frontiers | Frontiers in Computer Science

TYPE Conceptual Analysis
PUBLISHED 17 March 2026
DOI 10.3389/fcomp.2026.1686763

Check for updates

OPEN ACCESS

EDITED BY
Anna Xambó,
Queen Mary University of London,
United Kingdom

REVIEWED BY
Jenn Kirby,
University of Liverpool, United Kingdom
Sahar Sajadieh,
Massachusetts Institute of Technology,
United States

*CORRESPONDENCE
Jonathan Chaim Reus
✉ j.reus@sussex.ac.uk

RECEIVED 15 August 2025
REVISED 23 December 2025
ACCEPTED 19 January 2026
PUBLISHED 17 March 2026

CITATION
Reus JC (2026) The data-driven voice-body in performance: AI voices as materials, mediators, and gifts.
Front. Comput. Sci. 8:1686763.
doi: 10.3389/fcomp.2026.1686763

COPYRIGHT
© 2026 Reus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The data-driven voice-body in performance: AI voices as materials, mediators, and gifts

Jonathan Chaim Reus[1,2]*

[1]Sussex Digital Humanities Lab, University of Sussex, Brighton, United Kingdom, [2]Intelligent Instruments Lab, School of Humanities, University of Iceland, Reykjavík, Iceland

Data-driven, realistic and identity-bearing AI voice technologies have proliferated in recent years. Voice, a multiply embodied phenomenon situated within and across human bodies in space and time, is deeply disrupted by the disembodying tendencies of AI voice technologies and their processes of data collection and data creation, resulting in the need for a re-evaluation of perceptual, cognitive and cultural factors. This article addresses this need by synthesizing ideas from embodied cognition, voice studies, and material anthropology to analyze real-time, AI-mediated voice as a form of embodied cognition that is an intersubjective, extended, materially and socially distributed phenomenon. Through the case study of the live performance iː gəʊ weɪ, this article makes three contributions: (1) it articulates AI-mediated vocal identity as a process of continual reconfiguration across human and machine agencies; (2) it foregrounds audience perception as an active force in stabilising and destabilising emergent voice–body assemblages; and (3) it proposes a speculative ethical framework for vocal data practice grounded in the notion of voice as gift.

KEYWORDS

4E cognition, AI ethics in the arts, AI voice synthesis, embodied music interaction, vocal embodiment, voice perception

1 Introduction

Machine learning algorithms capable of producing realistic human voice conversion in near-realtime have proliferated widely in recent years. These include open source architectures such as RVC and RAVE, as well as commercial products from companies like Yamaha and Dreamtonics (Caillon and Esling, 2021; Masuda, 2023; Dreamtonics Co., Ltd., 2024). Such systems are generally referred to as voice conversion (VC) in the technical literature, or more colloquially as voice transfer. Such technologies can be considered identity-bearing, in that they function upon identity (transforming a source voice into one that bears the markers of another personhood) and literally carry the traces of living human bodies in the recorded voices of their training data. While acknowledging that performing artists have long experimented with live technological voice transformations (Tompkins, 2011), and that artistic practices such as puppetry, voice impersonation, animation voice acting, and ventriloquism offer important precedents for the separation of voice and body in performance (Connor, 2000; Bell, 2016); I will argue that AI voice transfer introduces a distinct condition posing novel design and ethical challenges.

Within performance and media art, the voice has long served as a site for exploring mediation, embodiment, and the destabilisation of identity through technology. As documented in foundational work on vocal aesthetics in digital arts and media (Neumark et al., 2010), artists have repeatedly used amplification, synthesis, fragmentation, and dislocation of voice to probe the boundaries between body, sound, and subjectivity. Notable examples of voice identity

transformation in live vocal performance include the artist Laurie Anderson, who has used formant shifters to perform on stage as her male alter-ego Fenway Bergamot (Novak, 2014). Artists like Imogen Heap and Justin Vernon have defined their styles around vocoders and harmonizers, while a wide range of 21st century music genres from Afrobeat to Hyperpop would be unimaginable without Auto-Tune (Reynolds, 2018). Experimental vocalists such as Franziska Baumann (Figure 1) or Alex Nowitz go further, developing custom gestural sensor controllers to wildly manipulate and layer their voices live on stage (Baumann, 2023). However, unlike mathematically autonomous voice modulation tools such as vocoders and auto-tune, AI voice transfer systems are dependent on recordings of a target voice, and encode the salient features of that person's vocal identity as a statistical pattern that can be re-synthesized (John, 2023; Walczyna and Piotrowski, 2023; Bargum and Erkut, 2024). This technical dependence on recorded traces and identity introduces embodied and ethical considerations for an artist, who becomes a handler and caretaker of bodily traces as well as a puppeteer of them, an embodied anchor point of plural vocal identities. This shift also implicates audiences as perceivers, who witness an absent body's voice animated and anchored by a present body. Such cyborg-like vocal bodies resonate with broader posthuman and queer theoretical accounts of cyborg subjectivity, becoming, and hybrid embodiment (Haraway, 1991; Jarman-Ivens, 2011; Hayles, 2024). Besides producing potentially sublime, grotesque, or unclassifiable experiences of a voice-body, this anchoring can also obscure the plurality of bodies, labors, and histories involved in producing the voice (Bentivegna, 2025; Crowdy and Leach, 2025). Thus the data-driven voice-body, as well as this being a perceptual phenomenon, is also a socially consequential one.

The identity-bearing nature of AI voice transfer systems also marks a necessary shift in design thinking. As technologies which depend on human traces to determine their sonic possibilities, design and engineering processes must become more social and situated. Here, "human traces" refers not only to identifiable vocal attributes associated with individual persons, but also to the embodied labor, stylistic practices, and culturally situated forms of expression through which voices are trained, performed, and recorded. These traces may be biometric, stylistic, or cultural in nature, and are inseparable from the histories of training, authorship, and mediation that produce them. Thus, like the artist using identity-bearing tools, designers and engineers are also handlers, caretakers and anchor points for multiple bodies. They must become archivists as well as engineers, which implies, by ethical necessity, the fostering of relational encounters between designer/engineers and the desires, wishes and lived experiences of those whose traces they handle. For artists, one's voice is often the result of years of training, exploration and practice. In the music industry, an artist's voice is further commodified as part of their brand. The creation of AI voice replicas therefore carry the risk of devaluing and alienating the very source of their training data: vocal artists, cultures and communities, whose creative labor and cultural identity are extracted as a fungible resource, a data-commodity (Almeda et al., 2025). Once extracted, such fungible voices are easily exploited, as many recent cases attest (Vallace, 2022; Coscarelli, 2023; Derico, 2024; Almeda et al., 2025), and risks reducing arts and cultural production to "mere inputs in AI development" (Creamer, 2023). Voices from an indigenous singing tradition might also be scraped into a vocal dataset and used out of context, raising issues of cultural appropriation. Meanwhile, voice recordings are classified in many legal jurisdictions as sensitive biometric data, encoding attributes like gender, age, ethnicity and emotional state (Hutiri et al., 2024). These examples offer a glimpse at how voice recordings are an especially fraught basis upon which to build a technology.

To develop an adequately voice-oriented conceptual analysis, this article weaves together concepts from cultural theory and cognitive science into a syncretic theory of AI-mediated vocal embodiment. Jody



FIGURE 1
Franziska Baumann performing live processed voice using her custom wearable glove interface. Photo credit Francesca Pfeffer. (©) With permission from Francesca Baumann.

Kreiman, a seminal figure in voice studies, has advocated the need for such voice-first conceptual frameworks, arguing that empirical models of voice fail in part because they "treat voices as knowable, fixed objects rather than as dynamic, socially derived thick objects" (Kreiman, 2022, p. 338). In other words, voice is inherently contextually dependent, complex and dynamic - emergent even. Much like Christopher Small's reframing of music as musicking - emphasising music as a situated activity rather than an object - voice is best understood here as a verb rather than a fixed object. This follows Nina Sun Eidsheim's account of voice as a "thick" vibrational event - emerging through embodied production and encultured listening rather than existing as a stable, fully knowable sonic object (Eidsheim, 2015). Rather, voice emerges as a contextually embedded and relational phenomena, what vocal identity researcher Carolyn McGettigan describes as a social signal (McGettigan, 2015). A voice-first conceptual framework, as Kreiman calls for, respects and builds from this ontology.

There are many potential ways to conceptually frame real-time voice transfer. Voice modulation technologies in general could be framed as classic technical extensions of the body, where a voice transformation technique is incorporated as part of the sensori-motor loop of a performer, or instrumentalized to extend the body with new potentials. This view aligns with much of the work on extended and embodied cognition in music technology and beyond (Dourish, 2001; Clark, 2008; Tuuri et al., 2017; Varela et al., 2017; Baumann, 2023). Another useful analytical frame for voice is that of somaesthetics, which focuses on understanding the first-person experience of mediated body schema and felt vibrations - two important aspects of vocal embodiment (Tarvainen, 2019; Cotton et al., 2021; Freire and Reed, 2024). However, returning to the insights of Kreiman, McGettigan and Eidsheim, we must take a moment to acknowledge that these approaches take human coupling with tools, such as hammers and notebooks, and the individual subjective body, producing and feeling vibration, as their starting points, not voice. Our goal is to use *i gəʊ weɪ* as a case study to raise useful questions for thinking through the design of AI mediated vocal identity, bringing together concepts of extended vocal embodiment that do not stop at the boundaries of the singular body, the fixed signal, or tool-like models of human-technology coupling.

These considerations lead to the questions that this paper aims to address: How does real-time AI voice transfer reshape vocal embodiment in live performance, particularly where we must account for a voice that is unstable, distributed, and partially autonomous? How do audiences perceive and make sense of these shift voice-body assemblances? What role does perceptual instability and resolution of the voice-body play in the attribution of agency, presence and identity? Finally, what social and ethical responsibilities arise when voice is mobilised in such performances, and how might practice-led approaches form alternatives to extractive, obscurant or purely instrumental models of vocal technology?

I address these questions through analysis of the live performance *i gəʊ weɪ*. Using this artistic work as a case study that is anchored in methodologies from arts-based research and critical-technical practice in human-media interaction, whereby the creation of a technological artwork, and the first-person embodied experiences of a performer-researcher, leads inquiry and knowledge generation (Agre, 1998; Green, 2014; Candy and Edmonds, 2018).

2 The case study of iː gəʊ weɪ

This section describes ***iː gəʊ weɪ*** (Figure 2)[1], a solo experimental voice performance practice in development since 2022, where I process my own voice through custom AI voice conversion models in real time.[2] To better appreciate the performance, I suggest the reader

1 The title is written in International Phonetic Alphabet notation, and is pronounced "ee-go-way", or "I go away".

2 iːgəʊ weɪ was first performed at Mengi, an experimental music venue in Reykjavik, developed within a residency at the Intelligent Instruments Lab in the summer of 2022.



FIGURE 2
iː gəʊ weɪ performance impression, see VIDEO 1 in the supplemental material. Pictured: Solo performer, Jonathan Reus, with microphone, laptop, audio interface, and physical controller. Photo credit Flora Reznik.

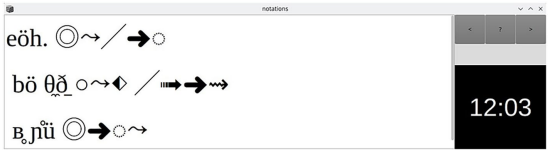


FIGURE 3
Score display interface. Showing three lines from the current score written in a custom vocal notation system called UniVoNo, combining phonetic notations and abstract symbols for expression and articulation. The interface is shown at 12:03 minutes into a performance.

to view the audiovisual documentation (Video 1) provided in the Supplementary materials and extended materials archived on the Catalogue of Artistic Research[3] (Reus, 2025). As stated by the artist, the performance is an ongoing practice of hybrid voice work, of inhabiting and embodying other voices, with the ultimate goal of "unraveling voice as a marker of individuality and identity connected to a specific body … dissolving and coalescing into multi-human, polyphonic, choral and alien forms" (Reus, 2022b, 2023).

2.1 Performance structure and on-stage dynamics

The live performance *i gov weɪ* is a solo vocal work with a typical duration of 25 min, presented as a continuous, uninterrupted performance. It is structured as a score-based improvisation, combining pre-written materials with real-time musical and performative decision-making. Each performance uses a newly prepared score, written in advance, consisting of vocal utterances written in the International Phonetic Alphabet (IPA), coupled with a bespoke Unicode-based notation system I have invented for extended vocal articulations (Figure 3). The score functions as a temporal and expressive scaffold rather than a fixed composition, allowing me to navigate between stability and breakdown while remaining responsive to the system's output and the arc of the performance.

Throughout the performance, the audience encounters three dominant vocal configurations, which shift in varying proportions. At certain moments, my unprocessed acoustic voice is presented directly through the sound system. At others, the voice is fully mediated through the voice conversion system, such that breath, exertion, and articulation are audible primarily through their effects on the generated voice and my visible bodily actions rather than as direct acoustic signals. A third configuration places these two states in antiphonal or delayed relation, with the transformed voice responding to or echoing my acoustic voice and converging into synchronicity. These alternating modes are used to destabilise the attribution of vocal identity to the visible body on stage, while also maintaining a fragile plausibility of voice-body alignment.

The software system processes audio continuously through four voice transfer models running in parallel. At any given time, I may morph

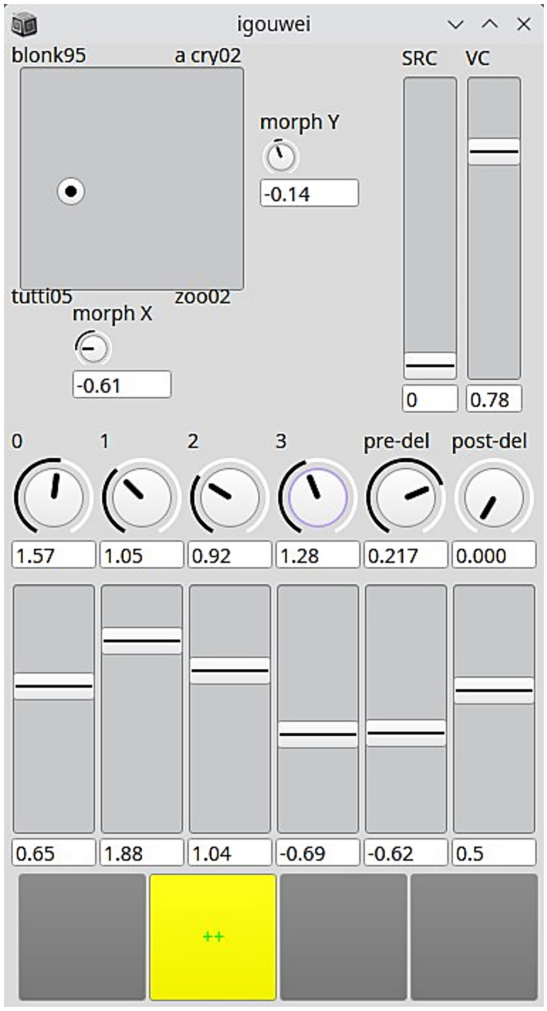


FIGURE 4
Software GUI, providing a visualization of parameters controlled by physical interface. Software state shows morph position roughly halfway between all four models, biased somewhat towards Blonk and Tutti. Source voice signal at zero volume while voice conversion (VC) signal is at 78% of full volume. Model 2 (A Cry) is selected as the current model for latent manipulation.

between these models using a two-dimensional control surface, implemented either via a MIDI controller or, in more recent performances, a wireless inertial motion sensor worn on the arm or hand. Each model occupies a corner of this 2D morphing grid, allowing smooth or abrupt transitions between distinct vocal characters (Figure 4). In addition to inter-model morphing, I may target a single active model for direct manipulation of its latent space. Model selection is typically guided by the current sonic focus of the performance or by preparatory actions when transitioning toward a different region of the morphing space.

Models are exchanged during the performance by hot-swapping one of the four active models for another, usually in response to a desire to explore a different vocal territory. For example, a model trained on my own voice may be replaced with a more unstable or noise-prone model when the focus shifts from self-resemblance to vocal distortion, plurality, or breakdown. These transitions are audibly seamless, with the combination of continuous morphing and model swapping often producing pronounced shifts in perceived vocal age,

3 Supplementary audiovisual documentation is archived at: researchcatalogue.net and https://www.researchcatalogue.net/view/3799783/3799784.

density, intelligibility, register, or plurality, with individual gestures occasionally expanding into choral textures before collapsing back into a single voice.

Visually, the my physical effort and exertion are clearly present throughout the performance, while the vocal outcome remains unstable and partially opaque. At times, sound continues despite minimal visible articulation; at others, articulation and sound appear closely coupled but never fully stabilised. Facial expressions, hand gestures, and bodily posture adapt continuously to the behavior of the system, making visible the labour required to sustain alignment between a present body and the complex, statistically generated voice. Rather than being focused on vocal mastery or transparency, the performance positions the audience as witnesses to an ongoing negotiation between bodily intention, machine response, and the construction of what ventriloquism scholar Steven Connor refers to as a vocalic body, or voice-body - the perceptual fusion of a perceived voice and body combination together with expectations of their plausibility.

The juxtaposition of voice and body is key to the aesthetics and poetics of the performance, as my source voice is computationally modulated into a transformed voice. These transformations range from subtle timbral shifts (e.g., male target voice of similar register) to utterly plural and non-human voices (e.g., animal calls, choral singers). In this active process of shaping the voice-body, I play with the divergence of expected voice and perceived voice, for example, by juxtaposing a differently gendered voice against my body size, gender and ethicity cues. These disjunctures are pushed to further extremes throughout the performance, with the intention of creating voice-body perceptions that defy categorical description. Meanwhile, I play with the visual alignment between bodily actions and transformed voice, creating both plausible illusory alignments and intentional misalignments - primarily through lip movements, facial and bodily gestures.

2.2 The socio-technical system of iː gəʊ weɪ

I will now briefly describe the hardware setup as well as voice transformation software and datasets of our case study. I will use the term *data ecology* to describe all the activities of voice recording, musical relationship building, data curating, collection and editing/processing that are intrinsic to the design of such a system.

Hardware: The hardware consists of a dynamic microphone (Sontronics Solo) with strong off-axis rejection, a low-latency USB audio interface, and a Linux laptop (10th gen Intel i7 CPU) running custom software. Synthesis parameters are manipulated via a Korg Nanokontrol2 MIDI controller, and through a bespoke wearable inertial movement sensor. The transformed voice is heard by an audience through the amplification system of the concert venue.

Software: iː gəʊ weɪ uses a real-time interactive music system capable of morphing between up to four voice transfer models at a given time, built within the open-source audio programming language *SuperCollider* (McCartney, 1996). The signal flow (Figure 5) begins with the microphone signal being processed with fixed compression and EQ which subtly enhance the source voice. This signal is then sent to a voice transfer sub-system with four voice model slots. Each slot includes a model-specific pre-processing block (e.g., some models require noise gating, or subtle noise added to the input signal) and post-processing block (e.g., model-specific reverb or compression). I may select one of the four models to intervene directly into its latent space by adding bias and scaling values. The outputs of the four models are then combined in a morphing block using selectable spectral morphing or cross-fading algorithms, with the blend of the four voice models controllable. The original microphone signal is then mixed with the output of the morphing block and given a final stage of compression, EQ and reverb. The mix of source voice and transformed voice are controllable, and both source and transformed voice signal have an adjustable delay that can be used to create antiphonal effects, or to compensate for model latency.[4] All unfixed parameters may be controlled through the physical control interfaces, and by directly modifying running synthesis routines through live coding (Blackwell et al., 2022).

Choice of model architecture: The author has worked extensively with RAVE since its introduction in 2021.[5] RAVE is a Variational AutoEncoder based architecture (Caillon and Esling, 2021) that is trained in an unsupervised manner on a dataset of voice recordings, learning a statistical mapping from arbitrary (out-of-domain) input voice audio to an intermediate, compressed numerical representation (the "latent space"), and then decoding that intermediate representation into reconstructed audio resembling the training data. While not specifically designed with speech features in mind (Bargum and Erkut, 2024), a trained RAVE model self-learns its own statistically optimal feature representations of the training voice dataset, and is quite capable of performing analysis-resynthesis timbre transfer from an input voice to a target voice. RAVE is especially suitable for live performance due to its integrations with interactive music workflows, real-time performance on a laptop-grade CPU, and open codebase.[6]

2.3 Voice models and their data ecology

The system offers a maximum of four actively generating models at a given time. This limitation is imposed to maintain real-time performance, and to create a sensible inter-model

4 An in-depth discussion of latency and delayed auditory feedback (DAF) are beyond the scope of this paper. It will suffice to say that during the early releases of RAVE (v2), end-to-end latency was close enough to the disruptive threshold for vocal production to motivate the performer to train specifically to adapt to DAF effects. Latency was also used creatively, and became the basis for developing antiphonal techniques later on. Latency issues have become negligible since the release of the RAVE v3 architecture, the option to run models at block-sizes below 2048, and the release of the low-latency architecture BRAVE (Caspe et al., 2025). For an introduction to the effects of DAF in fluent speakers see Yates (1963).

5 A prototype of iː gəʊ weɪ was created in response to the release of the original RAVE at the end of 2021 as part of the Neural Synthesis Hackathon (NASH). Other real-time voice transfer architectures have emerged, but the author has found that the ongoing open-source development, interoperability with interactive music workflows, and rich community around RAVE, as well the author's technical understanding of its idiosyncrasies, have proved worthwhile reasons for committing to a sustained practice with RAVE-like architectures. RAVE was also chosen because it can be trained on what was at the time relatively affordable consumer grade GPU hardware.

6 iː gəʊ weɪ is typically performed on a Linux laptop with an 10th Gen Intel i7-11800H @ 2.30GHz CPU. RAVE models were trained on a Linux desktop with NVIDIA TITAN X GPU with 12GB of VRAM.

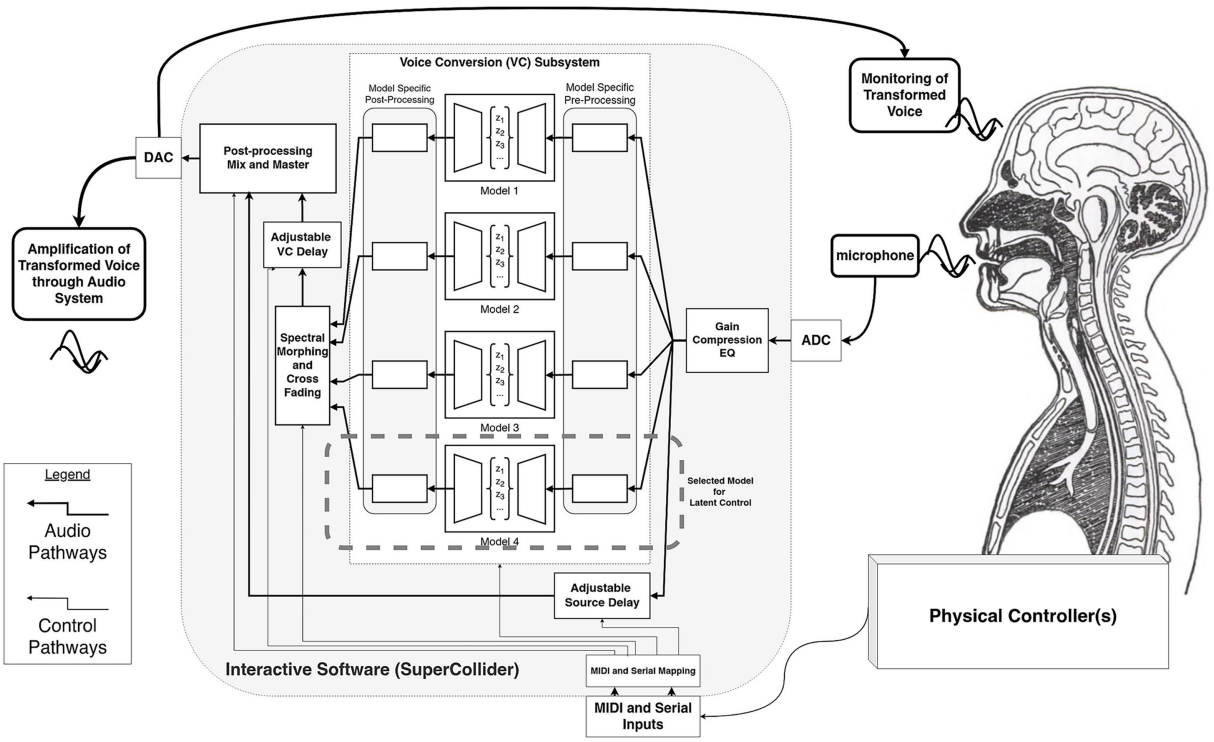


FIGURE 5
Block diagram showing system design and signal flow: performer Voice/Mic and physical controller p(MIDI/Serial) - > SuperCollider - > Parallel voice conversion subsystem p- > Post-processing/morphing - > Speakers/PA and monitoring for performer.

morphing system based on a 2-dimensional control surface (Figure 4). Despite this limitation, it is possible to use an arbitrary number of models during a single performance by hot-swapping new models into the four available slots. As of this writing, performances primarily use models trained on five target voice datasets that collectively scaffold the data ecology of the work. The selection of target voices was chosen in part to create a dynamic vocal playing field, allowing for a wide range and breadth of vocalizations that could be adjusted on-the-fly. Equally importantly, this selection of target voices was chosen to explore varied socio-technical relationships between the artist, training data and those who have contributed their voices with or without explicit knowledge. The following datasets reflect markedly different modes of vocal relation - ranging from collaborative and consent-based exchanges to field recordings and ethically fraught encounters - whose asymmetries are intentionally preserved rather than smoothed over.

This section is not intended as a technical description of model training parameters.[7] Rather, its purpose is to briefly describe a data ecology which will be later referenced to foreground how the data ecology influences my embodied relationships to the voice transfer models. In all cases, a given voice model implicates several iterations of model training, with incremental improvements made in dataset composition, model architecture and training parameters over a period of years.[8]

- **Self Voice:** A voice model trained on a dataset of my own biological voice, intended to cover a wide range of the my own (trained and untrained) vocal capabilities, which I have rehearsed and recorded over the period of a month during a residency at the Intelligent Instruments Lab, Reykjavik. The dataset includes vocal warm-ups and technical exercises in the genres of blues, gospel, death metal growling, and Appalachian field hollaring; paralinguistic expressions such as humming, breath sounds, and percussive lip smacking; speech sounds including a performance of Kurt Schwitter's abstract sound poem Ursonate (Schwitters, 2002, p. 217; Whissell, 2014), slow and fast readings of the International Phonetic Alphabet (IPA) (Ladefoged, 1990), and phonetically balanced English voice research scripts such as the Harvard Sentences (Rothhauser, 1969).

- **Blonk:** A voice model trained on the voice of Jaap Blonk, a prolific experimental vocal improviser and sound poet. Blonk is internationally recognized for his idiosyncratic and rigorous explorations of abstract vocal sounds, rooted in the 1970s live poetry scene and the Dadaist sound poetry tradition. In tandem with contributing his voice to this work, Blonk has collaborated with me as an artistic mentor and vocal coach over the years of its development. This mentorship was coupled with the process of dataset curation and model training for Blonk's

7 For a detailed description of training datasets, model evaluation criteria, and sound examples, see Supplementary materials archived at: https://www.researchcatalogue.net/view/3799783/3799784.

8 The first iteration models used a RAVE v2 architecture, and RAVE v3 in further iterations. The most recent models used in performance are a mixture of RAVE v3 and BRAVE models.

target voice, making the process deeply interpersonal and collaborative (Reus, 2024). The dataset includes approximately 2 h of recorded solo voice performances by Blonk.

• **Tutti:** A voice model trained on choral singing recordings created in collaboration with the MUSILON student choir during an artist residency at the Human-Media Interaction lab at the University of Twente. During the residency the I provided thesis mentorship for one of the lead choristers of the group, a student in Human-Media Interaction who was researching the choral phenomenon of "blend" in human-machine vocal collaboration. Through this mentorship the author was able to develop an intimate relationship with the choir over the course of a year, attending multiple rehearsals and building interpersonal relationships with a number of the choristers through a shared interest in music, voice and computing. The dataset includes approximately 1.5 h of recordings made with the choir using a combined near coincident and spaced microphone placement (Powell, 1989), and is augmented using a mixture of choral ensemble and soloist recordings from open choral singing datasets (Cuesta et al., 2019; Rosenzweig et al., 2020; Cuesta and Gómez, 2022,2022a).

• **Zoo:** A voice model trained on a wide variety of non-human vocal sounds by species that diverge by increasing degrees from homo-sapiens, recorded by the author over the past decade at various locations in the Netherlands and Ghana, including the Apenheul Primate Park in Apeldoorn, the Vondelpark in Amsterdam, and nature areas nearby the city of Kumasi. The dataset includes approximately 1 h of non-human vocalizations including gibbons, howler monkeys, capuchins, dogs, wild parrots, bee hives and cicadas. Unlike the human voice models described above, this dataset involves non-human vocalisations and therefore does not raise questions of consent or authorship in the same way, but instead foregrounds issues of alterity, imitation, and the limits of anthropocentric framings of voice and vocal rights.

• ***A Cry:*** One of the later voice models, created in late October 2023 as a response to the Al-Ahli Arab Hospital explosion in Gaza[9] on October 17th, and the increasingly graphic depictions of human suffering being streamed on social media networks (Abbas et al., 2024; Ghosh, 2024). The dataset includes 45 min of sound material from videos posted by peers and news sites on my personal Instagram feed from October 17–19, representing a collage of my personal media landscape during this period. The audio recordings are chosen for their notable emphasis on human voice, consisting of audible cries, screams, and pleas from Israeli and Palestinian citizens expressing grief, fear, desperation, and rage. Unlike the other voice models, the decision was made to intentionally underfit this model for performance, therefore creating a voice transfer model that is incapable of accurately reproducing audio from the domain of the training data. This prevents the model from being used to convincingly simulate, aestheticise, or repurpose voices associated with real, identifiable moments of suffering, while still allowing for engagement with the statistical residue of the dataset as a resistant and unstable material.

The model "A Cry" occupies a distinct and ethically fraught position within the data ecology of *i goo weɪ*. Unlike the other voice models, its dataset was not assembled through collaboration, mentorship, or shared artistic intent, but emerged from the author's personal exposure to graphic social media footage during an acute geopolitical crisis. Although the audio material was drawn from publicly accessible sources and processed such that individual speakers and social media accounts cannot be identified, the absence of consent or reciprocal relationship remains salient. The voices implicated in "A Cry" can be understood as part of what archival scholars describe as unwitting or unintentional archives, in which individuals become part of a record without anticipating its future circulation, reuse, or transformation (Thomson and Berriman, 2023). Indeed, much of the current social media landscape follows the logics of the unwitting archive, an ethically asymmetrical relationship between capture and later reanimation. This is particularly problematic when voice - unlike many other forms of data - carries embodied traces of distress, vulnerability, and address.

This complex situation led to the choice to intentionally underfit the model, producing a voice transfer system incapable of accurately reproducing the domain of its training data. Underfitting here functions not only as an aesthetic strategy, but as a form of ethical containment: a technological refusal to stabilise, extract, or re-animate voices produced under conditions of extreme vulnerability. In this sense, "A Cry" operates as a limit case within the artwork, exposing the breakdown of relational and gift-like logics that underpin other models in the system. The discomfort it produces - both for the author and potentially for the reader - is not incidental, but productive. It renders explicit the ethical asymmetries that are often obscured at industrial scales of data scraping, where distance, abstraction, and aggregation make such tensions easier to ignore. This contrast is taken up in later sections, particularly section 6 on the concept of the Vocal Gift.

3 Voice, body, and vibration: locating the AI-mediated voice

Now that I have introduced *i: goo weɪ*, its technical design and data ecology, I move on to developing this analysis from a foundational ontological question: *What is "voice"? Where, when, and how does it emerge when data-driven AI intervenes?* A central premise from vocal pedagogy is that voice and body are inextricably connected. Classical voice teachers and voice therapists alike emphasize breath support, resonance in the chest/head, and posture - all indicating that voice is a whole-body activity, not just coming from the mouth or vocal folds (Lehmann, 1924; Stengel and Strauch, 2000; Reed, 2022). This irreducible connection between voice and body is echoed widely by scholarship in interdisciplinary voice studies and music interaction scholarship (Symonds, 2007; Barthes, 2010; Kreiman and Sidtis, 2011; Cavarero and Langione, 2012; Eidsheim, 2015; Thomaidis and Macpherson, 2015; Deighton MacIntyre, 2018; Eidsheim and Meizel, 2019; Reed, 2023; Reed and McPherson, 2023; Cotton et al., 2024).

Adriana Cavarero, a prominent philosopher of voice, takes an even stronger ontological stance on voice, whereby voice signals something "more substantial than its sound" but articulates the embodied manifestation of a unique individual being, a "who." Voice, according to Cavarero, is also a relational phenomenon, and cannot be fully understood by analyzing it in isolation as an acoustic signal or carrier of language, but is understood in the encounter between one who emits and one who listens (Cavarero, 2005; Evans, 2018). Cavarero's conception of voice is both embodied and political, singular and plural, existing within processes of relation and separation between voices, forming from them the multivoiced body of society.

9 https://www.who.int/news/item/17-10-2023-who-statement-on-attack-on-al-ahli-arab-hospital-and-reported-large-scale-casualties

Such a definition of voice implicates an individual body, but also signifies how voice is a socially embedded, interrelational event connected to structures of power whereby some voices may be heard, while others may be silenced or hidden (Evans, 2018).

Complimentarily, the voice scholar Nina Sun Eidsheim frames voice from a deeply somaesthetic, material and situated perspective (Eidsheim, 2015; Tarvainen, 2019). Eidsheim describes voice and listening as an "*intermaterial vibrational practice*," emphasizing the rich "*thick*" phenomenon of embodied physical vibrations, social conditions, and the complexity of phenomena that must come together for us to hear and create the phenomenon of voice (Eidsheim, 2015). *Thickness* names the irreducibility of voice to any single dimension - acoustic signal, physiological mechanism, or symbolic meaning. From this perspective, voice does not pre-exist its perception, but is continuously co-produced by those who sound and those who listen. Eidsheim puts emphasis on acts of listening as the site where the phenomenon of voice emerges (Eidsheim, 2019). This implies that voice must include a perceiver, even if it is only being perceived by the one producing it. Thus voice is created through active interpretation, rather than it being an objective thing out in the world (Eidsheim, 2014, 2019; Camí and Martínez, 2022). When a voice becomes disembodied, such as with a vocal recording or AI voice clone, our socially evolved cognitive systems seem to have an innate desire to resolve what Eidsheim calls the acousmatic question: "*to whom does this voice belong?*" (Eidsheim, 2019; Kreiman, 2022; Sterne and Sawhney, 2022). In the disembodied case, voice emerges as a reflection of the listener, the one who "silently poses the question", rather than a truly "knowable" voice, one with the moral force of evoking a reply. Empirical research in cognitive science further supports Eidsheim's claim, showing that our cognitive systems, often unconsciously, project highly subjective, encultured, and often erroneous assumptions and identity categories like gender, age, ethnicity, race and emotional meaning into the perception of a vocal gestalt, which change based on level of familiarity with a known person (McGettigan, 2015; Scott and McGettigan, 2016; Camí and Martínez, 2022; Lavan and McGettigan, 2023). This projection of personhood happens remarkably fast, as work by Lavan et al. (2024) has shown, early representations emerge within 80ms of a perceived vocal stimulus, and continue to layer in complexity (gender, age, health, attractiveness) up to a half-second after that. Concurrent research has expanded upon existing knowledge of identity perception to demonstrate the existance of a dedicated 'person identity network' that activates during speech perception (Cordero et al., 2025). Recent work shows that similar mechanisms are at play with AI generated voices (Lavan et al., 2024; Rosi et al., 2025). Both Cavarero's voice as "an ontology of uniqueness" and Eidsheim's "voice is preceded by acts of listening" offer a two-sided coin of vocal theory - both relational and phenomenological - that is particularly resonant with frameworks of embodied, extended, embedded and enactive (4E) cognition (De Jaegher and Di Paolo, 2007; Clark, 2008; Varela et al., 2017; Carney, 2020).

This is especially true for the more recent action-perception theory of active inference, which describes the human organism (brain and body included) as a dynamic surprise-minimisation system in constant negotiation with a world in flux. This is done through the theoretical mechanism of hierarchical "generative models". The generative models *project* a world from the inside out that is in negotiation with the flux of sensory signals coming in from the world. They also assign importance to moments where the projections and incoming sensations run afoul of one another, a process of selective attention called *precision weighting*. Highly weighted incongruencies are the point where we are most sensitive to fine-grained cues and nuances, as well as the point where our models will adapt most readily to the unexpected. Where the projection meets the unexpected is, in some respects, what we might call lived experience (Hansen and Pearce, 2014; Clark, 2016; Constant et al., 2019). It is also important to note that generative models are not equivalent to fixed internal representations - they are living and adaptive. Rather think of them as complex dynamic systems coupled with body and world. Incompatibilities between the projected belief and the incoming sensory reality are propagated back up to the generative models, changing them, and therefore changing the nature of experience. Active Inference has proven to be a powerful theory in cognitive science, one that is both biologically plausible (Walsh et al., 2020), and has been implicated in research on speech production alongside other models of action-perception (Bradshaw et al., 2024).

The vocal ontology concept from voice studies offers an answer to "what is voice" that accounts for its embodied, encultured and intersubjective complexities. Meanwhile, Active Inference offers us a descriptive model of the embodied cognition, sensorimotor learning and encultured perception that is necessary for human bodies to take part in ontologically vocal worlds. Importantly, Active Inference is not meant to function here as a literal neurocomputational explanation of voice - Cavarero and Eidsheim already offer us accounts of vocal ontology necessary for this analysis. Rather, it offers an action-perception framework that elegantly conceptualises the questions of expression, relationality, and perception under uncertainty necessary for the voice-body to emerge. Rather than beginning with the body, active inference begins with volition and perception, and inherently resists reductions to fixed mechanistic descriptions or categories. This is especially true when considering intersubjective active inference, which has been studied in the context of dialogic, self-dialogic and joint behavior (Friston and Frith, 2015; Vasil et al., 2020; Bouizegarene et al., 2024; Maisto et al., 2024; Hinrichs et al., 2025). Voice, viewed through this synthesis of theories, is transformed from a channel of information transfer to a primary technology of social alignment and relational dynamics in a thick world.

4 Model of articulation and perception in iː goʊ weɪ

In this section I will describe the interactive vocal paradigm of *iː goʊ weɪ* in light of the theoretical lenses presented above. In *iː goʊ weɪ* I develop a voice in dialogue with a complex technological and cultural material (the AI system and its data ecology), and through engaging with this material over time develop new vocal techniques, tacit knowledge and expectations - in a similar way one might learn how to sing through repetition and training. This process involves multiple layers of practice: (a) *physical training* of the physiological voice - breath, musculature, resonances, etc., (b) *sensory attunement* - learning to feel one's bodily vibrations and *listen* to one's own output and the system's output and respond in a desired way, and (c) attunement to *cultural* priors - pre-existing encultured assumptions about vocality and bodily movement that are worked with and against, such as having an idea of what a "choir" should sound (and act) like, or the impulse to treat the voice of someone who is suffering with care

and empathy. These three elements are activated at points in the development and use of the system, forming dyadic and collective sense making arrangements at multiple scales:

4.1 Voice data creation

Earlier, I described how building the dataset for the *Self* model involved the rehearsal of vocal exercises across genres of voice practice, such as reading the phonetic alphabet and performing vocal warmups for death metal growling. In this case, I have used dataset recording actively as a practice of training and exploration of my own vocal capabilities. This exploration of the self voice is unique in that it is partly guided by a knowledge of what kind of structured data would result in a "good" statistical model. In the case of the *Self* model, the desired outcome was to have a model that would accept a wide range of my vocal utterances with minimal distortion, creating the illusion of there being no mediation at all, unless manipulated to become noticeable. Thus, the training repertoire was chosen so as to cover the widest possible range of phonemes, sung pitches, and paralinguistic vocal sounds like growls. Such repertoires for dataset creation have been common practice in speech research for decades (Fairbanks, 1960; Zhang, 2015), but are only recently becoming common in artistic communities with the development of open source voice synthesis tools like OpenUTAU and DiffSinger (Liu et al., 2022; Liu, 2025; StAkira, 2025). Much like a repertoire would serve to train and define a traditional singer's signature vocal style, the structured dataset becomes a training ground for the body and algorithm, carrying stylistic intentions. The practice of then listening back to these recordings, re-recording and refining them, as well as listening back to outputs from iterative versions of trained models shaped the refinement of technique, while suggesting additions to the training repertoire: such as the addition of phonetic alphabet readings when it became clear that *Ursonate* and *the Harvard Sentences* created a strong bias towards English-Germanic phonemic content.

4.2 Live performance and rehearsal with the system

During performance, I continuously modulate my vocalisations in response to the output of the system. For instance, when using a *Tutti* model, I have discovered that certain kinds of vocalizations excite the model's latent space predictably - such as strongly pitched vocalisations in different registers. Thus, I adjust (perhaps unconsciously) to *sing for the model*. Puppeteering *Tutti* through legato vowels and stable pitches, listening carefully and adapting my voice across its statistical thresholds. This is a classic integrated sensory-motor loop: voice action - > AI response - > new sound perception - > adjustments to action, where embodied skill forms over time. Such systems are the agreed upon cognitive mechanism for voice production and motor control across much of voice perception research. Examples include the DIVA model and others for speech (Tourville and Guenther, 2011; Bradshaw et al., 2024).

Described through the frame of active inference: I have learned somewhat reliable sensorimotor and listening expectations about how vocal gestures are likely to activate the system, but am also experiencing uncertainty in the moment - a heirarchy of surprise from small sonic nuances to the larger-scale domain of timbre and style. I refine my expectations through active listening and active vocal probing, seeking the unexpected at times, and working within the expected at others, achieving a dynamic equilibrium that brings the articulation of a transformed voice into being. This is much like how in singing one adjusts pitch by hearing oneself. In that case, if we notice that we are singing far off key, precision weights shift to make pitch a focal point for attention and adaptation until the deviation signal subsides. The voice conversion model itself adds a unique layer of the unexpected, a non-linear response to activation through the voice. For example, the *Tutti* models show a strong bias towards pitched, smooth and harmonically structured input. They are not universally accepting of all vocalisations: if you provide it with noisy or percussive input it might respond with reconstruction noise, or no sound at all. At the other extreme, *The Zoo* models, having probabilistically inferred the representational auditory features of monkeys shouting and parrots screeching, responds richly to raspy and non-tonal inputs with sharp dynamics, whereas providing it with a clean sustained pitch often produces silence or reconstruction artefacts. The performer learns these representational topologies in a similar way to a woodworker learns the grain of wood by carving, or a electric guitarist learns the feedback zones of an amplifier by playing and moving. This kind of in-the-moment vocal attunement has been shown to happen unconsciously at the very low levels of the heirarchy of surprise, in situations of subtle altered auditory vocal feedback (Behroozmand et al., 2016), but has not been thoroughly explored for less subtle alterations of voice like those of this work. For this, a comparison may be made to the the exploratory babbling and sensori-motor adaptation of infants, which is a basis for the acquisition of speech (Tourville and Guenther, 2011).

Indeed, a kind of *vocal prompting* is necessary to establish a relationship with a model within the *iː ɡoʊ weɪ* system. Even after having practiced for months with a single model, small adjustments to biases in the latent space may create dramatic shifts in the field of response, requiring that I use my voice as a probe to remap my understanding of the current possibility space of the model's responses - what is referred to in the predictive processing literature as *epistemic foraging* (Clark, 2018). Once the grammar of this new vocal space is roughly understood, I may proceed more confidently, with a stronger incorporated relationship to the transformed voice. Each voice transfer model thus has a material, agentive role, in training the performed vocal body - it invites certain techniques and discourages others in an ongoing dynamic process of call and response, alterity and incorporation.

I have discussed how a performer comes into unstable states of incorporation with a transformed voice. The result is a peculiar form of simultaneously dialogic and embodied vocal selfhood, one that is not about perfect prediction, but about working together to do the next best thing without fully knowing what the other will do. As we have stated earlier, committing to a voice-first ontology means there are relational and intersubjective dynamics which are also part of this story. Taking into account Eidsheim's acousmatic question and Cavarero's articulation of a who leads us to two realizations: 1) that the alterity of the AI system, to begin with, carries within it the articulation

of other beings which warrant a dimension of care, and 2) that in encountering these articulations the performer is perpetually navigating his own embodied answer to the acousmatic question. At all points on the spectrum between incorporation and alterity, call and response, are laden with identities that the performer has an imaginary of, and in one way or another colors his voice and body movements towards those identities.

4.3 Voice-body perception on the part of the audience

As Eidsheim argues, the disembodied voice becomes an identity through acts of encultured listening. We may extend this argument to ask how the *voice-body* emerges in partnership with the encultured multi-modal perception on the part of the audience. While the previous sections can be verified by my own first-person account as a performer, empirical evidence on audience experience is at this point limited to informal feedback, and remains a future subject of study. The goal of this section is therefore only to present a reasonable hypothesis for the enactive perception of the audience and their engagement in intersubjective relationships of expectation and prediction.

The perception of voice-body precepts have been studied in cognitive science through illusions and face-voice matching experiments (Bishop and Miller, 2011; Smith et al., 2016; Lavan et al., 2021). For example, in the McGurk effect, mismatched lip movements and speech sounds cause listeners to hear illusory syllables that appear to struggle in reconciling the visual cues of mouth and facial movements with sound (Rosenblum, 2019). In the ventriloquist illusion, a voice's apparent location and body is "captured" by a visible puppet presenting appropriately aligned visual speech cues (Halley, 2018; Bruns, 2019). Although, this capture is not stably associated with purely the visual or the sonic, but varies depending on the perceived reliability of each modality (Alais and Burr, 2004). As mentioned in the introduction, audiences have long encountered audio-visual vocal disjunctions through vocoders, harmonizers, looping, and other live electronic techniques. Such systems already allow a single performer to sound "other" or "more-than-one," and have trained listeners to accept technologically extended vocal bodies. However, real-time AI voice transfer is fundamentally aimed towards the perceptual effect of identity substitution, displacement and transformation, in much the same way an expert lip sync performer grounds a known voice in an unexpected body. The perceptual dynamics of the voice-body gestalt have been shown to vary specifically based on familiarity with known identities (Lavan et al., 2021; Lavan and McGettigan, 2023). We may even say that our cognitive-perceptual systems actively work to create voices and bodies that *stick to one another* - an act of inferring common causality (a shared personhood - see Magnotti and Beauchamp (2017)). Yet it is also true that this sticking together does not completely favor either voice or body, but rather creates a precept that is something different from both (Bonath et al., 2007; Bishop and Miller, 2011; Rosenblum, 2019). The voice-body gestalt is therefore both highly multi-modal and more than the sum of its perceptual parts, a dynamic state of the dyadic relationships between audience and performer rather than a fixed internal representation.

From the audience's perspective, a performer may present one apparent body, while a technologically modulated voice is heard that must be resolved with the apparent body. In the first moments there may be confusion about the conflict of visual and auditory cues. This could be thought of as the onset of something like Eidsheim's *acousmatic question*, whereby an audience member's perceptual-cognitive system is seeking a resolution to the "who" of this voice-body. Predictive processing describes the resolution to this dilemma through the lens of Bayesian causal inference, whereby the perceptual system seeks out the most likely cause from numerous assumptions (called "top-down priors" in the predictive processing literature). This causal inference might be abstracted as something like hypothesis testing, whereby their cognitive system combines prior beliefs (e.g., one human cannot produce the sound of a choir) with sensory evidence (there is only one person on stage, the person's mouth and body movements reflect the articulation of the sound). This dynamic has clear historical precedents in ventriloquism, where what Steven Connor terms the *vocalic body* emerges through the alignment of voice, gesture, and theatrical labor rather than through anatomical origin alone (Connor, 2000). Crucially, believability in such cases depends not on concealment of the body, but on the visible work of breath, timing, and expressive coordination.

Drawing on predictive processing, one can understand audience perception as an active process in which listeners continuously negotiate expectations about the relationship between the visible body on stage and the voice they hear. We may further suppose that, as the piece progresses, the audience dynamically *learns new models* of the voice-body gestalts presented to them, forming new provisional expectations about the range of vocal identities and timbres that can be articulated by the performer. Predictive processing would describe this as an ongoing process of reduction of prediction error over time - the same kind of *learning to listen* we do all the time in our culturally formed vocal worlds.

Rather than fully resolving the voice-body relationship, this process may involve holding multiple, partially incompatible interpretations in tension - for example, between hearing the voice as an extension of the performer and as the articulation of other vocal bodies. Importantly, deeply enculturated priors related to identity categories such as gender or humanness may resist rapid revision, resulting not in stable re-learning but in sustained perceptual ambiguity. While the nature of these dynamics are in part a question of the emotional and perceptual priors of the audience (Hardy III, 2025), the responsibility is also largely in the hands of the performer to keep up a certain level of dynamism and unpredictability such that the audience is not allowed to settle too comfortably into a stable gestalt.

5 The radically extended voice–data–body and the question of "who"

Having analyzed embodiment in *i: gou wei* at the micro-level (perceptual gestalts, action-response loops, and localized prediction error resolution), I will now zoom out to more thoroughly address the question of how to include the wider network of human bodies and voices that are part of the data ecology. Throughout all of the experiences described in the previous section, there exist the implicated bodies of those whose voices are present but bodies are not.

Such absent presences in data-driven AI music lead some scholars to describe AI music as *hauntological*, which describes a "plundering of the recorded past" that flattens out and hides the complexity of its history (Rubinstein, 2020). This perspective has begun to find form in the field of interactive music AI through the method of *hauntographic analysis* developed by Nicola Privato (Privato and Magnusson, 2024) and in investigations by Kelsey Cotton of how music practitioners are "bringing back" the invisibilized bodies of AI voice systems (Cotton et al., 2024).

But is the body, and its presence, truly lost? In her classic text *How We Became Posthuman*, N. Katherine Hayles makes a useful distinction between *pattern* and *presence*. Hayles argues that our current information cultures increasingly privilege pattern - abstract, transferable, and repeatable structure - over presence. Importantly, Hayles emphasizes that embodiment is not eliminated but rendered contingent rather than essential (Hayles, 2024). Traditional voice recordings, as opposed to AI generated voice, in this sense function as a residual trace of a prior presence, whereas a trained AI voice model operates as a patterned potential that can be instantiated repeatedly in new plausible forms. Performance with real-time voice transfer thus involves the performer supplying breath, effort, and expressive intention to animate a vocal pattern, producing a relational voice that emerges only through the coupling of computational structure and embodied presence.

Historically, recorded voice has been understood as an *indexical trace*: a sonic imprint of a past bodily event whose authority derives from having once been produced by a particular body at a particular moment (Sterne, 2003). Media theorists have argued how recordings gain much of their cultural force from this tethering of voice to prior presence, giving rise to experiences of vocal disembodiment and haunting even as the originating body is absent (Connor, 2000). AI voice transfer, by contrast, does not replay a vocal trace but generates new utterances according to learned statistical regularities. In Hayles's terms, the AI voice model operates as an informational *pattern* that can be instantiated across different bodies and moments without being identical to any one of them (Hayles, 2024).

A vocal ontological perspective might reframe the relationship between past recordings and present AI-mediated music making as not one of flattening and hiding of past bodies, but one rich with the potential for the patterns and traces of past bodies to make themselves known through acts of articulation and relationship. From this perspective, I do not experience myself as merely the puppeteer of the flattened abstraction of a ghostly choir, but rather as being in an extended embodied dialogue, reintroducing thickness to the patterns of young vocalists whom I have known and worked intimately with in the past. When a soprano's voice emerges through this dialogue, I know this person, while simultaneously exerting the labor of breath and articulation to keep this person present in some sense, engendering a unique sense of simultaneous singing-with and singing-through. When I vocalise through the model trained on my own voice data (recorded in 2022), I am singing-through and singing-with a snapshot of my past self. With all the models, the vocal assemblage of voice, body, AI system and training data creates a vocal phenomenon that is more-than-one by design. This complicates notions of embodiment, authorship and agency that rely on ideas of a fixed, immutable, individual self. The creative output of this vocal performance, and its embodiment, is the result of a network of interactions over time and between individuals. In cognitive terms, we can think of this as a form of distributed cognition or distributed agency, the idea that cognitive processes can be spread across people and tools in a system (Hollan et al., 2000). Here, the "cognitive system" that produces the singing is not bounded by my body or even the real-time human-machine-audience relationships of the performance, but extends to include the prior contributions of dataset creators, model designers, etc.

I will use the term *voice-data-body* to include the manifestation of voice and body that emerges on stage at the intersection of the performer's body, their voice, and other voices and other bodies who have contributed through dialogs across the medium of recorded sound and machine learning. This term emphasizes that for the voice-body on stage, *data* (recorded voices) acts as a crucial intersubjective mediator between voices, as well as a memory that allows voice to be reanimated and reimagined. This way of thinking about voice data and AI voice aligns strongly with the framing of a data-driven AI system as a kind of archive, or a *memory of voices*, bringing with it all the dimensions of care, transmission and heritage that are so core to the study and use of archives (Thomson and Berriman, 2023). This memory is by nature incomplete and only becomes real through use and context, such as is attested to in the reanimating archives work of sociologist Rachel Thomson (Thomson and McGeeney, 2024), in the work on data-mediated research through design by Elisa Giaccardi (Giaccardi, 2019), and other artistic work by the author such as *In Search of Good Ancestors / Ahnen in Arbeit* (Reus, 2022a) and *DadaSets* (Reus, 2024).

So, who is the "who" behind the voice we witness in *i: gov wei*? The performer on stage is the immediate one (and the one who gets the applause), but when considering an embodied ethics of voice AI, that attribution of recognition becomes unstable. One might argue that acknowledgement also belongs partly to the voices offstage - those who knowingly (or unknowingly, in the case of Zoo and A Cry) contributed vocal traces that shape the data ecology. Here, "who" does not refer to authorship alone, but to the plurality of human traces - of creative labor, vocal practice, cultural history and bodily exertion - that are activated and mediated through and with my own articulation. This *voice-data-body* enacts a vivid form of what Cavarero describes as the *vocal ontology of uniqueness*, in which voice is ontologically both singular and plural at the same time. To design voice technological systems without this awareness would be, according to such an ontology, a fundamental misunderstanding.

From a design perspective, acknowledging the *voice-data-body* means acknowledging that one should treat dataset creation and curation not as a purely technical task, but as a social task and an opportunity to build relationships and forms of dialogue. Engineers and designers of data-driven AI voice technologies may add to their duties the transmission and exegesis of context, such that a vocalic relationship may emerge across time and space. As predictive processing would tell us, such context informs experience at a granular level, priming us to experience sensory experience in a particular way.

In *i: gov wei*, the stories of how the datasets were gathered (and who they involve) are included as part of the performance, and in program notes provided to the audience, in order to provide this

context. Contextualisation operates as an infrastructural commitment of the work, even as its performative articulation remains situational and responsive. The integration of contextual exegesis and priming into the live performance itself remains adaptive and experimental. In some performances, I explicitly address the data sources in the performance itself and in post-performance commentary - acknowledging mentors, collaborators, or the trauma associated with certain vocal materials - while in others this contextualisation remains implicit or deferred to written materials depending on the differing affordances of performance settings. In this sense, the question of "who" is less about assigning exclusive authorship than about recognising and honoring the distributed conditions under which voice, agency, and value come to be articulated.

This finally brings us to the key ethical problem posed by *i: gov wei*: a live performance that works in two directions simultaneously. Inviting listeners to hear a singular voice where there are many, while requiring the performer to inhabit a distributed vocal body that exceeds any single origin. The ethical tension does not lie solely in perceptual illusion, but in how agency, responsibility, and presence are unevenly shared across this relation. There are many thorny problems here beyond the scope of this short paper, such as the politics and technological design of consent, authorship and attribution, of appropriation and data sovereignty. However I that many of these problems arise from a fundamental misunderstanding of voice, and a need to address the acousmatic question through design choices that actively resist human perceptual biases that seek singular identity where there is none.

6 Voice as gift: towards an ethical design paradigm

In this final section, I shift focus from a conceptual and perceptual analysis of *i: gov wei* to a speculative ethics of AI-mediated voice. If AI voice transfer systems are understood as embodied and relational phenomena, then ethical considerations must extend beyond usability and technical performance to include questions of conviviality, respect, and reciprocity (Fistetti, 2016). To articulate such an orientation, this section proposes the metaphor of *voice as gift* as a design orientation for AI voice systems and ecologies of practice, drawing on traditions of gift exchange in material anthropology and cultural theory (Mauss, 2004; Hyde, 2019). The ethical questions developed here emerge directly from the concrete practices of dataset creation, model training, and live performance described in the preceding sections, rather than from an abstract or universal theory of AI ethics.

Research in critical data studies has increasingly emphasized that datasets are socially constructed artefacts, much like archives, that are shaped by power dynamics and practices of selection, abstraction and interpretation (Thompson, 2017; Jo and Gebru, 2020; Orr and Crawford, 2024; Owen, 2024). From this perspective, ethical concerns arise not only from how data are used, but from how they are gathered, contextualised, and stripped of relational meaning during dataset creation. Voice presents a particularly acute case, as vocal recordings simultaneously encode biometric identity, creative labour, and culturally situated forms of expression, while contemporary machine learning pipelines frequently treat voice as a fungible, low-context resource - a pure signal which can be owned, annotated, and recombined independently of the embodied, social, cultural and relational conditions which make it possible (Hutiri et al., 2024; McGettigan, 2024). Recent ethnomusicological work on AI voice likewise cautions against treating voice as a stable object of ownership, instead foregrounding its relational, embodied, and culturally contingent character (Leach and Crowdy, 2025).

This logic is evident in the proliferation of voice cloning, synthesis and conversion services that advertise the ability to reproduce "any voice" from minimal audio input, often framing consent as a technical hurdle rather than a relational obligation (Longpre et al., 2024a; McGettigan, 2024). Documented cases of voice actors and public figures discovering their voices embedded in commercial systems without authorisation further illustrate how voice circulates as an exchangeable commodity rather than as an extension of personhood. By "low-context," I refer not only to missing metadata, but to the erasure of the conditions of voicing itself - who is speaking, to whom, under what circumstances, and with what expectations of listening and response. Such an approach is notably unsustainable for a functioning convivial creative economy (Pyyhtinen, 2016; Fourcade and Kluttz, 2020). What, then, would it mean to treat voice not purely as an extractable resource, but as an enactment of gifts?

Drawing on Mauss' theory of the gift, voice can be understood not as a resource but as an act that establishes conditions. For Mauss, a gift is never free: it binds giver and receiver through ongoing expectations of care, reciprocity, and restraint (Pyyhtinen, 2016). Lewis Hyde extends this logic to artistic practices, arguing that creative practices flourish when acts of creation circulate as gifts, lest their conditions of arisal be driven to exhaustion (Hyde, 2019). Applied to voice, this framing shifts the ethical problem of data-driven AI away from abstract forms of compliance, such as authorship rights and ownership schemes, toward ongoing vocal relational commitments and knowable systems of vocal transmission and exegesis. A gift always carries some of the giver's identity and intentions with it, and an intention of cyclical self-investment, what Mauss poetically describes as "the spirit of the gift". We may understand some of what is at stake in Mauss' warning, that "to retain the thing would be dangerous and mortal", as "the thing given is not inactive. Invested with life, often possessing individuality", it seeks to return to its "place of origin" to produce, on behalf of the community and site from which it sprang an "equivalent to replace it" (Mauss, 2004, p. 16).

Voice as gift does not imply a return to pre-digital moral economies, but it does entail concrete design commitments. In practical terms, voices should be contributed voluntarily and with informed understanding of how they will be used and how they will travel, establishing trust between contributors and system designers. Those who curate and deploy voice models assume the position of gift recipients, where receipt establishes a continuing bond, and incurring obligations to reciprocate within the appropriate context of the gift economy. System design choices can also embed care, restraint, and knowledge of fulfilment of obligations by maintaining traceable relations between voices and their conditions of use, even when outputs are anonymised or the composite of many voices. Reuse should be constrained in ways

that align with contributors' values rather than maximising generalisation or scalability alone. Importantly, this framing does not imply a return to pre-digital moral economies or a rejection of markets. Hyde is explicit that gifts and commodities are not mutually exclusive: the distinction lies not in whether exchange occurs, but in whether circulation preserves or exhausts relational obligation.

The ideal commodity circulates without regard to its source, destination or relationship between the parties involved - whereas a gift establishes a relation that cannot be fully discharged (Hyde, 2019). AI voice system designers risk increasingly approximating commodity logic by valuing vocal recordings precisely for their fungibility and freedom from contextual obligation. By contrast, live vocal performance can be understood as opening a space for gift-like vocality insofar as it functions as an act of high-context address. In such cases, the ethical responsibility of the artist working with the generative voices of others is not a matter of ownership alone, but regards an ongoing obligation to those whose voices animate and sustain the creative tools. Many of the models in *iː ɡoʊ weɪ* point to such a convivial function of data creation, but the "A Cry" model in particular renders visible at the individual level the same breakdown of relational obligation that large-scale data extraction systems suffer from. Despite its grounding in a personal emotional reaction to the training dataset, the ethical tension exposed by this dataset is symptomatic of broader dynamics in which voice circulates in the digital economy, where the conditions of gift relationships are vague at best. Such tensions in the fabric of the data economy points to the social archetype of "the parasite," one who takes but does not reciprocate the gift. Other issues with the online digital platform economy have led to an acceleration of what Pyyhtinen calls the "erosion of the gift," a situation where one gives with no clear idea of the obligations owed in return, or receives gifts without any clear origin or referent, and therefore without obligation to anyone or anything (Pyyhtinen, 2016).

Meanwhile, calls for a paradigm shift in AI ethics increasingly emphasize power, care, and situated accountability over neutrality and scale (D'Ignazio and Klein, 2020; Crawford, 2021; Creamer, 2023; Papakyriakopoulos et al., 2023; Longpre et al., 2024b). Much of this work foregrounds structural asymmetries, extractive infrastructures, and failures of governance at scale, focusing on how harm emerges through aggregation and institutional/interpersonal distance. The present argument complements these perspectives by focusing on voice as a site where ownership often overshadows the issues of relational address that are acutely at stake, and by reframing ethical responsibility not only in terms of rights or compliance, but in terms of obligations that persist beyond a singular act of capture, consent, or anonymisation.

The achievements of the open source and open internet movements are proof that convivial and gift-like technological futures are possible, and elements of gift-oriented approaches can already be observed in community-led voice datasets that foreground voluntary contribution and open circulation (Ardila et al., 2020; Müller and Kreutz, 2024), and in artist-developer collaborations that address attribution, consent and circularity, even as such initiatives remain constrained by prevailing logics of ownership, licensing, and platform governance (Cotton and Tatar, 2023; Ivanova and Ding, 2025). Regulatory frameworks are also beginning to require explicit consent for certain forms of voice synthesis, suggesting future arrangements in which contributing one's voice may involve both contractual agreements and an acknowledgement of ongoing obligations. Recent discourse across research, policy, and industry suggests that voice is increasingly being framed as a form of likeness requiring explicit governance, consent, and compensation within generative AI systems. While such models may address certain concerns around consent and remuneration, they primarily resolve ethical tensions through ownership and enclosure, privileging scalable market solutions over relational or embodied considerations.

Furthermore, ethical approaches to AI voice that rely solely on pre-emptive principles or governance frameworks risk overlooking how ethical tensions arise through situated, embodied practice - particularly in live performance contexts where perception, agency, and responsibility are dynamically negotiated. Recent work in artistic research has similarly argued that ethical questions surrounding AI cannot be fully resolved through such frameworks alone, but often emerge through artistic practice itself (Sunde et al., 2025). From this perspective, ethical responsibility is understood as processual and ongoing, unfolding through artistic decisions, constraints, and encounters rather than preceding them as a fixed evaluative framework. While no single approach resolves the ethical challenges posed by AI voice systems, *voice as gift* offers a way to orient design and practice toward sustaining the necessary thickness, to use Eidsheim's term, of vocal art.

7 Conclusion

In this paper I have explored the live performance *iː ɡoʊ weɪ* as a case study for conceptual analysis, addressing a set of interrelated questions concerning how real-time AI voice transfer systems reshape vocal embodiment, how listeners come to attribute voice and agency in such performances, and what ethical responsibilities emerge when voice is treated as data within live artistic practice. Through the case study of *iː ɡoʊ weɪ*, combining system design, performance analysis, and theoretical reflection, the analysis suggests that vocal embodiment in AI-mediated performance is not displaced or replaced, but redistributed across performer, model, and interface; that audience perception hinges on instability, effort, and misalignment rather than seamless imitation; and that ethical considerations cannot be fully resolved through consent, ownership schemes or anonymisation alone, but require practice-led approaches attentive to sustaining relational context. While these findings do not close the questions they raise, they clarify the conditions under which AI voice can be understood as an embodied, relational, and ethically charged phenomenon in performance.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

JR: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research has been made possible as part of The Leverhulme Trust's Doctoral Scholarships scheme, as part of the University of Sussex's "From Sensation and Perception to Awareness" interdisciplinary doctoral training programme. The development of the performance iː goʊ weɪ has received additional financial and in-kind support from the Intelligent Instruments Lab, University of Iceland, which is supported by the European Research Council (ERC) as part of the Intelligent Instruments project (INTENT), under the European Union's Horizon 2020 research and innovation programme (Grant agreement no. 101001848).

Acknowledgments

The author would like to acknowledge his musical collaborators, Jaap Blonk and the Musilon Student Choir, for their excitement and willingness to work together on the artwork described in this article. I would also like to thank the Intelligent Instruments Lab, University of Iceland, and the Sussex Digital Humanities Lab, University of Sussex, for their support in providing studio space and performance opportunities for developing this performance practice since 2022.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2026.1686763/full#supplementary-material

References

Abbas, S. O., Soharwardi, S. M. N., and Ameer, A. (2024). The influence of social media platforms on public opinion: a case study of the Gaza-Israel conflict. *Insights: Journal of Humanities and Media Studies Review* 1, 27–41. doi: 10.63290/a4hrn338

Agre, P. (1998) "Toward a critical technical practice lessons learned in trying to reform AI," in Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work. New York: Psychology Press. doi: 10.4324/9781315805849-8

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Almeda, S., et al. (2025) "Labor, Power, and Belonging: The Work of Voice in the Age of AI Reproduction," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25: The 2025 ACM Conference on Fairness, Accountability, and Transparency*, Athens Greece: ACM, 1238–1249.

Ardila, R., et al. (2020) "Common Voice: A Massively-Multilingual Speech Corpus." arXiv. Available online at: http://arxiv.org/abs/1912.06670 (accessed October 21, 2022).

Bargum, A. R., and Erkut, C. (2024). "RAVE for speech: efficient voice conversion at high sampling rates" in 27th international conference on digital audio effects (DAFx24).

Barthes, R. (2010). The grain of the voice. New York, NY: Penguin Random House.

Baumann, F. (2023). Embodied Human–Computer Interaction in Vocal Music Performance. Cham: Springer International Publishing.

Behroozmand, R., Sangtian, S., Korzyukov, O., and Larson, C. R. (2016). A temporal predictive code for voice motor control: Evidence from ERP and behavioral responses to pitch-shifted auditory feedback. *Brain Res.* 1636, 1–12. doi: 10.1016/j.brainres.2016.01.040

Bell, J. (2016). American puppet modernism: Essays on the material world in performance. Cham: Springer.

Bentivegna, F. (2025). Artificially voiced intelligences: Voice and the myth of AI. London: AI & SOCIETY, 1–12.

Bishop, C. W., and Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PLoS One* 6:e24016. doi: 10.1371/journal.pone.0024016

Blackwell, A. F., et al. (2022). Live Coding: A User's Manual. London: MIT Press.

Bonath, B., Noesselt, T., Martinez, A., Mishra, J., Schwiecker, K., Heinze, H. J., et al. (2007). Neural basis of the ventriloquist illusion. *Curr. Biol.* 17, 1697–1703. doi: 10.1016/j.cub.2007.08.050

Bouizegarene, N., et al. (2024). Narrative as active inference: an integrative account of cognitive and social functions in adaptation. *Front. Psychol.* 15. doi: 10.3389/fpsyg.2024.1345480

Bradshaw, A., Press, C., and Davis, M. H. (2024). "Active inference and speech motor control: a review and theory" in Centre for Open Science. doi: 10.31234/osf.io/eq4kh_v3

Bruns, P. (2019). The ventriloquist illusion as a tool to study multisensory processing: an update. *Front. Integr. Neurosci.* 13. doi: 10.3389/fnint.2019.00051

Caillon, A., and Esling, P. (2021). RAVE: a variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*.

Camí, J., and Martínez, L. M. (2022). "3 the conception of reality: we are our memories" in The Illusionist Brain: The Neuroscience of Magic. eds. J. Camí and L. M. Martínez (Princeton NJ: Princeton University Press), 32–42.

Candy, L., and Edmonds, E. (2018). Practice-based research in the creative arts: foundations and futures from the front line. *Leonardo* 51, 63–69. doi: 10.1162/leon_a_01471

Carney, J. (2020). Thinking avant la lettre: a review of 4E cognition. *Evol. Stu. Imag. Cult.* 4, 77–90. doi: 10.26613/esic/4.1.172

Caspe, F., et al. (2025). Designing neural synthesizers for low-latency interaction. *arXiv.* doi: 10.48550/arXiv.2503.11562

Cavarero, A. (2005). For more than one voice: Toward a philosophy of vocal expression. Redwood City, CA: Stanford University Press.

Cavarero, A., and Langione, M. (2012). The vocal body: extract from a philosophical encyclopedia of the body. *Qui Parle* 21, 71–83. doi: 10.5250/quiparle.21.1.0071

Clark, A. (2008). Supersizing the mind: Embodiment, action, and cognitive extension. Oxford: Oxford University Press (Philosophy of mind).

Clark, A. (2016). Surfing uncertainty: Prediction, action, and the embodied mind. Oxford: Oxford University Press.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenol. Cogn. Sci.* 17, 521–534. doi: 10.1007/s11097-017-9525-z

Connor, S. (2000). Dumbstruck: A cultural history of ventriloquism. Oxford: Oxford University Press.

Constant, A., et al. (2019). Regimes of Expectations: An Active Inference Model of Social Conformity and Human Decision Making. *Front. Psychol.* 10:679. doi: 10.3389/fpsyg.2019.00679

Cordero, G., et al. (2025). Perceiving speech from a familiar speaker engages the person identity network. *PLOS ONE_* 20:e0322927. doi: 10.1371/journal.pone.0322927

Coscarelli, J. (2023) "An a.I. Hit of fake 'drake' and 'the Weeknd' rattles the music world," The New York Times, 19 April. Available online at: https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html (accessed June 29, 2023).

Cotton, K., De Vries, K., and Tatar, K. (2024) "Singing for the missing: bringing the body Back to AI voice and speech technologies," in *Proceedings of the 9th International Conference on Movement and Computing. MOCO '24: 9th International Conference on Movement and Computing*, Utrecht Netherlands: ACM, pp. 1–12

Cotton, K., and Tatar, K. (2023) "Caring Trouble and Musical AI: Considerations towards a Feminist Musical AI," AIMC 2023. Available online at: https://aimc2023.pubpub.org/pub/zwjy371l/release/1 (accessed September 7, 2023).

Cotton, K., Sanches, P., Tsaknaki, V., and Karpashevich, P. (2021). "The body electric: a NIME designed through and with the somatic experience of singing" in International conference on new interfaces for musical expression. NIME 2021 (PubPub).

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. London: Yale University Press.

Creamer, E. (2023). MPs criticise UK government's handling of copyright policy related to AI. *The Guardian, 30 August*. Available at: https://www.theguardian.com/books/2023/aug/30/mps-criticise-uk-governments-handling-of-copyright-policy-related-to-ai (Accessed March 18, 2025).

Crowdy, D., and Leach, J. (2025). "CH 10 questions of voice in AI music" in Cultural technologies: Robots and artificial intelligence in the performing arts. 1st ed (New York, NY: Routledge).

Cuesta, H., and Gómez, E. (2022) "ESMUC choir dataset." Zenodo. Available online at: https://zenodo.org/records/5848990 (accessed July 20, 2025).

Cuesta, H., and Gómez, E. (2022a). Cantoría Dataset. *Zenodo.* Available online at: https://zenodo.org/records/5878677 (Accessed July 20, 2025).

Cuesta, H., et al. (2019). Choral Singing Dataset. *Zenodo. Available at*. doi: 10.5281/zenodo.2649950

D'Ignazio, C., and Klein, L. F. (2020). Data Feminism. New York, NY: MIT Press.

Deighton MacIntyre, A. (2018). The signification of the signed voice. *J. Interdiscip. Voice Stud.* 3, 167–183. doi: 10.1386/jivs.3.2.167_1

De Jaegher, H., and Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenol Cognit. Sci.* 6, 485–507. doi: 10.1007/s11097-007-9076-9

Derico, B. (2024). "A tech firm stole our voices - then cloned and sold them," BBC News, 1 September. Available online at: https://www.bbc.com/news/articles/c3d9zv50955o (accessed September 2, 2024).

Dourish, P. (2001). Where the action is: The foundations of embodied interaction. New York, NY: MIT Press.

Dreamtonics Co., Ltd. (2024) "Vocoflex | Dreamtonics." Available online at: https://dreamtonics.com/vocoflex/ (accessed July 20, 2025).

Eidsheim, N. S. (2014). The micropolitics of listening to vocal timbre. *Postmod. Cult.* 24:256.

Eidsheim, N. S. (2015). Sensing sound: Singing & listening as vibrational practice. Durham, N.C: Duke University Press.

Eidsheim, N. S. (2019). The race of sound: Listening, timbre, and vocality in African American music. Durham: Duke University Press.

Eidsheim, N. S., and Meizel, K. (2019). The Oxford handbook of voice studies. New York, NY: Oxford University Press.

Evans, F. (2018). Adriana Cavarero and the primacy of voice. *J. Speculative Philos.* 32, 475–487. doi: 10.5325/jspecphil.32.3.0475

Fairbanks, G. (1960). Voice and articulation drillbook. 2nd Edn. New York, NY: Harper & Row.

Fistetti, F. (2016). Convivialism, the 'counter-movement' of the 21st century. *Rev. MAUSS* 48, 247–258.

Fourcade, M., and Kluttz, D. N. (2020). A maussian bargain: accumulation by gift in the digital economy. *Big Data Soc.* 7:2053951719897092. doi: 10.1177/2053951719897092

Freire, R., and Reed, C. N. (2024). "Body Lutherie: co-designing a wearable for vocal performance with a changing body," in *Proceedings of the International Conference on New Interfaces for Musical Expression 2024. New Interfaces for Musical Expression (NIME)*, Utrecht, The Netherlands.

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405. doi: 10.1016/j.concog.2014.12.003

Ghosh, C. (2024). "The impact of social media on conflict perception: case studies of Russia- Ukraine and Gaza conflicts." doi: 10.13140/RG.2.2.17178.25285

Giaccardi, E. (2019). Histories and futures of research through design: from prototypes to connected things. *Int. J. Dsign* 13, 1–12.

Green, O. (2014). "Nime, musicality and practice-led methods" in Proceedings of the international conference on new interfaces for musical expression, 1–6.

Halley, C. (2018). "How Ventriloquism Tricks the Brain," JSTOR Daily, 11 July. Available online at: https://daily.jstor.org/how-ventriloquism-tricks-the-brain/ (Accessed July 21, 2025).

Hansen, N. C., and Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Front. Psychol.* 5:1052. doi: 10.3389/fpsyg.2014.01052

Haraway, D. J. (1991). Simians, cyborgs, and women: The reinvention of nature. London: Free Association Books.

Hardy III, J. H. (2025). Curiosity is the key to the future of learning and development. *Ind. Organ. Psychol.* 18, 134–138. doi: 10.1017/iop.2024.64

Hayles, N. K. (2024). How we became Posthuman: Virtual bodies in cybernetics, literature, and informatics. Chicago, IL: University of Chicago Press.

Hinrichs, N., et al. (2025). Geometric Hyperscanning of affect under active inference. *arXiv.* doi: 10.48550/arXiv.2506.08599

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.* 7, 174–196. doi: 10.1145/353485.353487

Hutiri, W., Papakyriakopoulos, O., and Xiang, A. (2024). "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, 359–376.

Hyde, L. (2019). The gift how the creative Spirit transforms the world. 3rd Edn. London: Vintage Books.

Ivanova, V., and Ding, J. (2025). Choral data "trust" experiment white paper. London: Serpentine Arts Technologies.

Jarman-Ivens, F. (2011). Queer voices: Technologies, Vocalities, and the musical flaw. Cham: Springer.

Jo, E.S., and Gebru, T. (2020). "Lessons from archives: strategies for collecting sociocultural data in machine learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery 306–316.

John, J. (2023). "SoftVC VITS Singing Voice Conversion." Available online at: https://github.com/justinjohn0306/so-vits-svc-4.0-v2 (Accessed September 6, 2023).

Kreiman, J. (2022). Who is speaking? An empirical perspective on the acousmatic question. *Kalfou* 9:256.

Kreiman, J., and Sidtis, D. (2011). Foundations of voice studies: An interdisciplinary approach to voice production and perception. 1st Edn. New York, NY: Wiley.

Ladefoged, P. (1990). The revised international phonetic alphabet. *Language* 66, 550–552. doi: 10.2307/414611

Lavan, N., and McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Commun. Psychol.* 1:152. doi: 10.1038/s44271-023-00001-4

Lavan, N., Rinke, P., and Scharinger, M. (2024). The time course of person perception from voices in the brain. *Proc. Natl. Acad. Sci.* 121:e2318361121. doi: 10.1073/pnas.2318361121

Lavan, N., Smith, H., Jiang, L., and McGettigan, C. (2021). Explaining face-voice matching decisions: The contribution of mouth movements, stimulus effects and response biases. *Atten. Percept. Psychophys.* 83, 2205–2216. doi: 10.3758/s13414-021-02290-5

Lavan, N., et al. (2024). Voice deep fakes sound realistic but not (yet) hyperrealistic. New York, NY: OSF.

Leach, J., and Crowdy, D. (2025). "Ch 10 questions of voice in AI music" in Cultural technologies: Robots and artificial intelligence in the performing arts. 1st ed (New York: Routledge), 146–159.

Lehmann, L. (1924). How to sing. London: Courier Corporation.

Liu, J. (2025). "MoonInTheRiver/DiffSinger." Available online at: https://github.com/MoonInTheRiver/DiffSinger (accessed August 15, 2025).

Liu, J., et al. (2022). DiffSinger: singing voice synthesis via shallow diffusion mechanism. *arXiv.* doi: 10.48550/arXiv.2105.02446

Longpre, S., Mahari, R., Lee, A., et al. (2024a). Consent in crisis: the rapid decline of the AI data commons. *Adv. Neural Inf. Process. Syst.* 37, 108042–108087.

Longpre, S., Mahari, R., Obeng-Marnu, N., et al. (2024b). "Position: data authenticity, consent, & provenance for AI are all broken: what will it take to fix them?" in Proceedings of the 41st international conference on machine learning (PMLR: International Conference on Machine Learning), 32711–32725. Available online at: https://proceedings.mlr.press/v235/longpre24b.html (Accessed January 28, 2026).

Magnotti, J. F., and Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent Audiovisual speech. *PLoS Comput. Biol.* 13:e1005229. doi: 10.1371/journal.pcbi.1005229

Maisto, D., Donnarumma, F., and Pezzulo, G. (2024). Interactive inference: a multi-agent model of cooperative joint actions. *IEEE Trans Syst Man Cybern Syst* 54, 704–715. doi: 10.1109/TSMC.2023.3312585

Masuda, N. (2023). "State-of-the-art Singing Voice Conversion methods," 16 October. Available online at: https://medium.com/qosmo-lab/state-of-the-art-singing-voice-conversion-methods-12f01b35405b (Accessed August 6, 2025).

Mauss, M. (2004). The gift: The form and reason for exchange in archaic societies. London: Routledge.

McCartney, J. (1996). "SuperCollider, a new real time synthesis language" in Proceedings of the 1996 international computer music conference (The International Computer Music Association), 257–258.

McGettigan, C. (2015). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Front. Hum. Neurosci.* 9:129. doi: 10.3389/fnhum.2015.00129

McGettigan, C. (2024). Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities. London: OSF.

Müller, T., and Kreutz, D. (2024). Thorsten-Voice Available online at: https://github.com/thorstenMueller/Thorsten-Voice (accessed August 2, 2024).

Neumark, N., Gibson, R., and Van Leeuwen, T. (eds.) (2010). VOICE: vocal aesthetics in digital arts and media. *The MIT Press.* doi: 10.7551/mitpress/9780262013901.001.0001

Novak, J. (2014). Singing corporeality: reinventing the vocalic body in postopera Available online at: https://www.academia.edu/6960918/Singing_corporeality_reinventing_the_vocalic_body_in_postopera (accessed July 21, 2025).

Orr, W., and Crawford, K. (2024). The social construction of datasets: on the practices, processes, and challenges of dataset creation for machine learning. *New Media Soc.* 26, 4955–4972. doi: 10.1177/14614448241251797

Owen, W. (2024). "What makes a good archive? A researcher's point of view on service and conservation," museum without walls, 6 November. Available online at: https://wdowen.substack.com/p/what-makes-a-good-archive-a-researchers (Accessed December 20, 2025).

Papakyriakopoulos, O., et al. (2023) "Augmented Datasheets for Speech Datasets and Ethical Decision-Making," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '23), pp. 881–904.

Powell, S. (1989). Recording your choir. *Choral J.* 30, 15–19.

Privato, N., and Magnusson, T. (2024). "Querying the ghost: AI Hauntography in NIME" in Proceedings of the International Conference on New Interfaces for Musical Expression (Zenodo), 432–438.

Pyyhtinen, O. (2016). The gift and its paradoxes: Beyond Mauss. London: Routledge.

Reed, C. N. (2022). "Communicating across bodies in the voice lesson" in 2022 CHI workshop on tangible interaction for well-being (New York: ACM).

Reed, C. N. (2023). "As the luthiers do: designing with a living, growing, changing body-material" in ACM CHI workshop on body X materials (New York, NY: ACM CHI Workshop on Body X Materials).

Reed, C.N., and McPherson, A.P. (2023). "The body as sound: unpacking vocal embodiment through auditory biofeedback," in *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction. TEI '23: Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*, Warsaw Poland: ACM, 1–15.

Reus, J. C. (2022a) "In search of good ancestors / Ahnen in Arbeit," in *Nordic Human-Computer Interaction Conference*. New York, NY: Association for Computing Machinery

Reus, J. C. (2022b). "Performing Voices Without Bodies." RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, 21 November. Available online at: https://www.uio.no/ritmo/english/news-and-events/events/workshops/2022/embodied-ai/p/jonathan-chaim-reus.html (accessed July 29, 2024).

Reus, J. C. (2023). "iː goʊ weɪ," in Proceedings of AI Music Creativity Conference 2023. University of Sussex. Available online at: https://aimc2023.pubpub.org/pub/hpy32yre/release/5 (accessed December 8, 2024).

Reus, J. (2024). Datasets. Available online at: https://jonathanreus.com/portfolio/dadasets/ (Accessed August 16, 2025).

Reus, J. (2025). iː goʊ weɪ by JC Reus [Musical Performance]. Available online at: https://www.researchcatalogue.net/view/3799783/3799784 (Accessed August 12, 2025).

Reynolds, S. (2018). How Auto-Tune Revolutionized the Sound of Popular Music, Pitchfork. Available online at: https://pitchfork.com/features/article/how-auto-tune-revolutionized-the-sound-of-popular-music/ (accessed August 6, 2025).

Rosenblum, L. D. (2019). "Audiovisual speech perception and the McGurk effect" in Oxford research encyclopedia of linguistics. ed. L. D. Rosenblum (Oxford University Press).

Rosenzweig, S., et al. (2020). Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing. *Trans. Int. Soc. Music Inf. Retrieval* 3, 98–110. doi: 10.5334/tismir.48

Rosi, V., Soopramanien, E., and McGettigan, C. (2025). Perception and social evaluation of cloned and recorded voices: effects of familiarity and self-relevance. *Comput. Hum. Behav. Artif. Hum.* 4:100143. doi: 10.1016/j.chbah.2025.100143

Rothhauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* 17, 225–246. doi: 10.1109/tau.1969.1162058

Rubinstein, Y. (2020). Uneasy listening: towards a Hauntology of AI-generated music. *Resonance* 1, 77–93. doi: 10.1525/res.2020.1.1.77

Schwitters, K. (2002) PPPPPP: Poems performances pieces proses plays poetics. Translated by J. Rothenberg and P. Joris. Cambridge, Mass: Exact Change.

Scott, S., and McGettigan, C. (2016). "The voice: from identity to interactions" in APA handbook of nonverbal communication. eds. D. E. Matsumoto, H. C. Hwang and M. G. Frank (Washington, DC: American Psychological Association), 289–305.

Smith, H. M. J., et al. (2016). Concordant Cues in Faces and Voices: Testing the Backup Signal Hypothesis. *Evol. Psychol.* 14:1474704916630317. doi: 10.1177/1474704916630317

StAkira (2025). stakira/OpenUtau Available online at: https://github.com/stakira/OpenUtau (Accessed August 15, 2025).

Stengel, I., and Strauch, T. (2000). Voice and self: A handbook of personal voice development therapy. London: Free Association Books.

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

Sterne, J., and Sawhney, M. (2022). The acousmatic question and the will to Datafy: otter.Ai, low-resource languages, and the politics of machine listening. *Kalfou* 9:617.

Sunde, E. K., Bartlett, V., and Pfefferkorn, J. (2025). Decentring ethics: AI art as method. New York, NY: Open Humanites Press.

Symonds, D. (2007). The corporeality of musical expression: the grain of the voice and the actor-musician. *Studies in Musical Theatre* 1, 167–181. doi: 10.1386/smt.1.2.167_1

Tarvainen, A. (2019). Music, sound, and voice in Somaesthetics: overview of the literature. *The Journal of Somaesthetics* 5.

Thomaidis, K., and Macpherson, B. (2015). Voice studies: Critical approaches to process, performance and experience. London: Routledge.

Thompson, P. (2017). The voice of the past: Oral history. London: Oxford University Press.

Thomson, R., and Berriman, L. (2023). Starting with the archive: principles for prospective collaborative research. *Qual. Res.* 23, 234–251. doi: 10.1177/14687941211023037

Thomson, R., and McGeeney, E. (2024). Reanimating data: working with archives to revitalise young sexualities, past and present. *Health Educ. J.* 22:00178969241304725. doi: 10.1177/00178969241304725

Tompkins, D. (2011). How to wreck a Nice Beach: The vocoder from world war II to hip-hop, the machine speaks. London: Melville House.

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cognit. Proc.* 26, 952–981. doi: 10.1080/01690960903498424

Tuuri, K., Parviainen, J., and Pirhonen, A. (2017). Who controls who? Embodied control within human–technology choreographies. *Interact. Comput.* 29, 494–511. doi: 10.1093/iwc/iww040

Vallace, C. (2022). "Actors launch campaign against AI 'show stealers,'" BBC News, 21 April. Available online at: https://www.bbc.com/news/technology-61166272 (accessed February 25, 2023).

Varela, F. J., Thompson, E., and Rosch, E. (2017). The embodied mind, revised edition: Cognitive science and human experience. London: MIT Press.

Vasil, J., et al. (2020). A world unto itself: human communication as active inference. *Front. Psychol.* 11. doi: 10.3389/fpsyg.2020.00417

Walczyna, T., and Piotrowski, Z. (2023). Overview of voice conversion methods based on deep learning. *Appl. Sci.* 13:3100. doi: 10.3390/app13053100

Walsh, K. S., et al. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Ann. N. Y. Acad. Sci.* 1464, 242–268. doi: 10.1111/nyas.14321

Whissell, C. (2014). The expressive force of primitive masculine sounds in Schwitters' sonata/poem Ursonate. *International Journal of Language and Linguistics* 2:343. doi: 10.11648/j.ijll.20140206.11

Yates, A.J. (1963). "Delayed auditory feedback," *Psychol. Bull.* 60, 213–232. doi: 10.1037/h0044155

Zhang, S. (2015). "The 'Harvard Sentences' Secretly Shaped the Development of Audio Tech," Gizmodo, 9 March. Available online at: https://gizmodo.com/the-harvard-sentences-secretly-shaped-the-development-1689793568 (Accessed August 15, 2025).