
# Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies

# Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies

Kelsey Cotton
Chalmers University of Technology
Göteborg, Sweden
kelsey@chalmers.se

Katja de Vries
Uppsala University
Uppsala, Sweden
katja.devries@jur.uu.se

Kıvanç Tatar
Chalmers University of Technology
Göteborg, Sweden
tatar@chalmers.se

## ABSTRACT

Technological advancements in deep learning for speech and voice have contributed to a recent expansion in applications for voice cloning, synthesis and generation. Invisibilised stakeholders in this expansion are numerous absent bodies, whose voices and voice data have been integral to the development and refinement of these speech technologies. This position paper probes current working practices for voice and speech in machine learning and AI, in which the bodies of voices are "invisibilised". We examine the *facts* and *concerns* about the voice-Body in applications of AI-voice technology. We do this through probing the wider connections between voice data and Schaefferian listening; speculating on the consequences of missing Bodies in AI-Voice; and by examining how vocalists and artists working with synthetic Bodies and AI-voices are 'bringing the Body back' in their own practices. We contribute with a series of considerations for how practitioners and researchers may help to 'bring the Body back' into AI-voice technologies.

## CCS CONCEPTS

• **Applied computing** → **Performing arts**; *Law*; *Sound and music computing*.

## KEYWORDS

musical AI; voice; AI; body; artificial intelligence; STS

**ACM Reference Format:**
Kelsey Cotton, Katja de Vries, and Kıvanç Tatar. 2024. Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies. In *9th International Conference on Movement and Computing (MOCO '24), May 30–June 02, 2024, Utrecht, Netherlands.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3658852.3659065

## 1 INTRODUCTION

The Body is fundamental to the production of human vocalised sound, directly impacting how our voices are shaped, produced and shared with the world. As Indonesian vocalist Rully Shabara puts it: *"Your body [has] already decided what sound you make"* [12]. How can our Bodies still decide the sounds we make, when there is not yet a functional place for them [108] in current implementations of AI-voice and speech? We see this question as increasingly urgent, given recent global strike action in North America [1] and court proceedings in China [145, 161] concerning applications of AI to media and Arts domains.

The Body in AI voice and speech has gone missing, silenced somewhere beneath the roaring advancement of artificial intelligence (AI) tools for synthesis and generation [18, 72]. A poetic statement, and one we will start unraveling by clarifying why 'Body' and not 'body'. We use 'Body' to refer explicitly to a literal or conceptual source or origin point of a sound. We frame this term as different from 'body', which we use to refer to a physical or physiological form which may encompass human, robotic, or other biological morphologies with the capacity for vocalisation. When we speak about what a 'voice' is, we are positioned in an interdisciplinary intersection of definitions. Legal perspectives frame voice as an attribute one's self [17, 91, 124] whilst voice researchers frame it as a "technology of selfhood". [28, 29]. We take a composite stance by incorporating both the legal and voice community understanding of what a voice is. How did this Body go missing (and what do we mean by "missing")? By missing, we mean that the Body has been factually dis-entangled in the capture of voice and speech data, while keeping implicit connotations such as voice characteristics of an individual. And silenced? That the technological advancement of AI tools has rendered discussion of Body in relation to voice and speech as non-urgent. In this paper, we narrow our focus down to the latest technology advancements on deep learning based techniques for voice and speech recognition (ASR), text to speech synthesis (TTS), transformations of voice and vocoders in our position. In those advancements, we observe that the pace of AI progress is turning down the volume on discussions around the body politics of human voice and speech data.

The discussion of Body is a recurring point within larger discussions of technology [51, 59, 123, 149], and is a frequent point of focus within movement computing research. Our cursory examination into publications within the MOCO community, revealed few texts that actively engaged with applications of AI technology for generating or synthesising voice or speech, **and** which addressed the role of Body. We acknowledge here the excellent work in probing the significance of the Body and voice connection [9, 108]; and the utilisation of gesture and Bodily movement during singing as an interactive or performative tool [8, 13, 109]. Further exciting research at the cross-hairs of voice and AI include error detection in Byzantine chant [74] indicates that the time is ripe to plant the seeds for more research into the landscape of AI for voice and Body.

This paper seeks to plant those seeds by first fertilising the soil with ideas and questions. Our interdisciplinary fertiliser is branched from science and technology studies, philosophy of technology, critical technology studies in feminism, AI and machine learning, sound



*MOCO '24, May 30–June 02, 2024, Utrecht, Netherlands*

ACM ISBN 979-8-4007-0994-4/24/05
https://doi.org/10.1145/3658852.3659065

studies, and musical practices. From this interdisciplinary perspective we probe *how*, and indeed *why*, the Body is made invisible and separate from voice within applications of AI voice and speech synthesis tools. We see this topic as crucial and timely in the increasing migration [3, 94, 139, 158] of interactive AI systems to voice-based modes of interaction. Further, the evolving sophistication and capabilities of AI technologies to generate and synthesise voice material poses unique challenges. Our concern here is in establishing new practices for bringing back the Body in voice-Bodies, whilst also nurturing a space for the constructive development with participation of practitioners, informed use and consensual deployment of generative voice and speech technology. Our interest is informed by our respective backgrounds and artistic practices as vocalists, musicians, technologists and critical user-explorers of the same technologies we critique. To that end, we therefore constrain our discussion around *human* body and *human* voice.

This paper offers the following contributions. Firstly, we unearth the connections of Bodily absence to electroacoustic music compositional theory and practice [26, 128, 162]. Further, we introduce the use of conceptual tools from general and feminist science and technology studies (STS) as potential devices for recognising Bodily presence and evaluating relationships between a technology and the implications of it's use [20, 24, 46, 81]. We also introduce the use of perspectives and values from feminist data ethics and feminist AI [46, 49] for critiquing what is 'visibilised' and 'invisibilised' in AI-voice and -speech. We provide a general view of how voice is implicated by choices made during technology use and development. Finally, we provide a series of recommendations for how to *'take the Body back'* when engaging with AI speech and voice technologies based upon similar progress in adjacent media fields. [19, 116, 162]

The structure of the paper is deliberate in that we have made certain narrative groupings to reflect our engagement with theory from STS and to methodically examine the *facts* and *concerns* of voice and Body in relation to cloning and generative AI technologies. We draw from Latour's *matters of fact and concern* [81], and separately examine the physiology of voice (framed as a *fact*); voice technologies (a *fact* in its own right); and theories of sound and listening when it comes to voice usage in AI technologies (framed as a *concern*).

In the forthcoming Section 2, we briefly summarise theoretical perspectives and core concepts relevant to this paper's inquiry. We outline the physiology of voice in Section 3, emphasising the importance of Body in the production of vocal sound and further contextualise the relationship between them. Section 3.2 briefly summarises existing AI and ML-based voice technologies, focusing on speech recognition and speech synthesis. We make connections between the treatment of voice in ML and AI domains and Schaefferian and post-Schaefferian theories of listening, sound-material and sound-objects in Section 4. Section 4 also examines novel actions undertaken within experimental and popular music offering novel approaches to voice copyright and proprietorship (see 4.4). We further examine how artists implementing AI-voice and -speech technologies have approached Body (see 4.3) Lastly in Section 5, we propose some approaches to "making present" the Body in AI-voice technologies. We outline concepts and values that we wish to see more intentionally implemented in both the development, but also in the culture of *use* of these AI models "in the wild".

## 2 BACKGROUND

In this section we give a brief summary of the theoretical perspectives and core concepts that are relevant to this paper's inquiry, and clarify our usage of certain terminology.

### 2.1 The Body versus the body?

The Body and the body occupy varying positions within sound, movement, phenomenology, and voice studies [11, 26, 28, 30, 31, 50, 60, 101, 136, 137]. In this paper, we approach both of these terms based on the following premises. Firstly, that the physicality of the Body and it's profile of movement informs, defines and contributes to the properties of the sound it produces.[31] Secondly, that the movement and spatialisation of sound itself constitutes a **moving sound-Body** [136]. We therefore understand 'Body' in our usage of the term: i) as reflecting *Body-as-source* of sound; ii) as reflecting *Body-as-origin* of a sound; iii) that the Body itself fundamentally shapes the movement of the *sound-Body* that it produces; and iv) that Body itself also serves as a *medium* through which an origin-Body may be experienced by another.

We note that this paper's scope leaves the discussions on robotic Body, or robot vocality to future work due to size constraints.

### 2.2 Sound, Movement and Body

As discussed further in the forthcoming Section 3.1, bodily movement is integral to vocal sound production and has a significant role in shaping and influencing voice. This is also applicable to more general discussions on sound and musical practice, and has been extensively explored in embodiment research [42, 43, 65, 66, 99]. Gesture is frequently used as an explanatory term to discuss the movement of both sound, Body and sound-Bodies [65, 82, 136]. We acknowledge here that the term 'gesture' carries a lot of baggage, and means many different things across differing disciplines. To clearly communicate precisely what we mean, we have elected to **not** utilise this particular term, given and its myriad of meanings and usage. We also further constrain our scope in concentrating on human vocality. When we use the terms 'singing', 'phonation', and 'vocalisation' we are strictly referring to a human's singing, phonation and vocalisation.

### 2.3 Sound Discourses

As this paper addresses the usage of voice, it is therefore necessary to ground our position in relation to theories of listening from sound studies [71]. We specifically reference Schaefferian-thinking and conceptualisations of sound and listening and post-Schaefferian thinking on sound. Pierre Schaeffer was a French engineer and musician who formulated a philosophy of listening–*écoute réduite* (*reduced listening*)–and developed an approach to music-making with electronics- *musique concrète* (concrete music). In their *Traité des objets musicaux* (Treatise on Musical Objects) [128], Schaeffer developed Edmund Husserl's phenomenological notion of reduction [58]. Husserl's reduction separates information considered peripheral to the object that is being perceived from the object itself, to

describe the object. Intrigued by the potentials this afforded to listening processes, Schaeffer devised a series of 4 listening modes which sought to separate the sound object (*objet sonore*) from its notation, its provenance, and from the listener's perspectives on the sound. [71, 128] These modes are: *écouter*, *ouïr*, *entendre* and *comprendre*. Each mode is structured to facilitate an approach to listening that first prioritises indicative listening (écouter); that attends to the physiology of listening (ouïr); attending to selective listening or hearing with attention (entendre) and then listening to identify and contextualise (comprendre).

Post-Schaefferian understandings and theories of sound and listening view sound as "contain[ing] references to its actual or perceived origins, to some external association, or to some combination of the two" [26]. "Sound, in other words, is a sign that indicates something beyond itself and as such can never exist as a pure abstraction." [26]. We understand that post-Schaefferian perspectives on sound are connected with the sound's origin: it's sound-Body (see Section 2.2). Further discourses on listening emphasises it as an "action-oriented" and "intentional" activity [148], where different modes of listening are correlated to the acoustic action and the listener's intention when listening.

## 2.4 Copyright, IP issues and Unstable Rights

In the discourse of this paper, it is fundamental to clarify the terms: copyright, "property rights", "private/publicity rights" and "personality rights". These have become increasingly significant points of discussion within AI applications to media and Arts. We view the discussion on copyright and various rights protections as connected with our discussion on missing Bodies in AI-voice. Recent advancements and ease of access in AI technologies radically destabilise media industries–such as film, radio, TV, voice-acting, and music–which depend on and use voice. Historical protections that have been the norm within these same industries are being challenged by the new reality brought by AI models with imitation capabilities in voice generation. Here, the main concerns expressed by artists are largely those around copyright; vocal proprietorship; and the potential economic impact of vocal-AI tools on their livelihoods.

Legal protections for voice and voice rights can differ substantially from country to country. In regions such as North America, voice is *not* recognised as intellectual property, yet *is* considered (in some states) as a transferable property right [124]. This ability to licence, sell or to have one's voice appropriated by others implies that there is–on some level–a recognition of ineffable qualities to the voice that we have yet to formally quantify. Further, that we can readily identify a particular voice as being synonymous with a particular person. Within a US context, existing historical cases centring on unauthorised usage of vocal likeness [100] have concentrated on preserving the legal right to control commercial usage of identifiable aspects of an individual's persona or likeness, or to protect an individual's right to privacy in non-consensual collection and dissemination of vocal material [91]. These differences in rights protections are largely positioned around the protection of "privacy/publicity rights" [124] and "property rights" [17, 77]

This differs somewhat from the continental European legal context, in which "personality rights" are the cornerstone of rights protections. Further, "*the human voice from a judicial perspective is [understood as] one of the ways by which a human being can express herself/himself, thusly allowing her/him to bring forth her/his individuality, physiologically as well as psychologically, thereby totally fulfilling herself/himself as a person*"[4] Overall, we would like to highlight two important aspects of the European legal context: firstly, that physiology, and it's manifestation, is judicially perceived as connected with–and indeed as a result of–the voice as **moving sound Body**. Secondly, that the voice is considered as a fulfillment of one's self - which we understand as constituting the Bodily self.

This is also a legal concern in China, with a recent initiation of court proceedings in the Beijing Internet Court in China by a voice-over artist (known only by the surname Yin). Yin is suing five companies, who are accused of recording her voice for non-consensual cloning of her voice in digital audio books. [161] This case marks the first instance in China of an AI voice rights case, with the defendant companies presenting the argument that the "AI-processed voice was not same as Yin's original voice, and the two should be distinguished". The Civil Code in China provides legal protections for an individual's voice under "portrait rights", which prevents the forgery, exploitation and defacing of an individual's voice through technology. [145] Presiding Judge Zhao Ruigang has indicated that the court's ruling will be forthcoming, as the case concerns both the protection of portrait and personality rights, but also technological development. [161]

What is abundantly clear is that there is no single approach to how we legally define and protect voice on a global scale. We view this as important and critical grounds for future work.

## 2.5 Matters of Fact, Concern and Care

Our methodological framing of our discussion of voice and speech in applications draws from Bruno Latour's concepts of *'matters of fact'* and *'matters of concern'* [80]. Latour establishes a relation between *fact* and *concern* as an act of positioning the objective in relation to the "whole scenography" of its contextual environment. We apply Latour's *facts* by first examining the objective and factual aspects of voice and speech in AI: what is the physiology and how is voice within these domains? What precisely is the technology? How is the voice-data positioned and utilised in applications with AI? We then apply Latour's *concerns* by examining how the *facts* have shaped and informed the treatment of voice and speech: what is conveyed about the role of Body in relation to voice? We frame this examination of the *facts* and *concerns* of AI voice as a *'matter of care'*, which is a notion from Maria Puig de la Bellacasa [24]. de la Bellacasa defines this process as an engagement with how matters of fact and concern come to be. We position this paper as a further enactment of care, in it's examination of the *facts* and the *concerns* of AI voice, and the relationship between them.

# 3 VOICE AS FACT

In the following, we provide brief overviews in the areas of the physiology of singing; and an overview of the architectures, datasets and applications of AI technologies for voice and speech. We frame this section within Latour's notion of *matters of fact* (see Section 2.5), and examine the objective or factual of voice physiology and AI-voice [80]. This will create a factual foundation for the discussions of societal implications in Section 4 and 5.

## 3.1 Vocal anatomy and physiology

The human voice harnesses 3 different physiological sub-systems: the respiratory system, the phonatory system and the resonance system [142]. The respiratory system is composed of the organs, muscle structures and bone structures which facilitate the passage of air into and out of the Body. This includes the lungs, ribcage, intercostal muscles, the diaphragm and the trachea. In a healthy voice, the organs and other structures work together to help the lungs inflate and to expel air from the Body. When air passes out of the Body during singing, it passes through the phonatory subsystem - encompassing the vocal folds (located in the larynx). The human Body has two main vocal folds which are protected and kept moist by two auxiliary folds (these are called "false" folds). During speech or singing, an increase of pressure upon the vocal folds causes them to open and close in a cyclic fashion. This fold closure disrupts the flow of air, producing a buzz which is shaped and amplified by the resonance subsystem [168]. This resonance subsystem encompasses the vocal tract, the oral cavity, the sinus cavity and the bones within the face. The manipulation of soft tissues in these regions also directly affects the timbre or tone colour of the produced sound. The production of a sustainable vocal sound therefore demands a nuanced kinaesthetic understanding of how a singer may manipulate her physiology across these sub-systems. She also masters physiological changes that enable 'on-the-fly' adjustments to dynamically respond to the acoustics of the environment around her, such as through micro-changes to her diction and articulation. [28, 30, 60]

## 3.2 Voice and AI Technologies

The current AI technologies and approaches for Voice, framed within the conceptual notion of *matters of fact* [81], can be categorised roughly in three main threads: architectures, algorithms, and approaches for Voice; voice datasets, and AI voice applications.

*3.2.1 Architectures, algorithms, and approaches for Voice.* Historically, the synthesis of voice and speech has previously relied on physical modelling and simulations. [10, 36, 93, 105, 110, 162] Recently, the advancement of deep learning for speech and voice systems has led to significant breakthroughs for the synthesis and generation of voice and the development of speech processing tools. This can largely be categorised into several key areas: automatic speech recognition (ASR) [106, 155]; text-to-speech synthesis (TTS) [69, 84, 120, 146]; and transformation of voice and speech [70, 118, 147]; and audio and speech generation. [103, 154, 160]

Automatic Speech Recognition is the processing of human speech into text. This is achieved by the transformation of audio waveforms into token sequences; then the extraction of speech features; and then mapping of input speech features and speech tokens to text. Common architectures within deep learning pipelines for ASR include Connectionist Temporal Classification (CTC), Listen-Attend-Spell (LAS) and Recurrent Neural Networks (RNN).

Text-to-Speech Synthesis (TTS) is the synthesis of human speech from text input to audio output. Currently, deep neural networks (DNN) are utilised to achieve more natural-sounding speech. A TTS pipeline typically has two stages. The input text is converted to mel-spectrogram form. The mel-spectrogram is then converted to an audio waveform. WaveRNN [69], Tacotron2 [133], WaveGlow [115] and MelGAN [78] are popular networks for synthesising audio from mel-spectrograms. Current platforms for text-to-speech synthesis include subscription-based options such as Lyrebird [27]; Resemble.AI [121]; and Eleven Labs [79] as well as free and open-source toolkits such as SpeechBrain [140], NeMo [34] and SpeechT5 [5] to name a few. Other applications for the transformation of speech and voice using AI include style transfer with Transformers [2], and voice conversion [165, 166].

*3.2.2 Voice Datasets.* Voice and speech datasets are a crucial component in training machine learning and AI models. This data can encompass many different contextual cases of voice and speech; including recorded phone conversations [38, 114, 134], recorded interviews, extracted audio from video or film, or can be specifically recorded to build a new voice or speech data corpus. Documentation conventions for voice datasets include the labelling of the dataset with metadata. [25, 126, 167] This metadata provides additional, functional information about the audio file.[54] Common metadata labels, such as those utilised in Mozilla's Common Voice dataset [95], can include the length of each recording, file format, the speaker's sex, the context of what is discussed, as well as the language or accent of the speaker. [40, 73, 83, 85] Often, a transcription of each recording file is kept alongside it's corresponding audio file. Some examples of well known and commonly used datasets include: AudioMNIST [138], Common Voice [95], GigaSpeech [141], LibriSpeech [104], LibriTTS [164], LJ Speech [63], VoxCeleb [98] and Acappella [61].

*3.2.3 AI Voice Applications.* AI technologies for voice serve a broad range of functions. One example is in "hands-free" interaction with digital devices. Increasingly, applications of AI technologies that engage with the voice have been concentrated towards the development and distribution of voice-based AI agents. [44, 53, 57, 112] Common examples of technologies "in-the-wild" include AI Voice Assistants such as Apple's Siri, Amazon's Echo and Alexa, Google's Voice Assistant and Meta's deepfake celebrity chatbots; text-to-speech (TTS) generators; and voice cloning systems. Across these various platforms and technologies, we can observe an intentional disembodiment of the voice (both real and synthesised) from the Body it inhabits. Our concern here is the pathway such disembodiment opens (and has historically opened) for questionable activities [23, 35, 37, 67, 87, 150] as well as enabling unfair, uncompensated or non-consenting usage of the voice. We discuss this issue at greater length in the forthcoming Section 4.2.

# 4 MATTERS OF CONCERN AROUND VOICE

In this section, we discuss our concerns about the missing link between Body and voice within AI. To do this, we utilise Latour's notion of *matters of concern* (see Section 2.5). [81] To do this, we build on our earlier established knowledge of the *matters of fact* of voice (see Section 3) and look at the 'scenography' of the current practices of voice and speech treatment in AI systems. Our non-Latour concerns are: how technological necessity positions voice as an *objet sonore*. (See Section 4.1) We speculate on the consequences of missing Bodies in AI voice in Section 4.2 and discuss how post-Schaefferian perspectives (see Section 4.3) offer insights into the

Body in some recent artistic work. We further outline some artist-led approaches towards copyright and voice ownership in Section 4.4.

## 4.1 The Voice as *objet sonore* in ML and AI

When we contextualise usage of voice in AI within theory from sound studies, we can observe that the (singing) Body's presence is reduced within the data set. This is due to the single modality of data in the digital domain. That is, audio, video, and sensor data are positioned to exist independent of each other in computational approaches, until they are connected with additional means. Thus, the voice and voice data is reduced to its audio content as a result of the recording process: the link between voice and Body is broken. Through the recording process, the Body is made absent whilst identity-related components of the singer remain. Here, we understand this framing of voice solely as its recorded audio as akin to Schaeffer's notion of *objet sonore*: that it is a recorded "acoustic action" or sounding object. [128]

We acknowledge that there may be a fundamental functional requirement to making the Body missing from it's voice data (see Section 3.2.2) in the case of a singular modality of digital audio data. Incorporating the Body requires significant additional provisions to purely audio-based models designed for the synthesis and generation of voice and speech. This undoubtedly adds extra labour; demanding additional technical work when it comes to cultivating and working with a voice dataset; and bringing in computational complexities of working with multi-modal data.

The Body carries vital contextual information [56, 131, 132, 151, 152] about the identity of the singer. [28, 31]. We have previously discussed in Section 2.4 how voice is legally considered an integral and identifying part of Body. Invisibilising the Body from its voice data, by breaking it's connection to the identity of the singer, positions voice purely as an *objet sonore*. We view this as a process-dependent consequence of the technology we work with for building voice models. However, neglecting this information for the sake of pure functionality 'brackets out' the expressive, communicative and contextual Body and the richness of information it provides about *who* the sound has come from. In this, we include the affective impact of the Body's movement and physicality, as well as the movement and diffusion of the *moving sound Body* it is producing (ie. the vocalised sonic output) [136].

We speculate if this absence of Body is further exacerbated by an auditory context-of-use in which we wish to perceive sound. [151, 152] That is, we expect to "ha[ve] a more or less transparent relation to the properties of the sounding Body we see before us." [22]. By this we mean that our focus on the quality of sound output is deemed more important than 'seeing' the Bodies it is born from. Indeed, we question how a transparent relationship is possible in a context-of-use in which the voicing of Body (via it's physiological changes) is made missing, or absent.

Although we acknowledge that this invisibilisation is a consequence of the technical demands of formatting voice data for the development of voice models (see, we question if this may be inadvertently positioning an AI model to perform the listening modes of *entendre* and *comprendre* (see Section 2.3) whilst depriving it of important contextual information. Practitioners engaging with voice and speech generation and synthesis should not forget "[the] body [has] already decided what sound you make". [12] We must we find a way to assist AI voice technology development to 'bring the Body back' so as to help further develop the range of sounds AI models are capable of making.

## 4.2 Consequences of *missing* the Body

Voice carries the residues of the Body it is produced within, and the bodies it has touched in its production. Voice manifests Body [60] and we argue Body in turn manifests voice (see Section 2.4). Currently, applications of AI for voice and speech are destabilising this manifestation: there is often no **clear** Body present. Young observes, "The mortal, carnal, fleshly Body is bypassed entirely in the machine's rendering of a disembodied, omnipresent, devine or perfect ideal." [162] Although Young is speaking about humanoid speech, we see similar weight in their statements when we replace the word "machine" with "AI generated voice".

An example of the consequences of the missing link between Body and voice is in the historical case of voice-over artists Susan Bennett and Jon Briggs. Both Bennett and Briggs provided voice recordings for GM Voices, which were later licensed to ScanSoft. Their voice datasets were then later allegedly used to build the voice of the American Siri (Bennett) and British Siri (Briggs) through speech concatenation. [89, 107, 119, 127, 156]. Apple has never confirmed, nor denied whether they utilised Bennett's concatenated speech data, nor Briggs'. In the case of Bennett, audio forensics expert Ed Primeau studied recordings of Siri and blind recordings of Bennett's voice and presented his the conclusion of his analysis that "*They are identical – a 100 % match.*" [119] Both Bennett and Briggs have publicly spoken about being the original voices of Siri, and expressed a wish to have been more acknowledged by Apple in contributing to such a globally significant application of voice technology. [107] There are a number of consequences in the missing link between the Body and voice in this example of Bennett and Briggs. Firstly, there is the consequence of both Bennett and Briggs not having the opportunity to consent to the use of their voices in Siri. Secondly, neither Bennett and Briggs have been financially compensated by Apple for the use of the originally recorded speech datasets. [89, 107, 119, 127, 156]

A more recent example discussed earlier in Section 2.4 is the current legal case in China concerning non-consensual AI voice cloning for profit. The litigant, a voice-over artist known only by the surname Yin, is suing five digital audio-book companies and an AI Voice Cloning Platform (which has not been named) [161]. Yin is suing on the basis of the unauthorised recording, cloning and licensing of her voice model in the sale of audiobooks. Yin did not sign a contract, authorising the recording of her voice, nor did she financially benefit in any way from the sale of audiobooks that used a voice model of her likeness. [163] She is suing under Chinese "portrait rights" protections, which provide protections for the forgery, exploitation and defacement of an individual's voice. [145] The defendants in the case have counter-argued that the voice model is not the same as Yin's original voice and that the two voices should be distinguished separately. [161] Here, we see a profound consequence in the missing link between Body and voice: that the non-consensual implementation of digital technologies

such as voice cloning has seriously violated a person's autonomy. An additional consequence is in how (and if) we formulate a legal difference between the voice produced from a human body and the synthetic voice produced from an AI voice model.

We see the consequences from these example as indicators that a change of approach is needed. Keeping the Body missing from voice data heralds a range of legal problems, but also raises important questions regarding how consent and autonomy are navigated in the application of AI voice technologies. What immediate concerns arise from keeping AI models for voice and speech naive to the Body? (See 4.3) What do we miss when we *miss* the Body in voice and speech AI? (See 4.2)

## 4.3 Post-Schaefferian Considerations for AI Voice-Bodies

In this section we examine how post-Schaefferian perspectives of listening may provide potential directions as to the immediate concerns of Body-naïve AI voice and speech models. As discussed previously in 2.3 several criticisms on the Schaeffer's notion of *objet sonore* have been put forward in sound studies [26]. For example, post-Schaefferian listening is framed as an embodied and intentional activity, whilst Schaeffer's is a reflective practice (see Section 2.3). Further, the post-Schaefferian considers sound as "indicat[ing] something beyond itself" [26]. The post-Schaefferian approaches can be the guiding light in 'bringing the Body back': the emphasis on embodied-ness and intentionality may prove beneficial in revealing how voice is *controlled* through the *absence* or the *making absent* of Body. In making the Body absent and positioning AI-synthesised voice and speech as separate from a Body, this enables voice to be appropriated and utilised in ways that might be morally "fuzzy". If a voice doesn't belong to a Body, and is "*without the distraction of the human 'grain'*" [162] it may be considered acceptable and permissible to use it for *any* purpose. When the significance of Body is absent, it becomes considered acceptable to *use* the voice. We can see this concern reflected in the actions of SAG-AFTRA to come to an agreement with AMPTP on acceptable usage of performer's likeness. We can observe similar patterns of object-ification of the Body in performance art of the 20th century [14, 68, 90]. When we consider the ramifications of considering the voice as *objet sonore*, concerns about use, copyright, and ambiguity of the boundaries between human and non-human voices emerge. [51, 102, 111, 130]

But, we are hopeful. One domain where we see voice begin to transcend it's framing as *objet sonore* within the wider landscape of AI voice is in the context of experimental music composition and even within mainstream popular music. This can be seen primarily through the usage of additional technologies such as virtual reality (VR); augmented reality (AR); or using deep generative visuals (or deepfakes) to construct a Body for applications of AI-voice/-speech.

One example within experimental music is the work of British-Iranian artist-performer-software humanist Ashkan Kooshanejadin, namely in their creation of and artistic activities with their the synthetic performer named 'Yona'. Yona is described as "first generation 'Auxiliary Human" [76, 97, 122], and is frequently visually presented in a humanoid-esque form in more static images, and in a holographic form during live performances. It utilises a generative pre-trained transformer model (GPT) and an autoregressive language model for poetry and lyric generation. Yona's poetry and lyrics are 'voiced' through a text-to-speech model which is pushed through a melodic filter, encoded and then decoded into more 'sung' output. [15] The Body of Yona is significant in terms of how it's morphology is presented, and what this communicates about it's vocality. Yona is Bodily presented throughout purely digitally-based technologies in the form of CGI and coded visuals and moving imagery from Isabella Winthrop. [62] The Body of Yona is never fixed, but is instead a *moving sound Body* that shifts morphology to occupy first screen-based domains and later the experiential domain through holographic form. [75] There is clearly an embrace of novel technologies to bring the Body of Yona dynamically into a context where it's physical presence can be more pervasively felt and experienced. It becomes 'real' to us as an audience through it's co-located inhabitance of the same space. This is not to say, however that its real-ness also refers to the apparent visual aesthetic of it's Body. Rather, the Body of Yona appears to consistently reflects a glitchy, highly synthesised and processed visual aesthetic. We hear this in Yona's voice also. It sings in a very text-oriented fashion, with heavily articulated phrases, charmingly stilted spoken syntax and a disjointed pace of vocal production. We hear a noisy-buzz and auto-tune-like timbral quality when Yona sings- a byproduct of the TTS pipeline that Kooshanejadin has used to give Yona it's voice.

An example from popular music is the usage of deepfakes to both provide a Body, and to generate a suitably convincing human vocal sound. [6, 55, 88, 96] VAVA is an AI artist produced from a collaboration between T-Town Digital Studio, PRO-toys, and Drive iGency. [143] We found limited information regarding the technical assemblage of VAVA, with sources only describing VAVA as being built with "AI technology" and not describing precisely what *form(s)* of AI-technology. [33] Regardless of the accessibility of details around it's technological composition, VAVA has a significant online presence. It is prominently featured on the T-Town Digital Studio YouTube channel, it has its own Instagram account and TikTok channel. The Body of VAVA has a very specific visual aesthetic, which may be a consequence of the technology used to realise it. Based on our subjective experience and listening, we speculate that motion capture technology is used to transpose an AI-generated face and facial movements onto the 'original' body. Watching VAVA perform in it's videos, it's Bodily engagement with the surrounding space is almost *too* human-like. It's Bodily movement is fluid, smooth, at a believably human pace and demonstrates minimal glitch (aside from the obvious post-production visual effects). VAVA begins to push the borders of "uncanny valley" territory [129], it is almost "hyper-real". Further, VAVA's vocal sound is very present within the overall mix, with a boosted warmth that helps it to "pop" against the backing instrumentals. The sound profile, to our ears, is reminiscent of the early 2000s female pop vocalist sound, but with heavy usage of reverb and filtering. We suspect that VAVA's mid-range has also been generously EQ-d as it is very difficult to hear the undertones in the voice.

In both the example of Yona and VAVA, we can observe that the voice-Body or origin sounding Body is being mediated by an auxiliary Body. Further, the auxiliary Bodies in turn become an origin point, or Body, in their own right. This is turn enables a more solid grounding, or connection, between the voice as a **moving sound**

**Body** in its own right, and begins to establish more solid terrain of the voices we hear as being born *from* a Body. Here we see two examples of artists and organisations producing vocal sound with AI tools actively demonstrating concepts and understandings from post-Schaefferian sound. That is: their usage of mediating technologies to produce a sound-Body constructs "references to...actual or perceived origins" of the voice. [26]

### 4.4 Legal Considerations and Current Actions in AI-Voice

The Post-Schaefferian perspectives in the previous section provide a philosophical framing in how we can conceptually bring the Body back together with voice. Still, there is an immediate need in discussions of policies towards entangling the Body and voice in public discourses and artistic practices.

As it stands currently, there is ambiguity and a general lack of clarity as to how one's voice or speech is included in the protections afforded by copyright in the age of generative AI (see Section 2.4). How *do* we establish *acceptable* difference between one voice quality profile compared to another? Does this mean we would need to trademark our voice or speech mannerisms? Potential answers to these questions may lie in the unfolding novel attitudes and approaches already taking place within the field of musical performance and composition, where vocalists working with tools for AI voice seem to be "leading the way".

On the front-lines of vocal proprietorship, vocalists such as Canadian artist Grimes and American artist Holly Herndon have opted for progressive approaches which actively trouble the notion of vocal ownership and copyright. As an example, Grimes has actively encouraged open-usage of her vocal likeness on AI-generated songs [48], and has publicly expressed their support of "killing copyright" [47]. An alternative approach is Herndon's *Holly+* voice model and accompanying *Holly+DAO* [45], which has previously been critiqued using feminist STS and interdisciplinary methods [20]. Herndon's approach to vocal ownership is to distribute proprietorship and guardianship of their voice model, and enable participants in the *Holly+DAO* to financially share in the profits of usage of the model.

Our stance here is to find a middle ground between complete abolition and distributed guardianship of vocal proprietorship in the age of vocal AI. To achieve this, we see that the core concerns in this regard need to incorporate values and perspectives which prioritise stewardship, management and the *who* (and their Bodies!) in voice data.

Some important progress made in this regard can be seen in the tentative agreement made by the Screen Actors Guild-American Federation of Television and Radio Artists (SAG-AFTRA) in their strike resolution with the Alliance of Motion Picture and Television Producers (AMPTP). [7, 19, 157] The tentative agreement specifically outline protocols and establishes compensation and rights protections of human performers whose likenesses–including their vocal likeness–are to be duplicated through generative AI for usage within film, television and radio broadcasts. Specifically, we highlight their emphasis on "clear and conspicuous" consent [1]; and the clarity on the conditions under which consent is the resultant replicas may be 'adjusted' in post-production.

Throughout the entirety of the agreement there is a continual reinforcement of informed and specific consent as one of the foremost obligations during contract negotiations. We point out that consent is largely presented in the agreement [1] as: *"[a]n endorsement or statement in the performer's employment contract that is separately signed or initiated by the performer or in a separate writing that is signed by the performer"* That is, that the contracted performer is responsible for establishing the details regarding *what they themselves have determined is acceptable* and permissible for the construction of their replica and any terms under which it is to be used. We view this as a potential learning to bring across to the music and sound domain: that the conditions and grounds of use of AI technologies applied to duplicate or replicate a performer should be established by the performer themselves, and with appropriate and accessible legal counsel However, we are concerned about some exceptions to the manipulation of performer voice and vocality in the agreement. As an example, exceptions to consent for alterations on non-background performers recorded performances encompass: noise reduction; timing; continuity of pitch; clarity; the addition of sound effects or filters; and even adjustments in dialogue [1]. Further exceptions to consent include the alteration of facial and body movements, as well as the voice itself, for adaptation to a different language. In our view, these manipulations are not insignificant, and may indeed dramatically change the overall affect of the performer's **moving sound Body** (see Sections 2.1,3.1 and 2.2). One potential avenue to counteract the implications and consequences of such (potentially) dramatic manipulation of voice is to examine how theories and perspectives from a post-Schaefferian view (see Section 2.3) may inform new approaches for 'bringing the Body back' .

The cases of Grimes and Herndon are two examples in a historical pile of artists leading new technology in its amalgamation to the society. Artists historically tend to be the earliest adopters of novel technologies [16, 32, 41] and establish the trends and directions of how such technologies may grow in future. Artist's engagement with new technologies to create, produce and distribute their work has led to the birth of significant cultural movements, such as internet art, software art and non-fungible tokens (NFTs). [144, 153] As early adopters, we speculate that the needs of artists in their usage of these technologies also provides indications for the construction of legal structures concerning the usage of AI.

## 5 NEW TERRAINS

It could be argued that the existing terrain of AI voice is primarily concerned with and defines an AI-model's success in terms of it's accuracy, it's speed and it's computational cost [92], with other significant factors and considerations such as the human labour and bodies which have contributed to the model's construction, it's data thrown by the wayside. We need to '*bring the Body back*' into the discussion when we talk about AI voice and speech generation and synthesis.

What are the consequences of *not* doing so? Sustaining a continued invisibilisation of Body in voice and speech AI applications launches a tsunami of formidable sociocultural issues and questions. We view the risks as constituting a continued devaluation of Bodily rights and labour (see Section 2.4); the normalisation of prioritising

technological progress over people; and an avoidance of asking ourselves and others sticky and squirmy questions. We ask them now: How do we protect human voice and vocality? How do we protect human voice-Bodies? How do we dismantle current modes and practices of generating and synthesising voice and speech with AI to '*bring the Body back*'?

Our intention in asking the sticky and squirmy questions is to provoke, to trouble [52], and to begin the process of imagining *new* terrains, systems and practices of working with AI technologies that constructively contribute to innovative and informed use in artistic contexts. What might a terrain for AI-voice technologies that *actively includes* the Body look like? And how might we as practitioners cultivate and navigate this new terrain? We have several propositions here.

## 5.1 Clear(er) Voice Body(ies) and Rights

Our first core proposition is to make voice-Body(ies) clearer. We envisage this to be done in the following ways. Firstly, by emphasising the connection of Body to voice in applications of AI-voice and -speech technology. Secondly, emphasising the connection of voice to Body. And thirdly, by asserting the connection of voice-Body to voice-data.

In adjacent media domains, we can see the beginnings of novel approaches to asserting the connection of Body to voice in the terms of the SAG-AFTRA agreement (see Section 4.4). Here, the establishment of consent and the conditions for permissible use is established by the performer themself. We do however believe that the issue of whether constitutional rights should be permitted to take precedence over an individual's consent are an important topic of public discussion. From these collective examples of voice rights "in action", we can derive 2 initial sub-propositions. Firstly, the establishment of clear, unambiguous and forward-looking copyright, privacy rights, property rights and publicity rights need to be a priority topic. This is especially and urgently needed within artistic contexts, and most particularly in the disem-Bodied usage of generated or synthetic voice and speech. Secondly, the boundaries of acceptable use should be *people*-led and -centred, not profit-centred or progress- driven. This may call for new models of voice proprietorship or stewardship, or even an examination on the suitability of current legal protections for voice and speech.

We have seen the clarity of voice to Body demonstrated within artistic contexts in the earlier examples of Kooshanejadin's creation Yona and VAVA (see Section 4.3). In those examples, we have clearly seen the impact that a mediating or auxiliary Body has in grounding the voice within a physicalised or digital morphology.

We further see positive assertion of the connection of voice-Body to voice-data in the terms of usage for the recently released VocalNotes voice dataset [116]. We specifically refer to their outlined requests in their Dataset Access Request Form: [117] *"The VocalNotes Dataset contains audio that includes sensitive religious and ritual recordings of living musicians and communities. Please treat the recordings with respect as you would treat the performers recorded in them, and do not share them on social media or disseminate them otherwise."* Voice-data and voice-Bodies are expected to be treated with equivocal respect: *"Please treat the recordings with respect as you would treat the performers recorded in them"*. We do not view this as an attempt to anthropomorphise audio recordings. We see this as an assertive positioning of the direct relation from the voice-data to the voice-Bodies. In requesting the same respectful treatment of data and the recorded performers, there is an acknowledgement of the important and sacredness of the labour and physical voice cultures captured in VocalNotes.

## 5.2 Make Space for the Human 'Grain'

The second core proposition is to make a space for the human 'grain' in AI-technologies for voice and speech. As Young observes, the historical development of speech and voice has been to rebuild" *the voice object, in its pure form, without the distraction of the human 'grain'."* [162]. We see this pursuit of purity as both problematic and uninspired. As technologies such as TTS continue to advance in sophistication and are increasingly normalised, we run the risk of manufacturing–and normalising–a vocal Uncanny Valley [21, 39, 64, 135]. As Ihde puts it, "Sounds are 'first' experienced as sounds of things " [60], and indeed the Bodies they are born from. The advocacy for, and inclusion of 'grain' is therefore imperative to remake a space for Body when it comes to synthesised vocal sound.

Our suggestions here are embryonic, but are towards pushing back against dichotomous ideals of the perceived 'imperfection' of the human fleshy Body and the coveted, idealised 'perfection' of machinistic or technological bodies [86, 125] . We advocate for embracing the glitch in the "mortal... fleshly Body" and the possibilities this affords musically and creatively in disturbing the "machine's rendering of a disembodied, omnipresent, devine or perfect ideal." [162]

## 5.3 Trouble with Care

The third proposition is to include process and procedures of 'Caring Trouble', and *matters of care* [24] into our development and implementation of AI-technologies for voice and speech. 'Caring Trouble' has previously been presented in [20] as an analytical approach to exploring how formal computational structures inform–and are in turn–informed by how an AI artefact is presented, used and shared. This analytical approach actively troubles the expectation that AI is, or should (still) be, a "black box" [113, 159] by outlining a scaffold-ed approach to examining the connections between AI form and function. One of the core tenements of 'Caring Trouble' is to "*critically examine what is 'visibilised' ... so that we may in turn be able to critically address the components ... that appear 'invisibilised'*" [20]. This is positioned as in line with de la Bellacasa's call to engage with how matters of fact and concern come to be. [24]

# 6 CONCLUSION

This work casts a critical eye on current practices in the usage of human voice and speech within applications of AI-voice and -speech technologies. To assist this critical evaluation, we established connections to methods and conceptual tools from general and feminist science and technology studies. We engaged with feminist data and ethics principles in probing what contributing factors have led to the 'invisibilisation' of the Body in AI-voice. We have drawn connections between the treatment of voice and voice-data as *objet sonore* with AI-voice, and speculated on the implications

this has upon copyright and legal protections for voice. We have examined novel directions in copyright and voice proprietorship within the domains of experimental and popular music, and further how auxiliary technologies assist in the formation of **moving sound Bodies**. Finally, we have contributed with a series of considerations for 'making present' the absent bodies which contribute to AI-voice technologies: to make space for the human 'grain' and to enact processes of 'Caring Trouble' to critically examine what is in-/visibilised in our implementation of AI tools and technologies for voice and speech.

This position paper has explored a rich and deep sea of interconnected domains. Throughout this paper are a range of exciting directions for further work into AI-voice and AI voice-Bodies. We imagine future research as encompassing the following areas and directions. Firstly, we urgently require more concrete definitions and universally implementable best legal practices when it comes to protecting voice and voice-Bodies in the continual advancement of generative AI. This, secondly, requires interdisciplinary discussions and conversations on how to practically achieve this whilst also ensuring these protections also nurture a space for the constructive development, informed use and consensual deployment of generative voice and speech technology. Thirdly, we see exciting potential in further clarifying the research field of collaborative Human and AI-Vocality. Future exploration in this area may further contribute to the development of novel frameworks and methods for evaluating artistic human-AI collaboration. And fourthly, that there is critical work needed with regards to further analysing and deconstructing power structures within the field of AI-voice, with ample consideration into how to dismantle the linguistic, social and digital barriers of access which concern AI research.

## ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marcus and Amalia Wallenberg Foundation.

## REFERENCES

[1] SAG AFTRA. 2023. Summary of 2023 Tentative Successor Agreement to the 2020 Producer-SAG-AFTRA Codified Basic Agreement ('Codified Basic Agreement') and 2020 SAG-AFTRA Television Agreement ('Television Agreement'). https://www.sagaftra.org/files/sa_documents/TV-Theatrical_23_Summary_Agreement_Final.pdf

[2] Shrutina Agarwal, Sriram Ganapathy, and Naoya Takahashi. 2022. Leveraging Symmetrical Convolutional Transformer Networks for Speech to Singing Voice Style Transfer. arXiv:2208.12410 [cs.SD]

[3] AIContentfy. 2023. The future of content creation for voice assistants. https://aicontentfy.com/en/blog/future-of-content-creation-for-voice-assistants

[4] Julia Ammerman Yebra. 2018. The Voice of the Opera Singer and Its Protection: Another Look at the Maria Callas Case. In *Law and Opera*, Filippo Annunziata and Giorgio Fabio Colombo (Eds.). Springer International Publishing, Cham, 253–267. https://doi.org/10.1007/978-3-319-68649-3_17

[5] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. https://doi.org/10.48550/arXiv.2110.07205 arXiv:2110.07205 [cs, eess].

[6] Srishreya (Shreya) Arunsaravanakumar. [n. d.]. Deepfake music sends ripples across the music industry. https://thewildcattribune.com/17528/ae/deepfake-music-sends-ripples-across-the-music-industry/ Section: Arts & Entertainment.

[7] Will Bedingfield. 2023. Hollywood Writers Reached an AI Deal That Will Rewrite History. *Wired* (2023). https://www.wired.com/story/us-writers-strike-ai-provisions-precedents/ Section: tags.

[8] Grégory Beller. 2014. The Synekine Project. In *Proceedings of the 2014 International Workshop on Movement and Computing (MOCO '14)*. Association for Computing Machinery, New York, NY, USA, 66–69. https://doi.org/10.1145/2617995.2618007

[9] Greg Beller. 2015. Sound space and spatial sampler. In *Proceedings of the 2nd International Workshop on Movement and Computing (MOCO '15)*. Association for Computing Machinery, New York, NY, USA, 156–159. https://doi.org/10.1145/2790994.2791010

[10] Elise Jozefa Bikker. 2021. *Mind over matter: the thinking and speaking machine in fiction of the long nineteenth century*. phd. University of York. https://etheses.whiterose.ac.uk/31783/

[11] Carolyn Birdsall and Anthony Enns. 2008. *Sonic Mediations: Body, Sound, Technology - Cambridge Scholars Publishing*. Cambridge Scholars Publishing. https://www.cambridgescholars.com/product/9781847188397

[12] Phil E. Bloomfield. 2021. Without Limits or Lyrics: The Human Voice as Instrument. https://daily.bandcamp.com/lists/human-voice-as-instrument-list Section: Lists.

[13] Courtney Brown. 2020. Lament: An Interactive Cabaret Song. In *Proceedings of the 7th International Conference on Movement and Computing (MOCO '20)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3401956.3404249

[14] Samantha Bruce. 2016. The Female Façade: How Performance Artists Are Changing the Way Patriarchal Pressures Objectify.... https://medium.com/@SamanthaBruce/the-female-fa%C3%A7ade-how-performance-artists-are-changing-the-way-patriarchal-pressures-objectify-c3b288fa35e4

[15] Henry Bruce-Jones. 2020. Ash Koosha Presents: YONA Part I - (Under Your Skin). https://www.factmag.com/2020/11/25/ash-koosha-presents-yona-part-i/

[16] Linda Candy, Ernest Edmonds, and Fabrizio Poltronieri. 2018. *Explorations in Art and Technology* (2 ed.). Springer London. https://link.springer.com/book/10.1007/978-1-4471-7367-0

[17] District of Columbia) Cato Institute (Washington (Ed.). 2023. *Cato Handbook for Policymakers* (9th edition ed.). Cato Institute, Washington.

[18] Devin Coldewey. 2023. VALL-E's quickie voice deepfakes should worry you, if you weren't worried already. https://techcrunch.com/2023/01/12/vall-es-quickie-voice-deepfakes-should-worry-you-if-you-werent-worried-already/

[19] Kevin Collier. 2023. Actors vs. AI: Strike brings focus to emerging use of advanced tech. *NBC News* (July 2023). https://www.nbcnews.com/tech/tech-news/hollywood-actor-sag-aftra-ai-artificial-intelligence-strike-rcna94191

[20] Kelsey Cotton and Kıvanç Tatar. 2023. Caring Trouble and Musical AI: Considerations towards a Feminist Musical AI. *AIMC 2023* (aug 29 2023). https://aimc2023.pubpub.org/pub/zwjy371l.

[21] Trevor Cox. 2019. The uncanny valley: does it happen with voices? http://trevorcox.me/the-uncanny-valley-does-it-happen-with-voices

[22] John Croft. 2007. Theses on liveness. *Organised Sound* 12, 1 (April 2007), 59–66. https://doi.org/10.1017/S1355771807001604 Publisher: Cambridge University Press.

[23] Cassandra Cross. 2022. Using artificial intelligence (AI) and deepfakes to deceive victims: the need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety* 24, 1 (March 2022), 30–41. https://doi.org/10.1057/s41300-021-00134-w

[24] María Puig de la Bellacasa. 2017. *Matters of Care: Speculative Ethics in More Than Human Worlds*. University of Minnesota Press. https://libgen.li/ads.php?md5=3dec273eb9043ae8b1a7140b1120c759

[25] Robbie De Sutter, Stijn Notebaert, and Rik Van de Walle. 2006. Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries. In *Research and Advanced Technology for Digital Libraries (Lecture Notes in Computer Science)*, Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco (Eds.). Springer, Berlin, Heidelberg, 220–231. https://doi.org/10.1007/11863878_19

[26] Joanna Teresa Demers. 2010. *Listening through the noise: the aesthetics of experimental electronic music*. Oxford University Press, Oxford ; New York. OCLC: ocn435918247.

[27] Descript. 2023. Lyrebird. https://www.descript.com/lyrebird

[28] Nina Eidsheim. 2008. *Voice as a technology of selfhood: Towards an analysis of racialized timbre and vocal performance*. Ph. D. Dissertation. University of California, San Diego. https://www.academia.edu/657536/Voice_as_a_technology_of_selfhood_Towards_an_analysis_of_racialized_timbre_and_vocal_performance

[29] Nina Eidsheim, Katherine Meizel, Nina Eidsheim, and Katherine Meizel (Eds.). 2019. *The Oxford Handbook of Voice Studies*. Oxford University Press, Oxford, New York.

[30] Nina Sun Eidsheim. 2011. Sensing Voice: Materiality and the Lived Body in Singing and Listening. *The Senses and Society* 6, 2 (July 2011), 133–155. https://doi.org/10.2752/174589311X12961584845729

[31] Nina Sun Eidsheim. 2015. *Sensing sound: singing & listening as vibrational practice*. Duke University Press, Durham.

[32] Jacques Ellul and Daniel Hofstadter. 1979. Remarks on Technology and Art. *Social Research* 46, 4 (1979), 805–833. http://www.jstor.org/stable/40970814

[33] Thailand Posts English. 2022. VAVA, Thailand's first female artist Ai who looks like a real human. ready to go through music and reality shows. https://thailand.postsen.com/local/85090/VAVA-Thailand%E2%80%99s-first-female-artist-Ai-who-looks-like-a-real-human-ready-to-go-through-music-and-reality-shows.html Section: Local.

[34] NVIDIA AI Enterprise. [n. d.]. NVIDIA NeMo. https://nvidia.github.io/NeMo/

[35] Jeong Eui-seok and Lim Jong-in. 2019. Study on Intelligence (AI) Detection Model about Telecommunication Finance Fraud Accident. *Journal of the Korean Institute of Information Security and Cryptology* 29, 1 (2019), 149–164.

[36] Zsuzsanna Fagyal. 2001. Phonetics and speaking machines: On the mechanical simulation of human speech in the 17th century. *Historiographia Linguistica* 28, 3 (Jan. 2001), 289–330. https://doi.org/10.1075/hl.28.3.02fag Publisher: John Benjamins.

[37] Emily Flitter and Stacy Cowley. 2023. Voice Deepfakes Are Coming for Your Bank Balance. *The New York Times* (Aug. 2023). https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html

[38] Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shayna Gardiner, Pooja Hiranandani, and Shashi Bhushan TN. 2022. Entity-level Sentiment Analysis in Contact Center Telephone Conversations. arXiv:2210.13401 [cs.CL]

[39] Gaby Gayles. 2019. Voice Assistants & the Uncanny Valley: The More Lifelike, the Less "Real". https://medium.com/voice-tech-podcast/voice-assistants-the-uncanny-valley-the-more-lifelike-the-less-real-fb0bab2755d1

[40] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[41] Luís Miguel Girão and Céu Santos, Maria. 2019. *The historical relationship between artistic activities and technology development.* Publications Office, LU. https://data.europa.eu/doi/10.2861/961315

[42] Rolf Inge Godøy. 2009. Gestural Affordances of Musical Sound. In *Musical Gestures*. Routledge. Num Pages: 23.

[43] Rolf Inge Godøy. 2010. Images of Sonic Objects. *Organised Sound* 15, 1 (April 2010), 54–62. https://doi.org/10.1017/S1355771809990264 Publisher: Cambridge University Press.

[44] Saurabh Goorha and Raghuram Iyengar. 2020. *Voice Analytics and Artificial Intelligence: Future Directions for a post-COVID world.* White Paper Wharton AI & Analytics for Business. Wharton University of Pennsylvania. https://aiab.wharton.upenn.edu/white-paper/voice-analytics-and-artificial-intelligence-future-directions-for-a-post-covid-world/

[45] Decrypt / Will Gottsegen. 2021. Holly Herndon Launches DAO-Controlled Vocal Deepfake Platform 'Holly+'. https://decrypt.co/75958/holly-herndon-launches-dao-controlled-vocal-deepfake-platform-holly/ Section: News.

[46] Joanne Gray and Alice Witt. 2021. A feminist data ethics of care for machine learning: The what, why, who and how. *First Monday* (Dec. 2021). https://doi.org/10.5210/fm.v26i12.11833

[47] Grimes [@Grimezsz]. 2023. I think it's cool to be fused w a machine and I like the idea of open sourcing all art and killing copyright. https://twitter.com/Grimezsz/status/1650304205981089793

[48] Grimes [@Grimezsz]. 2023. I'll split 50% royalties on any successful AI generated song that uses my voice. Same deal as I would with any artist i collab with. Feel free to use my voice without penalty. I have no label and no legal bindings. https://t.co/KIY60B5uqt. https://twitter.com/Grimezsz/status/1650304051718791170

[49] Lelia Marie Hampton. 2021. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/3442188.3445929

[50] Jungmin Grace Han. 2019. The Somaesthetics of Musicians: Rethinking the Body in Musical Practice. *The Journal of Somaesthetics* 5, 2 (Dec. 2019). https://journals.aau.dk/index.php/JOS/article/view/2200 Number: 2.

[51] Donna Haraway. 2006. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late 20th Century. In *The International Handbook of Virtual Learning Environments*, Joel Weiss, Jason Nolan, Jeremy Hunsinger, and Peter Trifonas (Eds.). Springer Netherlands, Dordrecht, 117–158. https://doi.org/10.1007/978-1-4020-3803-7\protect\discretionary{\char\hyphenchar\font}{}{}4

[52] Donna J. Haraway. 2016. *Staying with the trouble: making kin in the Chthulucene.* Duke University Press, Durham.

[53] Rajibul Hasan, Riad Shams, and Mizan Rahman. 2021. Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research* 131 (July 2021), 591–597. https://doi.org/10.1016/j.jbusres.2020.12.012

[54] Diane I. Hillmann, Rhonda Marker, and Chris Brady. 2008. Metadata Standards and Applications. *The Serials Librarian* 54, 1-2 (May 2008), 7–21. https://doi.org/10.1080/03615260801973364 Publisher: Routledge _eprint: https://doi.org/10.1080/03615260801973364.

[55] Thomas Hobbs. 2021. 'It's Fan Fiction For Music': Why Deepfake Vocals of Music Legends Are on the Rise. https://www.billboard.com/pro/deepfake-music-imitations-history/

[56] Kristina Höök, Sara Eriksson, Marie Louise Juul Søndergaard, Marianela Ciolfi Felice, Nadia Campo Woytuk, Ozgun Kilic Afsar, Vasiliki Tsaknaki, and Anna Ståhl. 2019. Soma Design and Politics of the Body. In *Proceedings of the Halfway to the Future Symposium 2019* (Nottingham, United Kingdom) *(HTTF 2019)*. Association for Computing Machinery, New York, NY, USA, Article 1, 8 pages. https://doi.org/10.1145/3363384.3363385

[57] Kuo-Liang Huang, Sheng-Feng Duan, and Xi Lyu. 2021. Affective Voice Interaction and Artificial Intelligence: A Research Study on the Acoustic Features of Gender and the Emotional States of the PAD Model. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.664925

[58] Edmond Husserl. 2012. *Ideas: General Introduction to Pure Phenomenology.* Routledge. https://www.routledge.com/Ideas-General-Introduction-to-Pure-Phenomenology/Husserl/p/book/9780415519038

[59] Don Ihde. 1990. *Technology and the Lifeworld: From Garden to Earth.* Indiana University Press.

[60] Don Ihde. 2007. *Listening and Voice* (2nd edition ed.). State University of New York Press, New York. https://sunypress.edu/Books/L/Listening-and-Voice2

[61] IPCV. [n. d.]. Acappella. https://ipcv.github.io/Acappella/acappella/

[62] IPNHK. 2019. Isabella Winthrop. https://medium.com/@IPNHK/isabella-winthrop-d07814ab4ba0

[63] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[64] Dennis Jansen. 2019. *Discovering the uncanny valley for the sound of a voice.* Ph. D. Dissertation. Tilburg University, Netherlands. http://arno.uvt.nl/show.cgi?fid=149554

[65] Alexander Jensenius, Marcelo Wanderley, Rolf Godøy, and Marc Leman. 2009. Musical Gestures: concepts and methods in research. *Musical Gestures: Sound, Movement, and Meaning* (Jan. 2009). https://doi.org/10.4324/9780203863411

[66] Alexander Refsum Jensenius. 2007. *Action-sound : developing methods and tools to study music-related body movement.* Doctoral thesis. University of Oslo, Norway. https://www.duo.uio.no/handle/10852/27149 Accepted: 2013-03-12T12:01:35Z.

[67] Eui-seok Jeong and Jong-in Lim. 2019. Study on Intelligence (AI) Detection Model about Telecommunication Finance Fraud Accident. *Journal of the Korea Institute of Information Security and Cryptology* 29, 1 (2019), 149–164. https://doi.org/10.13089/JKIISC.2019.29.1.149 Publisher: Korea Institute of Information Security and Cryptology.

[68] Amelia Jones. 1998. *Body Art/Performing the Subject.* University of Minnesota Press. https://www.upress.umn.edu/book-division/books/body-art-performing-the-subject

[69] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient Neural Audio Synthesis. http://arxiv.org/abs/1802.08435 arXiv:1802.08435 [cs, eess].

[70] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks. http://arxiv.org/abs/1806.02169 arXiv:1806.02169 [cs, eess, stat].

[71] Brian Kane. 2007. L'Objet Sonore Maintenant. *Organised Sound* 12, 1 (April 2007), 15–24. https://doi.org/10.1017/S135577180700163X Publisher: Cambridge University Press.

[72] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. 2021. How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey. arXiv:2111.14203 [cs.SD]

[73] Keunhyoung Luke Kim, Jongpil Lee, Sangeun Kum, Chae Lin Park, and Juhan Nam. 2020. Semantic Tagging of Singing Voices in Popular Music Recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 1656–1668. https://doi.org/10.1109/TASLP.2020.2993893

[74] K. Kokkinidis, A. Stergiaki, and A. Tsagaris. 2016. Error prooving and sensorimotor feedback for singing voice. In *Proceedings of the 3rd International Symposium on Movement and Computing (MOCO '16)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/2948910.2948952

[75] Ash Koosha. 2019. Meet Yona, a first generation 'Auxiliary Human' | Facebook. https://www.facebook.com/ashkoosha/posts/meet-yona-a-first-generation-auxiliary-human-who-uses-artificial-intelligence-an/10157017594454400/

[76] Ashkan Kooshanejad. [n. d.]. Yona. https://theyona.bandcamp.com

[77] Kimberlee Kruesi. 2024. Tennessee just became the first state to protect musicians and other artists against AI. https://artscanvas.org/arts-culture/tennessee-just-became-the-first-state-to-protect-musicians-and-other-artists-against-ai

[78] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. arXiv:1910.06711 [eess.AS]

[79] Eleven Labs. 2023. Eleven Labs: Text to Speech & AI Voice Generator. https://elevenlabs.io

[80] Bruno Latour. 2014. What Is the Style of Matters of Concern? In *The Lure of Whitehead*, Nicholas Gaskill and A. J. Nocek (Eds.). University of Minnesota Press, 92–126. https://doi.org/10.5749/minnesota/9780816679959.003.0004

[81] Bruno Latour and Peter Weibel. 2005. *Making Things Public.* MIT Press, Cambridge, Massachusetts. https://mitpress.mit.edu/9780262122795/making-things-public/

[82] Marc Leman. 2007. *Embodied music cognition and mediation technology.* MIT Press, Cambridge, Mass. OCLC: ocm74915535.

[83] Mingkuan Liu, Chi Zhang, Hua Xing, Chao Feng, Monchu Chen, Judith Bishop, and Grace Ngapo. 2021. Scalable Data Annotation Pipeline for High-Quality Large Speech Datasets Development. https://doi.org/10.48550/arXiv.2109.01164 arXiv:2109.01164 [cs, eess].

[84] Rui Liu, Berrak Sisman, and Haizhou Li. 2021. StrengthNet: Deep Learning-based Emotion Strength Assessment for Emotional Speech Synthesis. http://arxiv.org/abs/2110.03156 arXiv:2110.03156 [cs, eess].

[85] Ioannis E. Livieris, Emmanuel Pintelas, and Panagiotis Pintelas. 2019. Gender Recognition by Voice Using an Improved Self-Labeled Algorithm. *Machine Learning and Knowledge Extraction* 1, 1 (March 2019), 492–503. https://doi.org/10.3390/make1010030 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[86] Casey R. Lynch. 2022. Glitch epistemology and the question of (artificial) intelligence: Perceptions, encounters, subjectivities. *Dialogues in Human Geography* 12, 3 (Nov. 2022), 379–383. https://doi.org/10.1177/20438206221102952

[87] Kathryn Mannie. 2023. AI kidnapping scam copied teen girl's voice in $1M extortion attempt - National | Globalnews.ca. *Global News* (April 2023). https://globalnews.ca/news/9629883/ai-kidnapping-scam-teen-girl-voice-cloned-extortion-arizona-jennifer-destefano/

[88] Clovis McEvoy. [n. d.]. Vocal AI deepfakes of major artists are cropping up everywhere – should artists be worried? https://musictech.com/features/music-deepfakes-ai-drake-grimes-weeknd/

[89] Heidi A. McKee and James E. Porter. 2019. *Professional Communication and Network Interaction: A Rhetorical and Ethical Approach* (1st edition ed.). Routledge. https://www.routledge.com/Professional-Communication-and-Network-Interaction-A-Rhetorical-and-Ethical/McKee-Porter/p/book/9780367888398

[90] Anna McNay. 2015. The Body as Language: Women and Performance. https://www.studiointernational.com/the-body-as-language-women-and-performance-review-richard-saltoun

[91] Edwin F. McPherson. 2003. Voice Misappropriation In California - Bette Midler, Tom Waits, and Grandma Burger. https://mcpherson-llp.com/articles/voice-misappropriation-in-california-bette-midler-tom-waits-and-grandma-burger/

[92] Mohammad I. Merhi. 2023. An evaluation of the critical success factors impacting artificial intelligence implementation. *International Journal of Information Management* 69 (April 2023), 102545. https://doi.org/10.1016/j.ijinfomgt.2022.102545

[93] Mara Mills. 2012. Media and Prosthesis: The Vocoder, the Artificial Larynx, and the History of Signal Processing. *Qui Parle* 21, 1 (2012), 107–149. https://doi.org/10.5250/quiparle.21.1.0107 Publisher: Duke University Press.

[94] Keyaan Minhas. 2023. The-Rise-of-Voice-Assistants:-Changing-the-Way-We-Interact-with-Technology. https://medium.com/@keyaanminhas/the-rise-of-voice-assistants-changing-the-way-we-interact-with-technology-d613a1063929

[95] Mozilla. 2017. Mozilla Common Voice. https://commonvoice.mozilla.org/

[96] Madhumita Murgia and Anna Nicolaou. 2023. Google and Universal Music negotiate deal over AI 'deepfakes'. https://www.ft.com/content/6f022306-2f83-4da7-8066-51386e8fe63b

[97] MUTEK. [n. d.]. YONA featuring Ash Koosha. https://mutek.org/en/artists/yona-featuring-ash-koosha

[98] A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH.*

[99] Kristian Nymoen, Rolf Inge Godøy, Alexander Refsum Jensenius, and Jim Torresen. 2013. Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception* 10, 2 (June 2013), 9:1–9:22. https://doi.org/10.1145/2465780.2465783

[100] US Court of Appeals. 1988. Midler v. Ford Motor Co., 849 F.2d 460 (9th Cir. 1988). https://law.justia.com/cases/federal/appellate-courts/F2/849/460/37485/

[101] Linda O'Keefe and Nogueira. 2022. *The Body in Sound, Music and Performance: Studies in Audio and Sonic Arts* (1st ed.). Focal Press. https://www.routledge.com/The-Body-in-Sound-Music-and-Performance-Studies-in-Audio-and-Sonic-Arts/O-Keeffe-Nogueira/p/book/9780367441944

[102] Arlen Olsen. 2023. Voice Cloning Technology and its Legal Implications: An IP Law Perspective - Schmeiser, Olsen & Watts, LLP. https://iplawusa.com/voice-cloning-technology-and-its-legal-implications-an-ip-law-perspective/

[103] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. http://arxiv.org/abs/1609.03499 arXiv:1609.03499 [cs].

[104] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech ASR. http://www.openslr.org/12

[105] David Pantalony. 2009. Hermann von Helmholtz and the Sensations of Tone. In *Altered Sensations: Rudolph Koenig's Acoustical Workshop in Nineteenth-Century Paris*, David Pantalony (Ed.). Springer Netherlands, Dordrecht, 19–36. https://doi.org/10.1007/978-90-481-2816-7_2

[106] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019.* 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680 arXiv:1904.08779 [cs, eess, stat].

[107] Hannah Jane Parkinson. 2015. Hey, Siri! Meet the real people behind Apple's voice-activated assistant. *The Guardian* (Aug. 2015). https://www.theguardian.com/technology/2015/aug/12/siri-real-voices-apple-ios-assistant-jon-briggs-susan-bennett-karen-jacobsen

[108] Stella Paschalidou, Tuomas Eerola, and Martin Clayton. 2016. Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical Indian singing. In *Proceedings of the 3rd International Symposium on Movement and Computing (MOCO '16).* Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2948910.2948914

[109] Vesna Petresin. 2016. Extending Methods of Composition and Performance for Live Media Art Through Markerless Voice and Movement Interfaces: An Artist Perspective. In *Proceedings of the 3rd International Symposium on Movement and Computing (MOCO '16).* Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2948910.2948920

[110] Roberto Pieraccini. 2012. *The Voice in the Machine: Building Computers That Understand Speech.* MIT Press. Google-Books-ID: 3NjxCwAAQBAJ.

[111] Carlos Pinheiro. 2023. Voice Cloning Technology: The Benefits, Risks, and Ethical Considerations. https://medium.com/@ocarlospinheiro/voice-cloning-technology-the-benefits-risks-and-ethical-considerations-2e1f737a4722

[112] Valentina Pitardi and Hannah R. Marriott. 2021. Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing* 38, 4 (2021), 626–642. https://doi.org/10.1002/mar.21457 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21457.

[113] Rhett Power. [n. d.]. No Black Boxes: Keep Humans Involved In Artificial Intelligence. https://www.forbes.com/sites/rhettpower/2023/01/15/no-black-boxes-keep-humans-involved-in-artificial-intelligence/ Section: Entrepreneurs.

[114] Sathvik Prasad, Elijah Bouma-Sims, Athishay Kiran Mylappan, and Bradley Reaves. 2020. Who's calling? characterizing robocalls through audio and metadata analysis. In *Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20).* USENIX Association, USA, Article 23, 18 pages.

[115] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2018. WaveGlow: A Flow-based Generative Network for Speech Synthesis. https://doi.org/10.48550/arXiv.1811.00002 arXiv:1811.00002 [cs, eess, stat].

[116] Polina* Proutskova, John M. McBride, Yuto Ozaki, Gakuto Chiba, Yukun Li, Yu Zhaoxin, Wei Yue, Miranda Crowdus, Gabriel Zuckerberg, Olga Velichkina, Yulia Nikolaenko, Yannick Wey, Lawrence Shuster, Patrick E. Savage, Elizabeth Phillips, and Andrew Killick. 2023. The VocalNotes Dataset. In *Proceedings of the First MiniCon Conference.* 3. https://ismir2023program.ismir.net/lbd_354.html Conference Name: Ismir 2023 Hybrid Conference.

[117] Polina* Proutskova, John M. McBride, Yuto Ozaki, Gakuto Chiba, Yukun Li, Yu Zhaoxin, Wei Yue, Miranda Crowdus, Gabriel Zuckerberg, Olga Velichkina, Yulia Nikolaenko, Yannick Wey, Lawrence Shuster, Patrick E. Savage, Elizabeth Phillips, and Andrew Killick. 2023. VocalNotes Dataset Access Form. https://docs.google.com/forms/d/e/1FAIpQLSfWn7fh2pTUnrpwlURzwyCxrxeWDpdTQIq7unLKVE1td_KKsg/viewform

[118] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. http://arxiv.org/abs/1905.05879 arXiv:1905.05879 [cs, eess, stat].

[119] Jessica Ravitz. 2013. 'I'm the original voice of Siri' | CNN Business. https://www.cnn.com/2013/10/04/tech/mobile/bennett-siri-iphone-voice/index.html

[120] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. http://arxiv.org/abs/2006.04558 arXiv:2006.04558 [cs, eess].

[121] ResembleAI. 2023. ResembleAI: AI Voice Generator with Text to Speech and Speech to Speech. https://www.resemble.ai/

[122] rewire. 2019. Virtual singer Yona joins Ash Koosha live at Rewire 2019. https://www.rewirefestival.nl/artist/https:/www.rewirefestival.nl/artist/yona

[123] Robert Rosenberger and Peter P. C. C. Verbeek. 2015. A field guide to postphenomenology. *Postphenomenological Investigations: Essays on Human-Technology Relations* (2015), 9–41. https://research.utwente.nl/en/publications/a-field-guide-to-postphenomenology Publisher: Lexington Books.

[124] Jennifer E. Rothman. 2018. *The right of publicity: privacy reimagined for a public world.* Harvard University Press, Cambridge, Massachusetts.

[125] Legacy Russell. 2020. *Glitch Feminism.* Verso. https://www.penguinrandomhouse.com/books/646946/glitch-feminism-by-legacy-russell/

[126] M. Sano, H. Sumiyoshi, M. Shibata, and N. Yagi. 2005. Generating metadata from acoustic and speech data in live broadcasting. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 2. ii/1145–ii/1148 Vol. 2. https://doi.org/10.1109/ICASSP.2005.1415612 ISSN:

2379-190X.
[127] Vlad Savov. 2011. British voice of Siri only found out about it when he heard himself on TV. https://www.theverge.com/2011/11/10/2551519/british-voice-of-siri-only-found-out-about-it-when-he-heard-himself
[128] Pierre Schaeffer. 1966. *Traité des objets musicaux , Pierre Sc...* Éditions du Seuil., Paris, France. https://www.seuil.com/ouvrage/traite-des-objets-musicaux-pierre-schaeffer/9782020026086
[129] Simon Schreibelmayr and Martina Mara. 2022. Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology* 13 (2022). https://www.frontiersin.org/articles/10.3389/fpsyg.2022.787499
[130] Hardik Shah. 2023. Exploring the Pros and Cons of AI Voice Cloning. https://medium.com/@shahhardik2905/exploring-the-pros-and-cons-of-ai-voice-cloning-f4bb15514284
[131] Maxine Sheets-Johnstone. 2011. *The Primacy of Movement*. John Benjamins Publishing. Google-Books-ID: 2EDgXzWMfuwC.
[132] Maxine Sheets-Johnstone. 2015. *The Corporeal Turn: An Interdisciplinary Reader*. Andrews UK Limited. Google-Books-ID: RXPZCgAAQBAJ.
[133] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:1712.05884 [cs.CL]
[134] Siegert and Ohnemus. 2015. A new Dataset of Telephone-Based Human-Human Call-Center Interaction with Emotional Evaluation. In *Proc. of the 1st International Symposion on Companion Technology (ISCT 2015)*. Ulm, Germany, 143–148.
[135] Matt Simon. 2019. The Uncanny Valley Nobody's Talking About: Eerie Robot Voices. *Wired* (2019). https://www.wired.com/story/uncanny-valley-robot-voices/ Section: tags.
[136] Denis Smalley. 1997. Spectromorphology: explaining sound-shapes. *Organised Sound* 2, 2 (Aug. 1997), 107–126. https://doi.org/10.1017/S1355771897009059
[137] Alexis B. Smith. 2019. Resounding in the Human Body as the 'True Sanskrit' of Nature: Reading Sound Figures in Novalis' The Novices of Sais. *The Journal of Somaesthetics* 5, 2 (Dec. 2019). https://doi.org/10.5278/ojs.jos.v5i2.3344 Number: 2.
[138] soerenab. 2024. AudioMNIST. https://github.com/soerenab/AudioMNIST original-date: 2018-06-29T16:31:21Z.
[139] Jae Yung Song, Anne Pycha, and Tessa Culleton. 2022. Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition. *Frontiers in Communication* 7 (2022). https://www.frontiersin.org/articles/10.3389/fcomm.2022.995475
[140] SpeechBrain. [n. d.]. SpeechBrain: A PyTorch Speech Toolkit. https://speechbrain.github.io/
[141] SpeechColab. 2024. GigaSpeech. https://github.com/SpeechColab/GigaSpeech original-date: 2021-03-03T06:36:25Z.
[142] Johan Sundberg. 1989. *Science of the Singing Voice*. Northern Illinois University Press, Dekalb, Ill.
[143] Eric E. Surbano. 2023. VAVA, Thai pop's first AI artist, has dropped her first single. https://www.lifestyleasia.com/bk/tech/vava-ai-artist/
[144] Katherine Thomson-Jones and Shelby Moser. 2021. The Philosophy of Digital Art. In *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, N/A.
[145] The Strait Times. 2023. China's court hears nation's first AI voice rights case. *The Straits Times* (Dec. 2023). https://www.straitstimes.com/asia/east-asia/china-s-court-hears-nation-s-first-ai-voice-rights-case
[146] Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2019. Exploring Transfer Learning for Low Resource Emotional TTS. http://arxiv.org/abs/1901.04276 arXiv:1901.04276 [cs, eess].
[147] Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2020. Laughter Synthesis: Combining Seq2seq modeling with Transfer Learning. http://arxiv.org/abs/2008.09483 arXiv:2008.09483 [cs, eess].
[148] Kai Tuuri and Tuomas Eerola. 2012. Formulating a Revised Taxonomy for Modes of Listening. *Journal of New Music Research* 41, 2 (June 2012), 137–152. https://doi.org/10.1080/09298215.2011.614951 Publisher: Routledge _eprint: https://doi.org/10.1080/09298215.2011.614951.
[149] Peter-Paul Verbeek. 2008. Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences* 7, 3 (Sept. 2008), 387–395. https://doi.org/10.1007/s11097-008-9099-x
[150] Pranshu Verma. 2023. They thought loved ones were calling for help. It was an AI scam. *Washington Post* (March 2023). https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/
[151] M.M. Wanderley and P. Depalle. 2004. Gestural Control of Sound Synthesis. *Proc. IEEE* 92, 4 (April 2004), 632–644. https://doi.org/10.1109/JPROC.2004.825882
[152] Marcelo M Wanderley, Bradley W Vines, Neil Middleton, Cory McKay, and Wesley Hatch. 2005. The Musical Significance of Clarinetists' Ancillary Gestures: An Exploration of the Field. *Journal of New Music Research* 34, 1 (March 2005), 97–113. https://doi.org/10.1080/09298210500124208
[153] Vivian Wang and Dali Wang. 2021. The Impact of the Increasing Popularity of Digital Art on the Current Job Market for Artists. *Art and Design Review* 9, 3 (June 2021), 242–253. https://doi.org/10.4236/adr.2021.93019 Number: 3 Publisher: Scientific Research Publishing.
[154] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. http://arxiv.org/abs/1703.10135 arXiv:1703.10135 [cs].
[155] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. http://arxiv.org/abs/1804.03209 arXiv:1804.03209 [cs].
[156] Matt Warman. 2011. The voice behind Siri breaks his silence. https://www.telegraph.co.uk/technology/apple/8879705/The-voice-behind-Siri-breaks-his-silence.html
[157] Angela Watercutter. 2023. Hollywood Actors Strike Ends With a Deal That Will Impact AI and Streaming for Decades. *Wired* (2023). https://www.wired.com/story/hollywood-actors-strike-ends-ai-streaming/ Section: tags.
[158] Oskar M. Wiklund. 2023. Unveiling the Future: The Power of Voice in AI Interactions. https://www.multiply.co/multiply-blog/unveiling-the-future-the-power-of-voice-in-ai-interactions
[159] Chloe Xiang. 2022. Scientists Increasingly Can't Explain How AI Works. https://www.vice.com/en/article/y3pezm/scientists-increasingly-cant-explain-how-ai-works
[160] Ryuichi Yamamoto, Reo Yoneyama, and Tomoki Toda. 2023. NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit. http://arxiv.org/abs/2210.15987 arXiv:2210.15987 [cs, eess].
[161] Cao Yin. 2023. Chinese court hears nation's first AI voice rights case. https://asianews.network/chinese-court-hears-nations-first-ai-voice-rights-case/
[162] Miriama Young. 2016. *Singing the Body Electric: The Human Voice and Sound Technology*. Routledge, London. https://doi.org/10.4324/9781315609164
[163] Eileen Yu. 2023. China mulls legality of AI-generated voice used in audiobooks. *ZDNET* (Dec. 2023). https://www.zdnet.com/article/china-mulls-legality-of-ai-generated-voice-used-in-audiobooks/
[164] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv:1904.02882 [cs.SD]
[165] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. 2019. Sequence-to-Sequence Acoustic Modeling for Voice Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 3 (March 2019), 631–644. https://doi.org/10.1109/TASLP.2019.2892235 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[166] Mingyang Zhang, Yi Zhou, Li Zhao, and Haizhou Li. 2021. Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1290–1302. https://doi.org/10.1109/TASLP.2021.3066047 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[167] Shikun Zhang, Omid Jafari, and Parth Nagarkar. 2021. A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data. https://doi.org/10.48550/arXiv.2109.03784 arXiv:2109.03784 [cs].
[168] Zhaoyan Zhang. 2016. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America* 140, 4 (Oct. 2016), 2614–2635. https://doi.org/10.1121/1.4964509