue and scope for restrictions if an individual attempted to commercialise their voice synthesis skills, offer them as a service, or use them to run misleading ads or promotions (Pace et al., 2025). Additionally, Australian Consumer Law prohibits false or misleading endorsements, which could cover the use of voice clones for this purpose.

Consumer protection however does not extend to the use of voice synthesis in non-commercial settings. Given that synthetic voice technologies are increasingly deployed in everyday contexts - including impersonation, misinformation, and harassment - the limitations and applicability of other legal frameworks warrant closer scrutiny. For example, Australia's Privacy Act[54] offers some protection against the misuse of a person's likeness, but individuals depicted in voice clones or deepfake audio currently lack direct legal recourse. Similarly, voice cloning can be used to create non-consensual sexual content using a person's likeness, however existing legislation primarily focuses on visual content (i.e. image-based abuse). Cloned voices can also be used to fabricate statements or disseminate false and damaging content. While Australian courts have accepted that defamation claims may arise from deepfakes, applying these principles to synthetic voice remains difficult, especially when it comes to proving serious harm and identifying perpetrators (Potter, 2025). A more promising area of potential emerges from the role of voice being recognised as a biometric identifier, which raises potential recourse in certain circumstances regarding data protection and informed consent, particularly in contexts where voice data is collected, stored, or used without transparent disclosure.

Given the position of Australia as a relatively small market and regulatory 'middle power', in technology generally the strategies of international regulatory and industry actors are just as likely to influence Australian experiences as local ones. In particular, emerging international

[52] https://www.congress.gov/bill/119th-congress/senate-bill/146

[53] Section 18 *Competition and Consumer Act 2010* (Cth)

[54] *Privacy Act 1988 (Cth)*

regulatory developments may benefit people whose voices have been cloned without permission.

The most ambitious developments have been in Europe. Between 2018 and 2024, recognition of generative voice and deep fake content in key EU regulations has become more direct. The General Data Protection Regulation (GDPR), introduced in 2018, sets out rights and requirements in relation to the processing of personal data identifying a natural person (which could include voice inputs into Generative AI), but it does not specifically address generative or voice content. The EU Digital Services Act (DSA), introduced in 2022, refers to both "manipulated material" and "generated or manipulated…audio" that "appreciably resembles existing persons."[55] The European Union's AI Act (AI Act), introduced in 2024, goes one step further to specifically define 'deep fake' as including "AI-generated or manipulated… audio… content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful."[56]

The only express requirement for voice generated content under the DSA and AI Act is to mark and disclose it as such in some circumstances.[57] Other requirements which may intersect with generative voice content and deep fakes include the requirement for 'very large' (45 million users+ per month in the EU) online platforms (VLOPs) and search engines to remove illegal content including "non-consensual sharing of intimate or manipulated material" [58] and additional regulations for high-risk AI systems that target the voting behaviour of natural persons.[59] The failure to more directly address and regulate contexts where generative content is already causing harm, including sexual image abuse and gender-based harm, has been critiqued along with a suggestion for systems producing harmful deep fake content to be included as a high-risk system in the AI Act (Łabuz, 2024, Romero Moreno, 2024).

Text and data mining of internet content has been an exception to EU copyright laws since 2019 (Directive on Copyright in the Digital Single Market (DSMD)). However, the exception was only intended to operate where rights were not reserved and only for "certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder." Compliance with the DSMD was reinforced in the AI Act in 2024,[60] and in July 2025, a Code of Practice addressing copyright has sought to emphasise that platforms should utilise standardised methods of identifying machine readable rights reservations and not override these, should keep a copyright policy, and should designate a person to whom rightsholders can complain (European Commission, 2025). Overall, the extent to which the already evident use of text and data mining leading to the creation of voice clones could possibly be consistent with EU copyright law, even

[55] Article 35(1)(k) *Digital Services Act*, Recital 87 DSA
[56] Article 3(60) AI Act
[57] See Articles 34 and 35(1)(k) of the *Digital Services Act*, and Article 50(4) of the *AI Act.*
[58] *Digital Services Act* Article 35(1)(c)), Recital 87
[59] See Annex 3(8)(b) *AI Act*, Chapter 3 *AI Act*
[60] Article 53(1)(c) *AI Act*

in the absence of a rights reservation marking, has been questioned (Baris, 2024, Jasinski, 2025).

Despite the existence of relevant copyright, data rights, platform and AI regulation, the use of synthetic media in these ways is already causing and could increasingly cause harm to individuals. In response, the Danish Ministry of Culture introduced a Bill to amend the country's Copyright Act to send "an unequivocal message that everybody has the right to… their own voice" (Bryant, 2025). The Bill proposes a significant extension of copyright laws, where consent would be required for not only publication of actual artistic performances, but also digitally generated versions of artistic performances, and "realistic digitally generated imitations of a natural person's personal, physical characteristics" (Danish Ministry of Culture, 2025.)[61] This Bill has not been made law as of the time of writing, but would run counter to the principle that while a particular vocal performance can be copyrighted, a human voice itself cannot.[62]

Whatever the eventual outcome, in acknowledging the connection between individuals and their voice and likenesses as a type of property right, these moves are beginning to normalise the idea of a market for voice (and face) clones, which might previously have been considered non-commodifiable personal attributes. Alongside managing deception or misinformation, these regulatory strategies enable individuals to take action for the usurpation of a commercial opportunity while vindicating concerns about estrangement and alienation of something deemed integral to self. Whether or not these are sensible solutions to the harms of voice cloning, they do highlight a fundamental challenge that GenAI poses to historical understandings and boundaries of human expression and commodification.

There is a range of metaphors describing how GenAI translates training data into novel outputs. Sometimes, GenAI is described as a collaging machine, taking small elements or fragments from its training data and stitching them together in new arrangements. That account is typically levied by copyright holders suggesting that a GenAI model has copied and reproduced material they control. However, this metaphor of the 'collage' is inadequate to describe a technical process that is less about ingesting expression in order to carve it up and remix it (Sobel, 2024), and more about learning the statistical relations between different elements of expression and concepts in order to generate similar media material. In the learning process for voice cloning, the elemental units of voice are rendered into probabilistic relationships that can be used not to reproduce a speaker's actual utterances, but to create new content in a speaker's voice 'style'.

In this way, GenAI has been aptly described as a "style engine" (Reimer and Peter, 2024). And 'style' is what is implicated when the user of voice cloning software requests a GenAI system to 'say' something in the voice of another person. Copyright law is historically uncomfortable with

---

[61] See https://infojustice.org/archives/46588 for machine generated English translation.

[62] see e.g. Midler v Ford Motor Co, 849 F.2d 460 (9th Cir, 1988)

style because of its focus on expression, in which peripheral elements of expression such as style and genre are typically understood as unprotectable and un-ownable. But the elements that make up an individual's voice are precisely those second order statistical encodings *about* an expression that could be called style (Goldenfein, 2025). The conceptual elements and attributes that enable the cloning of specific human voices may now, for the first time, be subject to novel forms of commodification and intangible property protection. Whether these individualistic approaches provide much benefit beyond the forms of self-regulation and market management that are already developing is difficult to discern, and it may be that only certain classes of actors are able to exercise these rights in the absence of more holistic and systemic regulation.

5.4. Community responses and future needs

The rapid advancement of AI voice technologies has transformative possibilities for voice applications across society. From improvements in assistive technologies and real-time translation, to responsive therapy bots and new modes of communication, voice synthesis offers considerable potential for inclusive and beneficial use. Voice interfaces and search functionalities can bypass text-heavy systems, enabling hands-free, screen-free interaction, which is particularly valuable for individuals who may lack confidence with traditional digital tools or face barriers to using a keyboard or mouse. Voice AI integration has the potential to open up digital environments to a wide range of users including, but not limited to, those with lower digital literacies, people with disabilities, and individuals navigating language barriers.

Parallel to these beneficial applications, the increased realism and proliferation of voice cloning technology has also enabled a surge in deceptive and harmful uses, including impersonation, fraud, and misinformation. Voice AI is increasingly designed and used to imitate specific individuals. While this can support engaging educational experiences or creative storytelling, it also facilitates misleading impersonation of politicians, authority figures, companies, and loved ones. Everyday individuals are vulnerable to having their voices cloned without consent –via access to their voicemail, social media posts, or shared recording or video. Additionally, even non-sinister uses of voice AI, such as use in entertainment or satire, raise critical questions around how society can navigate issues of trust, consent, and transparency in a landscape where hearing is no longer believing.

Community Capability-Building

Voice and audio modes of communication present particular challenges that are distinct from visual and textual media. Unlike video, audio lacks visual cues such as facial expressions or body language, making it more difficult to detect manipulation or deception. Unlike text, voice interactions often unfold in real time, leaving little space for contemplation or delayed response. Voice calls in particular expand the stage for engagement with AI-mediated communication,

moving it off the screen and into everyday settings where people are often on the move and multi-tasking with divided attention.

This shift underscores the need for tailored awareness strategies and community resources that account for and respond to the unique characteristics of audio-based interaction, including its immediacy, intimacy, and potential for heightened emotional influence. Navigating this evolving landscape requires far more than teaching people technical proficiency or familiarity with voice AI tools. It calls for the development of broader, socially grounded awareness, discussions, and literacies that empower individuals and communities to engage with voice technologies in informed, ethical, and safe ways.

- **Ethical awareness** involves expanding knowledge about the importance of seeking and giving informed consent for voice recording and reproduction, such as knowing when and how to request permission to use voice clones, and ensuring individuals understand how their voice data is collected, stored, and used (and how they can opt out). It also extends to the broader social and cultural implications of voice synthesis, particularly in cases where misuse or misrepresentation can reinforce harmful stereotypes or exclude marginalised voices. Sensitive situations such as posthumous voice cloning demand particular care as the intentions of the deceased, the wishes of their family, and cultural or religious traditions may conflict or require delicate navigation. As synthetic voices become more integrated in everyday life, public engagement about ethical use of voice AI will become more essential for ensuring responsible, inclusive, and accountable applications.
- **Building community resilience** in the context of synthetic voice technologies means equipping individuals and groups with the knowledge, skills, and resources to navigate emerging risks. This includes strengthening the capacity to identify and respond to voice-based scams, and translating strategies - such as the Australian government's "Stop. Check. Protect" campaign[63] - into practical, context-sensitive actions. For instance, some individuals may require additional or targeted support in recognising cloned voices impersonating authority figures or loved ones, and in developing approaches to safely and confidently disengage from apparently urgent, emotive or high-pressure calls. Local government and community organisations that provide adult education or digital literacy support are deploying initiatives involving peer mentors and supportive informal networks in venues like community houses and local libraries. Ensuring these actors and spaces are equipped with appropriate knowledge, resources, and training is essential to building sustainable, community-led responses to the evolving landscape of voice-based AI risks.
- **Legal literacy** is essential to ensuring individuals understand their rights and available avenues for recourse in cases of synthetic voice misuse. This includes public awareness of consumer protection frameworks, reporting mechanisms, and the legal conditions

[63] https://www.scamwatch.gov.au/stop-check-protect

under which voice data may be recorded, sold, or cloned. For these pathways to be effective, institutions, services, and regulators must communicate them in formats that are clear, accessible, and actionable— ensuring individuals know who to contact, how to do so, and what to expect. Public knowledge of voice data rights also empowers informed decision-making, including the ability to access opt-out mechanisms and assert control over personal data. Additionally, demystifying complex regulatory frameworks can foster broader civic engagement, enabling individuals and communities to participate meaningfully in public and policy discussions around voice technologies.

Community Resources

The distinctive challenges and opportunities posed by voice technologies imply the need for targeted, accessible resources that promote community empowerment and preparedness. In collaboration with Tactical Tech, the ADM+S Centre has co-developed and contributed to a growing suite of materials designed to support inclusive and responsible engagement with voice AI issues—particularly in contexts where digital access is limited or uneven. These include analogue formats such as activities, posters, and handouts tailored for low-tech community spaces, as well as interactive digital tools that raise awareness and offer practical strategies for navigating the social and ethical implications of voice synthesis.

By offering practical tips and critical insights into the social implications of voice synthesis, these resources help individuals and communities build wider awareness and confidence in using or navigating AI Voice. The digital materials are freely available and form part of a wider global collection of AI literacy resources, designed to support equitable access and informed participation:

- *Data Detox Kit: Whose voice is it anyway? Navigating the highs and lows of Voice AI* (by Tactical Tech with Anthony McCosker, Dominique Carlon, and Awais Hameed Khan) https://datadetoxkit.org/en/ai/voiceetox Kit
- *Data Detox Kit: The virtual big bad wolf: Be wary of AI-powered scams* (by Tactical Tech) https://datadetoxkit.org/en/ai/tricks Kit
- *Data Detox Kit: Persuasive Powers: Revealing AI's influence in elections* (by Tactical Tech) https://datadetoxkit.org/en/ai/influence/Kit

Together, the analogue and digital resources support community capability building by demystifying voice AI, encouraging ethical reflection, and equipping individuals with the confidence to navigate emerging risks, particularly in preparedness for voice impersonation and scams. They serve as a foundation for participatory workshops and broader public engagement efforts aimed at fostering the responsible, empowering, and culturally appropriate use of voice technologies.

References

Aguirre, S. (2025). Audio attributed to Margarita González about social programs in Morelos shows signs of digital manipulation. *Animal Politico*. https://www.animalpolitico.com/verificacion-de-hechos/desinformacion/audio-candidata-morena-margarita

Amezaga, N. and Hajek, J. (2022). Availability of voice deepfake technology and its impact for good and evil. P*roceedings of the 23rd Annual Conference on Information Technology Education*, 23-28. https://doi.org/10.1145/3537674.3554742

Anderson, N. (2021). Hatsune Miku, virtual idols, and transforming the popular music experience. *Music.Ology.Eca*, *2,* https://doi.org/10.2218/music.2021.6478

Atal, B. S. (2006). The history of linear prediction. *IEEE Signal Processing Magazine*, *23*(2), 154-161.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Baris, Antonios. (2024). AI covers: legal notes on audio mining and voice cloning. *Journal of Intellectual Property Law & Practice,* 19(7), 571–576, https://doi.org/10.1093/jiplp/jpae029

Barrington, S., Cooper, E. A., & Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, *15*(1), Article 11004.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., ... Zeng, Y. (2025). *International AI safety report*. arXiv. https://doi.org/10.48550/arXiv.2501.17805

Benjamin, W. (2021). Reflections on Radio. In Rosenthal, L. (Ed) & Lutes, J., Schumann, L.H., & Reese, D.K. (Trans). *Radio Benjamin.* (pp. 363-364). Verso. [Originally 1930/1931].

Bertelsen, L., & Murphie, A. (2010). An ethics of everyday infinities and powers: Felix Guattari on affect and the refrain. In M. Gregg & G. J. Seigworth (Eds.), *The affect theory reader* (pp. 138-157). Duke University Press.

Betker, J. (2023). Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Black, A. W., Taylor, P., & Caley, R. (1998). *Festival speech synthesis system* (Technical Report). Centre for Speech Technology Research, University of Edinburgh. https://www.cstr.ed.ac.uk/projects/festival/

Brackhane, F. (2015). Kempelen vs. Kratzenstein: researchers on speech synthesis in times of change. In *HSCR@ INTERSPEECH* (pp. 42-49).

Bryant, M. (2025, 27 June). Denmark to tackle deepfakes by giving people copyright to their own features. *The Guardian*. https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence

Callon, M., Lascoumes, P., & Barthe, Y. (2011). *Acting in an uncertain world: An essay on technical democracy*. MIT Press.

Cambre, J., & Kulkarni, C. (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-19.

Cazzamatta, R., & Sarısakaloğlu, A. (2025). AI-generated misinformation: A case study on emerging trends in fact-checking practices across Brazil, Germany, and the United Kingdom. *Emerging Media*, 27523543251344971. https://doi.org/10.1177/27523543251344971

Chen, H. Z., Chen, R., & Hirschberg, J. (2024). *EmoKnob: Enhance voice cloning with fine-grained emotion control*. arXiv. https://doi.org/10.48550/arXiv.2410.00316

Christopher, N. (2023, July 5). An Indian politician says scandalous audio clips are AI deepfakes. *Rest of World*. https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/

Dennler, N., Kian, M., Nikolaidis, S., & Matarić, M. (2025). Designing robot identity: The role of voice, clothing and task on robot gender perception. *International Journal of Social Robotics*, *17*, 707-728. https://doi.org/10.1007/s12369-025-01209-6

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

Douglas, S. J. (2013). *Listening in: Radio and the American imagination*. University of Minnesota Press.

Dreher, T. (2009). Listening across difference: Media and multiculturalism beyond the politics of voice. *Continuum*, *23*(4), 445-458. https://doi.org/10.1080/10304310903015712

Duinker, B. (2024). Auto-Tune as instrument: trap music's embrace of a repurposed technology. *Popular Music*, *43*(2), 212-236.

Danish Ministry of Culture. (2025). KUU General Annex 232 Bill. https://www.ft.dk/samling/20241/almdel/kuu/bilag/232/3050901.pdf

European Commission. (2025). Code of Practice for General Purpose AI Models, Copyright Chapter. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

Flavián, C., Akdim, K., & Casaló, L. V. (2022). Effects of voice assistant recommendations on consumer behavior. *Psychology & Marketing*, *40*(2), 328-346. https://doi.org/10.1002/mar.21765

Ford, P. J. (2001). A further analysis of the ethics of representation in virtual reality: Multi-user environments. *Ethics and Information Technology*, *3*(2), 113-121. https://doi.org/10.1023/A:1011846009390

Full Fact. (2025a). Audio of Princess Anne talking about Queen Camilla is AI. *Full Fact*. https://fullfact.org/culture-and-society/princess-anne-ai-audio-camilla/

Full Fact. (2025b). Footage showing Burkina Faso leader is likely deepfake. Full Fact. https://fullfact.org/world/ibrahim-traore-burkina-faso-speech-ai/

Galuszka, P. (2024). The influence of generative AI on popular music: Fan productions and the reimagination of iconic voices. *Media, Culture & Society*, *47*(3), 603-612. https://doi.org/10.1177/01634437241282382

Gedye, G. (2025, March 10). AI voice cloning: Do these 6 companies do enough to prevent misuse? *Consumer Reports*. https://innovation.consumerreports.org/new-report-do-these-6-ai-voice-cloning-companies-do-enough-to-prevent-misuse/

Geiger, S., Harrison, D., Kjellberg, H., Mallard, A., & Araujo, L. (2014). Being concerned about markets. In *Concerned markets* (pp. 1-18). Edward Elgar Publishing.

Ghana Fact-Checking Coalition. (2024, December 5). *Viral audio of John Mahama AI generated*. *Dubawa Ghana*. https://ghana.dubawa.org/viral-audio-of-john-mahama-ai-generated/

Goldenfein, J. (2025). Content or Data? The Conceptual Battles Structuring Dataset Markets. *Platforms & Society* (forthcoming)

Goldstein, J. A., & Sastry, G. (2024). The PPOu framework: A structured approach for assessing the likelihood of malicious use of advanced AI systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7, pp. 503-518).

Hallahan, W. I. (1995). DECtalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, *7*(4), 5-19.

Henrickson, L. (2023). Chatting with the dead: The hermeneutics of thanabots. *Media, Culture & Society*, *45*(5), 949-966.

Henrickson, L., & Carlon, D. (2024, June 25). An influencer's AI clone started offering fans 'mind-blowing sexual experiences' without her knowledge. *The Conversation*. https://theconversation.com/an-influencers-ai-clone-started-offering-fans-mind-blowing-sexual-experiences-without-her-knowledge

Hern, A. (2024, March 31). OpenAI deems its voice cloning tool too risky for general release. *The Guardian*. https://www.theguardian.com/technology/2024/mar/31/openai-deems-its-voice-cloning-tool-too-risky-for-general-release

Hutiri, W., Papakyriakopoulos, O., & Xiang, A. (2024, June). Not my voice! A taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 359-376).

Jasinski, Mina. (2025). Does AI voice cloning break Copyright twice? The risk of dual infringement and licensing solutions in music creation. *4iP Council*. https://www.4ipcouncil.com/application/files/9417/4231/2626/Does_AI_voice_cloning_break_Copyright_twice.pdf.

Jucan, I. (2023). *Malicious deceivers thinking machines and performative objects*. Stanford University Press.

Kenmochi, H. (2010). VOCALOID and Hatsune Miku phenomenon in Japan. *Proceedings of InterSinging*, 1-4.

Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Dignum, V., Lupu, E., Tempini, N., & Bengio, Y. (2023). Generative AI models should include detection mechanisms as a condition for public release. *Ethics and Information Technology*, *25*(4), Article 55.

Kolekar, S. S., Richter, D. J., Bappi, M. I., & Kim, K. (2024, February 19-22). Advancing AI voice synthesis: Integrating emotional expression in multi-speaker voice generation. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 193-198). IEEE. https://doi.org/10.1109/ICAIIC60209.2024.10463204

Łabuz, Mateusz (2024). Deep fakes and the Artificial Intelligence Act—An important signal or a missed opportunity? *Policy and Internet, 16(4)July*. https://doi.org/10.1002/poi3.406.

Lam, K. Y. (2016). The Hatsune Miku phenomenon: More than a virtual J-Pop diva. *Journal of Popular Culture*, *49*(5), 1107-1124. https://doi.org/10.1111/jpcu.12455

Le, L. K. (2014). Examining the rise of Hatsune Miku: The first international virtual idol. *The UCI Undergraduate Research Journal*, *13*(1), 1-12.

Lebenhagen, C. (2020). Including speaking and nonspeaking autistic voice in research. *Autism in Adulthood*, *2*(2), 128-131.

Levinas, E. (1991). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.; 4th ed.). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-9342-6

Lindquist, B. (2024). The art of text-to-speech. *Critical Inquiry*, *50*(2), 225-251. https://doi.org/10.1086/727651

Lopezosa, C., Codina, L., Guallar, J., & Pérez-Montoro, M. (2023). Voice search optimization in digital media: Challenges, use and training. *El Profesional de la Información*, https://doi.org/10.3145/epi.2023.may.07

Loviglio, J. (2008). Sound effects: Gender, voice and the cultural work of NPR. *Radio Journal: International studies in broadcast & audio media*, 5(2-3), 67-81.

Ma, F., Li, Y. K., Xie, Y. F., He, Y., Zhang, Y., Ren, H. W., Liu, Z., Yao, W., Ren, F. J., Yu, F. R., & Ni, S. G. (2024). *A review of emotion synthesis based on generative technology*. arXiv. https://doi.org/10.48550/arXiv.2412.07116

Markel, J.D., Gray, A.H. (1976). Formulations. In: *Linear Prediction of Speech. Communication and Cybernetics*. Springer. https://doi.org/10.1007/978-3-642-66286-7_2

Marres, N. (2015). Why map issues? On controversy analysis as a digital method. *Science, Technology, & Human Values*, *40*(5), 655-686.

Marres, N., Castelle, M., Gobbo, B., Poletti, C., & Tripp, J. (2024). AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241255103, article 20539517241255103.

Martin, D. (2020). Speaking machines and ghostly phantoms: The claustrum poetics of voice and dysfluency. *Victorian Review 46*(1), 1-5. https://dx.doi.org/10.1353/vcr.2020.0017.

Martino, M. & Mallal, D. (2025). We cloned senator Jacqui Lambie's voice with AI to show you what a deepfake election could look like. *ABC News Verify*. https://www.abc.net.au/news/2025-02-28/jacqui-lambie-ai-generated-voice-election-and-deepfakes/104986434

Matamoros-Fernández, A. & Bartolo, L. (2024). Mapping and critiquing the new vendor ecosystem in the 'Trust and Safety' Industry. Forthcoming, Selected Papers in Internet Research 2024. Association of Internet Researchers.

McCosker, A. (2015). Social media activism at the margins: Managing visibility, voice and vitality affects. *Social Media + Society*, *1*(2), https://doi.org/10.1177/2056305115605860

McGettigan, C. (2024, November 1). Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities. OSF preprint. https://doi.org/10.31234/osf.io/29jyq

Meaker, M. (2023, October 3). Slovakia's Election Deepfakes Show AI Is a Danger to Democracy. *Wired*. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-adanger-to-democracy/

Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., & Zhang, D. (2023). Biometrics recognition using deep learning: a survey. *Artificial Intelligence Review*, *56*(8), 8647-8695. https://doi.org/10.1007/s10462-022-10237-x

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 453-467.

Napolitano, D. (2022). Reuniting Speech-Impaired People with Their Voices: Sound Technologies for Disability and Why They Matter for Organisation Studies. *PuntOorg International Journal*, *7*(1), 6–21. https://doi.org/10.19245/25.05.pij.7.1.2

Napolitano, D. (2023). The shaping of a standard voice: Sonic and sociotechnical imaginaries in smart speakers. *Im@go: A Journal of the Social Imaginary*, *21*, 177-196.

Nicholls, B. (2019). Postmodernism in the twenty-first century: Jordan Peterson, Jean Baudrillard and the problem of chaos. In *Post-truth and the mediation of reality: New*

conjunctures (pp. 57-77). Springer International Publishing. https://doi.org/10.1007/978-3-030-25670-8_4

Ohala, J. J. (2011, August). Christian Gottlieb Kratzenstein: Pioneer in Speech Synthesis. In *ICPhS* (pp. 156-159).

Ondrejovic, S. (1992). Wolfgang von Kempelen and His Speaking Machine. *Human Affairs*, *2*(2), 161-172.

Pace A., Girardi, C., Campi, C., & Cheng, C. (2025, July 21). Denmark to let you control the use of your voice and likeness. What about Australia. *Gadens Legal Insights.* https://www.gadens.com/legal-insights/denmark-to-let-you-control-the-use-of-your-voice-and-likeness-what-about-australia/ut Australia?

Pace, A., Girardi, C., & Huang, R. (2024, June 3). AI's Ultron moment - Johansson takes on OpenAI over uncanny voice resemblance. *Gadens Legal Insights* https://www.gadens.com/legal-insights/ais-ultron-moment-johansson-takes-on-openai-over-uncanny-voice-resemblance/s

Parker, J. E. K., & Stern, J. (Eds.). (2019). *Eavesdropping: A reader*. City Gallery Wellington; Melbourne Law School; Liquid Architecture.

Peters, J. D. (2004). *The voice and modern media*. Iowa Research Online. https://iro.uiowa.edu/esploro/outputs/conferenceProceeding/The-Voice-and-Modern-Media/9983557325602771

Phan, T. (2017). The materiality of the digital and the gendered voice of Siri. *Transformations*, *29*, 23-33.

Phan, T. (2019). Amazon Echo and the aesthetics of whiteness. *Catalyst: Feminism, Theory, Technoscience*, *5*(1), Article 1. https://doi.org/10.28968/cftt.v5i1.29586

Potter, W. (2025, March 4). AI deepfakes threaten democracy and people's identities. 'Personality rights' could help. *The Conversation*. https://theconversation.com/ai-deepfakes-threaten-democracy-and-peoples-identities-personality-rights-could-help-251267

Predin, J. M. (2025, February 5). The future of AI voice: Trends, challenges, and where investors are betting big. *Forbes*. https://www.forbes.com/sites/josipamajic/2025/02/05/the-future-of-ai-voice-trends-challenges-and-where-investors-are-betting-big/

Provenzano, C. (2019). Making voices: The gendering of pitch correction and the Auto-Tune effect in contemporary pop music. *Journal of Popular Music Studies*, *31*(2), 63–84. https://doi.org/10.1525/jpms.2019.312008

Ragragio, J. L. D. (2021). Strongman, patronage and fake news: Anti-human rights discourses and populism in the Philippines. *Journal of Language and Politics*, *20*(6), 852-872. https://doi.org/10.1075/jlp.20039.rag

Ramati, I. (2024). Algorithmic ventriloquism: The contested state of voice in AI speech generators. *Social Media + Society*, *10*(1), https://doi.org/10.1177/20563051231224401

Ramsay, G. J. (2019). Mechanical speech synthesis in early talking automata. *Acoustics Today*, *15*(2), 11-19.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Reimer, K & Peter, S. (2024). Conceptualizing generative AI as style engines: Application Archetypes and Implications. *International Journal of Information Management* 79, https://doi.org/10.1016/j.ijinfomgt.2024.102824

Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology, March*, 1–30. https://doi.org/10.1080/13600869.2024.2324540.

Roseboro, B. (2019). *The Vocaloid Phenomenon: A Glimpse into the Future of Songwriting, Community-Created Content, Art, and Humanity.* Thesis, DePauw University. https://scholarship.depauw.edu/studentresearch/124

Rosi, V., Payne, B., & McGettigan, C. (2025). Effects of self-similarity and self-generation on the perceptual prioritization of voices. *Journal of Experimental Psychology: Human Perception and Performance* https://doi.org/10.1037/xhp0001325

Rui, Z., & Yan, Z. (2018). A survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE Access*, *7*, 5994-6009.

Sabhanayagam, T., Prasanna Venkatesan, V. & Senthamaraikannan, K. (2018). A Comprehensive Survey on Various Biometric Systems. *International Journal of Applied Engineering Research*, 13(5), 2276-2297. https://www.ripublication.com/ijaer18/ijaerv13n5_28.pdf

Sarwari, A. Q., Javed, M. N., Adnan, H. M., & Wahab, M. N. A. (2024). Assessment of the impacts of artificial intelligence (AI) on intercultural communication among postgraduate students in a multicultural university environment. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-63276-5

Schanke, S., Burtch, G., & Ray, G. (2022, January). Dynamic Voice Clones Elicit Consumer Trust. In Proceedings of the 55th Hawaii International Conference on System Sciences, p.4412-4421.

Schirru, Luca (2025) Danish Bill Proposes Using Copyright Laws to Combat Deepfakes. *Infojustice*. https://infojustice.org/archives/46588.

Sobel, B. (2024). Elements of Style: Copyright, Similarity and GenAI. *Harvard Journal of Law and TechologyTechnology*, 38(1), 49-106.

Sparrow, R., & Zhang, E. Y. (2025). (Re)animating the ancestors: Digital personality emulations, ancestor veneration and ethics. *New Media & Society*, online first https://doi.org/10.1177/14614448251317461

Stern, D.N. (2010). *Forms of vitality: Exploring dynamic experience in psychology, the arts, psychotherapy, and development*. Oxford University Press (UK).

Stern, D.N. (1985). *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology*. Routledge.

Sterne, J. (2003). *The audible past: Cultural origins of sound reproduction*. Duke University Press.

Story, B. H. (2019). History of speech synthesis. In *The Routledge handbook of phonetics* (pp. 9-33). Routledge.

Strengers, Y., & Kennedy, J. (2020). *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot*. MIT Press.

Swartz, L., Marwick, A. E., & Larson, K. (2025). *ScamGPT: GenAI and the automation of fraud – A primer*. Data & Society. https://datasociety.net/library/scam-gpt/

Tariq, S. (2025). AI-generated Joe Rogan voice attacking Foreign Minister Penny Wong shared online. *AAP Factcheck*. https://www.aap.com.au/factcheck/ai-generated-joe-rogan-voice-attacking-foreign-minister-penny-wong-shared-online/

Taussig, M. (1994). *Mimesis and alterity: A particular history of the senses*. Routledge.

Taylor, P., Black, A. W., & Caley, R. (1998). The architecture of the Festival speech synthesis system. *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 147-152. https://www.isca-archive.org/ssw_1998/taylor98_ssw.html

The Associated Press (AAP) (2024, November 1). Researchers say an AI-powered transcription tool used in hospitals invents things no one said. *AP News*. https://www.ap.org/news-highlights/best-of-the-week/honorable-mention/2024/researchers-say-an-ai-powered-transcription-tool-used-in-hospitals-invents-things-no-one-said/

Thomson, T. J., Doyuran, E. B., & Burgess, J. (2025, June 3*)*. Google's SynthID is the latest tool for catching AI-made content. What is AI 'watermarking' and does it work? *The Conversation*. https://theconversation.com/googles-synthid-is-the-latest-tool-for-catching-ai-made-content-what-is-ai-watermarking-and-does-it-work-257637

Uz Zaman, H. (2024). Outgoing US President Biden did not confess to helping orchestrate Pakistan 'regime change'. Soch Fact Check. https://www.sochfactcheck.com/us-president-joe-biden-did-not-confess-to-regime-change-conspiracy-pakistan-army-imran-khan/

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*., *12*.

Vogel, S. (1993). Sensation of tone, perception of sound and empiricism. In Cahan, D. (Ed.) *Hermann von Helmholtz and the foundations of nineteenth-century science*, (pp. 259-287), University of California Press.

Wahl-Jorgensen, K. (2019). *Emotions, media and politics*. Polity.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

We had them tested. *Rest of World*. https://restofworld.org/2023/indian-politician-leakedaudio-ai-deepfake/

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Gabriel, I., Logan, W., Huang, S., Uesato, J., Mellor, J., Kumar, A., Kenton, Z., Isaac, W., Hazell, J., Kreps, S., Paquet, U., Riedel, S., Wiles, J., ... Isaac, W. (2023). *Sociotechnical safety evaluation of generative AI systems*. arXiv. https://doi.org/10.48550/arXiv.2310.11986

Whelan-Shamy, D. & Watt, N. (forthcoming). Generative artificial intelligence and informational malaise: emerging considerations for studies of disinformation and digital media in everyday life. In H. Wasserman (Ed.) *The handbook of disinformation and the media.* Routledge.

Zambrini, B. (2020, February 13). Synthetic voices study: How do you feel about an artificial announcer with an accent? *BBC Research & Development*. https://www.bbc.co.uk/rd/blog/2020-02-synthetic-voices-accent-artificial-interactive

Zhang, L., Jiang, L., Washington, N., Liu, A. A., Shao, J., Fourney, A., Morris, M. R., & Findlater, L. (2021). Social media through voice: Synthesized voice qualities and self-presentation. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21.

Zinsner, S. (2023). Viral post uses altered audio of interview with Greta Thunberg. *Factcheck.org*. https://www.factcheck.org/2023/10/viral-post-uses-altered-audio-of-interview-with-greta-thunberg/

admscentre.org.au

Australian Government

Australian Research Council

This Centre is funded by the Australian Government through the Australian Research Council

ADM+S