Thesis for The Degree of Licentiate of Engineering

Musical AI Voices: Facts, Concerns and Experimental Musical Practices with AI Voice Tools

Kelsey Cotton

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2024

Musical AI Voices: Facts, Concerns and Experimental Musical Practices with AI Voice Tools

Kelsey Cotton

Department of Computer Science and Engineering
Division of Data Science and AI
AI in Computational Arts, Music, and Games Research Group
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2024.

for Angela and Jameson ❤️

Musical AI Voices: Facts, Concerns and Experimental Musical Practices with AI Voice Tools

Kelsey Cotton

Department of Computer Science and Engineering

Abstract

As early adopters of technologies, artists are uniquely positioned to probe and explore the potentials and problematics that these tools afford and pose to their artistic practice. This is especially applicable to AI voice tools, which present unique challenges to the "value" of human voice and voice data; to our understandings of human vocality in the age of deep learning innovation; and how we work with AI voices in an experimental musical practice. This thesis investigates the intersections of AI voice tools and experimental musical practices, examining how artists critically engage with–and are implicated by–AI voice tools that clone, parse and synthesise human voice and speech.

Within this topic, we address the technological facts and the societal implications and concerns of generative AI voice tools–encompassing deep learning voice models and speech toolkits–which offer unique artistic potentials to work with a feasibly unending palette of generated vocal sounds. The fidelity of these tools are continually advancing, and are increasingly being utilised within artistic practices.

Our motivations in this thesis are grounded in an exploration of AI voice tools' potentials and problematics, which span a range of both pragmatic technology facts and societal concerns, and navigate interdisciplinary boundaries. The nature and pace of deep learning developments is such that we presently lack methods of visibilising and critiquing the potentials and problematics of AI voice tools used in musical contexts. Further, we are in a unfolding period of investigating the wider socio-technical implications that are constructed through these potentials and problematics within musical practice. This thesis therefore explores the following research questions: *What methodologies assist in visibilising the multifaceted potentials and/or problematics of AI voice tools used in musical contexts?*; *What wider socio-technical implications occur through these potentials and problematics within musical practice?* and; *What shifts occur within an experimental musical practice when critically exploring AI voice and speech tools?*

Seeking to answer these questions, this thesis develops methodologies for–and chronicles–interdisciplinary practical and theoretical engagements with AI voice and speech models. Further, it discusses and formulates practical methods for feminist and interventionist analysis, and the development of–and performance with–AI voice and speech tools in experimental musical settings. Questions on how to visibilise the potentials and problematics of AI voice tools in musical contexts are foregrounded, alongside explorations into the shifts that occur within experimental musical practices when engaging with such tools.

This thesis contributes with: 1) a novel analytical method for the critical analysis of artworks featuring musical AI voice tools; 2) the establishment of interdisciplinary perspectives as integral to understanding the *use*, cultures-of-use and implications of voice and speech AI tools in musical applications; 3) a Research-through-Design account of developing and performing with a series of AI voice models in a live music performance; and 4) a research stance on experimental musical practices as enabling the formation of new understandings of human and AI-mediated human vocality.

Keywords

musical AI, voice, AI vocality, musical AI performance

Abbreviations

- AI: Artificial Intelligence
- ML: Machine Learning
- STEM: science, technology, engineering and mathematics
- STS: Science and Technology Studies
- ASR: Automatic Speech Recognition
- CTC: Connectionist Temporal Classification
- LAS: Listen-Attend-Spell
- TTS: Text-to-Speech Synthesis
- CNN: Convolutional Neural Networks
- RNN: Recurrent Neural Networks
- LSTM: Long Short Term Memory
- DNN: Deep Neural Networks
- GAN: Generative Adversarial Networks
- LPC: Linear Predictive Coding
- HCI: Human-Computer Interaction

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] Kıvanç Tatar**[1], Petter Ericson**, **Kelsey Cotton****, Paola Torres Núñez Del Prado, Roser Batlle-Roca, Beatriz Cabrero-Daniel, Sara Ljungblad, Georgios Diapoulis, Jabbar Hussain, *A Shift in Artistic Practices through Artificial Intelligence*
Leonardo 2024; 57 (3): 293–297. doi:https://doi.org/10.1162/leon_a_02523.

[**Paper II**] **Kelsey Cotton** and Kıvanç Tatar, *Caring Trouble and Music AI: Considerations Towards a Feminist Musical AI*
In Proceedings of the AI Music Creativity Conference. Sussex, 2023. Retrieved from https://aimc2023.pubpub.org/pub/zwjy371l.

[**Paper III**] **Kelsey Cotton**, Katja de Vries, Kıvanç Tatar, *Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies Movement Computing 2024).* https://doi.org/10.1145/3658852.3659065

[**Paper IV**] **Kelsey Cotton**, Kıvanç Tatar, *Sounding out extra-normal AI voice: Non-normative musical engagements with normative AI voice and speech technologies*
In Proceedings of the AI Music Creativity Conference. Oxford, London, September 2024. Retrieved from https://aimc2024.pubpub.org/pub/extranormal-aivoice.

[**Paper V**] **Kelsey Cotton**, *glemöhnic*
In Proceedings of the AI Music Creativity Conference. Oxford, London, September 2024. Retrieved from https://aimc2024.pubpub.org/pub/glemonic.

[1] The ** denotes equal first authorship

Acknowledgement

“Acknowledgement” doesn’t quite seem like a big or meaningful enough word to fit all of my feelings, gratitude and sheer wonder to have had so many talented, inspiring and kind people so graciously share so much of their passion, wisdom and care with me. You know who you are 😉. You have all made this more gentle ground for me to tread upon. I am endlessly thankful.

An especially large thank you to my supervisor Dr Kıvanç Tatar, for your belief in this work, enthusiasm and for patiently listening to me rant about voice and singing.

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

Contents

Part I

The Compilation

Chapter 1

Introduction

Continual technological progress within the domain of deep learning has had significant implications upon sound and audio generation techniques, which have revolutionised how we synthesise, clone, parse and generate human voices. Using deep learning models, we can clone a voice rapidly [1], in high fidelity [2], with more human-like [3] and emotive prose [4] and affect. We can further generate non-textually-centred voice material such as coughs and laughter [5], providing rich artistic opportunities to use this material in musical compositions, soundscapes and performances.

The usage of non-textually-centred voice material (coughs, burps, laughter, etc) in musical contexts has a rich tradition spanning across the 20th and 21st centuries. Pioneering vocalists of the mid 20th century—such as Cathy Berberian [6], Joan La Barbara [7] and Meredith Monk (to name a few examples) [8]—explored the usage of alternative, experimental and unconventional sounds within their own vocal practices, compositions, and performances (see Section 2.1.2).[1]

Collectively, alternative and exploratory approaches to human vocalisation have been described as "extended vocal techniques" [9]. A more recent conceptualisation of experimental vocal practices and vocalisation instead use the more inclusive term "extra-normal" [10]. In light of the generative turn that AI voice tools have initiated within musical contexts, we may begin to understand these tools as similarly affording a new sonic space for extra-normal human-AI voice mediations (see Section 2.1.2 and 2.4.2). The exploratory research processes of artists working within experimental and electroacoustic music can be seen as modes of adopting, appropriating and accomplice-ing novel technologies to produce and distribute art. With regards to musical AI, this can be clearly seen in the work of vocalists and artists such as Holly

[1] Suggested listening:
Early Immersive Music of Joan La Barbara;
Tras un retazo del olvido, Movement 4 - Demian Rudel Rey, sung by **Kelsey Cotton**;
Visage - Luciano Berio, sung by Cathy Berberian; and a personal favourite:
Voice is the original instrument- Joan La Barbara, also sung by the composer.

Herndon, Claire Boucher[2], Ashkan Kooshanejadin[3] and Jennifer Walshe, who respectively work with or develop voice models in their artistic work [11]–[14]. [4]

By virtue of the technology itself, and how it in turn shapes and reshapes musical vocal practices, the domain of musical AI voice lies at an interdisciplinary junction. At this junction are research domains such as Musical Composition and Performance, Art and Technology, Science and Technology studies (STS), Human-Computer Interaction (HCI), Sound Studies, and Machine Learning (ML) and Artificial Intelligence (AI). This thesis approaches this interdisciplinary junction through engaging with these various domains, informed by a research stance that seeks to determine the technological facts of AI voice tools, and the concerns they establish within musical vocal contexts (see Sections 2.1.2, 2.4.2 and 2.6).

This thesis addresses a range of models and architectures which have been utilised–or are useful–in assisting artistic exploration of synthetic voice potentials. These predominantly encompass automatic speech recognition (ASR) models [15]–[17], text-to-speech synthesis (TTS) and neural-based vocoder approaches [18]–[20]. Of particular note are the "all-in-one" speech toolkit SpeechBrain [21], OpenAI's Whisper model [22], Mozilla's DeepSpeech [23], Meta's Wav2vec [24] and Kaldi [25]. In the domain of text-to-speech synthesis, we consider cloning and generation architectures including NVIDIA's comprehensive speech toolkit NeMo [26], TorToiSe [27], MatchaTTS [28], CoquiTTS [29] and Tacotron [30](See Section 2.4).

Broadly speaking, artistic implementation of these model architectures and toolkits affords opportunities to construct novel timbral and textural sounds which might otherwise be physiologically dangerous, physically taxing, or outside of the technical expertise or voice type of even the most experienced and technically proficient vocalists (see Section 2.1.2). Further, these models enable the recognition and transcription of sounds and vocal gestures, which may provoke reflection as to how experimental, textural and timbral vocalisations are mediated or "understood" by speech recognition models. As put by Kockelkoren in [31],

> *Contrary to generally held views that ascribe to the artist an almost innate autonomous position over and against cultural processes in which new technologies are adopted, artists actually tend to be accomplices to these social developments. Artists have always played a leading role in appropriating the new ways of looking and hearing that innovative technologies have offered. Technologies that open up new forms of experience have been domesticated and made manageable by artists. It is not an entirely innocent process.*

The new ways of hearing and singing with synthetic voices that we are presently learning and discovering in our usage of deep learning models are

[2] Known by the moniker Grimes

[3] Publishing under the name Ash Koosha

[4] Suggested listening:
URSONATE%24 by Kurt Schwitters, Jennifer Walshe;
C - by Yona

indeed forged in un-innocent ways, offering us as much artistic “potentials” as they do “problematics”. In this thesis, these terms refer specifically to applications of AI voice and speech tools which afford novel opportunities for areas of artistic exploration (“potentials”); or alternatively, to practical concerns or areas of consideration which necessitate further critique and scrutiny “problematics” (see Section 2.8).[5]

What then are the artistic potentials and problematics of deep learning voice and speech technologies? These potentials and problematics span the pragmatic (how we construct, curate and share our voice as a dataset) to the legal [32] (how do we protect human voice rights whilst nurturing AI voice model development) to the creative (how do we make art with a synthetic voice). So what are these “potentials” and “problematics” of deep learning voice and speech models, especially within the particular context of musical vocal practices?

First, the “potentials”. Advancements to deep learning architectures have enabled notable improvements to the reconstruction and synthesis of hi-fidelity human (or human-like) voice with relatively low computation time [33] from textual, symbolic or acoustic features such as mel-spectrograms [34]. These potentials afford the creative potential of generating and utilising hi-fidelity generated audio material from samples of one’s own voice; re-constructing one’s childhood voice—or a loved one’s—from historical recordings; or generating a projection of one’s voice to sound more nasal, resonant, or emotive. Such potentials offer novel approaches in re-forming an understanding of how we may use these powerful tools in creative ways within a musical context. Largely, this can be viewed as becoming equipped with novel compositional and performative possibilities to explore voice-centric composition and performance; making new voices to sing in harmony (or dissonance) with; and the opportunity to build and develop alternative versions of our **own** voices. We begin to uncover the potentials of what these models may offer to question and trouble our own understandings of what a human vocality is—or could become—when considered through the lens of an AI-mediated vocality. Perspectives from Human-Computer Interaction (HCI) on self-knowledge, and auto-ethnographic practices as research method [35]–[39] assert the importance of investigating how one relates to, shapes and is shaped in turn by an engagement with technology (see Section 2.9). Approached with this viewpoint in mind, the potentials of AI voice are especially applicable within my own vocal practice, which centres on an exploratory and auto-ethnographic experimentation with differing vocal timbres and textures; challenging the physiological limits of how I produce vocalised sound; and the integration of various technologies to sample, process, amplify and manipulate my voice.

However, with potentials of AI voice and speech tools, comes the “problematics”. Within the general domain of AI voice and speech technology development, the primary problematics identified in the literature encompass a number of areas of concern, which encompass data; governance and deployment; and analysis-related concerns. With respects to data, the concerns within the

[5] It should be noted, that these two terms are non-independent and can exist simultaneously.

field are centred on current practices of data curation, which largely involve non-consensual scraping of audio material [40], [41] from sites such as YouTube [42], Freesound [43] and SoundCloud [44]. This treatment of voice material as an object or resource to be **used** or consumed without the knowledge or consent of the vocalists reinforces a treatment of data as a material that is kept intentionally separate from the bodies who are implicated by web scraping practices [45]. This subsequently raises concerns regarding linguistic and accent diverse characteristics within current openly accessible voice datasets. Greater diversity in the form of wider linguistic variance, a wider spectrum of gender-diverse vocal characteristics and greater accent diversity is crucial for the improving the flexibility of AI models trained for speech recognition [46], [47].

Like my own voice, these technologies are dynamically and iteratively changing, and necessitate interdisciplinary approaches and methodologies to critique the impact (or the functional consequences) of AI model design upon vocal practices and the artworks produced through these practices. In this thesis, we focus primarily on the artistic usage of—and the associated sociocultural implications of—deep learning voice models implemented within experimental vocal practices. We argue that examining voice deep learning models in their usage in musical settings–and especially via an experimental musical practice that actively and critically engages and confronts technology–affords a more nuanced starting point for evaluating the role these technologies can (and do) play in the art-making process, and the wider correspondences they invite with society and culture. However, the technological developments within deep learning applications in musical and artistic contexts has accelerated at a pace that research into how we analyse and critique the implications of using these systems in our art has yet to catch-up. This necessitated the formulation of practical analytical methods for analysing *how* an artist has elected to integrate a voice model—or indeed a musical AI-agent—into their artwork. Further, there was a dearth of research into the implications—the problematics—of vocal AI integrations have upon wider music industry economic structures, human and synthetic voice legal protections and understandings of human- and AI-mediated vocality. This thesis, and the papers included in Part II therefore enter into this void.

1.1 Research Questions

The over-arching goal of this thesis is to highlight the multifaceted potentials and problematics in applications of AI tools for voice and speech, whilst simultaneously grounding these artistic dimensions within a critical and interdisciplinary analytical discourse. Taking this overarching aim as a central point of departure, this thesis therefore engages with first-person and Research-through-Design methods; artistic practices; and a wide spectra of interdisciplinary theory and perspectives from sound and music studies, voice studies and experimental voice practices, copyright law, science and technology studies, post-phenomenology and Human-Computer Interaction. Guided by these aims, methods and theoretical grounding, the following research questions are foregrounded:

> **RQ1:** What methodologies assist in visibilising the multifaceted potentials and/or problematics of AI voice tools used in musical contexts?
>
> **RQ2:** What wider socio-technical implications occur through these potentials and problematics within musical practice?
>
> **RQ3:** What shifts occur within an experimental musical practice when critically exploring AI voice and speech tools?

As has been previously established earlier in the Introduction, these research questions are intended towards: 1) the formalisation of a novel method(s) for the critical analysis of artworks featuring musical AI agents; 2) the development and implementation of interdisciplinary perspectives as integral to understanding the *use* and cultures-of-use implications of voice and speech AI tools; 3) exploring the development of and performance with AI voice models in a live music performance; and 4) leveraging experimental musical practices to enable the formation of new understandings of human and AI-mediated human vocality.

1.2 Motivations

This research is motivated by the concern for the growing research dearth into the how the potentials and problematics of AI voice tools implicate not just musical voice cultures and practices, but wider users and implicated parties who speak to, or are listened to, by implementations of these architectures and models. The mainstream utilisation of these AI voice tools informs how they are functionally shaped, which further implicates artistic decision- and music-making with these same tools when they are "appropriat[ed]...domesticated and made manageable by artists" [31]. To provide greater context as to the scope of this area, voice-based interactions have seen a marked incline across sectors such as advertising and marketing [48]–[50], the voice acting industry [51], public healthcare [52], education [53], automotive [54]–[60] and customer services [61]. Accompanying the development and utilisation of AI voice tools across these sectors are datasets, cultures of use and operational values. When appropriated into musical contexts these tools have implications upon how artists work, how artists navigate their vocal identity in the age of deepfakes, and upon their art itself. We have already begun to witness these implications within the music industry, as web-scrapping cultures in gathering voice data, the increasing ease of deepfake voice generation, the lack of clear "cultures of acceptable use" and legislative protections around voice and vocal identity pose formidable challenges [32], [62]–[73] to vocalists.

These challenges are confronted in a musical vocal practice, provoking questions around how "valuable" a voice (and voice data) is to others and how its use by others may be protected or regulated once published; how minority or marginalised accents are/are not understood by ASR models; how certain tessituras, harmonics and vowel placements in the voice generate particular gender-ed outputs in multilingual TTS models. Speaking personally, as a vocalist who has devoted more than half of my life to learn how to effectively work *with* my body and *with* my voice to produce a wide spectra of textural and timbral sound, the potentials and problematics of AI voice and speech offer a rich new sonic and performative terrain to explore.

1.3 Research Contributions

This thesis is a compilation, containing a series of papers which form the main contributions of my research work. These appended papers can be found in Part II of this text. In this section I provide a holistic overview of my research contributions and outline my positioning of these contributions with respect to my methodological framing and engagements with the respective AI voice and speech tools. The contributions in Part I can be summarised in the table below.

Chapter	Contributions	Venue
1.	- Introduction to AI for Voice and Speech - Research Questions - Motivations - Contributions	-
2.	- Overview of vocal anatomy and physiology - Overview of pertinent music and sound studies discourses - Overview of AI Tools for Voice and Speech Synthesis - Overview of Science and Technology Studies - Overview of Post-phenomenology	-
3. Appended Papers		
3.1	- Discussion around the issue of consent, the exploitation of artists' labor, and the incorporation of artists' intellectual property into training datasets or as "prompts" (**RQ2**) - Propositions for new needs regarding novel copyright structures to promote artistic data sovereignty; new measures for traceability, regulation, and accountability (**RQ2**)	Leonardo Journal, Volume 57, Issue 3 (June 2024)

Continued on next page

Chapter	Contributions	Venue
3.2	- Outlined how content generation capabilities have initiated a cultural shift regarding the capital “value” of Art (**RQ2**) - Development of Caring Trouble analytical method (**RQ1**) - Application of Caring Trouble analytical method to a case study of musical AI artwork featuring a voice model (**RQ1**) - Established future research directions within musical AI as encompassing further research into decentralisation processes for data and identity; management and reclamation; and legacy (**RQ2, RQ3**)	AI and Music Creativity 2023
3.3	- Established the research dearth on how current practices and cultures-of-use of AI voice tools implicitly and explicitly implicate human stakeholders with regards to data collection and processing protocols (**RQ2**) - Established connections between current functional requirements of voice data and historical perspectives from electroacoustic composition and practice, and listening and sound studies (**RQ1**) - Established the sociocultural consequences of AI voice development which does not proactively critique its use of data as separate to its original sound-Body (**RQ2**) - Introduced several propositions for returning an understanding of the importance of Body to AI voice and speech research (**RQ1**)	Movement Computing Conference 2024

Continued on next page

Chapter	**Contributions**	**Venue**
3.4	- An applied Research-through-Design process to building and developing a pipeline for the parsing of non-text-based improvised vocal gestures through a connected ASR and TTS model to generate sample banks of cloned vocal gestures for use in a live music coding environment. (**RQ1, RQ2, RQ3**) - Methodological contribution centred in the significance of utilising exploratory-, experimental- and artistic practice centred methods to trouble and destabilise normatively trained AI voice and speech models. (**RQ3**)	AI and Music Creativity 2024
3.5	- Utilisation (**RQ3**) of the pipeline presented in **Paper IV**. This has been presented as a demonstration paper, and in performances at (respectively):	- AI and Music Creativity 2024 (UK) - halffloor 2024 (SE) - Koami Art Festival 2024 (SE) - Göteborg Fringe Festival 2024 (SE)[6]

[6] see page 10 of the program booklet

Collectively, this thesis makes the following contributions:

- The thesis demonstrates that the advancements of AI tools and architectures implemented within musical contexts and practices have wider sociocultural connections and communicate implicit values around artistic labour; data; the cultural value of labour; and shifting values around the cultural value of artworks (**RQ2**).

  Addressed in Paper I

- We demonstrate that the utilisation of AI tools in the creation and distribution of music have implications for, and are in correspondence with, wider sociocultural discussions around musical copyright; data management; artistic legacies; processes of art-making; and power structures regulating the music industry (**RQ2**).

  Addressed in Paper II

- The thesis demonstrates that the incorporation of interdisciplinary perspectives and theories is fruitful in identifying and investigating the implications of design choices made with respects to datasets; architecture choices and AI model implementation in musical contexts (**RQ1**).

  Addressed in Paper II

- We present an interdisciplinary analytical method for critiquing the wider sociocultural connections between the usage of AI models in musical artworks, and the implications of these connections with respect to legacy; data management; and power (**RQ1, RQ2**).

  Addressed in Paper II

- This thesis examines the diminishing role of Body in relation to voice data as a potential functional consequence of mid 20th Century perspectives on listening and sound studies (**RQ2**).

  Addressed in Paper III

- We address the legal implications of regulating AI voice tools, with a particular focus upon differing legislation around voice rights and protections (**RQ2**).

  Addressed in **Paper III**

- We introduce novel approaches to re-establishing Body with voice-data in fields such as multimedia, experimental music and popular music, in which greater legal protections and auxiliary technologies are implemented (**RQ2**).

  Addressed in **Paper III**

- The thesis demonstrates the generative potentials of utilising auto-ethnographic and Research-through-Design methods to guide non-normative musical engagements with automatic speech recognition and text-to-speech synthesis models (**RQ3**).

  Addressed in **Paper IV** and **Paper V**

- We present how artistic and non-normative engagements with normative AI voice and speech tools afford creative potentials for extranormal AI vocality (**RQ3**).

  Addressed in **Paper IV** and **Paper V**

Individual Paper Contributions

The following table, based on the Contributor Roles Taxonomy (CRediT),[7] outlines my individual contributions within each of the appended papers presented in this thesis. I was a contributor for all papers, and was the sole contributor in Papers IV and V.

In Paper I, I was involved with *Conceptualisation*, *Data Curation* and *Formal Analysis.* I also contributed in *Writing- Original Draft* and *Writing - Review and Editing.* In Paper II, I led the *Conceptualisation*, *Data Curation* and *Formal Analysis*, with input and feedback from my co-author. I carried out the *Investigation* and co-developed the *Methodology* with my co-author. I wrote most of the initial draft for *Writing- Original Draft*, while my co-author added specific paragraphs and provided feedback on the research arguments. Additionally, I contributed to *Writing - Review and Editing* for the entire publication. In Paper III, I contributed to *Conceptualisation* and the *Formal Analysis.* I conducted the *Investigation*, and developed the *Methodology* in collaboration with my co-authors. I wrote the majority of the publication for *Writing- Original Draft*, with my co-authors contributing individual paragraphs and feedback on the formulation of the research arguments. My writing contributions also encompassed *Writing - Review and Editing* for the entire publication. In Paper IV, I carried out the *Conceptualisation*, *Data Curation*, *Formal Analysis*, *Investigation* and *Methodology.* I also implemented the *Software*, *Visualisation*, and was the sole author for *Writing- Original Draft* and *Writing - Review and Editing.* In Paper V, I was the sole contributor in *Conceptualisation*, *Data Curation*, *Formal Analysis*, *Investigation*, *Methodology* and *Project Administration.* Paper V implements the software framework developed in Paper IV. I was the sole author for *Writing- Original Draft* and *Writing - Review and Editing.*

Additional explanations for the respective CRediT contributions for each paper appended in Part II can be found in Chapter 3.

[7] https://credit.niso.org

Table 1.2: Individual Contributions to each paper using the Contributor Roles Taxonomy (CRediT)

CRediT Contribution	**Paper I**	**Paper II**	**Paper III**	**Paper IV**	**Paper V**
Conceptualization	✓	✓	✓	✓	✓
Data Curation	✓	✓		✓	✓
Formal Analysis	✓	✓	✓	✓	✓
Funding Acquisition					
Investigation		✓	✓	✓	✓
Methodology		✓	✓	✓	✓
Project Administration		✓	✓	✓	✓
Resources					
Software				✓	✓
Supervision					
Validation					
Visualization		✓	✓	✓	✓
Writing – Original Draft	✓	✓	✓	✓	✓
Writing – Review and Editing	✓	✓	✓	✓	✓

1.4 Thesis overview

In this section we provide an overview of the structure of this thesis.

In Chapter 1 we provided a general introduction to the research area and further established the prominent motivations, research questions and contributions of this thesis. Chapter 2 serves to provide an overview, an introduction to the various theoretical and analytical perspectives that are pertinent to and largely inform this thesis' methodology and the motivation for the research work chronicled in the appended papers. Specifically, it provides an overview of research domains such as science and technology studies (STS) (see Section 2.5); sound and voice studies (see Section 2.1; and post-phenomenology (see Section 2.7).

Chapter 3 outlines a detailed summary of the primary contributions made in each of the appended papers upon which this thesis is based. Within chapter 3, we discuss five research papers. **Paper I** addresses the shifting paradigms with an artistic practices that are a consequence of ubiquitous usage and increasing access of AI tools. **Paper II** discusses the generative potential of applying feminist and interdisciplinary theoretical methods to the analysis of musical AI artworks. Further, **Paper II** demonstrates the application of these feminist and interdisciplinary methods to formulate the Caring Trouble Analytical Framework. This framework is then applied to a single case study of a voice-AI artwork as a proof of concept. **Paper III** discusses the rising legal implications and concerns of AI voice and speech tools in relation to the visibility of voice Bodies. This paper further looks to adjacent media industries and cultures to examine what new steps have been made with respects to better integrating the Body in relation to voice data and applications of AI voice and speech tools. **Paper IV** chronicles a Research-through-Design informed engagement with a series of pre-trained AI voice tools deployed within an experimental vocal practice and live music coding setting. **Paper V** is an artwork that deploys the system described previously in **Paper IV**. All of the afore-listed papers can be found in Part II of this thesis.

Chapter 4 summarises and discusses the primary contributions that are derived from this thesis, and further establishes future research directions for this work. In Chapter 5 we address the ethics dimensions pertinent to this compilation thesis.

Chapter 2

Background

In this section, we provide a general overview to the theoretical perspectives integral to this thesis, and the appended papers. We first provide a general overview of the physiological processes involved in voice production (often referred to in this thesis as 'vocalised sound', 'vocalisation', 'phonation', and 'voice').[1] We then provide a brief look into perspectives on sound and music, to provide a contextual basis for how extra-normal voice relates to timbrally organised sound. Following this is a brief historical overview of approaches to human-like voice production using hardware and software methods. We then summarise the current[2] approaches to synthesising, parsing and generating human-like voice using machine learning methods. This is accompanied by a general overview into theories and perspectives from Science and Technology Studies and Human-Computer Interaction, which provide assistive terms and perspectives that this thesis engages with. We then discuss correlations between humans and technology in Section 2.8, to provide a theoretical grounding for the complex and bidirectional entanglements between human voice and AI voice tools. We close this section with an overview of Research-through-Design, to give insight into a methodological approach that foregrounds knowledge-making through the act(ivities) of making, exploring and creating.

2.1 Body, Voice and Sound

In this section, we provide an overview of human vocal physiology; phenomenological perspectives on voice and pertinent views on sound.

[1] We note here, as this thesis is concerned with human voice and speech the usage of the term 'voice' and 'vocalised sound' is therefore solely referencing human-produced sonic material.

[2] At the time of writing

2.1.1 The Body

The notion of the Body[3] occupies much discourse within sound, movement, phenomenology, and voice discourse [74]–[83]. As this thesis is concerned with the interaction and transference between a human (vocalising) Body and the use of AI models to produce human-like vocalised sound, the usage of the term 'Body' is informed by the following premises.

The first premise- that the physicality of a Body and its unique physiological changes (both intentional and unintentional) actively contributes to and shapes the timbral qualities of the sound it is producing [82]. To give a concrete example grounded in human vocal physiology- the (intentional) manipulation of soft tissue structures and musculature in the oral cavity contributes to the timbral and resonant qualities of a vocalised sound. Similarly, (unintentional) aspects such as salivary drainage may produce salivary 'clicks' during phonation. The second premise- that the movement, diffusion or spatialisation of sound also constitutes a **moving sound-Body** [75]. To give an example, the usage of speakers to amplify a human voice throughout a space in effect constructs a sound-Body with the capacity to 'move' its perceived position. Our usage of the term 'Body' thus reflects an understanding of:

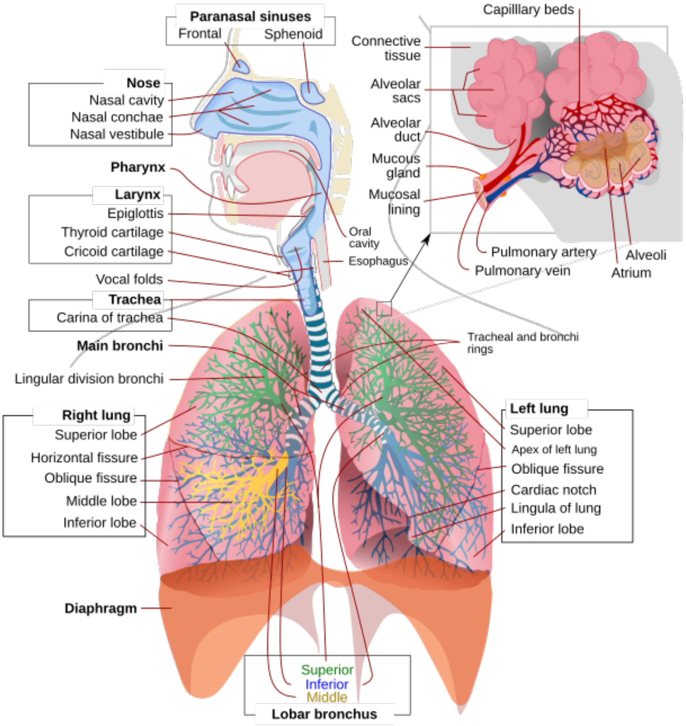
1. as reflecting *Body-as-source* of sound;
2. as reflecting *Body-as-origin* of a sound;
3. that the Body itself fundamentally shapes the movement of the *sound-Body* that it produces; and
4. that Body itself also serves as a *medium* through which an origin-Body may be experienced by another.

2.1.2 The Voice

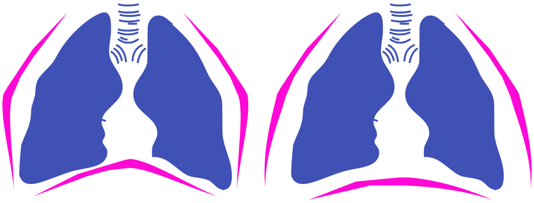
In this section, we provide a brief overview of human vocal physiology. The purpose of this section is to establish the depth of bodily knowledge involved in the production of vocalised sound. The human voice is an instrument which makes use of three physiological subsystems: the respiratory system, the phonatory system and the resonance system [84].

The respiratory subsystem regulates the movement of air into/out of the body, a process known as respiration. The primary organs and structures involved in respiration are the lungs, ribcage, intercostal muscles, diaphragm, the trachea, mouth and nose (see Figure 2.1(a)). During inspiration, air travels into the body via the nose/mouth and through the trachea into the lungs. The lungs inflate with air. During expansion, the ribcage expands laterally, causing the diaphragm to flatten (see Figure 2.1(b)). During exhalation, air leaves the lungs, travelling back through the trachea and out of the body via the nose/mouth. As this occurs, the lungs deflate and in turn the ribcage partially collapses as the intercostal muscles relax. The diaphragm relaxes also.

[3] The capitalisation is intentional

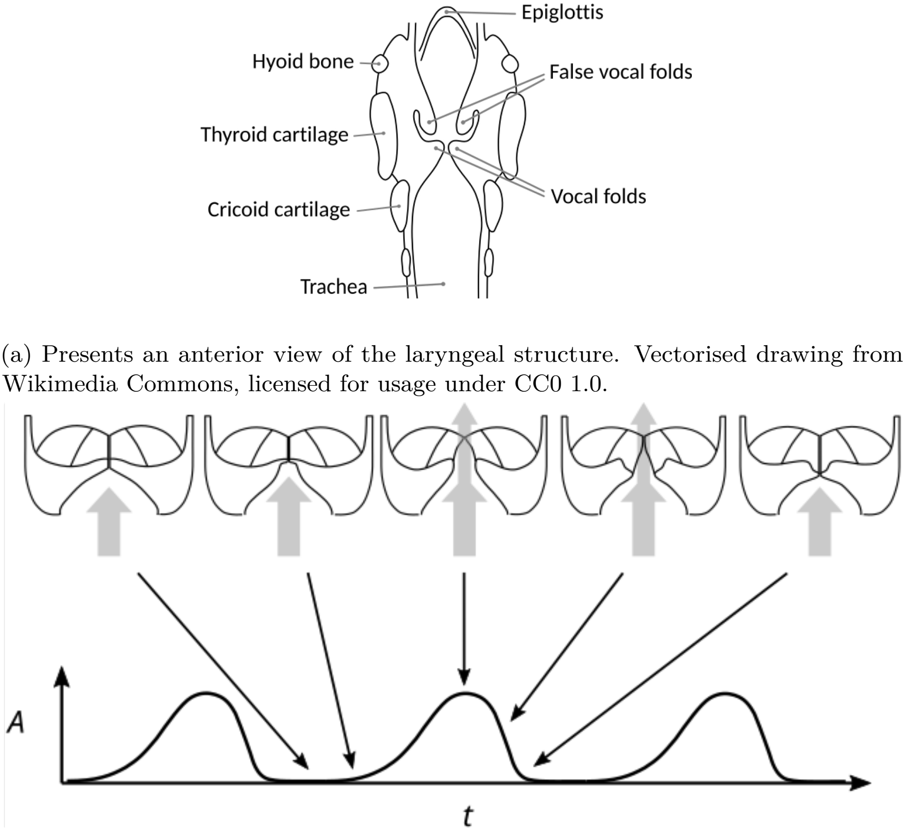


(a) Illustrated views of the entire respiratory subsystem. Figure depicts an anterior view of the oral and nasal cavity, trachea, lung structure and diaphragm. Image from Wikimedia Commons, image within public domain.



(b) Stylised views of the lungs and ribcage. Figure depicts a view of the lungs, trachea (in blue), ribcage and diaphragm (in magenta) at rest (left most image) and during inspiration (right most image). Base image from SVG SILH, magenta coloured regions have been added by Kelsey Cotton, licensed for usage under CC0 1.0.

When air passes into and out of the Body, it passes via the phonatory subsystem, which encompasses the larynx. The larynx is a cylindrical assemblage of cartilage, muscle and soft tissues and contains the vocal folds. The human larynx has 2 main vocal folds, which are protected by the vestibular fold (sometimes known as false folds, see Figure 2.2(a)). During speech and singing, the main vocal folds rapidly open and close (see Figure 2.2(b)). This action disrupts the flow of air and creates a buzzing sound [85]. This buzz is then shaped and amplified by the resonance subsystem, which is comprised of the vocal tract, oral cavity, sinus cavity and other bone structures in the face (see Figure 2.1(a)). Further manipulation of the soft tissue structures in this region (such as the tongue and lips, see Figure 2.3) affects the timbre or 'colour' of the produced vocal sound.



(a) Presents an anterior view of the laryngeal structure. Vectorised drawing from Wikimedia Commons, licensed for usage under CC0 1.0.

(b) Depicts an anterior view of the wave action of the vocal folds, correlated with changes in airflow (denoted by A) over time (denoted by t). Vectorised drawing from Wikimedia Commons, licensed for usage under CC BY-SA 4.0.

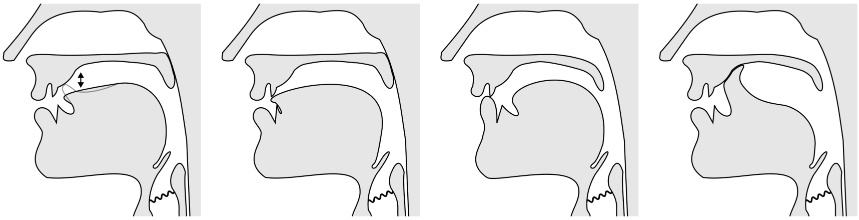


Figure 2.3: Illustrated views of the resonatory subsystem, depicting four midsagittal schematics of the following articulations (left to right): voiced alveolar trill [r]; voiced dental nasal [n̪]; voiced bilabial nasal [m]; and voiced retroflex nasal [ɳ]. Images from Wikimedia Commons, licensed for use under CC BY-SA 4.0.

Apart from the intricate knowledge required to manipulate the various vocal subsystems to produce sound, singers also further manipulate their vocal physiology to make 'on-the-fly' adjustments to dynamically respond to environmental acoustics, and to introduce more expressive affects to their singing [79], [81]. As an example, physiological adjustments in response to environmental acoustic may include over articulated diction of percussive sounds (such as labial and fricative consonants) in a highly reverberant acoustic space. The introduction of more expressive actions may encompass physical exaggeration of breathing or mouth movement, as well as gestural actions [86]. As discussed by Godøy,

> "*...[a] straightforward definition of gesture is that it is a movement of part of the Body, for example a hand or the head, to express an idea or meaning. In the context of musical performance, gestures are movements made by performers to control the musical instrument when playing a melodic figure, to coordinate actions among musicians (conducting gestures), or to impress an audience (for example, moving the head during a solo performance)*" [86].

Extra-normal Voice

In the previous paragraphs we outlined the conventional physiological production of vocalised sound as it occurs in a healthy voice and able body. There are of course, many ways of producing vocalised sound which do not adhere to the aforementioned processes. As an example, technological mediation of vibrations that might otherwise occur in the larynx, but do not due to vocal damage, trauma or medical interventions [87], [88] might utilise external devices such as an electrolarynx or other vibro-tactile stimulant [89]–[91]. Concerning vocalised sound production in a healthy voice and able body, other methods for producing sound are commonly described as "extended" vocal techniques [92], [93]. Though commonly utilised to describe alterations to one's playing of an instrument to produce different timbres, the usage of "extended technique" in relation to vocal practices has drawn criticism for its preoccupation of a definitive, singular vocal technique that is the universally accepted baseline

[10], [92]. Vocal technique and approaches to vocalised sound production are informed by the cultural and musical aesthetics they originate and are utilised within [94]–[98]. A contrasting term—“extra-normal”—has been suggested in [10], which instead foregrounds a wide spectrum of physiological changes to define “extra-normal” voice.

2.2 Sound and Music

Defining what constitutes “sound” is deceptively simple. There are an abundance of definitions for “sound”. A broad definition proposed by Truax in [99] establishes sound as “*Any vibration in the air or other medium, some types of which are able to cause a sensation of hearing.*” Truax’s definition is intentionally expansive, and accounts for a range of different perspectives from acoustics, psychoacoustics and soundscape studies as well as more holistic views of hearing. For the purposes of this thesis, it is necessary to establish that we are discussing sound within the range of **able** human hearing. This thesis further frames discussions of music within Varése’s proposition of “organised sound” [100]. That is, music can be understood as encompassing differing organisational systems for structuring of sound events either with respects to the temporal (i.e. events occurring and organised in time) or the timbral/textural (i.e. events that are organised by their timbral/textural qualities).

Of additional importance to this thesis are perspectives from sound studies, specifically post-Schaefferian thinking on sound. Before we can understand *Post*-Schaefferian thought on sound, it is important to first establish how this diverges from Schaefferian perspectives. Pierre Schaeffer was a French engineer-musician-composer renowned for his development of *musique concrète*[4]. Musique concrète was as approach to creating electroacoustic music utilising pre-recorded sounds, natural environments, synthesizers and digital signal processing. Accompanying the development of musique concrète, Schaeffer also developed a philosophy of listening- *écoute réduite*[5][101]. In [101], Schaeffer builds upon Edmund Husserl’s phenomenological notion of reduction [102], applying it in a sound-context in order to develop an understanding of the different ways that we may listen to and understand sound. From this phenomenological reduction of listening, Schaeffer proposed a series of modes that sought to separate a sound object from notation (how it is written); to separate a sound object from how the sound itself is created; and separates the sound object from the listener’s contextual connections that they make in the act of listening [103], [104].

Post-Schaefferian thought rejects the separation of a sound object from *reflective*, *denotative*, and *experiential* dimensions [105]. Sound is instead understood as “contain[ing] references to its actual or perceived origins, to some external association, or to some combination of the two” [83]. This is further argued in [83] that “Sound, in other words, is a sign that indicates something beyond itself and as such can never exist as a pure abstraction.”

[4] Translation: concrete music

[5] Translation: reduced listening

2.3 Making Synthetic Voice

In this section we provide an overview of historical developments in synthesising human(-like) voice using mechanical and digital methods.

The synthesis of human voice and speech as a research domain has an extensive history. Indeed, the first attempts to make or synthesise human voice utilised physical machines [106]. One such documented attempt to synthesise human voice was using a physical synthesizer model (see See Figure 2.4) designed by Christian Gottlieb Kratzenstein in 1730 as part of a contest [107]. Kratzenstein's machine used a configuration of resonant tools and a reed to produced five vowels, and differed substantially from the physiology of the human vocal subsystems.

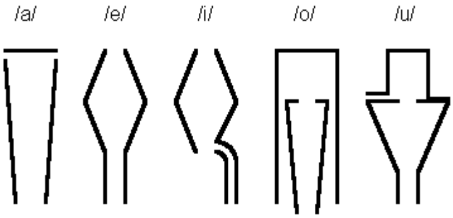


Figure 2.4: Kratzenstein's resonant tubes for the vowels: [a], [e], [i], [o] and [u]. From [108].

Subsequent developments in mechanical synthesis, such as Wolfgang von Kempelen's "Speaking Machine" (see Figure 2.5), saw the simulation of human vocal physiology via hand-operated machinery to produce individual sounds [109].

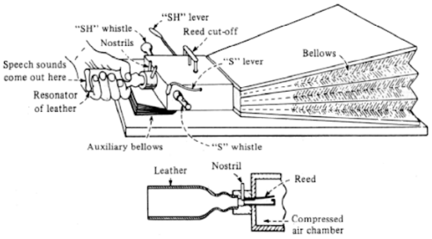


Figure 2.5: A sketch of von Kempelen's Speaking Machine. From [110]

Later attempts to synthesise human voice took the form of electrical synthesizers. The first successful attempt was achieved with the keyboard-operated synthesizer Voder (Voice Operating DEmonstratoR), designed by Homer Dudley and an engineering team at Bell Laboratories in 1937 [111]. Further research throughout the 20th Century prioritised the digital modelling of human vocal tracts [112]–[117]. Later, the development of Linear Predictive Coding (LPC) in the 1960s digitally encoded speech by mathematically modelling the vocal tract [118]–[120]. Formant synthesis also emerged during the 60s, and was centred on replicating the formants of the vocal tract which characterise differing speech sounds [121], [122]. One particularly notable implementation of formant synthesis was the DECtalk system, an assistive speech technology [123]. Speech synthesis developments during the early 21st Century utilised parametric synthesis models, based specifically on Hidden Markov Models (HMM) [32]. This marked a notable advancement in more "natural" sounding speech by modelling speech as a sequence of probabilistic states. They were, however, somewhat lacking in "expressive-ness".

The exploration of means to synthesise more 'natural" and "expressive" speech reached a turning point when deep learning techniques were applied to this task, enabling the direct modelling of speech waveforms. The development of DeepMind's WaveNet in 2016 marked a breakthrough in using deep convolutional neural networks to generate speech waveforms [20]. Such an approach afforded a novel level of natural and expressive synthetic speech material, and affirmed the generative potentials of applying neural-based methods to generate human voice and speech.

2.4 AI for Voice and Speech

There are many different definitions for what precisely AI *is*. In the most expansive sense, it is described as a field of research within computer science in which methods, software and approaches are developed to enable machines to perceive, learn from and act within an environment to achieve predefined goals [124]. Within the context of human voice and speech we are therefore referencing architectural approaches which facilitate the cloning, generation and synthesis of human-like voiced sound. These approaches[6] are discussed in the coming section.

2.4.1 Architectures, algorithms, and approaches for Voice

Presently, deep learning approaches for voice and speech can largely be described as encompassing several key areas: automatic speech recognition (ASR) [125], [126]; text-to-speech synthesis (TTS) [4], [127], [128]; voice conversion and style transfer [5], [129]–[134]; and audio and speech generation [20], [135], [136].

Automatic Speech Recognition is the processing of human speech by machine learning architectures into text. This is achieved by the transformation of audio waveforms into token sequences; the extraction of speech features from audio material; and the mapping of extracted features to tokens sequences that can be constructed into 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. Currently, deep neural networks (DNN) are utilised to achieve more natural sounding speech. A TTS pipeline typically has two stages. The input

[6] Current at the time of writing

text is converted to mel-spectrogram form. The mel-spectrogram is then converted to an audio waveform. WaveRNN [127], Tacotron [135], and WaveGlow [137] are popular networks for synthesising audio from mel-spectrograms. Current platforms for text-to-speech synthesis include subscription-based options such as Lyrebird; Resemble.AI; and Eleven Labs as well as free and open-source toolkits such as SpeechBrain, NeMo and SpeechT5 [138].

Voice conversion and style transfer utilises feature extraction from source and target audio material as input for deep neural networks [18], [131], [139], [140]. Deep neural networks learn the transformation of these speech features, which are then synthesised in the target voice using a vocoder.

Other applications for the transformation of speech and voice using AI include style transfer with Transformers [141], and voice conversion.

2.4.2 AI Voice Applications

AI tools for voice have contributed substantially to improving digital accessibility [142], [143] and revolutionising the field of speech therapy [144]–[146]. The development of voice-based AI agents—such as Apple's Siri and Amazon's Alexa—has been particularly notable [147]–[150]. Within the context of musical applications such as performance and composition, such development similarly affords the widening of creative potentials in the usage of this technology [151]–[157]. In parallel, the non-neutrality of technology in general [158], [159]—and especially AI—[160] prompts critical evaluation as to *what is further invited in* when these tools are utilised in an artistic praxis, or within an artwork. This concern is similarly applicable to applications concerning the synthesis, cloning or generation of our human voices: what do we invite in when we use AI tools for cloning and generating voice and speech?

2.5 Science and Technology Studies

Science and Technology Studies (STS) is an interdisciplinary research field that examines the wider contextual frame surrounding science, technology, engineering and mathematics (STEM). Specifically, it considers technological developments and advancements as informed by, and as a consequence of, wider historical, cultural and societal situatedness [161]. Within this field, researchers probe the societal and cultural impact of technology development and discovery, the ethics and the implications connected with certain advancements within STEM, and the complex relations between 'scientific' knowledge and societal value. Within the larger view of STS, there are several domains of study which foreground particular perspectives and critical views of technology in society.

2.5.1 Feminist Science and Technology Studies

One such foregrounded perspective is feminist science and technology studies (feminist STS). Like more general STS, it is an interdisciplinary research field. At its core, feminist STS explores the intersection of gender, science and technology [162]–[165]. The field is primarily concerned with developing a nuanced understanding of how factors such as gender, socio-economic status, socio-cultural connections, ethnicity, race, age, ability shape and are shaped by the influence of science and technology. The emergence of feminist STS was a response against the exclusion of marginalised groups within society [36], [166]–[173]. As a research field, it has strong connections with the aims and perspectives of third-wave feminism [164], [168], [174]–[178].

2.6 Matters of Fact and Matters of Concern

An additional perspective within the larger domain of STS is the framing of objective knowledge in relation to its wider and subjective contextual embeddedness. One notable concept proposed by proposed by French sociologist Bruno Latour is the complementary and connected notions of *matters of fact* and *matters of concern* [179]. These concepts were proposed as a mechanism for critiquing the systems that form knowledge, and further developing a contextual awareness of what this knowledge represents within the world [180]. Latour defines a *matter of concern* thus:

> A matter of concern is what happens to a matter of fact when you add to it its whole scenography, much like you would do by shifting your attention from the stage to the whole machinery of a theatre [181].

Latour elaborates on how shifting our focus is instrumental to our understanding of the relationships that unfold between actors/stakeholders and this scenography:

> Instead of simply being there, matters of fact begin to look different, to render a different sound, they start to move in all directions, they overflow their boundaries, they include a complete set of new actors, they reveal the fragile envelopes in which they are housed [181].

2.6.1 Matters of Care

de la Bellacasa proposes an addition to Latour's matters of fact and concern, considering care as further mediation between the becoming of matters of fact and concern. That is, intentionally considering how particular histories and values are presented or implicitly embedded within systems, knowledge and artefacts is an enactment of care.

> ”[C]are not only in thinking the processes of construction of socio-technical assemblages but as an ethico-political attitude in the everyday doing of knowledge practices.” [182]

That is, de la Bellacasa positions 'matters of care' as a further layer of consideration in the formation of knowledge (Latour's matters of fact) within a particular context (Latour's matters of concern) as ethico-politically framed.

2.7 Post-phenomenology

Post-phenomenology is a critical qualitative study of how artefacts and humans interact with and shape each other [183]–[185]. The “post” in post-phenomenology emerged in opposition to the prevailing phenomenological thought, which was more concerned with direct experience and an understanding of consciousness as intentional [186]. Post-phenomenology diverges from this, and instead considers human experience as occurring in relation to wider interconnected societal and cultural factors. As a research lens, post-phenomenology has been extensively applied to research into human interactions with and through technology [183]–[185], [187].

Post-phenomenologists propose different types of human-technology interactions. As suggested by Ihde, the nature of human-technology relations could be understood as encompassing four different forms: hermeneutic, alterity, background and embodiment. [185]. These four relations were later critiqued for the singular focus on “analyzing human-technology configurations in which technologies are *used*” [187]. But what happens when technologies *use* or are in *intimate correspondence* with the body? Verbeek argues for greater consideration into how expanding upon Ihde’s four human-technology relations inadvertently overlook technologies which blur the boundaries between human flesh and technological body [183]. As a research method, post-phenomenology has been incredibly generative within the field of Human-Computer Interaction (HCI) [188]–[190]. It has further been used as a reflective analytical tool [191]–[195].

2.8 Connecting Human-Technology Relations to Musical AI agents

With the understanding of human-technology relations as constituting complex, often bidirectional entanglements, we further establish connections between post-phenomenology and literature on multi-agent systems [196] and musical agents [157]. This is framed as ‘a simplified world of reality’ in which agent(s), object(s) and environment are co-located (either physically or digitally) and interact. Within the environment, agent(s) have a capacity and capability to perform various actions. These actions are influenced by the agents’ affordances, which constrain or enable action potentials that are possible. Similarly, object(s) have affordances that influence how actions are performed on, to, at, with, or around them. Actions carried out by agent(s) have **enacted** agency. We understand **enacted** agency as actions that are taken directly, or those that are available for us to take within a world or environment [197]–[199]. These are the decisions made by either a human musician or musical AI-agent. The environment in which the action takes place has **situated** agency, which is acted upon all agents and objects within the environment. **Situated** agency encompasses the available decisions that are possible and can be made *by us* or *for us* (the human actor), and/or *by* or *for* a technology (the AI actor) [200]–[205].

When we consider this construction of reality unfolding within the composition of an artwork or the “world” of a music performance, we look beyond an objective classification of how a technology is positioned in relation to a human (and vice-versa). Indeed, this positioning occurs via **enacted** and **situated** agencies: in that the agents (human and non-human) are bidirectionally impacting upon each others’ sonic contributions, and in turn establishing constraints which inform the decision-making possibilities within the “world” of the performance. These constraints are established either pre- or during performance, and alternatively by both the human and non-human agents. Their subsequent interactions within the “world” of the performance may unfold—dynamically—in differing ways.

2.9 Research-through-Design

Research-through-Design (RtD) is a methodological approach towards the formation of knowledge through the act or activity of designing, or making [206], [207]. Practically speaking, RtD emerged in the 1960s and 70s, in which the field of design progressively adapted research methodologies from the sciences and humanities. As a term, RtD came to prominence in the 1990s, introduced by Christopher Frayling in [208]. Frayling critiques the dichotomy between research as “scientific” or “un-scientific”. He contextualises the nature of knowledge creation occurring within arts and design, and establishes three different modes of research: research **for** art and design, research **into** art and design, and research **through** art and design. Presently, RtD is a prominent paradigm within fields such as human-computer interaction (HCI) [209]–[212]. Within the particular setting of musical performance and composition, mediated by AI voice and speech tools, RtD-informed approaches may yield insights into how the potentials and problematics are revealed—or implicate—artistic working processes.

Chapter 3

Summary of Included Papers

3.1 A Shift in Artistic Practices through Artificial Intelligence

This paper was an output resulting from a WASP-HS Community Reference Meeting on AI and Media, in which scholars and artistic practitioners gathered to discuss the (then) current state of the art and current state of concerns regarding AI development and its implications for cultural industries such as music, visual art and games. In the specific round-table discussion from which this paper was born, we discussed how various AI tools transform the production (and consumption) or visual arts, music and wider media and the further implications of this upon data; artistic labour; capital value; copyright; intellectual property; and the construction of new power structures.

In this position paper, we centre the discussion on the paradigm shift that we see as a consequence of increased access to AI models and tools within cultural industries. We examine specific examples of current (at the time of writing) technologies for image and video generation; approaches for symbolic music generation; approaches for audio signal processing; approaches for mapping symbolic music to audio materials; and cross-domain applications. These examples are contextualised in relation to a number of emergent concerns that we argue are a consequence of AI tools instigating a paradigm shift when it comes to the capital value of artistic labour.

3.1.1 Contributions

This journal paper contributes with an outline of the current landscape and cultures-of-use of AI tools involved in the generation of artistic content and media. It gives an overview of how large, commercial AI models have, and are, contributing to a revolution in content generation capabilities. Primarily, the discussion in this paper focuses on the ethical implications of infinite content generation with respects to consent, data and copyright. The paper further contributes with a survey of the widening power discrepancies between the conglomerates that build, scrape data and train these models, and the artistic communities that are negatively impacted by the non-consensual usage of their artistic material as training data. The paper further contributes with a stance towards prioritising informed consent from artists and content creators whose material is used as training data. Further, the paper calls for more pedagogical investment for both artists and technologists to facilitate and develop greater cross disciplinary exchange.

3.1.2 Author Contributions, using CRediT

The *Conceptualisation* of this journal paper was done by Kıvanç Tatar, Petter Ericson and **Kelsey Cotton**, who share equal lead authorship. The content of the paper was informed by discussions between Kıvanç, Petter, **Kelsey** and each of the co-authors (Paola Torres Núñez Del Prado, Roser Batlle-Roca, Beatriz Cabrero-Daniel, Sara Ljungblad, Georgios Diapoulis and Jabbar Hussain). These discussion occurred within a WASP HS Community Reference Meeting on A Shift in Culture Through AI, which was jointly organised by Kıvanç, Petter and **Kelsey**. The *Data Curation* process for this paper involved the collation of meeting minutes, discussion points, shared media and links - which *Kelsey* contributed to. The *Formal Analysis* of this data was conducted equally Kıvanç, Petter and **Kelsey**. The *Writing - Original Draft* stage of this paper was collaboratively done by Kıvanç, Petter and **Kelsey**, with subsequent *Writing - Review and Editing* collectively done by Kıvanç, Petter and **Kelsey**. The final preparation of the manuscript for publication was done by Kıvanç Tatar. Please see column 2 of Table 1.2 for an overview of Kelsey's individual contributions to **Paper I**.

3.2 Caring Trouble and Musical AI: Considerations towards a Feminist Musical AI

In this paper we establish that the increasing implementation of AI tools and technologies within artistic working practices has wider implications. This paper thus examines a range of perspectives and methodologies from AI-ethics, science and technology studies, feminism and Human-Computer Interaction to formulate an interdisciplinary method for examining the objective 'facts´ of a technology's use in relation to its wider situated societal and cultural concerns. The resultant methodology formulated in the paper encompasses an examination of feminist interventionist principles; a STS informed knowledge map of the technology across multiple dimensions; and a survey of data principles prevalent within the work. We deploy this methodology in the paper, conducting a case study of a musical AI artwork by new-media artist Holly Herndon: *Holly+*. Emergent from our analysis of *Holly+* are several "matters of concerns" which demonstrate the generative potential of applying critical and feminist lenses in the examination of AI models implemented in artistic work.

3.2.1 Contributions

This paper offer several contributions to the research domain of musical AI and AI tools implemented in musical settings. The main contribution is the development and implementation of an interdisciplinary analytical method for critiquing musical AI artworks. This analytical method—Caring Trouble—draws from a cross-section of interdisciplinary methods and perspectives from AI data ethics, feminist STS and Human-Computer Interaction. In this paper we further establish the importance of utilising interdisciplinary methods to formulate analytical frameworks for critically examining the multifaceted connections between the functional and the wider sociocultural impact of AI models utilised in musical work. We further contribute with a demonstration of applying the Caring Trouble Analytical method to a case study of a neural network for musical style transfer - the artwork *Holly+* by new-media artist Holly Herndon. We close the paper with several propositions for future research in this domain. The first of which is to further examine what industry power structures might be de-stabilised as a consequence of further usage of AI in the production and distribution of music. We further contribute with considerations for how artistic usage of AI tools may provoke novel developments with respect to the preservation of artistic legacy in posthumous contexts. Finally, we establish the urgent need for prioritising people-, process- or data-oriented principles pertaining to artist's usage of AI voice tools.

3.2.2 Author Contributions, using CRediT

The *Conceptualisation* of this conference paper was done by **Kelsey Cotton** and Kıvanç Tatar. The content of the paper was informed by discussions between **Kelsey** and Kıvanç. The *Data Curation* for this paper involved the sourcing of open access material, articles and quotes from the artist Holly Herndon, whose artwork *Holly+* was a case study in the paper- which **Kelsey** contributed to. The *Formal Analysis* of this data, and the *Investigation* was carried out by **Kelsey**. The *Project Administration* for this work encompassed the research activity planning and execution, and was carried out by **Kelsey** and Kıvanç. The *Visualization* component of this paper involved the formulation of the deployed Caring Trouble Method into an info-graphic, done by **Kelsey** with feedback from co-authors. The *Writing - Original Draft* was written by **Kelsey**, with feedback from co-authors. Subsequent *Writing - Review and Editing* was done by **Kelsey**. The final preparation of the manuscript for publication was done by **Kelsey**. Please see column 3 of Table 1.2 for an overview of Kelsey's individual contributions to **Paper II**.

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

In this paper we examine the rising phenomena of invisibilised voice Bodies as a functional consequence of current working practices with voice data. We speculate on the origins of this trajectory as connected with mid-20th century perspectives on sound and listening. Specifically, we critique the influence of Schaefferian listening perspectives to uncover possible sympathies between reductionist listening practices and the functional reduction of voice Bodies into single modality data for use in model training. From this connection, we examine some real-world consequences of invisibilised voice Bodies- surveying a series of notable legal cases in which AI voice and speech companies are defendants. We speculate on potential progressive directions which afford the ethical and mindful development of AI voice and speech tools whilst simultaneously emphasising the significance of the Body. To do so, we examine developments in adjacent media industries, looking to current approaches from film, television and radio as well as experimental and electronic music. Within this brief survey, we note the changes to legal definitions concerning voice rights; the use of particular mediating technologies; and the significance of artist-led actions in establishing the future landscape of legal structures concerning AI voice and speech.

3.3.1 Contributions

This paper offers a series of contributions. The first contribution is a contextualisation of the Body's absence with respects to voice data as in correspondence with perspectives of listening and sound from electroacoustic music compositional theory and practice [83], [101], [113]. Further, we contribute with the implementation of conceptual tools from general and feminist science and technology studies (STS) as mechanisms for evaluating relationships between a technology and the implications of it's use [171], [182], [213], [214]. We further contribute with a general overview of how voice and voice-Bodies are implicated by the functional needs of technology development. Finally, we contribute with a series of propositions for increasing the visibility of the Body during engagements with AI speech and voice tools, derived from our survey of progress in adjacent media fields and developments in legislation.

3.3.2 Author Contributions, using CRediT

The *Conceptualisation* of this conference paper was done by **Kelsey Cotton** and Kıvanç Tatar. Katja de Vries was invited by Kıvanç and **Kelsey** to contribute domain knowledge relevant to legal structures around copyright and international differences of voice rights and legal protections. The content of the paper was informed by discussions between **Kelsey**, Katja and Kıvanç. These discussions served as the material for the *Investigation*, which was conducted by **Kelsey**. The *Investigation* of legal structures around AI voice, historical perspectives on sound and listening, and the investigation of current approaches to voice data, was similarly carried out by **Kelsey** with suggestions from co-authors. The *Methodology* and emergent research perspective of the paper was formulated by **Kelsey**, with feedback from co-authors. The *Project Administration* for this work constituted the planning and execution of the *Investigation* and *Methodology* development, and was similarly carried out by **Kelsey** with feedback from co-authors. With regards to *Visualization*, **Kelsey** prepared and presented the published work at the Movement Computing Conference in Utrecht. The *Writing- Original Draft* was conducted by **Kelsey**, with feedback and suggestions from co-authors. Subsequent *Writing - Review and Editing* was done by **Kelsey**. The final preparation of the manuscript for publication was done by **Kelsey**. Please see column 4 of Table 1.2 for an overview of Kelsey's individual contributions to **Paper III**.

3.4 Sounding out extra-normal AI voice: Non-normative musical engagements with normative AI voice and speech technologies

This article present a Research-through-Design methodology which explores unconventional approaches to utilising AI voice tools, focusing on the specific use of speech synthesis and text-to-speech models like SpeechBrain, Whisper, and CoquiTTS. A connected model pipeline is developed to be utilised within an experimental musical practice, using live recorded and historical voice data to synthesise extra-normal voice sounds and vocal gestures. As a result of the extra-normal voice sounds, the utilised models are provoked to behave in unexpected ways - the outcomes of which are actively integrated into the author's vocal practice and used as material input for a live coded musical performance. By engaging with AI tools through auto-ethnography, the paper emphasises the generative potential of non-normative AI model usage in artistic practices. Such non-normative engagements are discussed in relation to sound poetry, framing voice-AI interactions as a form of digital *zaum*.

3.4.1 Contributions

This paper makes several significant contributions related to the utilisation of AI models for processing and synthesising human voice and speech. Firstly, the paper introduces a novel research perspective which emphasises the utilisation of non-normative modes of engagements with AI models developed for functional purposes. This is presented through the documentation of the Research-through-Design and auto-ethnographic process, which establishes an additional contribution. A further contribution is established in the paper's demonstration of non-normative engagements with normative AI voice and speech models as provoking new understandings and insights into human vocality when mediated by these models. Finally, the paper contributes with a case for exploring non-normative methods for provoking novel engagements with normative and functional models and architectures that are intended to process and handle human voice.

3.4.2 Author Contributions, using CRediT

The conceptualisation of this conference paper was done by **Kelsey Cotton**, with input from co-author Kıvanç Tatar. The *Methodology* of the paper was informed by discussions between **Kelsey** and Kıvanç on Research-through-Design and artistic research practices. The *Data Curation*–Kelsey's personal vocal sound library–was built, curated and is managed by **Kelsey**. The *Investigation* and *Formal Analysis* of the artistic engagements with the Whisper, SpeechBrain and CoquiTTS models was carried out by **Kelsey**. **Kelsey** also implemented the *Software* for parsing and generating audio samples using her sound library. The *Visualisation* component for this paper–encompassing the pipeline visualisation and research logs–was similarly carried out by **Kelsey**. **Kelsey** carried out the *Writing- Original Draft* and *Writing - Review and Editing*, with feedback and suggestions from co-authors. The final preparation of the manuscript for publication was done by **Kelsey**. Please see column 5 of Table 1.2 for an overview of Kelsey's individual contributions to **Paper IV**.

3.5 glemöhnic

The artwork "*glemöhnic*" explores the interaction between human vocality and AI by utilising SpeechBrain, Whisper, and CoquiTTS models to transcribe, synthesize, and clone live improvised and historical vocal gestures. Non-text and non-word based audio samples are used as input into automatic speech recognition models from SpeechBrain and Whisper to generate nonsensical and poetic text outputs. These text outputs are then used as input materials for the CoquiTTS text-to-speech synthesis model to generate sound banks of samples which are clones of the originally recorded sound. Both the original voice gesture recording and the synthesised clones are used as sonic material within a live music coding performance setup. Thematically, *glemöhnic* makes connections between sound art movements such as Dada and *zaum*, and using non-normative voice sounds as provocations to trigger unusual and novel outputs from normative and functional voice and speech architectures.

The system has been utilised in several performances at the Koami Art Festival 2024, the Gothenburg Fringe Festival, the 2024 AI Music Creativity conference, and has been featured by the Stockholm sound art organisation halffloor.

3.5.1 Contributions

The main contribution of this paper is in the harnessing of the different ASR models—SpeechBrain and Whisper—to reveal differing behaviours in how these respective models handle non-text based and non-word based speech. The differences in how the respective ASR models process voice gestures are further emphasised by the author's use of an additional CoquiTTS text-to-speech synthesis model which utilises the SpeechBrain/Whisper textual output as input material. The text is used as a prompt material, and paired with the original recorded voice gesture to generate a warped clone of the vocalist's sound. This cloned material is then used as a sample bank within a live music coding performance. An additional contribution are the various artworks that have bee produced using this pipeline.

3.5.2 Author Contributions, using CRediT

The *Conceptualisation* of this conference paper was done solely by **Kelsey Cotton**. The content of the paper was informed by **Kelsey's** first-person documentation of their research-through-design inspired engagement with specific ASR and TTS models within their artistic practice - forming the basis of the *Data Curation*, *Formal Analysis*, *Investigation* and *Methodology* contributions. The *Software* was implemented by **Kelsey**. **Kelsey** carried out the *Visualization*–the public performance and presentation–of the paper at (chronologically) Stockholm sound art organisation halffloor, the Koami Art Festival (2024), the Gothenburg Fringe Festival (2024) and the 2024 AI Music Creativity conference. **Kelsey** was sole author for *Writing- Original Draft*, *Writing - Review and Editing* and prepared the manuscript for publication. Please see column 6 of Table 1.2 for an overview of Kelsey's contributions to **Paper V**.

Photo by Angelika Thiele, used with permission

Chapter 4

Discussion and Future Work

In this thesis, we have demonstrated that the domain of Musical AI has far-reaching implications upon both the sociocultural dimensions around artistic labour; the capital value of artistic data; and the production and distribution of musical artworks. The findings—documented across the five papers appended to this text—illustrate significant transformations in artistic practices facilitated by AI tools. These transformations have been examined through various interdisciplinary lenses. This discussion contextualises these findings within the broader landscape of AI and the Arts, examines their implications, acknowledges limitations, and offers recommendations for future research and practice.

We have discussed the sociocultural implications of the rising use and advancing developments of AI utilized in Arts and Media in **Paper I**. In this paper, we addressed the paradigm shift occurring within arts and cultural industries due to increasing technological development and ubiquitous use of AI tools in the creation, distribution, and consumption of media and Arts. We proposed that greater technological and artistic literacy might be beneficial for artists and technologists actively working with, and developing AI tools in these domains. We further addressed the tensions around current cultures around data scraping and non-consensual usage of intellectual property, as adversely affecting artistic minorities. We emphasised that approaches to these issues should encompass more explicit consent mechanisms, as well as improving technical and artistic literacy to ”bridge the gap” between artists and technologists. **Paper I** directly relates to RQ1, through our discussion on more qualitative approaches to considering and evaluating the potentials and problematics of AI voice tools in musical contexts.

In **Paper II**, we discussed the importance of developing analytical methods for both identifying and critiquing the implicit sociocultural values communicated when AI architectures are used in the creation and distribution of musical artworks. We formulated and proposed a novel analytical method—Caring Trouble—which is informed by interdisciplinary perspectives from data ethics, feminist STS and Human-Computer Interaction. A case study of Holly Herndon's AI-artwork, *Holly+*, was analysed using Caring Trouble. In doing so, we demonstrated the generative potential of applying critical and feminist lenses to examine AI models in artistic work. This paper highlighted the rich insights that can be gained from incorporating interdisciplinary research perspectives to develop novel methods for examining musical applications of AI. With respects to the research questions, the discussions in this paper addressed both RQ1 and RQ2. We addressed RQ1 in this paper by exploring how the application of qualitative and interdisciplinary methods assist in visibilising the multifaceted nature of AI tools and architectures used in musical artworks. We further address RQ2 in our discussion of how Herndon's usage of AI tools in *Holly+* further established wider socio-technical implications through Herndon's utilisation of blockchain technologies to introduce distributed ownership in the profits from usage of her voice model.

Paper III examined the invisibilisation of voice bodies in AI voice and speech tools, connecting this phenomenon to historical perspectives on sound and listening. Our discussion highlighted the overlap between Schaefferian reductionist listening approaches and the current functional treatment and usage of voice data as correlated with the difficulties in legislation around fair and consensual usage of voice data in training voice synthesis models. We proposed that the visibility of the Body in AI voice tools may be improved through incorporating practices from adjacent media fields. We highlighted the necessity for more holistic approaches to working with voice data as a sound Body in its own right. The insights from **Paper III** are closely tied to RQ3, in our investigation of experimental musical practices as contributing novel approaches to repositioning the Body in relation to AI synthesised and generated voice.

The exploration of non-normative engagements with AI voice tools, as presented in **Paper IV** and **Paper V**, showcased the generative potential of creative and unconventional uses of AI voice and speech tools. Through a Research-through-Design methodology, the paper provided a novel perspective on how AI can be harnessed to explore new dimensions of human vocality. These papers emphasised the importance of experimental and interdisciplinary approaches in AI research, providing a decisive example of how non-normative artistic practices can be in challenging and revealing the novel capabilities of normative and functional AI voice and speech tools. This exploration connects with both RQ1 and RQ3, demonstrating how qualitative methods and critical exploration via experimental and voice-centred research methods impact the understanding and application of AI voice tools.

Examining all five of the appended papers with respects to the discourses that are pertinent to AI voice tools in musical settings, we can observe some emergent commonalities. These commonalities are: concerns regarding the ethics and power structures in AI voice tools used in musical contexts, and AI tools in multimedia more generally (discussed in **Papers I, II and III**); issues pertaining to data and labour (**Paper I**); interdependencies of voice data and identity (discussed in **Papers II** and **IV**); as well as functionally-centred conflicts between voice data-as-a-Body, and human voice-Bodies (in **Paper III**).

We further identified commonalities across the papers with regards to the embedding of values within how data (and especially voice data) is treated functionally within the training of AI voice models (**Paper III** and **Paper IV**); and that the social and cultural impact of AI voice tools has the potential to inform and re-form existing understandings of human voice and vocality in musical contexts (discussed in **Papers III**, **IV** and **V**). Further, that interdisciplinary methodologies enable novel approaches to critiquing and investigating musical decision-making processes when engaging with AI voice tools in an exploratory and experimental musical practice (**Papers II**, **IV** and **V**). Finally, we noted the generative potential of non-normative and experimental musical and vocal practices as pathways to discovering *how* one might engage a AI voice tools within ones respective musical practice (in **Papers IV** and **V**). Extending further, we note that the usage of AI voice tools in musical settings evokes wider implications - for both human and AI-generated voice-Bodies (discussed in **Paper III, IV** and **V**) and necessitates the potential usage of auxiliary technologies to visibilise or invisibilise.

Collectively, the findings discussed across these papers underscore the transformative impact of AI on artistic practices and the necessity of interdisciplinary and ethical approaches to understanding and guiding this transformation. Building on the insights gained from this thesis, future work should focus on refining our proposed Caring Trouble method, to further examine artistic research processes with AI tools and technologies, looking beyond a solely gestalt-centric analytical method. This will enhance understanding of the ethical, cultural, and socio-technical dimensions of working *with* AI in musical contexts. From a focus on analysing musical working and research processes *with* AI, we anticipate to contribute to the formation of more nuanced views of musical-AI Research-through-Design and how we may better formulate ethical guidelines and protocols to ensure the mindful deployment of AI tools to prioritise artists' rights and agency. Another important direction for future research involves investigating the long-term implications of AI tools on the cultural and economic landscapes of the arts. This includes examining how AI can disrupt existing power structures and exploring ways to mitigate negative impacts on artistic communities. Building on the themes in **Paper III** and **Paper IV**, further research may also explore how these disruptions could be managed and how artists can be supported in preserving their artistic legacies and affording their continued exploration with AI for their musical expression.

Chapter 5

Ethics Statement

In this section, we address some ethics dimensions of this work, informed by principles and codes of practice from venues engaging with musical AI research[1] The following headings reflect the dimensions from these principles which are most pertinent to this compilation thesis.

Accessibility: This thesis has been published digitally via the Chalmers Library, and is accessible as a PDF document. Further, the individual papers appended to this thesis are all open-access publications.

There are examples and sound excerpts that are featured in Papers IV and Paper V which, due to the PDF format are not accessible as rich media in the printed book. These materials are online and digitally accessible by following the links provided in the Appendix.

Data and Privacy: No data considered sensitive under the definitions establish in the current General Data Protection Regulations (GDPR) [215] was collected. The primary dataset used is Kelsey's own sound library, which is not publicly available. There are a number of sound samples and sound excerpts featuring Kelsey's sound library that are presented in Paper IV and Paper V. Kelsey retains the copyright of her own audio material, and these excerpts and samples are shared for illustrative purposes only.

Agency and Oversight: Paper I, appended in this thesis, has been submitted within the purview of a single-blind review process. Papers II, III, IV and V have been submitted within a double-blind review process. All of the appended papers have met the individual publication venues' ethics standards and requirements.

Societal and environmental well-being: This thesis has utilised open-sourced, pretrained AI models. The computational impact of working with these models is comparable to daily personal computer usage- the predicted environmental impact of this work as minimal.

This thesis has the intention to contribute to musical AI research, and to support future research within this community. There are no anticipated negative impacts upon societal well-being as a consequence of this thesis.

[1] https://www.nime.org/ethics/

Bibliography

[1] G. Neskovic, S. Majumdar, N. R. Koluguri, A. Arora, H. Xu, and E. Rrastorgueva, *NVIDIA Speech and Translation AI Models Set Records for Speed and Accuracy*, Mar. 2024. [Online]. Available: `https://developer.nvidia.com/blog/nvidia-speech-and-translation-ai-models-set-records-for-speed-and-accuracy/` (visited on 08/08/2024) (cit. on p. 3).

[2] Q. Liao, *Azure Neural TTS upgraded with HiFiNet, achieving higher audio fidelity and faster synthesis speed*, https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/azure-neural-tts-upgraded-with-hifinet-achieving-higher-audio/ba-p/1847860. (visited on 08/08/2024) (cit. on p. 3).

[3] H. Barakat, O. Turk, and C. Demiroglu, “Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 11, Feb. 2024, ISSN: 1687-4722. DOI: `10.1186/s13636-024-00329-7`. (visited on 08/08/2024) (cit. on p. 3).

[4] N. Tits, K. E. Haddad, and T. Dutoit, *Exploring Transfer Learning for Low Resource Emotional TTS*, en, arXiv:1901.04276 [cs, eess], Jan. 2019. [Online]. Available: `http://arxiv.org/abs/1901.04276` (visited on 11/10/2023) (cit. on pp. 3, 24).

[5] N. Tits, K. E. Haddad, and T. Dutoit, *Laughter Synthesis: Combining Seq2seq modeling with Transfer Learning*, en, arXiv:2008.09483 [cs, eess], Aug. 2020. [Online]. Available: `http://arxiv.org/abs/2008.09483` (visited on 11/10/2023) (cit. on pp. 3, 24).

[6] E. P. Karantonis, F. Placanica, A. Sivuoja, and P. Verstraete, *Cathy Berberian: Pioneer of Contemporary Vocality*, 1st Edition. Routledge, 2014, ISBN: 978-1-315-57107-2 (cit. on p. 3).

[7] *The Diverse Explorations and Inspirations of Joan La Barbara.* [Online]. Available: `https://daily.redbullmusicacademy.com/2016/06/joan-la-barbara-interview` (visited on 11/25/2020) (cit. on p. 3).

[8] R. Pym, “The voice as gesture in meredith monk’s atlas,” McGill University, 2002. [Online]. Available: `https://escholarship.mcgill.ca/concern/theses/gq67jr811` (visited on 09/25/2024) (cit. on p. 3).

[9] A. Kegerreis. “Lexicon of Extended Vocal Techniques.” (), [Online]. Available: `https://music.destinymanifestation.com/lexicon-of-extended-vocal-techniques/` (visited on 09/25/2024) (cit. on p. 3).

[10] M. E. Edgerton, *The 21st-century voice: contemporary and traditional extra-normal voice*, en, Second edition. Lanham: Rowman & Littlefield, 2015, ISBN: 978-1-4422-4824-3 978-0-8108-8840-1 (cit. on pp. 3, 22).

[11] D. W. Gottsegen, *Holly Herndon Launches DAO-Controlled Vocal Deepfake Platform 'Holly+'*, en-US, Section: News, Jul. 2021. [Online]. Available: `https://decrypt.co/75958/holly-herndon-launches-dao-controlled-vocal-deepfake-platform-holly/` (visited on 10/16/2023) (cit. on p. 4).

[12] Grimes [@Grimezsz], *I think it's cool to be fused w a machine and I like the idea of open sourcing all art and killing copyright*, en, Tweet, Apr. 2023. [Online]. Available: `https://twitter.com/Grimezsz/status/1650304205981089793` (visited on 04/24/2023) (cit. on p. 4).

[13] J. Monroe, *Grimes Unveils Software to Mimic Her Voice, Offering 50-50 Royalties for Commercial Use*, https://pitchfork.com/news/grimes-unveils-software-to-mimic-her-voice-and-announces-2-new-songs/, May 2023. (visited on 08/09/2024) (cit. on p. 4).

[14] *YONA featuring Ash Koosha*, en. [Online]. Available: `https://mutek.org/en/artists/yona-featuring-ash-koosha` (visited on 01/10/2024) (cit. on p. 4).

[15] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: IEEE, May 2014, pp. 2494–2498, ISBN: 978-1-4799-2893-4. DOI: `10.1109/ICASSP.2014.6854049`. (visited on 04/13/2023) (cit. on p. 4).

[16] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, *Conformer: Convolution-augmented Transformer for Speech Recognition*, May 2020. arXiv: `2005.08100 [cs, eess]`. (visited on 11/10/2023) (cit. on p. 4).

[17] A. Kamble, A. Tathe, S. Kumbharkar, A. Bhandare, and A. C. Mitra, *Custom Data Augmentation for low resource ASR using Bark and Retrieval-Based Voice Conversion*, Nov. 2023. arXiv: `2311.14836 [cs, eess]`. (visited on 12/01/2023) (cit. on p. 4).

[18] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, “Sequence-to-Sequence Acoustic Modeling for Voice Conversion,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, Mar. 2019, Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, ISSN: 2329-9304. DOI: `10.1109/TASLP.2019.2892235`. [Online]. Available: `https://ieeexplore.ieee.org/abstract/document/8607053` (visited on 11/13/2023) (cit. on pp. 4, 25).

[19] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, “A Review of Deep Learning Based Speech Synthesis,” *Applied Sciences*, vol. 9, p. 4050, Sep. 2019. DOI: 10.3390/app9194050 (cit. on p. 4).

[20] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, *WaveNet: A Generative Model for Raw Audio*, en, arXiv:1609.03499 [cs], Sep. 2016. [Online]. Available: http://arxiv.org/abs/1609.03499 (visited on 11/10/2023) (cit. on pp. 4, 24).

[21] *SpeechBrain: A PyTorch Speech Toolkit.* [Online]. Available: https://speechbrain.github.io/ (visited on 01/15/2024) (cit. on p. 4).

[22] *Whisper/model-card.md at main · openai/whisper*, en. [Online]. Available: https://github.com/openai/whisper/blob/main/model-card.md (visited on 02/29/2024) (cit. on p. 4).

[23] *Mozilla/DeepSpeech*, Mozilla, Aug. 2024. (visited on 08/08/2024) (cit. on p. 4).

[24] *Speech Recognition with Wav2Vec2 — Torchaudio 2.0.1 documentation.* [Online]. Available: https://pytorch.org/audio/stable/tutorials/speech_recognition_pipeline_tutorial.html (visited on 04/13/2023) (cit. on p. 4).

[25] *Kaldi-asr/kaldi*, Kaldi, Aug. 2024. (visited on 08/08/2024) (cit. on p. 4).

[26] E. Harper, S. Majumdar, O. Kuchaiev, *et al.*, *NeMo: A toolkit for Conversational AI and Large Language Models*, Aug. 2024. (visited on 08/08/2024) (cit. on p. 4).

[27] J. Betker, *TorToiSe text-to-speech*, Apr. 2022. (visited on 08/08/2024) (cit. on p. 4).

[28] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, “Matcha-TTS: A fast TTS architecture with conditional flow matching,” in *Proc. ICASSP*, 2024 (cit. on p. 4).

[29] G. Eren and The Coqui TTS Team, *Coqui TTS*, Jan. 2021. DOI: 10.5281/zenodo.6334862. (visited on 08/08/2024) (cit. on p. 4).

[30] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, *Learning latent representations for style control and transfer in end-to-end speech synthesis*, 2019. arXiv: 1812.04342 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1812.04342 (cit. on p. 4).

[31] P. J. H. Kockelkoren, “Art and Technology Playing Leapfrog: A History and Philosophy of Technoèsis,” in *Inside the Politics of Technology : Agency and Normativity in the Co-Production of Technology and Society*, Amsterdam University Press, 2005, pp. 147–167. (visited on 08/08/2024) (cit. on pp. 4, 7).

[32] M. Schröder, “Expressive Speech Synthesis: Past, Present, and Possible Futures,” in *Affective Information Processing*, J. Tao and T. Tan, Eds., London: Springer, 2009, pp. 111–126, ISBN: 978-1-84800-306-4. DOI: 10.1007/978-1-84800-306-4_7. (visited on 08/02/2024) (cit. on pp. 5, 7, 24).

[33] Y. Ren, C. Hu, X. Tan, *et al.*, *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*, en, arXiv:2006.04558 [cs, eess], Aug. 2022. [Online]. Available: `http://arxiv.org/abs/2006.04558` (visited on 11/10/2023) (cit. on p. 5).

[34] J. Shen, R. Pang, R. J. Weiss, *et al.*, *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*, en, arXiv:1712.05884 [cs], Feb. 2018. [Online]. Available: `http://arxiv.org/abs/1712.05884` (visited on 11/10/2023) (cit. on p. 5).

[35] D. Butz and K. Besio, "Autoethnography," en, *Geography Compass*, vol. 3, no. 5, pp. 1660–1674, 2009, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-8198.2009.00279.x, ISSN: 1749-8198. DOI: `10.1111/j.1749-8198.2009.00279.x`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2009.00279.x` (visited on 02/16/2024) (cit. on p. 5).

[36] A. Garry and M. Pearsall, Eds., *Women, knowledge, and reality: explorations in feminist philosophy*, en, 2nd ed. New York: Routledge, 1996, ISBN: 978-0-415-91796-4 978-0-415-91797-1 (cit. on pp. 5, 26).

[37] T. Schiphorst, "Self-evidence: Applying somatic connoisseurship to experience design," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11, New York, NY, USA: Association for Computing Machinery, May 2011, pp. 145–160, ISBN: 978-1-4503-0268-5. DOI: `10.1145/1979742.1979640`. [Online]. Available: `https://dl.acm.org/doi/10.1145/1979742.1979640` (visited on 02/29/2024) (cit. on p. 5).

[38] K. Höök, S. Eriksson, M. Louise Juul Søndergaard, *et al.*, "Soma Design and Politics of the Body," in *Proceedings of the Halfway to the Future Symposium 2019*, ser. HTTF 2019, New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 1–8, ISBN: 978-1-4503-7203-9. DOI: `10.1145/3363384.3363385`. [Online]. Available: `https://dl.acm.org/doi/10.1145/3363384.3363385` (visited on 02/29/2024) (cit. on p. 5).

[39] K. Hook, *Designing with the Body: Somaesthetic Interaction Design*, en. MIT Press, Nov. 2018, Google-Books-ID: 9oZ0DwAAQBAJ, ISBN: 978-0-262-03856-0 (cit. on p. 5).

[40] S. Javaid, *Audio Data Collection for AI: Challenges & Best Practices in 2024*, https://research.aimultiple.com/audio-data-collection/. (visited on 08/07/2024) (cit. on p. 6).

[41] D. Pastukhov, *How Broken Metadata Affects the Music Industry (And What We Can Do About It)?* Aug. 2019. (visited on 08/07/2024) (cit. on p. 6).

[42] R. Scott, "Data Scraping YouTube for the Study of Lieder Reception," *Nineteenth-Century Music Review*, vol. 19, pp. 1–13, Aug. 2022. DOI: `10.1017/S1479409822000143` (cit. on p. 6).

[43] *Freesound in the era of generative Artificial Intelligence | The Freesound Blog.* (visited on 08/07/2024) (cit. on p. 6).

[44] C. Thomé, *Carlthome/audioscrape*, Jul. 2024. (visited on 08/07/2024) (cit. on p. 6).

[45] A. King, *Meet Nightshade—A Tool for Fighting Back vs. AI Data Scraping*, Dec. 2023. (visited on 08/07/2024) (cit. on p. 6).

[46] O. Papakyriakopoulos and A. Xiang, "Considerations for Ethical Speech Recognition Datasets," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '23, New York, NY, USA: Association for Computing Machinery, Feb. 2023, pp. 1287–1288, ISBN: 978-1-4503-9407-9. DOI: 10.1145/3539597.3575793. (visited on 08/07/2024) (cit. on p. 6).

[47] M. K. Ngueajio and G. Washington, "Hey ASR System! Why Aren't You More Inclusive?" In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, J. Y. C. Chen, G. Fragomeni, H. Degen, and S. Ntoa, Eds., Cham: Springer Nature Switzerland, 2022, pp. 421–440, ISBN: 978-3-031-21707-4. DOI: 10.1007/978-3-031-21707-4_30 (cit. on p. 6).

[48] F. Efthymiou, C. Hildebrand, E. de Bellis, and W. H. Hampton, "The Power of AI-Generated Voices: How Digital Vocal Tract Length Shapes Product Congruency and Ad Performance," *Journal of Interactive Marketing*, vol. 59, no. 2, pp. 117–134, May 1, 2024, ISSN: 1094-9968. DOI: 10.1177/10949968231194905. [Online]. Available: https://doi.org/10.1177/10949968231194905 (visited on 09/26/2024) (cit. on p. 7).

[49] D. Grewal, A. Guha, E. Schweiger, S. Ludwig, and M. Wetzels, "How communications by AI-enabled voice assistants impact the customer journey," *Journal of Service Management*, vol. 33, no. 4/5, pp. 705–720, Jan. 1, 2022, ISSN: 1757-5818. DOI: 10.1108/JOSM-11-2021-0452. [Online]. Available: https://doi.org/10.1108/JOSM-11-2021-0452 (visited on 09/26/2024) (cit. on p. 7).

[50] X. Wang, Z. Zhang, and Q. Jiang, "The effectiveness of human vs. AI voice-over in short video advertisements: A cognitive load theory perspective," *Journal of Retailing and Consumer Services*, vol. 81, p. 104 005, Nov. 1, 2024, ISSN: 0969-6989. DOI: 10.1016/j.jretconser.2024.104005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969698924003011 (visited on 09/26/2024) (cit. on p. 7).

[51] S.-H. Lu, H. T. T. Tran, and T.-S. Ngo, "Are we ready for artificial intelligence voice advertising? Comparing human and artificial intelligence voices in audio advertising in a multitasking context," *Quality & Quantity*, Sep. 9, 2024, ISSN: 1573-7845. DOI: 10.1007/s11135-024-01967-x. [Online]. Available: https://doi.org/10.1007/s11135-024-01967-x (visited on 09/26/2024) (cit. on p. 7).

[52] R. N. Brewer, "“If Alexa knew the state I was in, it would cry”: Older Adults' Perspectives of Voice Assistants for Health," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22, New York, NY, USA: Association for Computing Machinery, Apr. 28, 2022, pp. 1–8, ISBN: 978-1-4503-9156-6. DOI: 10.1145/3491101.3519642. [Online]. Available: https://dl.acm.org/doi/10.1145/3491101.3519642 (visited on 09/26/2024) (cit. on p. 7).

[53] S. D. Craig and N. L. Schroeder, "Reconsidering the voice effect when learning from a virtual human," *Computers & Education*, vol. 114, pp. 193–205, Nov. 1, 2017, ISSN: 0360-1315. DOI: 10.1016/j.compedu.2017.07.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131517301653 (visited on 09/26/2024) (cit. on p. 7).

[54] H. Amoozad Mahdiraji, K. Hafeez, H. Kord, and A. Abbasi Kamardi, "Analysing the voice of customers by a hybrid fuzzy decision-making approach in a developing country's automotive market," *Management Decision*, vol. 60, no. 2, pp. 399–425, Jan. 1, 2022, ISSN: 0025-1747. DOI: 10.1108/MD-12-2019-1732. [Online]. Available: https://doi.org/10.1108/MD-12-2019-1732 (visited on 09/26/2024) (cit. on p. 7).

[55] M. Andrade, E. Wanderley, M. Azevedo, *et al.*, "A Voice-Assisted Approach for Vehicular Data Querying from Automotive IoT-Based Databases," in *2023 Symposium on Internet of Things (SIoT)*, Oct. 2023, pp. 1–5. DOI: 10.1109/SIoT60039.2023.10389856. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10389856 (visited on 09/26/2024) (cit. on p. 7).

[56] W.-H. Hsiao and T.-S. Chang, "Exploring the opportunity of digital voice assistants in the logistics and transportation industry," *Journal of Enterprise Information Management*, vol. 32, no. 6, pp. 1034–1050, Jan. 1, 2019, ISSN: 1741-0398. DOI: 10.1108/JEIM-12-2018-0271. [Online]. Available: https://doi.org/10.1108/JEIM-12-2018-0271 (visited on 09/26/2024) (cit. on p. 7).

[57] L. Lazzaroni, "The Facets of Edge AI in Automotive: Exploring Embedded Frameworks, Voice Assistants, and Deep Reinforcement Learning," Mar. 14, 2024. [Online]. Available: https://tesidottorato.depositolegale.it/handle/20.500.14242/68022 (cit. on p. 7).

[58] K. C. Majji and K. Baskaran, "Artificial Intelligence Analytics—Virtual Assistant in UAE Automotive Industry," in *Inventive Systems and Control*, V. Suma, J. I.-Z. Chen, Z. Baig, and H. Wang, Eds., Singapore: Springer, 2021, pp. 309–322, ISBN: 9789811613951. DOI: 10.1007/978-981-16-1395-1_24 (cit. on p. 7).

[59] J. Nascimento and A. Cessa, "Use of Industry 4.0 Concepts to Use the "Voice of the Product" in the Product Development Process in the Automotive Industry," in *Product Lifecycle Management (Volume 4): The Case Studies*, J. Stark, Ed., Cham: Springer International Publishing, 2019, pp. 223–232, ISBN: 978-3-030-16134-7. DOI: 10.1007/978-3-030-

16134-7_17. [Online]. Available: https://doi.org/10.1007/978-3-030-16134-7_17 (visited on 09/26/2024) (cit. on p. 7).

[60] M. Vernuccio, M. Patrizi, and A. Pastore, "Developing voice-based branding: Insights from the Mercedes case," *Journal of Product & Brand Management*, vol. 30, no. 5, pp. 726–739, Jan. 1, 2021, ISSN: 1061-0421. DOI: 10.1108/JPBM-08-2019-2490. [Online]. Available: https://doi.org/10.1108/JPBM-08-2019-2490 (visited on 09/26/2024) (cit. on p. 7).

[61] D. Buhalis and I. Moldavska, "Voice assistants in hospitality: Using artificial intelligence for customer service," *Journal of Hospitality and Tourism Technology*, vol. 13, no. 3, pp. 386–403, Jan. 1, 2022, ISSN: 1757-9880. DOI: 10.1108/JHTT-03-2021-0104. [Online]. Available: https://doi.org/10.1108/JHTT-03-2021-0104 (visited on 09/26/2024) (cit. on p. 7).

[62] G. Amato, M. Behrmann, F. Bimbot, *et al.* "AI in the media and creative industries." arXiv: 1905.04175 [cs]. (May 10, 2019), [Online]. Available: http://arxiv.org/abs/1905.04175 (visited on 09/26/2024), pre-published (cit. on p. 7).

[63] G. M. Arumsari, K. Lee, D. Kim, F. M. Lim, S. Almatari, and T. P. Vu, "From uncanny to unison: Entertainment professionals' perceptions of singing voice synthesis (svs) and singing voice conversion (svc) technologies," *HCI* , pp. 785–793, Jan. 2024. [Online]. Available: https://www.dbpia.co.kr (visited on 09/26/2024) (cit. on p. 7).

[64] M. Avdeeff, "Artificial Intelligence & Popular Music: SKYGGE, Flow Machines, and the Audio Uncanny Valley," *Arts*, vol. 8, no. 4, p. 130, 4 Dec. 2019, ISSN: 2076-0752. DOI: 10.3390/arts8040130. [Online]. Available: https://www.mdpi.com/2076-0752/8/4/130 (visited on 09/26/2024) (cit. on p. 7).

[65] A. Baris, "AI covers: Legal notes on audio mining and voice cloning," *Journal of Intellectual Property Law & Practice*, vol. 19, no. 7, pp. 571–576, Jun. 7, 2024, ISSN: 1747-1532. DOI: 10.1093/jiplp/jpae029. [Online]. Available: https://doi.org/10.1093/jiplp/jpae029 (visited on 09/26/2024) (cit. on p. 7).

[66] C. Broberg, I. Doshoris, and I. van de Haar, "How Artificial Intelligence is changing The Relationship between The Consumer and Brand in The Music Industry," (cit. on p. 7).

[67] E. Drott, "Copyright, compensation, and commons in the music AI industry," *Creative Industries Journal*, vol. 14, no. 2, pp. 190–207, Aug. 4, 2021, ISSN: 1751-0694. DOI: 10.1080/17510694.2020.1839702. [Online]. Available: https://doi.org/10.1080/17510694.2020.1839702 (visited on 09/26/2024) (cit. on p. 7).

[68] H. S. Josan, "AI and Deepfake Voice Cloning: Innovation, Copyright and Artists' Rights," (cit. on p. 7).

[69] K. Lee, G. Hitt, E. Terada, and J. H. Lee, “ETHICS OF SINGING VOICE SYNTHESIS: PERCEPTIONS OF USERS AND DEVELOPERS,” 2022 (cit. on p. 7).

[70] P. Patel, “AI Voice Enters the Copyright Regime: Proposal of a Three-Part Framework,” *Fordham Intellectual Property, Media & Entertainment Law Journal*, vol. 34, p. 451, 2023–2024. [Online]. Available: `https://heinonline.org/HOL/Page?handle=hein.journals/frdipm34&id=463&div=&collection=` (cit. on p. 7).

[71] T. d. A. R. Pinheiro, “How Is AI-created Music Being Commercialized Outside Of The Recording Industry?,” Jul. 19, 2021. [Online]. Available: `https://repositorio-aberto.up.pt/handle/10216/135615` (visited on 09/26/2024) (cit. on p. 7).

[72] I. Ramati, “Algorithmic Ventriloquism: The Contested State of Voice in AI Speech Generators,” *Social Media + Society*, vol. 10, no. 1, p. 20 563 051 231 224 401, Jan. 1, 2024, ISSN: 2056-3051. DOI: `10.1177/20563051231224401`. [Online]. Available: `https://doi.org/10.1177/20563051231224401` (visited on 09/26/2024) (cit. on p. 7).

[73] L. Shroff, “AI & Copyright: A Case Study of the Music Industry,” *GRACE: Global Review of AI Community Ethics*, vol. 2, no. 1, 1 Jan. 22, 2024, ISSN: 2996-5837. [Online]. Available: `https://ojs.stanford.edu/ojs/index.php/grace/article/view/3226` (visited on 09/26/2024) (cit. on p. 7).

[74] L. O’Keefe and Nogueira, *The Body in Sound, Music and Performance: Studies in Audio and Sonic Arts*, en, 1st. Focal Press, 2022, ISBN: 978-0-367-44194-4. [Online]. Available: `https://www.routledge.com/The-Body-in-Sound-Music-and-Performance-Studies-in-Audio-and-Sonic-Arts/O-Keeffe-Nogueira/p/book/9780367441944` (visited on 01/04/2024) (cit. on p. 18).

[75] D. Smalley, “Spectromorphology: Explaining sound-shapes,” en, *Organised Sound*, vol. 2, no. 2, pp. 107–126, Aug. 1997, ISSN: 13557718. DOI: `10.1017/S1355771897009059`. [Online]. Available: `http://www.journals.cambridge.org/abstract_S1355771897009059` (visited on 01/04/2024) (cit. on p. 18).

[76] C. Birdsall and A. Enns, *Sonic Mediations: Body, Sound, Technology - Cambridge Scholars Publishing*, en. Cambridge Scholars Publishing, 2008, ISBN: 978-1-84718-839-7. [Online]. Available: `https://www.cambridgescholars.com/product/9781847188397` (visited on 01/04/2024) (cit. on p. 18).

[77] J. G. Han, “The Somaesthetics of Musicians: Rethinking the Body in Musical Practice,” en, *The Journal of Somaesthetics*, vol. 5, no. 2, Dec. 2019, Number: 2, ISSN: 2246-8498. [Online]. Available: `https://journals.aau.dk/index.php/JOS/article/view/2200` (visited on 01/04/2024) (cit. on p. 18).

[78] A. B. Smith, “Resounding in the Human Body as the ‘True Sanskrit’ of Nature: Reading Sound Figures in Novalis’ The Novices of Sais,” en, *The Journal of Somaesthetics*, vol. 5, no. 2, Dec. 2019, Number: 2, ISSN: 2246-8498. DOI: `10.5278/ojs.jos.v5i2.3344`. [Online]. Available: `https://journals.aau.dk/index.php/JOS/article/view/3344` (visited on 01/04/2024) (cit. on p. 18).

[79] D. Ihde, *Listening and Voice*, en-US, 2nd Edition. New York: State University of New York Press, 2007, ISBN: 978-0-7914-7256-9. [Online]. Available: `https://sunypress.edu/Books/L/Listening-and-Voice2` (visited on 10/25/2023) (cit. on pp. 18, 21).

[80] N. Eidsheim, “Voice as a technology of selfhood: Towards an analysis of racialized timbre and vocal performance,” en, Ph.D. dissertation, University of California, San Diego, Jan. 2008. [Online]. Available: `https://www.academia.edu/657536/Voice_as_a_technology_of_selfhood_Towards_an_analysis_of_racialized_timbre_and_vocal_performance` (visited on 11/08/2023) (cit. on p. 18).

[81] N. S. Eidsheim, “Sensing Voice: Materiality and the Lived Body in Singing and Listening,” en, *The Senses and Society*, vol. 6, no. 2, pp. 133–155, Jul. 2011, ISSN: 1745-8927, 1745-8935. DOI: `10.2752/174589311X12961584845729`. (visited on 06/29/2023) (cit. on pp. 18, 21).

[82] N. S. Eidsheim, *Sensing sound: singing & listening as vibrational practice* (Sign, storage, transmission), en. Durham: Duke University Press, 2015, ISBN: 978-0-8223-6046-9 978-0-8223-6061-2 (cit. on p. 18).

[83] J. T. Demers, *Listening through the noise: the aesthetics of experimental electronic music*, en. Oxford ; New York: Oxford University Press, 2010, OCLC: ocn435918247, ISBN: 978-0-19-538765-0 978-0-19-538766-7 (cit. on pp. 18, 22, 32).

[84] J. Sundberg, *Science of the Singing Voice*, English. Dekalb, Ill: Northern Illinois University Press, Sep. 1989, ISBN: 978-0-87580-542-9 (cit. on p. 18).

[85] Z. Zhang, “Mechanics of human voice production and control,” *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2614–2635, Oct. 2016, ISSN: 0001-4966. DOI: `10.1121/1.4964509`. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5412481/` (visited on 11/03/2023) (cit. on p. 20).

[86] R. I. Godøy and M. Leman, *Musical Gestures: Sound, Movement, and Meaning*, en. Routledge, Feb. 2010, Google-Books-ID: lHaMAgAAQBAJ, ISBN: 978-1-135-18363-9 (cit. on p. 21).

[87] N. H. Kleinsasser, F. G. Priemer, W. Schulze, and O. F. Kleinsasser, “External trauma to the larynx: Classification, diagnosis, therapy,” en, *European Archives of Oto-Rhino-Laryngology*, vol. 257, no. 8, pp. 439–444, Sep. 2000, ISSN: 1434-4726. DOI: `10.1007/s004050000263`. [Online]. Available: `https://doi.org/10.1007/s004050000263` (visited on 06/18/2024) (cit. on p. 21).

[88] S. D. Schaefer, “Management of acute blunt and penetrating external laryngeal trauma,” en, *The Laryngoscope*, vol. 124, no. 1, pp. 233–244, 2014, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lary.24068, ISSN: 1531-4995. DOI: 10.1002/lary.24068. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.24068 (visited on 06/18/2024) (cit. on p. 21).

[89] R. Kaye, C. G. Tang, and C. F. Sinclair, “The electrolarynx: Voice restoration after total laryngectomy,” *Medical Devices: Evidence and Research*, vol. 10, pp. 133–140, Jun. 2017, Publisher: Dove Medical Press _eprint: https://www.tandfonline.com/doi/pdf/10.2147/MDER.S133225, ISSN: null. DOI: 10.2147/MDER.S133225. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/MDER.S133225 (visited on 06/18/2024) (cit. on p. 21).

[90] S. Khosravani, A. Mahnan, I.-L. Yeh, *et al.*, “Laryngeal vibration as a non-invasive neuromodulation therapy for spasmodic dysphonia,” en, *Scientific Reports*, vol. 9, no. 1, p. 17 955, Nov. 2019, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-019-54396-4. [Online]. Available: https://www.nature.com/articles/s41598-019-54396-4 (visited on 06/18/2024) (cit. on p. 21).

[91] H. Liu and M. L. Ng, “Electrolarynx in voice rehabilitation,” *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, Sep. 2007, ISSN: 0385-8146. DOI: 10.1016/j.anl.2006.11.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0385814606001751 (visited on 06/18/2024) (cit. on p. 21).

[92] C. Noble, “Extended from What?: Tracing the construction, flexible meaning, and cultural discources of ”Extended Vocal Techniques”,” en, Ph.D. dissertation, University of California, Santa Cruz, 2019. [Online]. Available: https://escholarship.org/content/qt6qn119zh/qt6qn119zh_noSplash_b604aff101759d9e818f6af5b5159091.pdf?t=ppqqs6 (cit. on pp. 21, 22).

[93] K. Warren, “Show More/Show Less: Extended Voice, Technology, and Presence,” en, Ph.D. dissertation, University of Virginia, Apr. 2017. DOI: 10.18130/V3KW8K. [Online]. Available: https://libraetd.lib.virginia.edu/public_view/9g54xh755 (visited on 08/07/2020) (cit. on p. 21).

[94] A. Schlichter and N. S. Eidsheim, “Introduction: Voice Matters,” *Postmodern Culture*, vol. 24, no. 3, 2014, Publisher: Johns Hopkins University Press, ISSN: 1053-1920. [Online]. Available: https://muse.jhu.edu/pub/1/article/589565 (visited on 06/18/2024) (cit. on p. 22).

[95] K. Meizel, “A Powerful Voice: Investigating Vocality and Identity,” en, *Voice and Speech Review*, vol. 7, no. 1, pp. 267–274, Jan. 2011, ISSN: 2326-8263, 2326-8271. DOI: 10.1080/23268263.2011.10739551. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/23268263.2011.10739551 (visited on 06/18/2024) (cit. on p. 22).

[96] C. A. Valentine and B. S. Damian, “Gender and culture as determinants of the ’ideal voice’,” en, vol. 71, no. 3-4, pp. 285–304, Jan. 1988, Publisher: De Gruyter Mouton Section: Semiotica, ISSN: 1613-3692. DOI: 10.1515/semi.1988.71.3-4.285. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/semi.1988.71.3-4.285/html (visited on 06/18/2024) (cit. on p. 22).

[97] S. A. Carter, “Forging a Sound Citizenry: Voice Culture and the Embodiment of the Nation, 1880-1920,” en, *The American Music research Centre Journal*, vol. 22, pp. 11–34, 2013. [Online]. Available: https://www.proquest.com/openview/0f6642c15544950ecd04f026844d3c19/1.pdf?pq-origsite=gscholar&cbl=9332 (visited on 06/18/2024) (cit. on p. 22).

[98] A. Weidman, “Anthropology and Voice,” en, *Annual Review of Anthropology*, vol. 43, no. Volume 43, 2014, pp. 37–51, Oct. 2014, Publisher: Annual Reviews, ISSN: 0084-6570, 1545-4290. DOI: 10.1146/annurev-anthro-102313-030050. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-anthro-102313-030050 (visited on 06/18/2024) (cit. on p. 22).

[99] B. Truax, *Handbook for Acoustic Ecology*, 2nd ed. Cambridge Street Publishing, 1999. [Online]. Available: http://www.sfu.ca/sonic-studio-webdav/handbook/index.html (visited on 06/11/2024) (cit. on p. 22).

[100] E. Varèse and C. Wen-chung, “The Liberation of Sound,” *Perspectives of New Music*, vol. 5, no. 1, pp. 11–19, 1966, Publisher: Perspectives of New Music, ISSN: 0031-6016. DOI: 10.2307/832385. [Online]. Available: https://www.jstor.org/stable/832385 (visited on 03/23/2023) (cit. on p. 22).

[101] P. Schaeffer, *Traité des objets musicaux , Pierre Sc...* en. Paris, France: Éditions du Seuil., 1966. [Online]. Available: https://www.seuil.com/ouvrage/traite-des-objets-musicaux-pierre-schaeffer/9782020026086 (visited on 10/23/2023) (cit. on pp. 22, 32).

[102] E. Husserl, *Ideas: General Introduction to Pure Phenomenology*, en. Routledge, 2012, ISBN: 978-0-415-51903-8. [Online]. Available: https://www.routledge.com/Ideas-General-Introduction-to-Pure-Phenomenology/Husserl/p/book/9780415519038 (cit. on p. 22).

[103] B. Kane, “L’Objet Sonore Maintenant,” en, *Organised Sound*, vol. 12, no. 1, pp. 15–24, Apr. 2007, Publisher: Cambridge University Press, ISSN: 1469-8153, 1355-7718. DOI: 10.1017/S135577180700163X. (visited on 10/23/2023) (cit. on p. 22).

[104] P. Schaeffer, C. North, and J. Dack, *Treatise on Musical Objects: An Essay across Disciplines*, 1st ed. University of California Press, 2017. [Online]. Available: https://www.jstor.org/stable/10.1525/j.ctt1qv5pqb (visited on 06/15/2022) (cit. on p. 22).

[105] K. Tuuri and T. Eerola, "Formulating a Revised Taxonomy for Modes of Listening," en, *Journal of New Music Research*, vol. 41, no. 2, pp. 137–152, Jun. 2012, ISSN: 0929-8215, 1744-5027. DOI: `10.1080/09298215.2011.614951`. [Online]. Available: `http://www.tandfonline.com/doi/abs/10.1080/09298215.2011.614951` (visited on 03/29/2016) (cit. on p. 22).

[106] P. Birkholz, P. Häsner, and S. Kürbis, "Acoustic comparison of physical vocal tract models with hard and soft walls," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8242–8246. DOI: `10.1109/ICASSP43922.2022.9746611` (cit. on p. 23).

[107] J. Ohala, "Christian Gottlieb Kratzenstein: Pioneer in Speech Synthesis," in *International Congress of Phonetic Sciences*, 2011. (visited on 08/08/2024) (cit. on p. 23).

[108] M. R. Schroeder, "A brief history of synthetic speech," *Speech Communication*, vol. 13, no. 1, pp. 231–237, 1993, ISSN: 0167-6393. DOI: `https://doi.org/10.1016/0167-6393(93)90074-U`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/016763939390074U` (cit. on p. 23).

[109] H. Dudley and T. H. Tarnóczy, "The speaking machine of wolfgang von kempelen," *Journal of the Acoustical Society of America*, vol. 22, pp. 151–166, 1949. [Online]. Available: `https://api.semanticscholar.org/CorpusID:120560845` (cit. on p. 23).

[110] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Berlin, Heidelberg: Springer, 1972, ISBN: 978-3-662-01564-3 978-3-662-01562-9. DOI: `10.1007/978-3-662-01562-9`. (visited on 08/09/2024) (cit. on p. 23).

[111] H. W. Dudley, "Signal transmission," pat. US2151091A, Mar. 1939. (visited on 08/08/2024) (cit. on p. 23).

[112] E. J. Bikker, "Mind over matter: The thinking and speaking machine in fiction of the long nineteenth century," en, phd, University of York, May 2021. [Online]. Available: `https://etheses.whiterose.ac.uk/31783/` (visited on 11/08/2023) (cit. on p. 23).

[113] M. Young, *Singing the Body Electric: The Human Voice and Sound Technology*. London: Routledge, Mar. 2016, ISBN: 978-1-315-60916-4. DOI: `10.4324/9781315609164` (cit. on pp. 23, 32).

[114] Z. Fagyal, "Phonetics and speaking machines: On the mechanical simulation of human speech in the 17th century," en, *Historiographia Linguistica*, vol. 28, no. 3, pp. 289–330, Jan. 2001, Publisher: John Benjamins, ISSN: 0302-5160, 1569-9781. DOI: `10.1075/hl.28.3.02fag`. [Online]. Available: `https://www.jbe-platform.com/content/journals/10.1075/hl.28.3.02fag` (visited on 11/08/2023) (cit. on p. 23).

[115] D. Pantalony, “Hermann von Helmholtz and the Sensations of Tone,” en, in *Altered Sensations: Rudolph Koenig’s Acoustical Workshop in Nineteenth-Century Paris*, D. Pantalony, Ed., Dordrecht: Springer Netherlands, 2009, pp. 19–36, ISBN: 978-90-481-2816-7. DOI: `10.1007/978-90-481-2816-7_2`. [Online]. Available: `https://doi.org/10.1007/978-90-481-2816-7_2` (visited on 11/08/2023) (cit. on p. 23).

[116] M. Mills, “Media and Prosthesis: The Vocoder, the Artificial Larynx, and the History of Signal Processing,” *Qui Parle*, vol. 21, no. 1, pp. 107–149, 2012, Publisher: Duke University Press, ISSN: 1041-8385. DOI: `10.5250/quiparle.21.1.0107`. [Online]. Available: `https://www.jstor.org/stable/10.5250/quiparle.21.1.0107` (visited on 11/08/2023) (cit. on p. 23).

[117] R. Pieraccini, *The Voice in the Machine: Building Computers That Understand Speech*, en. MIT Press, Mar. 2012, Google-Books-ID: 3Njx-CwAAQBAJ, ISBN: 978-0-262-30077-3 (cit. on p. 23).

[118] S. K. Jagtap, M. S. Mulye, and M. D. Uplane, “Speech Coding Techniques,” *Procedia Computer Science*, Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3’15), vol. 49, pp. 253–263, Jan. 2015, ISSN: 1877-0509. DOI: `10.1016/j.procs.2015.04.251`. (visited on 08/09/2024) (cit. on p. 24).

[119] D. O’Shaughnessy, “Linear predictive coding,” *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, Feb. 1988, ISSN: 0278-6648. DOI: `10.1109/45.1890`. (visited on 08/09/2024) (cit. on p. 24).

[120] A. Gersho, “Linear Prediction Techniques in Speech Coding,” in *Adaptive Signal Processing*, L. D. Davisson and G. Longo, Eds., Vienna: Springer, 1991, pp. 69–95, ISBN: 978-3-7091-2840-4. DOI: `10.1007/978-3-7091-2840-4_2`. (visited on 08/09/2024) (cit. on p. 24).

[121] D. J. Broad, “Formants in automatic speech recognition,” *International Journal of Man-Machine Studies*, vol. 4, no. 4, pp. 411–424, Oct. 1972, ISSN: 0020-7373. DOI: `10.1016/S0020-7373(72)80037-3`. (visited on 08/09/2024) (cit. on p. 24).

[122] J. N. Holmes, “Formant synthesizers: Cascade or parallel?” *Speech Communication*, vol. 2, no. 4, pp. 251–273, Dec. 1983, ISSN: 0167-6393. DOI: `10.1016/0167-6393(83)90044-4`. (visited on 08/09/2024) (cit. on p. 24).

[123] A. Pollack, “Technology; Audiotex: Data By Telephone,” *The New York Times*, Jan. 1984, ISSN: 0362-4331. (visited on 08/09/2024) (cit. on p. 24).

[124] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, en. Prentice Hall, 2010, ISBN: 978-0-13-604259-4 (cit. on p. 24).

[125] P. Warden, *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*, en, arXiv:1804.03209 [cs], Apr. 2018. [Online]. Available: `http://arxiv.org/abs/1804.03209` (visited on 11/10/2023) (cit. on p. 24).

[126] D. S. Park, W. Chan, Y. Zhang, *et al.*, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” en, in *Interspeech 2019*, arXiv:1904.08779 [cs, eess, stat], Sep. 2019, pp. 2613–2617. DOI: `10.21437/Interspeech.2019-2680`. [Online]. Available: `http://arxiv.org/abs/1904.08779` (visited on 11/10/2023) (cit. on p. 24).

[127] N. Kalchbrenner, E. Elsen, K. Simonyan, *et al.*, *Efficient Neural Audio Synthesis*, en, arXiv:1802.08435 [cs, eess], Jun. 2018. [Online]. Available: `http://arxiv.org/abs/1802.08435` (visited on 11/10/2023) (cit. on pp. 24, 25).

[128] R. Liu, B. Sisman, and H. Li, *StrengthNet: Deep Learning-based Emotion Strength Assessment for Emotional Speech Synthesis*, en, arXiv:2110.03156 [cs, eess], Oct. 2021. [Online]. Available: `http://arxiv.org/abs/2110.03156` (visited on 11/10/2023) (cit. on p. 24).

[129] Z. Wang, Q. Xie, T. Li, *et al.*, *One-shot Voice Conversion For Style Transfer Based On Speaker Adaptation*, arXiv:2111.12277 [cs, eess], Feb. 2022. DOI: `10.48550/arXiv.2111.12277`. [Online]. Available: `http://arxiv.org/abs/2111.12277` (visited on 06/19/2024) (cit. on p. 24).

[130] N. Takahashi, M. K. Singh, and Y. Mitsufuji, *Cross-modal Face- and Voice-style Transfer*, arXiv:2302.13838 [cs, eess], Mar. 2023. DOI: `10.48550/arXiv.2302.13838`. [Online]. Available: `http://arxiv.org/abs/2302.13838` (visited on 06/19/2024) (cit. on p. 24).

[131] Y. A. Li, C. Han, and N. Mesgarani, *StyleTTS-VC: One-Shot Voice Conversion by Knowledge Transfer from Style-Based TTS Models*, en, arXiv:2212.14227 [cs, eess], Dec. 2022. [Online]. Available: `http://arxiv.org/abs/2212.14227` (visited on 06/19/2024) (cit. on pp. 24, 25).

[132] E. A. AlBadawy and S. Lyu, “Voice Conversion Using Speech-to-Speech Neuro-Style Transfer,” en, in *Interspeech 2020*, ISCA, Oct. 2020, pp. 4726–4730. DOI: `10.21437/Interspeech.2020-3056`. [Online]. Available: `https://www.isca-archive.org/interspeech_2020/albadawy20_interspeech.html` (visited on 06/19/2024) (cit. on p. 24).

[133] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, *StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks*, en, arXiv:1806.02169 [cs, eess, stat], Jun. 2018. [Online]. Available: `http://arxiv.org/abs/1806.02169` (visited on 11/10/2023) (cit. on p. 24).

[134] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, *AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss*, en, arXiv:1905.05879 [cs, eess, stat], Jun. 2019. [Online]. Available: `http://arxiv.org/abs/1905.05879` (visited on 11/10/2023) (cit. on p. 24).

[135] Y. Wang, R. J. Skerry-Ryan, D. Stanton, *et al.*, *Tacotron: Towards End-to-End Speech Synthesis*, en, arXiv:1703.10135 [cs], Apr. 2017. [Online]. Available: `http://arxiv.org/abs/1703.10135` (visited on 11/10/2023) (cit. on pp. 24, 25).

[136] R. Yamamoto, R. Yoneyama, and T. Toda, *NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit*, en, arXiv:2210.15987 [cs, eess], Mar. 2023. [Online]. Available: `http://arxiv.org/abs/2210.15987` (visited on 11/10/2023) (cit. on p. 24).

[137] R. Prenger, R. Valle, and B. Catanzaro, *WaveGlow: A Flow-based Generative Network for Speech Synthesis*, arXiv:1811.00002 [cs, eess, stat], Oct. 2018. DOI: `10.48550/arXiv.1811.00002`. [Online]. Available: `http://arxiv.org/abs/1811.00002` (visited on 11/13/2023) (cit. on p. 25).

[138] J. Ao, R. Wang, L. Zhou, *et al.*, *SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing*, arXiv:2110.07205 [cs, eess], May 2022. DOI: `10.48550/arXiv.2110.07205`. [Online]. Available: `http://arxiv.org/abs/2110.07205` (visited on 11/13/2023) (cit. on p. 25).

[139] T. Walczyna and Z. Piotrowski, "Overview of Voice Conversion Methods Based on Deep Learning," en, *Applied Sciences*, vol. 13, no. 5, p. 3100, Jan. 2023, Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: `10.3390/app13053100`. [Online]. Available: `https://www.mdpi.com/2076-3417/13/5/3100` (visited on 06/19/2024) (cit. on p. 25).

[140] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1290–1302, 2021, Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, ISSN: 2329-9304. DOI: `10.1109/TASLP.2021.3066047`. [Online]. Available: `https://ieeexplore.ieee.org/abstract/document/9380685` (visited on 11/13/2023) (cit. on p. 25).

[141] S. Agarwal, S. Ganapathy, and N. Takahashi, *Leveraging symmetrical convolutional transformer networks for speech to singing voice style transfer*, 2022. arXiv: `2208.12410 [cs.SD]` (cit. on p. 25).

[142] Q. Yang, W. Jin, Q. Zhang, *et al.*, "Mixed-modality speech recognition and interaction using a wearable artificial throat," en, *Nature Machine Intelligence*, vol. 5, no. 2, pp. 169–180, Feb. 2023, Publisher: Nature Publishing Group, ISSN: 2522-5839. DOI: `10.1038/s42256-023-00616-6`.

[Online]. Available: https://www.nature.com/articles/s42256-023-00616-6 (visited on 06/18/2024) (cit. on p. 25).

[143] W. S. Yue and N. A. M. Zin, "Voice Recognition and Visualization Mobile Apps Game for Training and Teaching Hearing Handicaps Children," *Procedia Technology*, 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013, vol. 11, pp. 479–486, Jan. 2013, ISSN: 2212-0173. DOI: 10.1016/j.protcy.2013.12.218. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017313003721 (visited on 06/18/2024) (cit. on p. 25).

[144] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, ISSN: 1387-2877. DOI: 10.3233/JAD-200888. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7836050/ (visited on 06/18/2024) (cit. on p. 25).

[145] C. Deka, A. Shrivastava, A. K. Abraham, S. Nautiyal, and P. Chauhan, "AI-based automated speech therapy tools for persons with speech sound disorder: A systematic literature review," EN, *Speech, Language and Hearing*, May 2024, Publisher: Taylor & Francis. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/2050571X.2024.2359274 (visited on 06/18/2024) (cit. on p. 25).

[146] M. Calvano, A. Curci, A. Pagano, and A. Piccinno, "Speech Therapy Supported by AI and Smart Assistants," in *Product-Focused Software Process Improvement: 24th International Conference, PROFES 2023, Dornbirn, Austria, December 10–13, 2023, Proceedings, Part II*, Berlin, Heidelberg: Springer-Verlag, Dec. 2023, pp. 97–104, ISBN: 978-3-031-49268-6. DOI: 10.1007/978-3-031-49269-3_10. [Online]. Available: https://doi.org/10.1007/978-3-031-49269-3_10 (visited on 06/18/2024) (cit. on p. 25).

[147] S. Goorha and R. Iyengar, "Voice Analytics and Artificial Intelligence: Future Directions for a post-COVID world," en-US, Wharton University of Pennsylvania, White Paper Wharton AI & Analytics for Business, Aug. 2020. [Online]. Available: https://aiab.wharton.upenn.edu/white-paper/voice-analytics-and-artificial-intelligence-future-directions-for-a-post-covid-world/ (visited on 11/03/2023) (cit. on p. 25).

[148] K.-L. Huang, S.-F. Duan, and X. Lyu, "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*, vol. 12, 2021, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.664925. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.664925 (cit. on p. 25).

[149] R. Hasan, R. Shams, and M. Rahman, "Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri," *Journal of Business Research*, vol. 131, pp. 591–597, Jul. 2021, ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2020.12.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296320308419 (visited on 11/03/2023) (cit. on p. 25).

[150] V. Pitardi and H. R. Marriott, "Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence," en, *Psychology & Marketing*, vol. 38, no. 4, pp. 626–642, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21457, ISSN: 1520-6793. DOI: 10.1002/mar.21457. (visited on 11/03/2023) (cit. on p. 25).

[151] M. Alaeddine and A. Tannoury, "Artificial Intelligence in Music Composition," in Jun. 2021, pp. 387–397, ISBN: 978-3-030-79149-0. DOI: 10.1007/978-3-030-79150-6_31 (cit. on p. 25).

[152] D. Ando and H. Iba, "Real-time Musical Interaction between Musician and Multi-agent System," en, p. 9, (cit. on p. 25).

[153] R. Arcand, *The Artists Using Artificial Intelligence to Dream Up the Future of Music*, en-US, Jun. 2019. [Online]. Available: https://www.spin.com/2019/06/ai-music-artificial-intelligence-feature-holly-herndon-yacht/ (visited on 01/12/2024) (cit. on p. 25).

[154] K. Kumar, R. Kumar, T. de Boissiere, *et al.*, *MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis*, arXiv:1910.06711 [cs, eess], Dec. 2019. DOI: 10.48550/arXiv.1910.06711. [Online]. Available: http://arxiv.org/abs/1910.06711 (visited on 02/05/2023) (cit. on p. 25).

[155] E. Cackett, "The application of music technology in live performance to create new ways of musical expression and communication that engages the audience," en, [Online]. Available: https://www.academia.edu/7172652/The_application_of_music_technology_in_live_performance_to_create_new_ways_of_musical_expression_and_communication_that_engages_the_audience (visited on 08/07/2020) (cit. on p. 25).

[156] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, *Deep Learning Techniques for Music Generation* (Computational Synthesis and Creative Systems), en. Cham: Springer International Publishing, 2020, ISBN: 978-3-319-70162-2 978-3-319-70163-9. DOI: 10.1007/978-3-319-70163-9. [Online]. Available: http://link.springer.com/10.1007/978-3-319-70163-9 (visited on 03/17/2022) (cit. on p. 25).

[157] K. Tatar and P. Pasquier, "Musical agents: A typology and state of the art towards Musical Metacreation," en, *Journal of New Music Research*, vol. 48, no. 1, pp. 56–105, Jan. 2019, ISSN: 0929-8215, 1744-5027. DOI: 10.1080/09298215.2018.1511736. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/09298215.2018.1511736 (visited on 03/03/2022) (cit. on pp. 25, 27).

[158] M. Kranzberg, “Technology and History: ”Kranzberg’s Laws”,” *Technology and Culture*, vol. 27, no. 3, pp. 544–560, 1986, Publisher: [The Johns Hopkins University Press, Society for the History of Technology], ISSN: 0040-165X. DOI: `10.2307/3105385`. [Online]. Available: `https://www.jstor.org/stable/3105385` (visited on 10/11/2023) (cit. on p. 25).

[159] D. Ihde, *Existential Technics*, en-US. State University of New York Press, 1983, ISBN: 978-0-87395-687-1. [Online]. Available: `https://sunypress.edu/Books/E/Existential-Technics2` (visited on 06/18/2024) (cit. on p. 25).

[160] C. Stinson, “Algorithms are not neutral,” en, *AI and Ethics*, vol. 2, no. 4, pp. 763–770, Nov. 2022, ISSN: 2730-5961. DOI: `10.1007/s43681-022-00136-w`. [Online]. Available: `https://doi.org/10.1007/s43681-022-00136-w` (visited on 06/18/2024) (cit. on p. 25).

[161] D. Haraway, “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective,” *Feminist Studies*, vol. 14, no. 3, pp. 575–599, 1988, Publisher: Feminist Studies, Inc., ISSN: 0046-3663. DOI: `10.2307/3178066`. [Online]. Available: `https://www.jstor.org/stable/3178066` (visited on 09/26/2022) (cit. on p. 25).

[162] A. Bailey and C. J. Cuomo, *The Feminist Philosophy Reader*, en. Boston: McGraw-Hill, 2008, ISBN: 978-0-07-340739-5 (cit. on p. 25).

[163] J. Weber, “From Science and Technology to Feminist Technoscience,” in *Women, Science, and Technology*, 3rd ed., Routledge, 2013, ISBN: 978-0-203-42741-5 (cit. on p. 25).

[164] M. Mayberry, B. Subramaniam, and L. H. Weasel, *Feminist Science Studies: A New Generation*, en. Psychology Press, 2001, ISBN: 978-0-415-92696-6 (cit. on pp. 25, 26).

[165] S. Harding, “Postcolonial and feminist philosophies of science and technology: Convergences and dissonances,” *Postcolonial Studies*, vol. 12, no. 4, pp. 401–421, Dec. 2009, ISSN: 1368-8790. [Online]. Available: `https://doi.org/10.1080/13688790903350658` (visited on 02/28/2024) (cit. on p. 25).

[166] A. Garry, S. J. Khader, and A. Stone, *The Routledge companion to Feminist Philosophy* (Routledge philosophy companions), en. New York (N.Y.): Routledge, 2017, ISBN: 978-1-138-79592-1 (cit. on p. 26).

[167] C. Ramazanoğlu and J. Holland, *Feminist methodology: challenges and choices*, en, 1. publ. London Thousand Oaks, Calif: Sage, 2002, ISBN: 978-0-7619-5122-3 978-0-7619-5123-0 (cit. on p. 26).

[168] R. M. Schott, *Discovering feminist philosophy: knowledge, ethics, politics* (Feminist constructions), en. Lanham, Md.: Rowman & Littlefield, 2003, ISBN: 978-0-7425-1455-3 978-0-7425-1454-6 (cit. on p. 26).

[169] L. M. Hampton, “Black Feminist Musings on Algorithmic Oppression,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21, New York, NY, USA: Association for Computing Machinery, Mar. 2021, p. 1, ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445929. [Online]. Available: https://doi.org/10.1145/3442188.3445929 (visited on 07/04/2022) (cit. on p. 26).

[170] S. Carroll, I. Garba, O. Figueroa-Rodríguez, *et al.*, “The CARE Principles for Indigenous Data Governance,” *Data Science Journal*, no. 19, pp. 1–12, 2020. DOI: https://doi.org/10.5334/dsj-2020-042. (visited on 03/02/2023) (cit. on p. 26).

[171] J. Gray and A. Witt, “A feminist data ethics of care for machine learning: The what, why, who and how,” en, *First Monday*, Dec. 2021, ISSN: 1396-0466. DOI: 10.5210/fm.v26i12.11833. [Online]. Available: https://journals.uic.edu/ojs/index.php/fm/article/view/11833 (visited on 09/26/2022) (cit. on pp. 26, 32).

[172] J. E. Gray, *What can feminism do for AI ethics?* en, Feb. 2022. [Online]. Available: https://medium.com/mlearning-ai/what-can-feminism-do-for-ai-ethics-b7e401889441 (visited on 09/26/2022) (cit. on p. 26).

[173] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” en, *Scientific Data*, vol. 3, no. 1, p. 160 018, Mar. 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://www.nature.com/articles/sdata201618 (visited on 03/09/2023) (cit. on p. 26).

[174] S. Ahmed, “Embodying diversity: Problems and paradoxes for Black feminists,” *Race Ethnicity and Education*, vol. 12, no. 1, pp. 41–52, Mar. 2009, ISSN: 1361-3324. DOI: 10.1080/13613320802650931. [Online]. Available: https://doi.org/10.1080/13613320802650931 (visited on 07/04/2022) (cit. on p. 26).

[175] S. Ahmed, *No*, en-US, Publication Title: feministkilljoys Type: Blog, 2017. [Online]. Available: https://feministkilljoys.com/2017/06/30/no/ (visited on 03/10/2023) (cit. on p. 26).

[176] S. Gamble, *The Routledge Companion to Feminism and Postfeminism*, en. Routledge, 2001, Google-Books-ID: JKAUXu9vpn0C, ISBN: 978-0-415-24310-0 (cit. on p. 26).

[177] C. Beasley, *What Is Feminism?: An Introduction to Feminist Theory.* Australia: SAGE Publications, 1999, ISBN: 978-0-7619-6335-6 (cit. on p. 26).

[178] S. Bordo, “Feminism, Foucault and the politics of the body,” in *Up Against Foucault*, Num Pages: 24, Routledge, 1993, ISBN: 978-0-203-40868-1 (cit. on p. 26).

[179] B. Latour and P. Weibel, *Making Things Public*, en-US. Cambridge, Massachusetts: MIT Press, 2005, ISBN: 0-262-12279-0. [Online]. Available: `https://mitpress.mit.edu/9780262122795/making-things-public/` (visited on 02/24/2023) (cit. on p. 26).

[180] P. Stephan, "Designing 'matters of concern' (Latour): A future design task ?" In *Transformation Design – Perspectives on a New Design Attitude*, Birkhäuser, Dec. 2015, pp. 202–226, ISBN: 978-3-0356-0653-9 (cit. on p. 26).

[181] B. Latour, "What Is the Style of Matters of Concern?" en, in *The Lure of Whitehead*, N. Gaskill and A. J. Nocek, Eds., University of Minnesota Press, Oct. 2014, pp. 92–126, ISBN: 978-0-8166-7995-9. DOI: `10.5749/minnesota/9780816679959.003.0004`. [Online]. Available: `https://academic.oup.com/minnesota-scholarship-online/book/29369/chapter/244364332` (visited on 11/24/2022) (cit. on p. 26).

[182] M. P. de la Bellacasa, *Matters of Care: Speculative Ethics in More Than Human Worlds.* University of Minnesota Press, 2017, ISBN: 978-1-5179-0064-9. [Online]. Available: `https://libgen.li/ads.php?md5=3dec273eb9043ae8b1a7140b1120c759` (visited on 03/02/2023) (cit. on pp. 26, 32).

[183] P.-P. Verbeek, "Cyborg intentionality: Rethinking the phenomenology of human–technology relations," en, *Phenomenology and the Cognitive Sciences*, vol. 7, no. 3, pp. 387–395, Sep. 2008, ISSN: 1572-8676. DOI: `10.1007/s11097-008-9099-x`. [Online]. Available: `https://doi.org/10.1007/s11097-008-9099-x` (visited on 09/10/2022) (cit. on p. 27).

[184] R. Rosenberger and P. P. C. C. Verbeek, "A field guide to postphenomenology," English, *Postphenomenological Investigations: Essays on Human-Technology Relations*, pp. 9–41, 2015, Publisher: Lexington Books. [Online]. Available: `https://research.utwente.nl/en/publications/a-field-guide-to-postphenomenology` (visited on 09/10/2022) (cit. on p. 27).

[185] D. Ihde, *Technology and the Lifeworld: From Garden to Earth.* Indiana University Press, 1990 (cit. on p. 27).

[186] B. Pula, "Does Phenomenology (Still) Matter? Three Phenomenological Traditions and Sociological Theory," en, *International Journal of Politics, Culture, and Society*, vol. 35, no. 3, pp. 411–431, Sep. 2022, ISSN: 1573-3416. DOI: `10.1007/s10767-021-09404-9`. [Online]. Available: `https://doi.org/10.1007/s10767-021-09404-9` (visited on 04/08/2024) (cit. on p. 27).

[187] R. Rosenberger and P.-P. Verbeek, *Postphenomenological Investigations: Essays on Human–Technology Relations*, en. Lexington Books, 2015, ISBN: 978-0-7391-9437-9 (cit. on p. 27).

[188] C. Frauenberger, "Entanglement hci the next wave?" *ACM Trans. Comput.-Hum. Interact.*, vol. 27, no. 1, 2019, ISSN: 1073-0516. DOI: `10.1145/3364998`. [Online]. Available: `https://doi.org/10.1145/3364998` (cit. on p. 27).

[189] P. le Roux, C. van Staden, and K. Kraus, “Phenomenology: Excavating contextual practices from open distance learning tutoring experiences,” in *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*, ser. SAICSIT ’19, Skukuza, South Africa: Association for Computing Machinery, 2019, ISBN: 9781450372657. DOI: 10.1145/3351108.3351112. [Online]. Available: https://doi.org/10.1145/3351108.3351112 (cit. on p. 27).

[190] C. Frauenberger, J. Good, and W. Keay-Bright, “Phenomenology, a framework for participatory design,” in *Proceedings of the 11th Biennial Participatory Design Conference*, ser. PDC ’10, Sydney, Australia: Association for Computing Machinery, 2010, pp. 187–190, ISBN: 9781450301312. DOI: 10.1145/1900441.1900474. [Online]. Available: https://doi.org/10.1145/1900441.1900474 (cit. on p. 27).

[191] M. M. Jensen and J. Aagaard, “A postphenomenological method for hci research,” in *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, ser. OzCHI ’18, Melbourne, Australia: Association for Computing Machinery, 2018, pp. 242–251, ISBN: 9781450361880. DOI: 10.1145/3292147.3292170. [Online]. Available: https://doi.org/10.1145/3292147.3292170 (cit. on p. 27).

[192] F. Ohlin and C. M. Olsson, “Beyond a utility view of personal informatics: A postphenomenological framework,” in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ser. UbiComp/ISWC’15 Adjunct, New York, NY, USA: Association for Computing Machinery, Sep. 2015, pp. 1087–1092, ISBN: 978-1-4503-3575-1. DOI: 10.1145/2800835.2800965. [Online]. Available: https://doi.org/10.1145/2800835.2800965 (visited on 09/10/2022) (cit. on p. 27).

[193] W. Odom, J. Pierce, E. Stolterman, and E. Blevis, “Understanding why we preserve some things and discard others in the context of interaction design,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’09, New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 1053–1062, ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518862. [Online]. Available: https://doi.org/10.1145/1518701.1518862 (visited on 09/10/2022) (cit. on p. 27).

[194] D. Fallman, “The new good: Exploring the potential of philosophy of technology to contribute to human-computer interaction,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 1051–1060, ISBN: 9781450302289. DOI: 10.1145/1978942.1979099. [Online]. Available: https://doi.org/10.1145/1978942.1979099 (cit. on p. 27).

[195] J. J. Benjamin, A. Berger, N. Merrill, and J. Pierce, “Machine learning uncertainty as a design material: A post-phenomenological inquiry,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445481. [Online]. Available: https://doi.org/10.1145/3411764.3445481 (cit. on p. 27).

[196] M. Wooldridge, *An Introduction to MultiAgent Systems*, English, 2nd edition. Chichester, U.K: Wiley, Jun. 2009, ISBN: 978-0-470-51946-2 (cit. on p. 27).

[197] J. Pyysiäinen, “Sociocultural affordances and enactment of agency: A transactional view,” en, *Theory & Psychology*, vol. 31, no. 4, pp. 491–512, Aug. 2021, Publisher: SAGE Publications Ltd, ISSN: 0959-3543. DOI: 10.1177/0959354321989431. [Online]. Available: https://doi.org/10.1177/0959354321989431 (visited on 04/08/2024) (cit. on p. 27).

[198] E. Rietveld and J. Kiverstein, “A Rich Landscape of Affordances,” EN, *Ecological Psychology*, Oct. 2014, ISSN: 1040-7413. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/10407413.2014.958035 (visited on 04/08/2024) (cit. on p. 27).

[199] R. Withagen, D. Araújo, and H. J. de Poel, “Inviting affordances and agency,” *New Ideas in Psychology*, vol. 45, pp. 11–18, 2017, ISSN: 0732-118X. DOI: 10.1016/j.newideapsych.2016.12.002 (cit. on p. 27).

[200] K. Barad, “Getting Real: Technoscientific Practices and the Materialization of Reality,” *differences*, vol. 10, no. 2, pp. 87–128, Jul. 1998, ISSN: 1040-7391. DOI: 10.1215/10407391-10-2-87. [Online]. Available: https://doi.org/10.1215/10407391-10-2-87 (visited on 07/04/2022) (cit. on p. 27).

[201] K. Barad, “Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter,” en, *Signs*, vol. 28, no. 3, 2003. [Online]. Available: https://www.academia.edu/1857574/Posthumanist_performativity_Toward_an_understanding_of_how_matter_comes_to_matter (visited on 10/09/2023) (cit. on p. 27).

[202] C. Draude, ““Boundaries Do Not Sit Still” from Interaction to Agential Intra-action in HCI,” en, in *Human-Computer Interaction. Design and User Experience*, M. Kurosu, Ed., Cham: Springer International Publishing, 2020, pp. 20–32, ISBN: 978-3-030-49059-1. DOI: 10.1007/978-3-030-49059-1_2 (cit. on p. 27).

[203] R. Gondomar and E. Mor, “Understanding Agency in Human-Computer Interaction Design,” en, in *Human-Computer Interaction. Theory, Methods and Tools*, M. Kurosu, Ed., Cham: Springer International Publishing, 2021, pp. 137–149, ISBN: 978-3-030-78462-1. DOI: 10.1007/978-3-030-78462-1_10 (cit. on p. 27).

[204] J. McEneaney, “Agency Effects in Human-Computer Interaction,” *International Journal of Human-Computer Interaction*, vol. 29, pp. 798–813, Sep. 2013. DOI: 10.1080/10447318.2013.777826 (cit. on p. 27).

[205] P. Worthy, T. Hunter, B. Matthews, and S. Viller, "Musical agency and an ecological perspective of DMIs: Collective embodiment in third wave HCI," en, *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 797–807, Aug. 2021, ISSN: 1617-4917. DOI: `10.1007/s00779-020-01429-9`. [Online]. Available: `https://doi.org/10.1007/s00779-020-01429-9` (visited on 04/08/2024) (cit. on p. 27).

[206] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07, New York, NY, USA: Association for Computing Machinery, Apr. 2007, pp. 493–502, ISBN: 978-1-59593-593-9. DOI: `10.1145/1240624.1240704`. (visited on 04/12/2023) (cit. on p. 28).

[207] J. Zimmerman, E. Stolterman, and J. Forlizzi, "An analysis and critique of Research through Design: Towards a formalization of a research approach," in *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, ser. DIS '10, New York, NY, USA: Association for Computing Machinery, Aug. 2010, pp. 310–319, ISBN: 978-1-4503-0103-9. DOI: `10.1145/1858171.1858228`. (visited on 04/12/2023) (cit. on p. 28).

[208] C. Frayling, *Research in Art and Design*, en. Royal College of Art, 1993, Google-Books-ID: _Ar4QwAACAAJ, ISBN: 978-1-874175-55-1 (cit. on p. 28).

[209] K. Höök and J. Löwgren, "Strong concepts: Intermediate-level knowledge in interaction design research," *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 3, 23:1–23:18, Oct. 2012, ISSN: 1073-0516. DOI: `10.1145/2362364.2362371`. (visited on 03/22/2023) (cit. on p. 28).

[210] W. Gaver, J. Bowers, T. Kerridge, A. Boucher, and N. Jarvis, "Anatomy of a failure: How we knew when our design went wrong, and what we learned from it," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09, New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 2213–2222, ISBN: 978-1-60558-246-7. DOI: `10.1145/1518701.1519040`. (visited on 04/12/2023) (cit. on p. 28).

[211] W. Gaver, "What should we expect from research through design?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, New York, NY, USA: Association for Computing Machinery, May 2012, pp. 937–946, ISBN: 978-1-4503-1015-4. DOI: `10.1145/2207676.2208538`. (visited on 04/12/2023) (cit. on p. 28).

[212] W. Gaver, P. G. Krogh, A. Boucher, and D. Chatting, "Emergence as a Feature of Practice-based Design Research," ser. DIS '22, New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 517–526, ISBN: 978-1-4503-9358-4. DOI: `10.1145/3532106.3533524`. [Online]. Available: `https://doi.org/10.1145/3532106.3533524` (visited on 04/17/2023) (cit. on p. 28).

[213] K. Cotton and K. Tatar, “Caring Trouble and Musical AI: Considerations towards a Feminist Musical AI,” *AIMC 2023*, Aug. 2023 (cit. on p. 32).

[214] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*, en. Oxford University Press, Jul. 2005, ISBN: 978-1-383-03965-8. DOI: `10.1093/oso/9780199256044.001.0001`. [Online]. Available: `https://academic.oup.com/book/52349` (visited on 04/08/2024) (cit. on p. 32).

[215] P. O. o. t. E. Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, Apr. 27, 2016. [Online]. Available: `https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en` (visited on 09/23/2024) (cit. on p. 41).

Part II

Appended Papers

Paper I

A Shift in Artistic Practices through Artificial Intelligence

Kıvanç Tatar**[2], Petter Ericson**, Kelsey Cotton**, Paola Torres Núñez Del Prado, Roser Batlle-Roca, Beatriz Cabrero-Daniel, Sara Ljungblad, Georgios Diapoulis, Jabbar Hussain

Leonardo 2024; 57 (3): 293–297. doi:10.1162/leon4$_{a0}$2523

[2] ** denotes equal first authorship

GENERAL ARTICLE

A Shift in Artistic Practices through Artificial Intelligence

KIVANÇ TATAR, PETTER ERICSON, KELSEY COTTON, PAOLA TORRES NÚÑEZ DEL PRADO, ROSER BATLLE-ROCA, BEATRIZ CABRERO-DANIEL, SARA LJUNGBLAD, GEORGIOS DIAPOULIS, AND JABBAR HUSSAIN

ABSTRACT

The explosion of content generated by artificial intelligence (AI) models has initiated a cultural shift in arts, music, and media, whereby roles are changing, values are shifting, and conventions are challenged. The vast, readily available dataset of the Internet has created an environment for AI models to be trained on any content on the Web. With AI models shared openly and used by many globally, how does this new paradigm shift challenge the status quo in artistic practices? What kind of changes will AI technology bring to music, arts, and new media?

CURRENT AI APPLICATIONS IN ARTISTIC PRACTICES

Increasing interest and accessibility in artificial intelligence (AI) technologies have moved the societal discussion of AI to the mainstream. Ease in accessing computation resources, open-access AI knowledge, open-source AI software, and vast datasets in different modalities enable artists to explore new possibilities. The recent advancements in AI architectures have been applied to all artistic domains, including visual arts, video art, music and sound art, dance and performance arts, literature, and interdisciplinary arts.

AI technologies for artistic image and video generation [1] have had to build an entirely new vocabulary for describing images or frames in video. While image generation with autonomous painting was established decades ago (such as in Harold Cohen's AARON system), it is only relatively recently that AI in the visual domain has become accessible to vast communities thanks to open-source, open-access, and browser-based tools that require minimal local computation resources, such as Stable Diffusion, and generative adversarial networks such as BigGAN and StyleGAN.

In music and sound arts, there are four main tracks of AI applications [2] and interactive music systems [3]: approaches working with symbolic music; systems working with audio signals directly; systems that map a symbolic music domain to musical audio or vice versa; and approaches that connect musical domains to other domains such as bodily movement.

In performing arts, movement generation [4] has been one AI research track. The advances in this domain have created a ripple effect in generative dance and animation: AI systems have been used as a tool for assistance, choreography support, performance collaboration, and movement generation.

Text generation and conversational bots [5] have initiated recent discussions on how omnipresent large language models (LLMs) such as ChatGPT can change all aspects of society. Additionally, LLMs have been incorporated into interdisciplinary domains and interactive artworks with chatbots, as well as tool generation for other artistic domains such as audio synthesizer code or creative coding script generation examples.

There is currently no consensus on the definition and coverage of the term *artificial intelligence* in the literature [6]. Even though we mainly cover machine learning (ML) approaches utilizing training datasets here, these are often embedded in complex systems with sensing and action capabilities. These

Kıvanç Tatar (artist-technologist, musician, researcher), Chalmers University of Technology, Department of Computer Science and Engineering, Data Science and AI Division, Rännvägen 6B, Gothenburg, Sweden, 412 96. Email: tatar@chalmers.se. ORCID: 0000-0003-4133-8641.

Petter Ericson (researcher, hacker, musician), Department of Computing Science, Umeå University, MIT-huset, Campustorget 5, Umeå Universitet, Umeå, Sweden, 901 87. Email: pettter@cs.umu.se. ORCID: 0000-0002-8722-5661.

Kelsey Cotton (artist, musician, researcher), Chalmers University of Technology, Department of Computer Science and Engineering, Data Science and AI Division, Rännvägen 6B, Gothenburg, Sweden, 412 96. Email: kelsey@chalmers.se. ORCID: 0000-0002-2802-0244.

Paola Torres Núñez del Prado (transdisciplinary artist, creative technologist, researcher), Stockholm University of the Arts, Brinellvägen 34, Stockholm, Sweden, 114 28. Email: paola.torres@uniarts.se. ORCID: 0000-0002-5525-3684.

Roser Batlle-Roca (researcher, musician), Universitat Pompeu Fabra, Tanger, 122-144, Barcelona, Spain, 115 41. Email: roser.batlle@upf.edu. ORCID: 0000-0003-3591-9378.

Beatriz Cabrero-Daniel (researcher), Department of Computer Science and Engineering, University of Gothenburg, Hörselgången 5, Gothenburg, Sweden, 417 56. Email: beatriz.cabrero-daniel@gu.se. ORCID: 0000-0001-5275-8372.

Sara Ljungblad (researcher), Department of Computer Science and Engineering, University of Gothenburg, Interaktionsdesign, Kuggen, Lindholmspiren 1, Gothenburg, Sweden, 417 56. Email: sara.ljungblad@chalmers.se. ORCID: 0000-0002-2751-9801.

Georgios Diapoulis (music technologist, creative coder, researcher), Department of Computer Science and Engineering, Chalmers University of Technology, Interaktionsdesign, Kuggen, Lindholmspiren 1, Gothenburg, Sweden, 417 56. Email: georgios.diapoulis@chalmers.se. ORCID: 0000-0002-3101-1875.

Jabbar Hussain (researcher), Department of Applied IT, University of Gothenburg, Forskningsgången 6 Gothenburg, Sweden, 412 96. Email: jabbar.hussain@ait.gu.se. ORCID: 0000-0003-1170-9069.

See https://direct.mit.edu/leon/issue/57/3 for supplemental files associated with this issue.

https://doi.org/10.1162/leon_a_02523

approaches are alternatively referred to in the literature as multi-agent systems [7]. In the remainder of the paper, we use the term AI to include systems based on ML approaches, which may or may not have sensing and action capabilities. We call those systems applied artificial intelligence technologies—or simply AI—in the remainder of the paper.

INFINITE CONTENT ON THE BOUNDLESS INTERNET

There is a long history of digital art that generates infinite content in various guises [8]. New AI approaches such as Stable Diffusion and MusicLM have received widespread interest in public discussions due to their aesthetic possibilities when combined with LLMs. These AI models are trained, fine-tuned, and parameterized through a process involving many actors, including data creators, AI developers, and model architecture researchers. Artists have been joining all processes of AI technology—including data creation, developing novel AI architectures, training specific models, and utilizing readily available models. As the technical skill barrier in using AI technology has decreased due to ease of access for known and recent models, all digital platforms have been receiving a new wave of AI-generated content.

As a result of the increased accessibility of AI tools, values in artistic production have been shifting from favoring manually made content to automatically generated content. AI systems such as Stable Diffusion have been challenging conventional values in visual arts practices by generating infinite amounts of content. In some known cases, the training datasets of widely used AI content generators were gathered by scraping content from the Internet, even when the data was held within copyright. A recent Vox short documentary mentions how the American illustrator James Gurney has become a common style prompt entry in AI image generation. The artist raises the issue of consent within infinite content generators. Gurney has stated:

> I think it is only fair to people looking at this work that they should know what the prompt was and also what software was used. I think the artists should be allowed to opt-in or opt-out of having their work that they worked so hard on by hand be used as a dataset for creating this other artwork [9].

Gurney's comment highlights the current status quo: AI technology developers overlook the question of explicit consent from artists and content owners, and current regulations fall short of empowering artists in decisions on the use of their content. Furthermore, massive content generation in the style of a particular artist could devalue it, as in the case of Gurney, or increase it by social networking, as in the case of Holly Herndon [10], who has shared her voice as an ML model (vocal deepfake) for other artists to use in their performances. This shift in capital value on cultural practices calls for a new take on ethics for artistic practices in the era of ubiquitous AI.

LABOR IN ARTISTIC DATA

The increasing requirement of large datasets for mass generated high-fidelity content using AI has triggered a notable shift in practice of Internet ethics of copying, appropriating, and distributing art and music. Currently, the datasets that are used to train AI models such as DALL-E, Stable Diffusion, or GitHub CoPilot have been scraping data from the Web while ignoring the licenses and intellectual property of data. This is akin to a digital enclosure of the cultural commons.

The case of Gurney is one example of how the creators of these widely used, easily accessible AI models for content generation have been gathering data for their training datasets: by overlooking the consent of data creators. This directly exploits the artist's labor in creating the original content [11]. The data that is publicly shared on the Internet is approached as free to take and is transformed into a capital commodity through the process of training ML models, exhibiting similarities to historical practices of exploitatively ignoring cultural context and consent within the structures of colonialism and its enclosures. This current status quo in data gathering can be addressed through empowering the artist by incorporating their explicit consent into current structures. The consent can be manifested by the artist, as in the case of Stability.ai, where artists can choose an option so that their work is not used in the training of Stable Diffusion models. However, it is not clear to what extent this pledge can be honored, considering the multitude of remixes and format shifts of popular artists' works on the Internet. Other licenses in that discussion, such as Creative Commons, emphasize the importance of fair use for public good [12]. While both proposals highlight valuable aspects of the issue, we envisage that a future of copyright in artistic practices will require platforms and structures promoting artistic data sovereignty.

Although existing copyright licenses are legally recognized mechanisms for protecting intellectual property, the current gatekeeping mechanisms in accessing online platforms have implicit consent structures related to artistic work, such as terms of service agreements and cookie mechanisms. Those agreements are often produced without the participation of the artists, thereby benefiting the industry that is gatekeeping access to the digital platforms. Whether the copyright permissions of artists are respected through explicit consent is an issue of traceability, accountability, and regulation. Many practitioners in artistic domains are individuals or small business owners who do not have the power to counter the impositions of big tech such as OpenAI. We still need tools of traceability, third-party nonprofit organizations for proposing regulations, and public structures for accountability. Traceability, regulation, and accountability all require inclusive and participatory discussions of ethics for the foreseeable future of AI.

A SHIFT IN ARTISTIC PRACTICES

The emergence of accessible AI tools for artistic production caused a natural shift in artistic practices. The conception of artist as genius, where a single person produces a masterpiece, has been shifting toward communal production, where several actors are involved in the artistic production. Nonprofits such as EleutherAI, and for-profits such as Stability.ai, with involvement by other developers, have been pursuing open-sourcing the groundbreaking AI architectures, which has been benefiting artistic practices by initiating public discussions.

Even though AI technology employs the conventions of open-source and open-access tools, the question remains whether the power structures within these technologies are truly democratic. The majority of tools for AI development are provided, hosted, or maintained by a few technology companies, such as Google (TensorFlow, Collab), OpenAI (DALLE-2, GPT-3), Facebook (PyTorch), and Microsoft (GitHub). Although there is a shift toward communal production in artistic practices, with varying levels of contributions to the artwork production by different actors, these technology companies still wield decisive power and capital determining the use of AI tools for artistic practices and their computational resources. Decision-making in creation of AI tools for artistic production has yet to be democratized.

An artwork exists in close relation to its medium. Along with ubiquitous AI, the evolution of new art markets such as social media, streaming platforms, and nonfungible tokens initiated a cultural and monetary value reassignment by aggregating their status as major channels of artwork "storage" and distribution. The value of an artwork is typically affected by the ranking or curation mechanisms of the medium. Those mechanisms are directly related to the artwork's visibility and thereby its value. These platforms now have substantial influence in reshaping current social conceptions around economic and cultural capital. For example, the notion of added value through engagement appears in Herndon's AI system for voice synthesis, titled Holly+. Herndon trained the AI architecture on her own voice recordings, and Holly+ is an AI model through which users can synthesize recorded audio in the style of Herndon's voice. Herndon deploys a decentralized autonomous organization (DAO) blockchain to allow voting on the minting of artworks made using Holly+ and distributes tokens [13] to members of the DAO (and the creator of minted artworks) to share in any profits from the usage of Holly+ recorded on the relevant blockchain. Here, Herndon acknowledges the value put into Holly+ by each user when they engage with the AI system. By giving away DAO tokens, Herndon creates a platform for users to both acquire a voting stake in the Holly+ DAO and to fiscally benefit from the capital value produced by artistic work made with Holly+. This semi-decentralized governing structure for Holly+ is centered on connecting value to engagement processes: The more people who engage with the art, the more valuable the art itself is perceived to be.

SUSTAINABILITY, ETHICS, ACCESSIBILITY, AND INCLUSION

Discussion on societal aspects of AI is entangled with the concepts of sustainability, ethics, and values. A broader framing of these issues, focused on the evolution of culture and impact upon culture industries, has implications for the environmental and climate impact of leveraging AI technologies within artistic practices. Within discussions around sustainability, there are also broader concerns with social sustainability for artists and the sustainability of the role of art within society. A core concern in this regard is how advancements in AI technologies deployed within artistic contexts may adversely impact the cultural economy of artists in society, and further, how their role in society is affected by the wider usage of AI tools.

The appreciation of art is in the eyes of the beholder. A larger discussion in this regard is centered on the intrinsic nature of AI tool usage as originating from a desire to "be better" and thereby enable "better" artistic work. This highlights questions regarding what the nature of "better" actually is, who decides this, and if this continual pursuit of "better" is an enacted artistic Darwinism. While conclusive decisions on how power, relevancy, and what "better" means are yet to be navigated, it is plainly apparent that the positioning of interdisciplinary borders is significant in shaping value systems determining the role AI occupies within cultural industries. To this end, the rigidity (or fuzziness) of these boundaries needs to facilitate equitable exchange, promote artistic diversity, and encourage an artistic culture of mindful progress.

A world with more widely available AI art creation tools and globalized platforms, giving more people than ever the ability to create and share art, could lead to a greater chance to find a wider audience for previously unknown and/or marginalized creators, styles, and art forms. Currently, however, the same world and those same platforms are more often built to entrench the dominance of the hegemonic culture and existing artists. There is still more to be done to improve inclusion [14]. Geographical borders still matter in accessibility of the Internet and protection of rights in AI applications (Fig. 1). Therefore, communities must take action to address the active participation of underrepresented cultures beyond the passive appropriation of their imaginaries. Cultivating diversity requires ethical considerations in all processes of all stakehold-

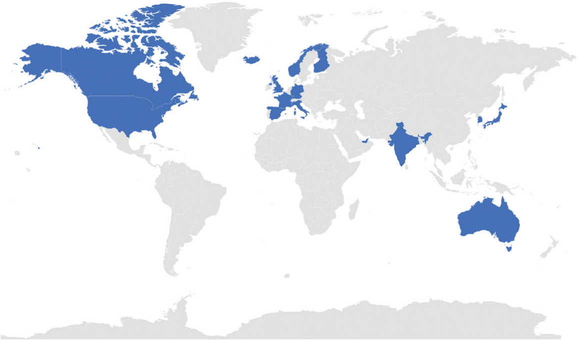


Fig. 1. Countries with at least one organization (e.g., government, university) issuing AI guidelines for responsible and ethical AI development and implementation (not necessarily laws, as of 2019): Australia, Canada, Finland, France, Germany, Iceland, India, Italy, Japan, Netherlands, Norway, Singapore, South Korea, Spain, U.A.E., U.K., and U.S.A. Populations of these countries (as of 2021): 2.8 billion, compared to the total world population of 7.9 billion. Therefore, these organizations make recommendations for 35% of the world's population [19]. Creative Commons 0 license—Public domain.

ers and actors of AI while reducing biases and analyzing stereotypes and representation of individuals and their creations.

Some ubiquitous AI models have been shown to reproduce problematic stereotypes in the data [15], which also may appear in artworks that are generated using these models. It is often hard for a single user or artist to be aware of those stereotypes in AI models, since they become apparent after many output generations on similar prompts and require statistical analysis on the generated content. In the case of tools such as DALL-E, Midjourney, and Stable Diffusion, it is worth noting the difference in the quality of the images generated when the prompted text used for generation refers to images, symbols, or art that belongs to a hegemonic culture (undoubtedly easier to access) when compared to the outputs of prompts that relate to more "obscure," underrepresented, or nonhegemonic cultural manifestations [16]. Cultural hierarchies could indeed be extrapolated from how much detail, quality, or "realism" (or lack thereof) the generated images finally contain.

A BETTER FUTURE FOR ARTISTIC PRACTICES WITH ARTIFICIAL INTELLIGENCE

From data production and curation to model design, implementation, training, configuration, and final use, a human is involved at every point in AI technologies. Ethical considerations emerge in the decisions of all stakeholders and actors, in addition to the artists. The conceptualizations and values of technology actors and stakeholders leave "residues" behind that influence the artworks and the culture and society that they are situated in. AI technology, with its immense power to shift culture and society globally, goes beyond proprietary software rights of a single for-profit company. Public discussions on what a future society with AI may look like are critical.

Primarily, a fundamental change in power and the distribution of power is necessary for inclusivity in the decision-making of all AI system designs. In the case of recommendation systems, inclusive decision-making processes can have significance in the obscuring and hypervisibilizing of artworks and artists, especially those from underrepresented and marginalized communities, cultures, and regions. Additionally, a crossover between art and technology within pedagogical contexts is beneficial, allowing artists to acquire the knowledge, skills, and access to make artistic use of new AI technologies, and technologists the time, training, and opportunity to explore artistic pursuits and their requirements.

Copyright is an important aspect of artistic practice in a digital world that is further highlighted and complicated by the introduction of widely available and powerful AI technologies. Exact future steps on consent, traceability, regulations, and accountability are not clear yet. However, the discussions within, for example, various pirate parties, digital rights groups, hacktivist organizations, and open-access global initiatives of technology replications (such as GPT-NEO, GPT-J, Stable Diffusion, and DALL-E 2 PyTorch replication) are significantly more instructive than the various rearguard actions by other industry organizations. It is evident that there is an immense need for reform of the current regime of might makes right, wherein large corporations can infringe on copyright with absolute impunity, shedding light on problematic power relations. The discrepancy between corporate copyright transgressions—in the creation of enormous datasets as input to AI systems with no knowledge or consent from the copyright holders—and the automated "copyright" takedowns of private individuals' meticulously researched fair-use remix and commentary works on media platforms highlights the differences in power and punishment for private versus industrial use. Legislation, communal guidelines, and ethical dimensions of AI technologies for artistic practices are ongoing societal discussions.

At the time of writing of this paper, Italy has banned the popular ChatGPT, citing security concerns for Italian citizens, in a move that may be followed by other countries, and will not allow access to this new technology until their data protection to-do list is implemented [17], whereas countries such as Japan are considering softer regulations in which content usage for AI model training is widely allowed [18]. As an alternative to draconian measures, it is long overdue, and this is the perfect moment in time to include artists and practitioners proactively in these discussions. Accommodating and fusing different voices and knowledge is a must for the reformation of equity, equality, and justice in AI technology creation. Art is for everyone, and the tools we use to make art, especially AI tools, should enable and empower just and equitable creation.

Acknowledgments

The first, second, and third authors share lead authorship. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

References and Notes

1 F.-A. Croitoru et al., "Diffusion Models in Vision: A Survey," arXiv (2022), doi: 10.48550/arXiv.2209.04747.

2 F. Carnovalini and A. Rodà, "Computational Creativity and Music Generation Systems: An Introduction to the State of the Art," *Frontiers in Artificial Intelligence* **3** (2020) p. 14, doi: 10.3389/frai.2020.00014.

3 K. Tatar and P. Pasquier, "Musical agents: A typology and state of the art towards Musical Metacreation," *Journal of New Music Research* **48**, No. 1, 56–105 (2019), doi: 10.1080/09298215.2018.1511736.

4 D. Bisig, "Generative Dance—a Taxonomy and Survey," *Proceedings of the 8th International Conference on Movement and Computing* (2022) pp. 1–10, doi: 10.1145/3537972.3537978.

5 T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University—Computer and Information Sciences*, **34**, No. 6, Part A, 2515–2528 (2022), doi: 10.1016/j.jksuci.2020.04.001.

[6] M. Haataja and J.J. Bryson, "Reflections on the EU's AI act and how we could make it even better," TechREG™ Chronicle (2022).

[7] Michael Wooldridge, *An Introduction to MultiAgent Systems* (West Sussex, U.K.: John Wiley & Sons, 2009).

[8] M.A. Boden and E. A. Edmonds, "What is generative art?" *Digital Creativity* **20**, No. 1, 21–46 (2009), doi: 10.1080/14626260902867915.

[9] Joss Fong, "AI Art, Explained," *Vox* (1 June 2022): https://www.youtube.com/watch?v=SVcsDDABEkM (accessed 16 April 2023).

[10] Holly Herndon, "Holly+": (13 July 2021): https://holly.mirror.xyz/54ds2IiOnvthjGFkokFCoaI4EabytH9xjAYy1irHy94 (accessed 12 April 2023).

[11] I. Arrieta-Ibarra et al., "Should We Treat Data as Labor? Moving beyond 'Free,' " *AEA Papers and Proceedings* **108** (May 2018) pp. 38–42, doi: 10.1257/pandp.20181003.

[12] Stephen Wolfson, "Fair Use: Training Generative AI," *Creative Commons* (17 February 2023): https://creativecommons.org/2023/02/17/fair-use-training-generative-ai/ (accessed 12 April 2023).

[13] Scott Fitsimones, "The DAO Handbook: How Internet Strangers Are Building Collective Movements," *The DAO Handbook* (19 January 2023): https://www.daohandbook.xyz/ (accessed 12 April 2023).

[14] A. Birhane et al., "Power to the People? Opportunities and Challenges for Participatory AI," *Equity and Access in Algorithms, Mechanisms, and Optimization,* Article No. 6, 1–8 (2022), doi: 10.1145/3551624.3555290.

[15] A. Luccioni et al., "Stable Bias: Analyzing Societal Representations in Diffusion Models," arXiv:2303.11408 (2023), doi: 10.48550/arXiv.2303.11408.

[16] Lorena O'Neil, "These Women Tried to Warn Us About AI," *Rolling Stone* (12 August 2023): https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367 (accessed 18 August 2023).

[17] Elvira Pollina, "ChatGPT can resume in Italy if meets data watchdog's demands.", *Reuters* (13 April 2023): https://www.reuters.com/technology/italy-lift-curbs-chatgpt-if-openai-meets-demands-by-end-april-data-protection-2023-04-12 (accessed 16 April 2023).

[18] "Intellectual Property Plan Signals Reversal on AI Policy," *Yomiuri Shimbun* (10 June 2023): https://japannews.yomiuri.co.jp/politics/politics-government/20230610-115423 (accessed 14 August 2023).

[19] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence* **1**, No. 9, Article 9 (2019): 10.1038/s42256-019-0088-2.

Manuscript received 13 December 2022.

KIVANÇ TATAR *is a musician, artist, technologist, and researcher who works in the intersection of machine learning, artificial intelligence, music, interactive arts, and design. The computational approaches developed through that interdisciplinary research have been integrated into musical and audiovisual performances, interactive artworks, and immersive environments including virtual reality. Tatar is currently an assistant professor and a WASP-HS fellow at Chalmers University of Technology, starting a new research group connecting art, music, technology, and artificial intelligence.*

PETTER ERICSON *is a postdoctoral fellow in the research group for Responsible AI at Umeå University in Sweden, working on graph problems and formal descriptions of structured data, with a strong interest in ethics, music, and society. His recent research interests include anti-capitalist artificial intelligence and circumventing political and structural barriers that bar AI from being used to support democratic and egalitarian values. Outside of academia, his musical interests have led him to everything from seedy late-night jam sessions at Copenhagen jazz clubs to organizing 24-hour hackathons around producing electronic music from ESA data.*

KELSEY COTTON *is a vocalist-artist-mover working with experimental music, musical artificial intelligence, and human-computer interaction. Passionate about somatic interaction, the potential for intersomatic experiences between fleshy and synthetic bodies, and first-person feminist perspectives of musical AI, Cotton is pursuing doctoral studies in interactive music and AI at Chalmers University of Technology.*

PAOLA TORRES NÚÑEZ DEL PRADO *is pursuing doctoral studies at the Stockholm University of the Arts, focusing on researching and developing hybrid interactive textile sound interfaces that include the use of AI systems.*

ROSER BATLLE-ROCA *is pursuing doctoral studies in transparent AI and music generation at Universitat Pompeu Fabra in Spain, in collaboration with JRC-EC and Sony.*

BEATRIZ CABRERO-DANIEL *is a postdoctoral fellow at Gothenburg University in Sweden, currently researching trustworthy AI.*

SARA LJUNGBLAD *is a researcher and senior lecturer at Gothenburg University and Chalmers University of Technology in Sweden, doing critical robotics, studying people's experiences and use of robotic products and autonomous systems in everyday settings in the field of human-robot interaction.*

GEORGIOS DIAPOULIS *is pursuing doctoral studies in gestural interaction with generative algorithms for machine musicianship at Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden.*

JABBAR HUSSAIN *is pursuing doctoral studies at Gothenburg University in Sweden in Informatics in the area of trustworthy/responsible AI.*

Paper II

Caring Trouble and Musical AI: Considerations towards a Feminist Musical AI

K. Cotton, K. Tatar

AI Music Creativity (August 2023)

https://aimc2023.pubpub.org/pub/zwjy371l

ABSTRACT

The ethics of AI as both material and medium for interaction remains in murky waters within the context of musical and artistic practice. The interdisciplinarity of the field is revealing matters of concern and care, which necessitate interdisciplinary methodologies for evaluation to trouble and critique the inheritance of 'residue-laden' AI-tools in musical applications. Seeking to unsettle these murky waters, this paper critically examines the example of *Holly+*, a deep neural network that generates raw audio in the likeness of its creator Holly Herndon. Drawing from theoretical concerns and considerations from speculative feminism and care ethics, we care-fully trouble the structures, frameworks and assumptions that oscillate within and around *Holly+*. We contribute with several considerations and contemplate future directions for integrating speculative feminism and care into musical-AI agent and system design, derived from our critical feminist examination.

Introduction

Growing concerns of algorithmic bias and oppression [1] [2] [3] [4] [5]; dataset ownership and data access [6]; and general lore [7] around what AI troubles in our capitalist society is of increasing concern within popular discourse [8] [9]. We see this as an urgent area of concern within musical applications and contexts, which see the integration and assimilation of 'residue-laden' AI systems into musical praxis, musical artworks, and even as synonymous with prominent practitioners.

Within the field of musical-AI, current discourse has examined development of novel tools and architectures for creation and performance [10], artistic potentials in friction and fallacy [11], and human-AI musical interaction [12][13]. Our motivation in this paper is to address longitudinal concerns around the embedding of values, and implications for musical futures. This has previously been alluded to in the literature [14], and we bring a broadening shift towards how AI technologies are changing the landscape of musical creativity [15]. Existing work [16] in evaluating and critiquing AI technologies deployed in performance and artwork contexts has argued for context-specific approaches to system evaluation, but with little exploration into inter-contextual framings of the technology. Emerging concerns regarding AI's influence in curating and shifting musical culture have been outlined in [15] [17], with propositions for policy interventions and the need for future research to examine alternative economic models, longitudinal study and greater diversity.

In this paper, our approach to addressing the myriad of concerns is to examine novel approaches of analysis, which contribute to the field by revealing pathways for workable and ethical practices in the design of musical-AI systems. To do so, our critical examination therefore requires the inclusion of knowledge from other disciplines [18]. We therefore draw together perspectives and methodologies from AI-ethics, Human-Computer Interaction (HCI), science and technology studies (STS), and feminism. The intention with such interdisciplinarity is to uncover and to situate the 'matters of concern' [19] particular to (and within) the field, as evident in our evaluation of our case study. We see interdisciplinarity in approaches to musical-AI as vitally necessary for the community to consider the implications of AI-artworks put out into wider society.

It's now time to look at what's beneath the murky surface of Musical-AI, and to unsettle the water. To assist in our exploration of how concerns are echo-ed between STS, HCI, and care ethics, we care-fully[1] trouble dimensions of Holly Herndon's artwork *Holly+*, a deep neural network that generates audio reminiscent of Herndon's unique vocal aesthetic. We have chosen *Holly+* as a precursory example of how an artist has navigated popular media discourses; collaborative approaches to identity concerns; and articulations of self-governance in the construction and presentation of the artwork. We see our care-full troubling of *Holly+* as an example in taking a step forward from the initial discussions we posed in [20]. In [21] we advocated and argued for conversations on data to expand beyond a generalist and larger-society-centric viewpoint, to elicit domain-specific conversations in the field of musical AI. *Holly+* is both a starting point for the conversations that need to be had, and an example of an artwork that is actively excavating these issues and walking across the borders.

Our contributions here are multifold. First, we contribute with an example of an interdisciplinary analysis, drawing from 3 methodologies in STS, HCI and AI ethics: speculative feminism [22][23]; matters of concern [19] and care [24]; and feminist data ethics [25]. Second, our interdisciplinary analysis reveals 'matters of concern' [19] in the case study, which we argue have implications on the musical-AI community in issues of data, management and legacy. This contribution is augmented by our inclusion of a Knowledge Map [26], to help visibilise the 'matters of concern' and the potential connections between concerns. Third, we outline considerations and future directions for embedding speculative feminism and care into musical-AI design. The stance that we occupy is a *first* step towards provoking change within our community, guided by our concerns as designers, artists, developers and users of the selfsame systems and technologies we critique.

To provide a structural outline of this paper: in the forthcoming Background section, we provide a brief summary of theoretical perspectives and core concepts critical to this paper's inquiry. Chief amongst these are: compounding stances of fact, concern and care; a glimpse into speculative feminist perspectives in STS and HCI; and an overview of current practices in feminist AI-ethics. Drawing upon this theoretical grounding, we then progress to our critical examination of *Holly+*. Extrapolating from our critique, we close by outlining important considerations for inviting speculative feminism into wider discourses on musical-AI and speculate on future directions.

Background

This paper draws heavily from a breadth of theoretical intersections and disciplines—forming the interdisciplinary foundation of our analysis—and which we will take a moment to address now.

Matters of Fact, Matters of Concern and Matters of Care

As we examine various dimensions within *Holly+*, Bruno Latour's notion of 'matters of fact' and 'matters of concern' provide assistive concepts [19]. 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. When we consider an AI-agent or AI-system structure as further constituting the objective (matters of fact) in relation to wider contexts (matters of concern), this necessitates a deliberate and care-full troubling [19]. de la Bellacasa unsettles matters of fact and matters of concern [27] with 'matters of care' [24]. Care is defined as engaging with the *becoming* of matters of fact and concern: an intentional *seeking out* of the histories and values present in systems. It is through *seeking out* the histories and contextual entanglements of technologies that *care* is enacted. We see the contribution of these theories as assistive in attending to the positioning of *Holly+* in relation to its intersecting contextual environments, and its becoming. We attend to one of these intersecting environments in our following discussion of feminism in STS.

Feminism in STS

Looking towards the *what* of what feminism *is*, this helps to formulate the *why* and to establish what feminism can do for musical AI in addressing the inequities, questionable practices and working cultures that are being developed within the music-technology community. This is especially pertinent to musical-AI, which often inherits and adopts models, algorithms and approaches which may carry residues of inequity and bias. Feminist principles can be of benefit here. Speaking broadly, feminism is a series of socio-political movements [28] [29] [30] which seek to address systems of oppression within society, which can encompass gender, political, economic, personal, and social inequalities [31] [32] [33] [34] [35]. Notable work in STS in this regard is the work of Donna Haraway [22] [23] [36], offering speculative feminist narratives of socio-cultural co-construction and fusion of synthetic and organic bodies. Of specific relevance to this paper is their Implosion Analytical Method [37] [38], which critically examines various dimensions of an artefact.

Feminist Stances and Methodologies in AI and Data Ethics

Looking into broader communities, there is ongoing discourse into the formation and implementation of feminist perspectives [39] [40] [41] in data ethics. Specifically, we will now draw attention to Carroll et al [42], and Gray and Witt [25], who examine critical concerns pertaining to issues around data sovereignty, AI ethics, care, and feminist research ethics within AI.

Carroll et al [42] formulate a care-centric data practice, building upon critical concerns pertaining to data sovereignty and self-management within Indigenous communities from Oceania, the United States and Canada. They offer a set of principles—CARE[2]— to complement an existing approach to data management [3] [43]. They articulate the objectives of the CARE principles as constituting people-; purpose-; and data-centric concerns. In our analytical approach, we specifically work with these as lenses for our analysis.

Gray and Witt [25] formulate a preliminary roadmap for integrating a feminist data ethics of care framework within the field of AI. They argue that the ambiguity around mainstream understandings of AI-ethics lends itself to 'fuzzy' definitions, enabling systematic failure in responsibility which in turn implicitly reinforces gender-power imbalances. Of particular note to this paper is their focused attention to both the actors (the 'who') and the practicalities (the 'how') in bringing feminist approaches and methods as a remedy-of-sorts to the principle-to-practice gap. They frame this as 'making interventions' into the economy of machine learning. They propose 5 interventionist principles for feminist data ethics and care. These encompass: 1) diversity with regards to representation and participation in the machine learning economy; 2) critique of positionality; 3) foregrounding human(s) throughout a machine learning pipeline; 4) ensuring the implementation of accountability and transparency measures; and 5) equitable distribution of responsibility. It is these 5 principles that we have identified as assistive lenses for our critique of *Holly+*.

Gray [44] expands upon their earlier paper with Witt, articulating perspectives on how the development and advancement of AI ethics will not see significant, positive change until all stakeholders take on the responsibility of engaging with ethical work and practices throughout the entirety of the economy of machine learning. They highlight that the current landscape is "dominated by a heteropatriarchal class of men", referring to the work by Chang in [45]. Gray underlines the burning need for people working within technology fields to radically change the existing culture of these fields, which they propose as key to "build[ing] capacity for care throughout the entire machine learning economy".

Concerning Matters and Care-fully Troubling *Holly+*

In this section, we draw together theory and methods of speculative feminism, care-ethics, and feminist data-ethics and utilise these as critical evaluative tools in analysing the artwork *Holly+* by Holly Herndon. We proceed by first providing a brief overview of what *Holly+* is, and then delve into our speculative feminist and care-ethics informed critique.

A few words about our approach to critique of *Holly+*

We conduct our critique from a particular point-of-access, in which we occupy a position as spectators and observers of the work. We see this as in coherence with the observable intent that Herndon wishes for their work to be experienced. Our occupancy of this stance is deliberate. We have utilised only publicly accessible information regarding the *Holly+* artwork, encompassing information regarding the general model structure, its governance, affiliated parties and Herndon's recorded statements regarding this artwork. This is so that we may critically examine what is 'visibilised' in and about the work, so that we may in turn be able to critically address the components of the work that appear 'invisibilised' [46]. Our understanding of these terms proposed by Hampton—'invisibilise' and 'visibilise'—as an active choice in what components of a system are seen versus unseen and acknowledging that these choices are (potentially) accompanied by harms. We see invisibilisation and visibilisation as core concerns in connection with Gray and Witt's 4th interventionist principle: accountability and transparency.

About *Holly+*

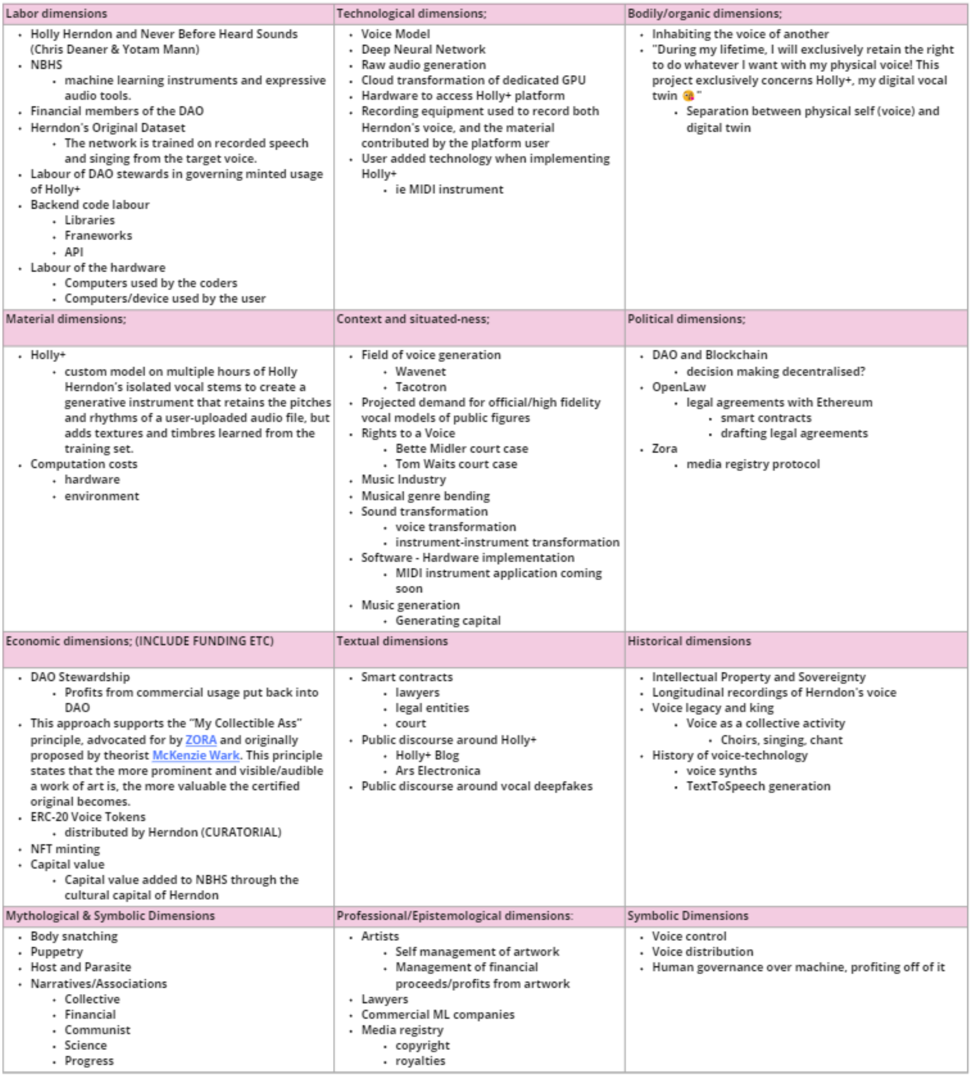
Created by artist Holly Herndon, the work *Holly+* is a voice model built in collaboration with Never Before Heard Sounds (NBHS hereafter), a music studio devoted to the construction, development and deployment of AI powered tools for browser-based musical production. Structurally, it is a custom deep neural network trained upon recorded voice data (constituting singing and/or speech) of Herndon, deployed as a browser-based tool where prospective users upload an audio file (presumably of their own, or publicly sourced recorded material). The model utilises pitches and rhythms from the uploaded audio file, adding additional components from the training data provided by Herndon [47]. The browser-based platform with which one can engage with *Holly+* is presented in Figure 1 below, and accessible for engagement via the following link.

Visit the web version of this article to view interactive content.

FIGURE 1. Holly+ Browser-Based Platform

Care-full Troubling

As a starting point for our critical evaluation, we drew inspiration from Haraway's Implosion [48] methodology as delineated by Dumit [38] to formulate a Knowledge Map and preliminary index (see Figure 2 below) of the various dimensions and structures oscillating within and around *Holly+*. We highlight, that of the 14 dimensions described by Dumit, our Knowledge Map below consists of 12 dimensions, delimited due to the scope of this paper.



Labor dimensions	Technological dimensions;	Bodily/organic dimensions;
• Holly Herndon and Never Before Heard Sounds (Chris Deaner & Yotam Mann) • NBHS • machine learning instruments and expressive audio tools. • Financial members of the DAO • Herndon's Original Dataset • The network is trained on recorded speech and singing from the target voice. • Labour of DAO stewards in governing minted usage of Holly+ • Backend code labour • Libraries • Framewroks • API • Labour of the hardware • Computers used by the coders • Computers/device used by the user	• Voice Model • Deep Neural Network • Raw audio generation • Cloud transformation of dedicated GPU • Hardware to access Holly+ platform • Recording equipment used to record both Herndon's voice, and the material contributed by the platform user • User added technology when implementing Holly+ • ie MIDI instrument	• Inhabiting the voice of another • "During my lifetime, I will exclusively retain the right to do whatever I want with my physical voice! This project exclusively concerns Holly+, my digital vocal twin 😘 " • Separation between physical self (voice) and digital twin
Material dimensions;	Context and situated-ness;	Political dimensions;
• Holly+ • custom model on multiple hours of Holly Herndon's isolated vocal stems to create a generative instrument that retains the pitches and rhythms of a user-uploaded audio file, but adds textures and timbres learned from the training set. • Computation costs • hardware • environment	• Field of voice generation • Wavenet • Tacotron • Projected demand for official/high fidelity vocal models of public figures • Rights to a Voice • Bette Midler court case • Tom Waits court case • Music Industry • Musical genre bending • Sound transformation • voice transformation • instrument-instrument transformation • Software - Hardware implementation • MIDI instrument application coming soon • Music generation • Generating capital	• DAO and Blockchain • decision making decentralised? • OpenLaw • legal agreements with Ethereum • smart contracts • drafting legal agreements • Zora • media registry protocol
Economic dimensions; (INCLUDE FUNDING ETC)	Textual dimensions	Historical dimensions
• DAO Stewardship • Profits from commercial usage put back into DAO • This approach supports the "My Collectible Ass" principle, advocated for by ZORA and originally proposed by theorist McKenzie Wark. This principle states that the more prominent and visible/audible a work of art is, the more valuable the certified original becomes. • ERC-20 Voice Tokens • distributed by Herndon (CURATORIAL) • NFT minting • Capital value • Capital value added to NBHS through the cultural capital of Herndon	• Smart contracts • lawyers • legal entities • court • Public discourse around Holly+ • Holly+ Blog • Ars Electronica • Public discourse around vocal deepfakes	• Intellectual Property and Sovereignty • Longitudinal recordings of Herndon's voice • Voice legacy and king • Voice as a collective activity • Choirs, singing, chant • History of voice-technology • voice synths • TextToSpeech generation
Mythological & Symbolic Dimensions	Professional/Epistemological dimensions:	Symbolic Dimensions
• Body snatching • Puppetry • Host and Parasite • Narratives/Associations • Collective • Financial • Communist • Science • Progress	• Artists • Self management of artwork • Management of financial proceeds/profits from artwork • Lawyers • Commercial ML companies • Media registry • copyright • royalties	• Voice control • Voice distribution • Human governance over machine, profiting off of it

FIGURE 2. Knowledge Map featuring our preliminary exploration of utilising Haraway's Implosion to reveal matters of concern in *Holly+*

We then troubled the artwork through extrapolating connections extending from our central matters of concern data and identity; management and reclamation; and preservation and protection of legacy. We connected these matters of concern to principles from Carroll et al's CARE Data Principles [42], and Gray and Witt's [25] Feminist Data Ethics praxis. With regards to CARE principles, we utilised their larger categorisations of people-, purpose- and data-oriented concerns to probe how our matters of concern revealed in *Holly+* may be motivated through these larger CARE categorisations. Similarly, Gray and Witt's feminist data ethics principles were engaged as critical lenses of how matters of concern in *Holly+* may or may not be coherent with a feminist data ethics. This can be seen in Figure 3 (below), which depicts our three-layer methodological approach to analysis.

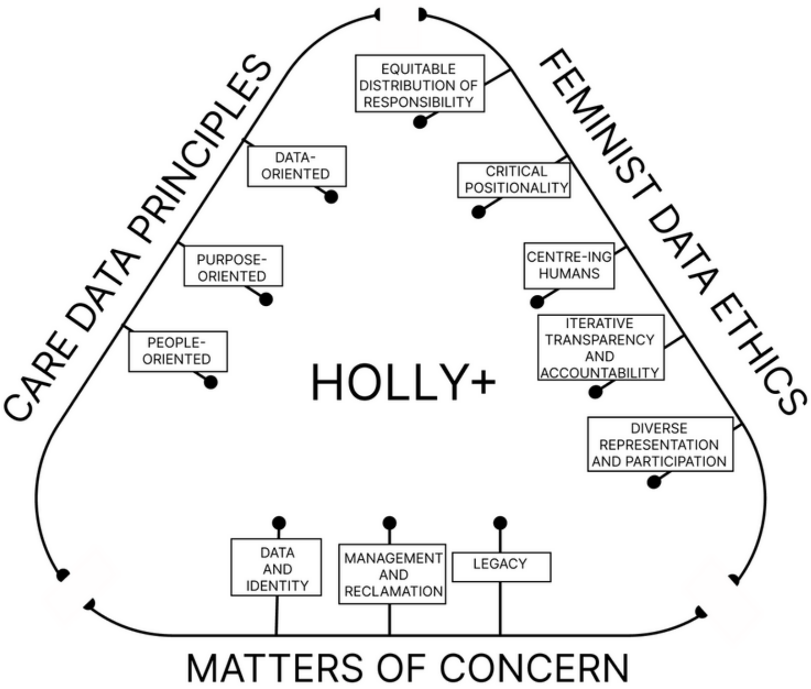


FIGURE 3. Our three-layer methodological approach to analysis, incorporating Feminist Data Ethics, Care Data Principles and Matters of Concern in *Holly+*

Our Haraway Index was highly generative in illustrating dimensions with multiple entanglements to our central matters of concern—data and identity; management and reclamation; and legacy. Of the 12 dimensions we evaluated, the most richly entangled were the technological; labour; political; and economic dimensions. We have utilised these 4 dimensions as additional Latour-informed scenography to our feminist and care-centric analysis.

Matters of Concern in *Holly+*

From our positionality outlined above, we identify three main pillars of concern pertaining to and within *Holly+*, concerning data and identity; management and reclamation; and legacy[4].

From Data and Identity to Management and Reclamation

One especially notable aspect of *Holly+* is the novel approach Herndon adopts in the management of artistic work engaging with the voice model as a generative tool. *Holly+* is a publicly accessible tool, and Herndon motivates her decision in open-access as an intention " …to decentralize access, decision making and profits made from my digital twin, Holly+ …" [47]. Here, we wish to cast a critical gaze over the particularities of how principles and modus operandi of speculative feminism and care ethics may (and may not) be embedded in the procedures, presentation and adjacent framings of this artwork.

It is clearly disclosed on Herndon's personal webpage that a Decentralised Autonomous Organisation (DAO) [49] stewards artistic work that deploys *Holly+*. For contextual grounding with respect to how we proceed with our critique in light of the DAO[5] stewardship, Herndon has previously engaged in discussion around decentralisation within AI-arts [50], and the reclamation of ownership of one's (literal) voice in an age of increasing concern of ethical implications of vocal deepfakes and voice synthesis [51] [52] [53] [54] [55].

They argue that the distribution of tools such as those offered through *Holly+* is in alignment with values pertaining to communality and commonality of voice. Further, they argue for DAO as a means to enable ethical, officially sanctioned and informed experimentation with another's vocal likeness and further enabling communal financial benefit in economic proceeds generated from the use of a voice model. We, however, argue that the deployment of the DAO is in fact not substantially decentralising decision-making. We argue that significant decision-making which has implications for how *Holly+* has been made and can be used, has clearly already been established by Herndon and NBHS in their design of the system, the means of interacting with *Holly+*, and the terms of agreement within the stewardship itself. The potential responsibility of stewards is thus delimited to governing 'fair-usage' of minted artworks created with *Holly+*, and not affording governance of the evolution of the architecture of *Holly+* over time. We therefore do not see the full scope of decision-making pertaining to *Holly+* as *fully* decentralised.

Legacy

Herndon describes one of their underlying motivations for the birth of *Holly+* as an act of futuring and "maintaining the value and reputation of [their] voice [rather] than the rights being passed down to someone less familiar with the values and standards associated with [their] work". Their justification for this is grounded in concerns that inherited rights—through a next-of-kin or other Western-centric inheritance tradition—offer less posthumous protections than a public and digital distribution of governance. We do not critique Herndon's expression of feeling more comfort in distributed ownership of her voice model, we do however note interesting and "sticky" concepts entangled with this pertaining to the matter of the public following of *Holly+* and the DAO stewardship.

The first "sticky" concept we wish to highlight is the entrance procedure of the DAO [56] stewardship. Herndon outlines how membership into the DAO is contingent on the distribution of ERC-20 VOICE tokens

[57] which are on the Ethereum blockchain [58]. These tokens represent voting shares in *Holly+ DAO*. These tokens will be "airdropped to collectors of my art, friends and family of the project, and other artists selected to participate in using the *Holly+* voice to create new works." We can therefore plainly assume, that *Holly+DAO* stewards, either already have a vested financial interest in Herndon's work (in the example of collectors), are already intimately familiar with Herndon and their work (friends and family of the project), or have been deemed by Herndon as possessing sufficient technical competency or musicality to create 'suitable enough' artwork using *Holly+* (other artists selected to participate in using the *Holly+* voice). We argue that the procedure for becoming a DAO stewards is highly selective, curatorial and holds the potential for exclusion based on cultural capital, digital accessibility and economic status.

This leads into our second concern, the preservation of legacy. The formation of culture does not take place in a vacuum, and there are (potentially) deeper issues in anticipating that one's values and standards may be preserved for the future production of artwork taking place in a future environment and context that we cannot yet imagine. How might the *Holly+DAO* stewards in 100-years' time be best suited or situated to make decisions that honour Herndon when living memory of Herndon as an artist may no longer exist? This assumption can be further troubled by speculating how applicable or relevant the cultural values or artistic standards of an artist may be in this selfsame future context. Stickiness and murkiness reside in the question around what constitutes an appropriate, or artistically relevant, usage of Herndon's voice especially when voting stewards may approve an offensive or uncharacteristic deployment of *Holly+* [59]. The premise of *Holly+* as an artwork is grounded—and indeed, dependent—in public interaction. The greater the engagement levels, the greater the social value that is attributed to the artwork. The DAO incentivisation scheme concretises this 'value of attention', distributing profits of artworks made with *Holly+* amongst stewards to encourage their decision-making in 'minting' usages of the voice model to increase the social capital (i.e. the visibility and distribution) of *Holly+*. We speculate on the potential stickiness of this in regards to Herndon's intention to preserve their artistic legacy. This proves especially troublesome a notion, especially when we acknowledge that the economic profits generated by any usage [60] may indeed subvert Herndon's own vision for fiscally incentivising DAO stewards to preserve her artistic legacy. Money talks, controversy sells, and the distinction between 'acceptable' and profitable are not necessarily kept apart [61].

The visibilised and the invisibilised

Our Haraway Implosion Index further revealed imbalances pertaining to what was visibilised and invisibilised in the artwork, which we wish to draw attention to. The first is the involvement with NBHS, which we understand as having been involved in the design and development of *Holly+*. Adjacent to this invisibilised element is the code, which has not been made accessible anywhere that we could locate. Presumably, any open-source access to the code has been dismissed by the immediate partners in *Holly+* (which we speculate encompasses Herndon and NBHS) in the protection NBHS's commercial interests as an organisation developing online AI music tools. Second, we note also that there is ambiguity as to the sourcing of the original dataset of Herndon's voice, and whether this dataset was compiled specifically *for* the training of the *Holly+* model or constitutes Herndon's historical vocal data.

CARE-ing values and *Holly+:* People-, purpose- and data-centric values

We turn now to our critical examination of people-, purpose- and data-centric values in *Holly+* in relation to our matters of concern. The larger structural design of *Holly+* as an artwork reflects values around collective interaction with identity, through the capacity to re-realise an audio recording through the consensual usage of Herndon's vocal likeness. Here, we understand that *Holly+* is prioritising both people- and purpose-centric values: by serving as a model for establishing working processes for consensual and collective engagement with identity play [62]. We understand this concentration of collective engagement and explorative play with *Holly+* as likewise reflecting principles of care- in 'demystifying' AI-tools through open-access play, and through Herndon's attention to how future demand of voice models must be informed by a system of governance that is in the best interests of the voice-origin.

Further, the legal and financial structures that govern the verified usage of *Holly+* reflect a concern towards the people-, purpose- and data-centric values proposed by Carroll et al. Herndon's vocal likeness is established as connected with their personhood[6] and image as an artist- and therefore Herndon as implicated in any future utilisation of *Holly+* in an artistic work. The DAO stewardship system appears to address these values by protecting the verified usage Herndon's vocal likeness and artistic legacy whilst enabling collective participation in the formation of musical subcultures that would utilise the usage of an artist's voice in a posthumous context [63]. We can understand this as establishing a prioritisation towards people (the stakeholders in *Holly+*; purpose (Herndon's view of future voice model usage); and data (Herndon's voice and artistic legacy as data).

However, when we further position the *Holly+* DAO in relation to matters of fact and concern, we observe a conflict of people-versus-purpose centric values. From a matters of fact position, Herndon is reclaiming ownership (of data; of voice; of their likeness) with NBHS, and enabling open-source access to their identity play. However, a matter of concern is that NBHS (and Herndon) are making very rigid decisions about how the model interacts with user-contributed material; the selectivity of what vocal material is added to the dataset the model is trained on; and how the model's verified usage may be distributed (through deploying a *Holly+DAO)*. Here is where the conflict lies: the conflict between the open-access intentionality of Herndon, with the closed-system development of the *Holly+* architecture.

Feminist Data Ethics and *Holly+*

When we take additional lenses from Gray and Witt's 5 feminist data ethics principles, we can further understand that the matters of concern in *Holly+* become somewhat more tremulous, or ambiguous. By this, we mean that our matters of concern pertaining to *Holly+* are troubled when examined through the 5 feminist principles formulated by Gray and Witt. This therefore requires our care and attention.

With regard to the first principle—equitable distribution of responsibility—although the structure of governance of *Holly+* through the DAO aims to equitably divide decision-making power, we do not see this distribution as *truly* equitable. As had previously been discussed, the distribution of stewardship tokens is contingent on either a financial and/or labour investment in Herndon's artistic work; a familial or network (by this we also presume a cultural capital) connection to Herndon; or bestowment of a token based on Herndon's assessment of the recipient's artistic merit or capacity. We argue that this limits access, by requiring technical capacity and familiarity with—and a secure financial position to invest in—cryptotech. Based on these grounds it can be argued that these pathways to stewardship are in fact not entirely equitable.

The second principle—critical positionality—is more clearly addressed. Herndon has continually articulated their views on the future usage of voice modeling, and the inclusion of Web3 technologies to safeguard the legal interests of artists. In the third principle—the centering of human(s) throughout the pipeline—this becomes more difficult to discern. Naturally there is a centering of Herndon's capacity to share their likeness and encourage collective and creative usage of their identity play. However, we were unable to ascertain specifically how a human-centered approach was applied with regards to data collection or system design. This invites further speculation as to how the valuation of collective play—an apparent concern addressed thematically in *Holly+*—may be more transparently addressed in the design of the model's architecture and its interactivity.

The fourth principle—transparency and accountability— proves similarly more problematic for us to assess. We observed a lack of transparency with regards to the particularities of the vocal model, and specifically to the availability of the code. We assume this as withheld due to the commercial interests of Herndon and NBHS, yet this withholding necessitates clarification. It must not go unacknowledged that *Holly+* is an artwork made in collaboration with 2 entities, one an artist and one an organisation, both with vested interests in preserving certain aspects of code as their intellectual property and holding cultural and financial value.

When we further consider transparency and accountability, it is also not clear how this is factored in with regards to the *Holly+DAO*. On one hand, there is transparency with regards to how the stewardship is implemented to govern verified usages of *Holly+*. On the other hand, we were unable to find any information regarding *who* specifically had been awarded DAO stewardship. This is an area of concern, as the transparency of this system is selectively visibilised and invisibilised- and contrary to Gray and Witt's proposed principle. There is also ambiguity as to the nature of accountability in regards to actions taken by DAO stewards, present and future. We had previously ruminated on potential implications of stewards having to make decisions regarding usage of Herndon's vocal likeness in a speculative future context with—potentially—markedly different values around culture than we have in our present reality.

In the fifth and final principle—diverse representation and participation— this is perhaps the most ambiguous to determine. On the development end, *Holly+* utilises libraries and frameworks which have presumably been constructed by a particular demographic [45], with shared or complimentary technical skillsets. With regards to user engagement, its browser-hosting enables widespread, international participation. However, access is contingent on computer or smartphone access and a reliable internet connection- which can be significant factors of exclusion.

Burning Concerns and Speculative Future Directions towards a Feminist Musical AI

Our critique reveals that AI-systems carry many residues: unintentional and intentional imprints from the datasets they have been trained on[7]; the algorithms they have been made with[8]; and the actors who then inherit or use these systems in varying applications. We see a critical need for these residues to be address, and we propose interdisciplinary feminist methods as a means to do so. This constitutes work in 3 categories: de-centralisation of management; preservation and protection of legacy; and the critical prioritisation of people-, process- or data-oriented principles.

With regards to de-centralisation, we see potential in exploring alternative structures to put ‘power’ back in the hands of people accessing and making the artworks, rather than a board of directors from a recording label [64]. We foresee this matter of concern as being a crucial area for future troubling of existing power structures within the music industry.

In terms of preservation and protection of legacy, we have seen the deployment of Web3 legal and financial technologies troubling current modus operandi protecting artistic legacy. Through engagement with novel systems of artwork stewardship, we foresee future ‘unsettling of the waters’ with how artists can protect or distribute ownership and management of their data[9] and artistic legacy. We foresee such engagement with AI technologies eliciting profound changes in the preservation and protection of the legacy of an artist.

A final[10] matter of concern is the critical prioritisation of people-, process- or data-oriented principles. We have seen these navigated in *Holly+* through the concern for the future of ethical voice model utilisation (people- and process-oriented) and subsequent implementation of DAO governance. We propose 2 central questions that researchers may utilise as a preliminary step in their implementation of feminist and care-full methods in musical-AI design: ‘*who and/or what is invisibilised?*’; and conversely, ‘*who and/or what is visibilised?*”

Future Work

This paper is an initial peek into implementations of feminism and care ethics into musical-AI. The scope of the matters of concern addressed in this paper is substantial, with important future work needed with regards to further analysis required of hegemonic power structures within the field of musical-AI, and evaluation of how barriers of access—linguistic, social and digital—are implemented throughout the economy of AI. We further anticipate that the presentation of practical examples of conscious engagement with matters of fact, concern and care will form the basis of our future work in this regard.

Conclusion

Within this paper, we have opened and stepped into a critical space within which we have troubled the waters of Musical-AI. We have outlined existing research work that engages with the implementation of feminist discourse, perspectives and methodologies across disciplines such as science and technology studies (STS), care ethics, and AI and Data Ethics. We have taken up a critical feminist lens of a musical-AI artwork—*Holly+*—and care-fully troubled the various dimensions within this artwork that we see as collective matters of concern across intersecting disciplines negotiating tensions around AI. Through our interdisciplinary analytical approach, we have revealed matters of concern which pose future troubling of power structures within the music industry. Further resulting from our preliminary critique through a speculative feminist lens, we address the burning matters of concern within Musical-AI and outline potential directions for future work in troubling inherited tools, systems, methodologies and lore around artistic and musical AI use.

Ethics Statement and Acknowledgements

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

Footnotes

[1] The hyphenation is intentional here, and throughout ↩

[2] Collective benefit; Authority to control; Responsibility; Ethics ↩

[3] The ‘FAIR Guiding Principles for scientific data management and stewardship’ ↩

[4] As indicated in Figure 3 ↩

[5] A DAO is a community-led entity that governs decision-making processes of a product or project its operating protocols are formulated into a smart contract which is written onto the blockchain. DAO members receive a distribution of profits resulting from the usage of the product , proportional to their financial investment. ↩

[6] By this we do not mean personhood in the biological sense, nor a quality of the voice (such as timbre). Instead, we are referencing Herndon’s own separation of her vocal self into 2: one occupying her own physical body, and the other (Holly+) as a digital vocal twin. Both are Holly, but occupying different forms. ↩

[7] Which further encompasses the unintentional and intentional leakages of values, assumptions and intentionalities of both the subject whose data has been utilised and the intentionalities of the data collectors ↩

8. Encompassing the unintentional and intentional leakages of values, assumptions and intentionalities of the developers and makers of said algorithms ↩

9. In this case, their artworks ↩

10. In the context of this paper at least. ↩

References

1. Hampton, L. M. (2021). Black Feminist Musings on Algorithmic Oppression. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 1. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445929 ↩
2. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR. Retrieved from https://proceedings.mlr.press/v81/buolamwini18a.html ↩
3. Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, *41*(1), 20–33. https://doi.org/10.1109/MAHC.2019.2897667 ↩
4. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375820 ↩
5. Offert, F., & Phan, T. (2022). *A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space*. arXiv. https://doi.org/10.48550/arXiv.2211.06323 ↩
6. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. arXiv. https://doi.org/10.48550/arXiv.2110.01963 ↩
7. Schwartz. (2018). “The discourse is unhinged”: how the media gets AI alarmingly wrong. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong ↩
8. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. (2021). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%253A52021PC0206 ↩
9. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act)*. (2020). Retrieved from https://eur-lex.europa.eu/legal-

content/EN/TXT/?uri=CELEX%253A52020PC0767↩

10. Carnovalini, F., & Rodà, A. (2020). Computational Creativity and Music Generation Systems: An Introduction to the State of the Art. *Frontiers in Artificial Intelligence*, *3*, 14. https://doi.org/10.3389/frai.2020.00014 ↩

11. Döbereiner, L. (2022). Artistic Potentials of Fallacies in AI Research. *Proceedings of the 3rd Conference on AI Music Creativity, AIMC.*, 5. https://doi.org/10.5281/zenodo.7088311 ↩

12. Trump, S. (2021). Musical Cyborgs: Human-Machine Contact Spaces for Creative Musical Interaction. *Proceedings of the 2nd Conference on AI Music Creativity*, 11. ↩

13. Dahlstedt, P. (2021). Musicking with Algorithms: Thoughts on Artificial Intelligence, Creativity, and Agency. In E. R. Miranda (Ed.), *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (pp. 873–914). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-72116-9_31 ↩

14. Born, G., & Devine, K. (2016). Gender, Creativity and Education in Digital Musics and Sound Art. *Contemporary Music Review*, *35*(1), 1–20. https://doi.org/10.1080/07494467.2016.1177255 ↩

15. Tatar, K., Ericson P., Cotton K., Núñez del Prado P. T., Batlle-Roca R. Cabrero-Daniel, B, Ljungblad S., Diapoulis G., Hussain J. *A Shift In Culture through Artificial Intelligence*. In press. ↩

16. Sturm, B., Monaghan, O., Collins, Ú., Et, D., Year, A., Sturm, B., … Pachet, F. (2018). Machine Learning Research that Matters for Music Creation: A Case Study. *Journal of New Music Research, In Press*. https://doi.org/10.1080/09298215.2018.1515233 ↩

17. Born, G., Morris, J., Diaz, F., & Anderson, A. (2021). *Artificial intelligence, music recommendation, and the curation of culture: A white paper* (p. 27) [Techreport]. University of Toronto; Schwartz Reisman Institute for Technology. ↩

18. Dignum, V., Casey, D., Cerratto-Pargman, T., Dignum, F., Fantasia, V., Formark, B., … Tucker, J. (2023). *On the importance of AI research beyond disciplines*. arXiv. https://doi.org/10.48550/ARXIV.2302.06655 ↩

19. Latour, B. (2014). *What Is the Style of Matters of Concern?* https://doi.org/10.5749/minnesota/9780816679959.003.0004 ↩

20. **Tatar, K.**, Ericson P., Cotton K., Núñez del Prado P. T., Batlle-Roca R. Cabrero-Daniel, B, Ljungblad S., Diapoulis G., Hussain J. *A Shift In Culture through Artificial Intelligence*. In press. ↩

21. Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066 ↩

22. Haraway, D. J. (2016). *Staying with the trouble: making kin in the Chthulucene*. Durham: Duke University Press. ↩

23. de la Bellacasa, M. P. (2011). Matters of care in technoscience: Assembling neglected things. *Social Studies of Science*, *41*(1), 85–106. https://doi.org/10.1177/0306312710380301 ↩

24. Gray, J., & Witt, A. (2021). A feminist data ethics of care for machine learning: The what, why, who and how. *First Monday*. https://doi.org/10.5210/fm.v26i12.11833 ↩

25. Dumit, J. (2014). Writing the Implosion: Teaching the World One Thing at a Time. *Cultural Anthropology*, *29*(2), 344–362. https://doi.org/10.14506/ca29.2.09 ↩

26. Latour, B., & Weibel, P. (2005). *Making Things Public*. Cambridge, Massachusetts: MIT Press. Retrieved from https://mitpress.mit.edu/9780262122795/making-things-public/ ↩

27. States, T. C. R. C. (2019). Monthly Review | A Black Feminist Statement. Retrieved from https://monthlyreview.org/2019/01/01/a-black-feminist-statement/ ↩

28. Coaston, J. (2019). The intersectionality wars. Retrieved from https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination ↩

29. Ahmed, S. (2017). No [Blog]. Retrieved from https://feministkilljoys.com/2017/06/30/no/ ↩

30. Gamble, S. (2001). *The Routledge Companion to Feminism and Postfeminism*. Routledge. ↩

31. Beasley, C. (1999). *What is Feminism?: An Introduction to Feminist Theory*. SAGE Publications. Retrieved from https://libgen.li/ads.php?md5=0c10b7721de68fd4d685b49257df0ce5 ↩

32. Weedon, C. (2002). Key Issues in Postcolonial Feminism: A Western Perspective. *Gender Forum: An Internet Journal for Gender Studies*, (1). Retrieved from https://web.archive.org/web/20131203002056/http://www.genderforum.org/issues/genderealisations/key-issues-in-postcolonial-feminism-a-western-perspective/ ↩

33. Frankenberg, R. (1993). Growing up White: Feminism, Racism and the Social Geography of Childhood. *Feminist Review*, (45), 51–84. https://doi.org/10.2307/1395347 ↩

34. Henry, N., Vasil, S., & Witt, A. (2022). Digital citizenship in a global society: a feminist approach. *Feminist Media Studies*, *22*(8), 1972–1989. https://doi.org/10.1080/14680777.2021.1937269 ↩

35. Haraway, D. (2006). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late 20th Century. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds.), *The International Handbook of Virtual Learning Environments* (pp. 117–158). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-3803-7_4 ↩

36. Haraway, D. J. (2018). *Modest_Witness@Second_Millennium. FemaleMan_Meets_OncoMouse: Feminism and Technoscience* (2nd ed.). Second edition. | New York, NY : Routledge, 2018. | The title is an email: Routledge. https://doi.org/10.4324/9780203731093 ↩

37. Dumit, J. (2014). Writing the Implosion: Teaching the World One Thing at a Time. *Cultural Anthropology*, *29*(2), 344–362. https://doi.org/10.14506/ca29.2.09 ↩

38. Bardzell, S. (2010). Feminist HCI: taking stock and outlining an agenda for design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1301–1310. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1753326.1753521 ↩

39. Bardzell, S., & Bardzell, J. (2011). Towards a feminist HCI methodology: social science, feminism, and HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 675–684. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1978942.1979041 ↩

40. Michelfelder, D. P., Wellner, G., & Wiltse, H. (2017). *Designing differently : toward a methodology for an ethics of feminist technology design*. Rowman & Littlefield International. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-134366 ↩

41. Carroll, S., Garba, I., Figueroa-Rodríguez, O., Holbrook, J., Lovett, R., Materechera, S., … Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, (19), 1–12. https://doi.org/https://doi.org/10.5334/dsj-2020-042 ↩

42. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18 ↩

43. Gray, J. E. (2022). What can feminism do for AI ethics? Retrieved from https://medium.com/mlearning-ai/what-can-feminism-do-for-ai-ethics-b7e401889441 ↩

44. Chang, E. (2018). *Brotopia: Breaking Up the Boys' Club of Silicon Valley*. Portfolio. Retrieved from https://www.amazon.com.au/Brotopia-Breaking-Boys-Silicon-Valley/dp/0735213534 ↩

45. Hampton, L. M. (2021). Black Feminist Musings on Algorithmic Oppression. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 1. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445929 ↩

46. Herndon, H. (2021). Holly+ □ □ □. Retrieved from https://holly.mirror.xyz/54ds2IiOnvthjGFkokFCoaI4EabytH9xjAYy1irHy94 ↩
47. Haraway, D. J. (2018). *Modest_Witness@Second_Millennium. FemaleMan_Meets_OncoMouse: Feminism and Technoscience* (2nd ed.). Second edition. | New York, NY : Routledge, 2018. | The title is an email: Routledge. https://doi.org/10.4324/9780203731093 ↩
48. Shuttleworth, D. (2021). What Is A DAO And How Do They Work? Retrieved from https://consensys.net/blog/blockchain-explained/what-is-a-dao-and-how-do-they-work/ ↩
49. Minsker, E. (2021). Holly Herndon’s AI Deepfake “Twin” Holly+ Transforms Any Song Into a Holly Herndon Song. Retrieved from https://pitchfork.com/news/holly-herndons-ai-deepfake-twin-holly-transforms-any-song-into-a-holly-herndon-song/ ↩
50. Khanjani, Z., Watson, G., & Janeja, V. (2021). *How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey*. ↩
51. Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., & Nie, L. (2022). *Voice-Face Homogeneity Tells Deepfake*. ↩
52. Coldewey, D. (2023). VALL-E’s quickie voice deepfakes should worry you, if you weren’t worried already. Retrieved from https://techcrunch.com/2023/01/12/vall-es-quickie-voice-deepfakes-should-worry-you-if-you-werent-worried-already/ ↩
53. Almutairi, Z., & Elgibreen, H. (2022). A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms*, *15*, 19. https://doi.org/10.3390/a15050155 ↩
54. Khanjani, Z., Watson, G., & Janeja, V. (2023). Audio deepfakes: A survey. *Frontiers in Big Data*, *5*, 1001063. https://doi.org/10.3389/fdata.2022.1001063 ↩
55. Decentralized autonomous organizations (DAOs). (2023). Retrieved from https://ethereum.org ↩
56. What Are ERC-20 Tokens on the Ethereum Network? (n.d.). Retrieved from https://www.investopedia.com/news/what-erc20-and-what-does-it-mean-ethereum/ ↩
57. What is Ethereum? (n.d.). Retrieved from https://ethereum.org ↩
58. Ars-Electronica. (2022). Holly+. Retrieved from https://starts-prize.aec.at/en/holly-plus/ ↩
59. Egan, M. (2020). The art world has a money laundering problem | CNN Business. Retrieved from https://www.cnn.com/2020/07/29/business/art-money-laundering-sanctions-senate/index.html ↩

60. Hahn, L. (2020). Artists and their controversies: a question of profit? Retrieved from https://artgateblog.altervista.org/artists-and-their-controversies-a-question-of-profit/ ↩

61. Live, U. (2022). *IP for the AI Era: Holly+ Presents Identity Play*. Unfinished Live 2022. Retrieved from https://live.unfinished.com/videolibrary/mainstage-i-ip-for-the-ai-era-holly-presents-ident ↩

62. Mbembé, J.-A., & Meintjes, L. (2003). Necropolitics. *Public Culture*, *15*, 11–40. ↩

63. Brook, T. (2023). Music, Art, Machine Learning, and Standardization. *Leonardo*, *56*(1), 81–86. https://doi.org/10.1162/leon_a_02135 ↩

Paper III

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

K. Cotton, K. de Vries, K. Tatar

Movement Computing (March 2024), 1-12

https://doi.org/10.1145/3658852.3659065

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

Paper IV

Sounding out extra-normal AI voice: Non-normative musical engagements with normative AI voice and speech technologies

K.Cotton, K.Tatar

AI Music Creativity (September 2024)

https://aimc2024.pubpub.org/pub/extranormal-aivoice/release/1

This paper contains rich media, which is accessible in the linked Zenodo record.

Author Keywords

AI voice, speech recognition, speech synthesis, AI research-through-design, musical AI

Abstract

How do we challenge the norms of AI voice technologies? What would be a non-normative approach in finding novel artistic possibilities of speech synthesis and text-to-speech with Deep Learning? This paper delves into SpeechBrain, OpenAI and CoquiTTS voice and speech models with the perspective of an experimental vocal practitioner. An exploratory Research-through-Design process guided an engagement with pre-trained speech synthesis models to reveal their musical affordances in an experimental vocal practice. We recorded this engagement with voice and speech Deep Learning technologies using auto-ethnography, a novel and recent methodology in Human-Computer Interaction. Our position in this paper actively subverts the normative function of these models, provoking nonsensical AI-mediation of human vocality. Emerging from a sense-making process of poetic AI nonsense, we uncover the generative potential of non-normative usage of normative speech recognition and synthesis models. We contribute with insights about the affordances of Research-through-Design to inform artistic processes in working with AI models; how AI-mediations reform understandings of human vocality; and artistic perspectives and practice as knowledge-creation mechanisms for working with technology.

Introduction

The human voice is capable of producing a vast range of complex sounds for the purposes of both communication and creative expression [1] [2] [3] [4]. It is simultaneously an instrument with a personal and intimate relation to the vocalist [5] [6]. The rapid development of AI toolkits for generating, synthesizing and cloning human voices has drawn much attention across both mainstream media [7] [8] [9] [10] and research communities [11] [12] [13] [14] [15] [16] [17] [18]. We contribute to this ongoing discussion by looking at what AI models do **for** our understanding of human vocality, not what they do **to** voice. We see this as a critical area of research, as the advancement of voice and speech cloning, synthesis and generation models continues to re-form our understandings of how human vocality is implicated by AI technologies [19]. We envisage that the creative and critical engagement with these technologies establishes novel relations and understandings of what a collaborative human and AI-vocality might mean (and become) [20] [21] [22]. Further, we speculate of this as affording new explorations of creative human vocality. This invites further rumination as to what novel cognitive and expressive understandings of human- and AI-vocality we uncover as a result of engaging with these tools in an artistic practice [23]. Accordingly, this paper takes the research question as a core point of departure: *how does the engagement of AI tools for voice and speech in an auto-ethnographic voice practice contribute to novel understanding of human- and AI-vocality?*

Taking this research question as a guide, we engage with exploratory Research-through-Design to uncover emergent affordances [24] when working with normative AI voice and speech models. Throughout this process, we probed the ways that such models may be forcibly glitched [25] [26] [27] when faced with input material that is contrary to how these models were trained: by using non-text based audio material as input for text-expectant models. This usage of non-text based audio is framed within an auto-ethnographic lens of the first author's[1] experimental vocal practice: how they understand and relate to their own voice data; and the subsequent re-formation of their relation to their own voice after it has been de- and re-constructed by speech recognition, synthesis and cloning models.

From this curious excavation of AI tools for voice and speech, we make a series of contributions. Foremost, we introduce a novel research position on AI voice. That is: critical and creative engagement with AI models in non-normative ways may assist in forming new understandings of human vocality. Second, we contribute with an example of non-normative engagement with normative ASR models. Third, we contribute with an example of auto-ethnographic engagement with a TTS model on a dataset of non-text based vocalisations. Fourth, we establish the importance of interdisciplinarity within musical AI research and demonstrate the generative potential of including perspectives and techniques from human-computer interaction, musical and artistic practices, sound studies, and social sciences. Finally, we contribute an example of an artistic practice that creates knowledge in AI vocality, by active engagement with technology through an artistic perspective and grounded within an artistic practice.

The remainder of this paper is organised as follows. In the **Background** section, we provide a brief theoretical introduction to the methodologies that we utilised for our research in AI tools for voice and speech. In **Methodology**, we provide an overview of the steps that we have taken in our research process. The **Research through Design with Speech AI** section, chronicles the first author's[2] auto-ethnography and RtD explorations with Deep Learning technologies for voice and speech. These encompass more artistic exploration of research problems and the utilisation of self knowledge and understanding as contributing methods for knowledge production. In **Research Findings**, we present our findings emerging from the research logs recorded during the RtD. We address in the **Discussion** how the RtD engagement with AI tools for voice and speech in an auto-ethnographic voice practice contribute novel understanding of human- and AI-vocality, establishing a series of research contributions within this domain.

Background

This section outlines the theoretical background of the research methodologies that we engage with in this paper and contextualises our usage of certain terms particular to vocal practice. We first introduce Research-through-Design, which frames artistic research activities as a form of knowledge creation. This was the research "through-line" of the first author's working process across the various AI models. We then discuss auto-ethnography, which utilises self knowledge [28] documentation techniques as forms of knowledge creation. Auto-ethnographic methods such as journalling and self-documentation have been critical to the first author's artistic process. We then discuss our understanding of 'experimental' voice and vocality, which connects to the overall research question informing this study: engaging with AI tools for voice and speech in an experimental vocal practice.

Research-through-Design[3] is a term introduced by Christopher Frayling in a seminal position paper which sought to contextualise the particular nature of knowledge generation within arts and design [23]. Frayling asserts that artistic research practices constitute their own modes of knowledge generation, and should be considered as a part of a larger academic research context. Frayling critiques the historical dichotomy between research conducted under presumably "scientific" domains, and research conducted in presumably "non-scientific" domains. In their text, Frayling frames "non-scientific" domains as those in connection to a particular craft or practice (such as art, design, etc). Three different forms of research are outlined and defined: research ***for*** art and design, research ***into*** art and design, research ***through*** art and design.

Auto-ethnography is a qualitative research methodology that examines a researcher's own subjective experience, which is analysed and critiqued in reference to wider social, cultural and historical contexts [29][30][31][32]. As a research methodology, it encompasses a range of reflective documentation techniques, including journalling, story-telling, the collection and documentation of mixed media forms (audio, video and photo) [33]. Auto-ethnography as a method has been utilised across a wide array of research domains, including human-computer interaction (HCI). The use of self-knowledge has informed the design and development of technological systems and artefacts that are in close connection to human bodies, or mediate experiences with technology [34] [35] [36].

Though a full overview of experimental voice practices is not feasible within the scope of this paper, we will briefly discuss what is meant by the 'experimental' voice. Our present understanding of what constitutes an "experimental" vocality is informed by singers' engagement of and with *extra-normal* vocalisation techniques [37]. Here, we intentionally avoid using the term 'extended vocal techniques'. The classification of certain vocal techniques (such as whistling, vocal fry, gurgling) as an 'extended' technique for vocalisation is dependent on the musical and cultural context of the voice practice [38]. Like Noble [38], we think that the concept of an 'extended' vocal technique implies a normative vocal technique, which can be rooted in a normative-European vocal technique and aesthetic. We instead utilise the term "extra-normal" as coined by Edgerton in [39], which broadly catalogues the physiological potentials of human vocality, contributing to a rich domain of vocal physiology [40] [41] [42], acoustics [43] and vocality-as-selfhood [44] [45] [46] [47]. Augmentations of voice via technology are further outlined in [48] [49] [50] [51]. As Eidsheim, Edgerton and others have established, the non-modularity and non-uniformity of human voice further establishes vocality as a "technology of selfhood" [44]. Eidsheim puts it succinctly in their thoughts on the self actualisation of voice: "*The production and dissemination of a particular vocal timbre is an act with an impact similar to a speech act. The emission of a particular vocal timbre is a self-presentation...*" [44]

Methodology

Our research methodology in this work consists of several steps in which we bring together Research-through-Design and auto-ethnographic research. Figure 1 gives an overview of the steps and branches that Kelsey followed in the Research-through-Design. Phase 1 is an data creation and engagement process for building an extra-normal voice dataset, which is personalized to the voice of the first author. Phase two dives into two Deep Learning models for automatic speech recognition (ASR) to explore how ASR models would interpret an extra-normal audio input. Phase 3 is a reinterpretation of the original voice dataset to highlight the discrepancies that are introduced by an AI pipeline with ASR and TTS. Our aim in the voice resynthesis in phase 3 is to visibilise the extra-normality in AI voice models. Phase 4 is the artistic interpretation of extra-normality in AI voice. We created an audio sample dataset generated by AI resynthesis of extra-normal voice, to explore the musical materiality of extra-normal AI voice in live coding performances. The details of the Research-through-Design can be found in the next section.

At each step of the Research-through-Design, we employed auto-ethnographic methods to document our engagement with the chosen AI voice technologies. Those engagements are collected into research logs, which we used later to find over-arching, interconnected, and entangled human-AI vocality. Thus, our findings are grouped accordingly:

- Data: Somatic Experience
- Automatic Speech Recognition: From Sound to Text to Poem
    - Visibilising Attention Mechanisms
    - Signs of Scraped Data and Attack of the Emojis
- Live coding as *digital zaum*

The section *Research Findings* gives the details on how we build the emerging concepts from the auto-ethnographic research records.

The Research through Design with Speech AI

In this section, we outline the research process which encompasses exploration with SpeechBrain's [52] and OpenAI's Whisper [53] models for automatic speech recognition; CoquiTTS' XTTS_V2[4] model for text synthesis and cloning; and real-time musical engagement with the resultant cloned and synthesised audio using strudelREPL[5]. The intention—as established in our research question—is to probe the sonic affordances of non-normative usage of normative-ly trained model within an experimental vocal practice. An outline of these working phases is provided below, and will systematically be discussed.

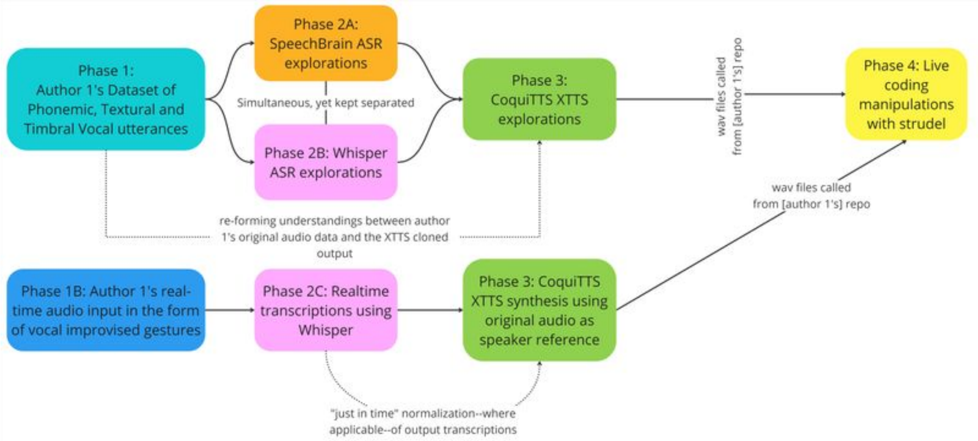

Phase 1: Author 1's Dataset of Phonemic, Textural and Timbral Vocal utterances
Phase 2A: SpeechBrain ASR explorations
Simultaneous, yet kept separated
Phase 2B: Whisper ASR explorations
Phase 3: CoquiTTS XTTS explorations
wav files called from [author 1'S] repo
Phase 4: Live coding manipulations with strudel
wav files called from [author 1's] repo
re-forming understandings between author 1's original audio data and the XTTS cloned output
Phase 1B: Author 1's real-time audio input in the form of vocal improvised gestures
Phase 2C: Realtime transcriptions using Whisper
Phase 3: CoquiTTS XTTS synthesis using original audio as speaker reference
"just in time" normalization--where applicable--of output transcriptions

Figure 1
Chronology of Kelsey's working phases

Phase 1A-B: Data

In this first phase (1A-B in Figure 1), Kelsey worked with her own sound library and real-time vocalisations. The choice to use her own data was motivated by the first author's concerns about intentionally and knowingly engaging with scraped data and the appropriation of others' bodily labour [54] [55]. To give a brief overview of the content of the dataset, the sound library consists of a wide range of timbral and textural vocal techniques, from the first author's own experimental voice practice. As an example, the recorded sounds include a wide range tongue, lip and palatal clicks; vocal multiphonics; vocal fry; burps; sounds produced with objects inside the mouth; clicking and tapping on the teeth; ingressive phonation [56][57] and phonemic sounds recorded across vocal registers and with varying vowel placements [58].

In this phase, the first author engaged with RtD activities pertaining to the curation of this dataset so that it reflected only word-less, textural and timbral vocal sounds: mainly cataloguing and mapping the scope of her vocal sounds. Methodologically, this was motivated by the understanding of these sounds as potentially disruptive to the normative speech recognition models, and intention to explore new phonemic combinations transcribed by the models' mediation of the non-text sounds. The first author engaged in auto-ethnographic methods such as research diaries to note the somatic context of the recordings, providing an artistic framing of this context as the 'wordless' in her dataset.

Phase 2A-C: ASR Model explorations

In this second stage (2A-C in Figure 1), Kelsey worked with speech recognition models to transcribe her curated sound library. The methodological intention of using speech recognition and transcription was to explore how these text-input models processed non-text input, with a larger aim of exploring the artistic affordances of any novel phonemic or syllabic output. The choice to use both SpeechBrain (an open-source model) and Whisper (also open-sourced, but with dataset ambiguity) was intentional. The first author was curious about the potential differences in transcriptions across Whisper and SpeechBrain, due primarily to the differing datasets these models have been trained on. SpeechBrain is pretrained on the LibriSpeech audio dataset [59], which is an English corpus. Kelsey comes from an English-speaking background, and was curious about whether her extra-normal vocalisations would trigger noticeable 'differences in acoustic salience' [60] in the output transcriptions. Whisper, in comparison, is pretrained on 680,000 hours of multilingual audio and correlating transcripts scraped from the internet[6][61][62]. Methodologically, Kelsey was curious about whether the multilingual capabilities of Whisper would yield emergent and unexpected mappings between her phonemic palette and multilingual transcriptions. She engaged in research explorations with both models to probe the sonic affordances, convergence speeds and exploring how each model transcribed the sound library as well as live-input vocal gestures. Kelsey catalogued these in comic strip format. Throughout Kelsey's auto-ethnography, she noted a disruption of her own understanding of—and relation—to both her own human voice (as sonic material), to transcriptions of her voice (as visual material). Further, Kelsey's auto-ethnographic engagement with these normative AI voice models using extra-normal vocal sounds prompted curiosity as to the scope of the phonemic palette her vocal sounds have, and how she might be able to investigate sensitive inclusion of phonemes from other language groups.

Phase 3: CoquiTTS Re-synthesis and Cloning

In phase 3 (see Figure 1), Kelsey separately parsed the SpeechBrain and Whisper transcriptions from phase 2. The output transcriptions from both were then collated into separate CSV files, and manually normalized. The normalized transcriptions were then used as prompt material for the CoquiTTS XTTS_V2[7] Voice Generation Model, and paired with her corresponding audio file from the original dataset. This yielded banks of re-synthesised audio files from the normalized ASR transcriptions of Kelsey's original dataset. In this phase, the RtD activities constituted comparative analysis between the original audio data and the cloned synthesis files, and experimenting with different text normalization approaches. Kelsey documented her improvisation practice with the original and cloned recordings in a series of live coding performances.

Research Findings

In this section, we discuss the findings emerging from Kelsey's RtD method.

Data - Somatic Experience

As addressed previously, Kelsey utilised her own data, motivated by her understanding of her sound library as a reduced catalogue of her vocal experiences, and as constituting more than just input for a model. As noted in a reflective journal entry: "*The source material used within this collection of poetry is my private sound* *seasonal depression. I had lost a* ***lot*** *of weight. Spending time in the recording studio, recording and documenting a small spectrum of what my voice does was a sort of holy communion with myself, and I clung to those hours in the studio like a life raft. You hear none of this when you listen through to my recordings. But I hear it…Listening back to my library, I hear and I feel everything that I was in April 2019."*

Kelsey further ruminated on the contextual connections in the somatic understandings of her voice from this time period, and her curiosity about possible emergent qualities of the "wordless" in the voice. From the same journal entry: "*…to me, at least—this [somatic history] was the wordless in my voice, in my data, in my communion with self*". Kelsey's somatic reflections on her sound library as a dataset and her speculations on "a wordless" that is uncovered with the assistance of automatic speech recognition tools recalls D'Ignazio and Klein's comments on "the process of converting life experiences into data always necessarily entails a reduction of that experience" [63]. Further, it re-affirms Kelsey's intention to connect her bodily experience of singing with the data concerning that experience, and the intimacy of data as a reflector of self [64].

ASR - From Sound to Text to Poem

This next phase of the research process comprised sense-making of the SpeechBrain and Whisper models' AI-mediated transcriptions of the sounds. The scope of the transcription output was overwhelming and alien to Kelsey, and they felt an urgency to connect and contextualise the ASR transcriptions this text as both a by-product and a continuation of the sound it was born from. She chronicles this urgency in a research diary:

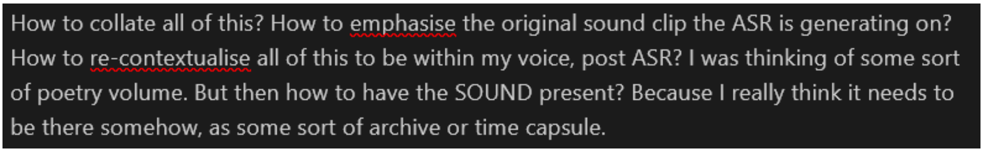
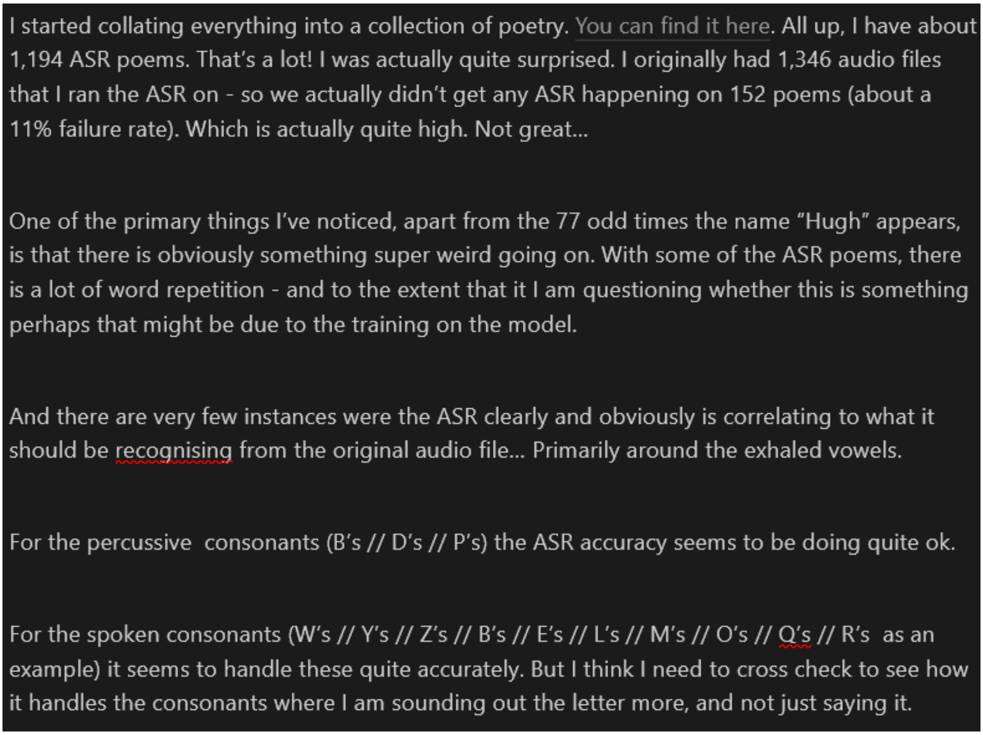

How to collate all of this? How to emphasise the original sound clip the ASR is generating on? How to re-contextualise all of this to be within my voice, post ASR? I was thinking of some sort of poetry volume. But then how to have the SOUND present? Because I really think it needs to be there somehow, as some sort of archive or time capsule.

Figure 2
A screenshot from the first author's research diary

The sheer volume of the text transcriptions led Kelsey to explore ways of re-framing the ASR output, to better connect the poetic AI nonsense with sound. Ultimately, this took the form of a poetry collection.



I started collating everything into a collection of poetry. You can find it here. All up, I have about 1,194 ASR poems. That's a lot! I was actually quite surprised. I originally had 1,346 audio files that I ran the ASR on - so we actually didn't get any ASR happening on 152 poems (about a 11% failure rate). Which is actually quite high. Not great...

One of the primary things I've noticed, apart from the 77 odd times the name "Hugh" appears, is that there is obviously something super weird going on. With some of the ASR poems, there is a lot of word repetition - and to the extent that it I am questioning whether this is something perhaps that might be due to the training on the model.

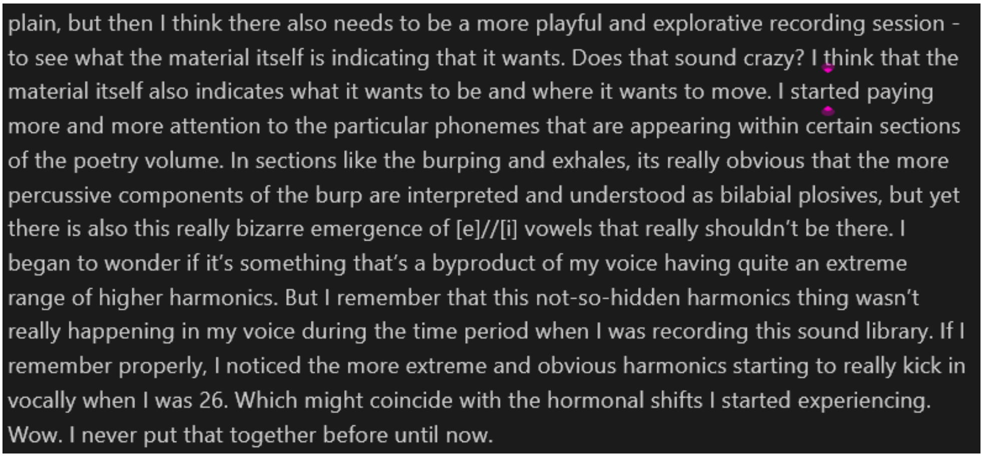
And there are very few instances were the ASR clearly and obviously is correlating to what it should be recognising from the original audio file... Primarily around the exhaled vowels.

For the percussive consonants (B's // D's // P's) the ASR accuracy seems to be doing quite ok.

For the spoken consonants (W's // Y's // Z's // B's // E's // L's // M's // O's // Q's // R's as an example) it seems to handle these quite accurately. But I think I need to cross check to see how it handles the consonants where I am sounding out the letter more, and not just saying it.

Figure 3
A screenshot from the first author's research diary

Here, we see an engagement with conventions of *zaum:* a 20th century Russian Cubo-Futurist experimental linguistic practice. [65] [66] [67] Created by Aleksei Kruchenykh, *zaum* was viewed as the "manifestation of a spontaneous non-codified language" [67]. Structurally, it is built with neologisms[8] which have no clear meaning, and syntactically organised by sonic patterns and rhythm. As an experimental sonic practice, it was highly influential upon avant-garde movements and Surrealism [68] [69] [70]. Working with the transcriptions as poems, Kelsey was able to re-contextualise them as text scores. Here, her conceptualisation of the ASR poems as a AI-generated *zaum* helped to bring forth observations about the recurrence of mismatched vowel phonemes in the SpeechBrain ASR transcriptions. From the SpeechBrain research logs:



plain, but then I think there also needs to be a more playful and explorative recording session - to see what the material itself is indicating that it wants. Does that sound crazy? I think that the material itself also indicates what it wants to be and where it wants to move. I started paying more and more attention to the particular phonemes that are appearing within certain sections of the poetry volume. In sections like the burping and exhales, its really obvious that the more percussive components of the burp are interpreted and understood as bilabial plosives, but yet there is also this really bizarre emergence of [e]//[i] vowels that really shouldn't be there. I began to wonder if it's something that's a byproduct of my voice having quite an extreme range of higher harmonics. But I remember that this not-so-hidden harmonics thing wasn't really happening in my voice during the time period when I was recording this sound library. If I remember properly, I noticed the more extreme and obvious harmonics starting to really kick in vocally when I was 26. Which might coincide with the hormonal shifts I started experiencing. Wow. I never put that together before until now.

Figure 4

A screenshot from the first author's research diary

Kelsey later self-published the SpeechBrain ASR and Whisper ASR poems as an online accessible eBooks. A sample of the first volume of the SpeechBrain ASR Poems can be found in Embedded Frame 1 below:

Visit the web version of this article to view interactive content.

Embedded Frame 1

An eBook of Kelsey's collation of the SpeechBrain ASR transcriptions into a volume of poetry. The embedded eBook contains volume 1.

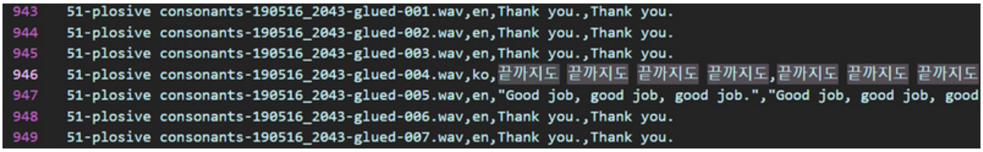
Visibilising Attention Mechanisms

Several of the poems, as seen in the above eBook, are curiously long.[9] When within the confines of a CSV file, the sheer length of the poems is invisibilised. However, when presented in a form that more clearly visibilises the length of the output text transcription, this frames the temporal context of SpeechBrain's attention mechanism in a far more accessible and immediately visible way. As Kelsey observed in her *zaum* of each of poems, the act of reciting these long poems aloud induces a combination of semantic satiation [71][72] and somatic estrangement [73]. Further, she noted that her *zaum* progressively divorced the words from the immediate sonic context of the respective raw audio files. The somatic estrangement triggered from this *zaum* afforded a '*making strange*' of both the original audio and the resultant ASR transcriptions [74] [75]. In turn, this provoked Kelsey to more deeply consider how the collation of the poems visibilised the temporal context of the ASR attention mechanisms and to further visually communicate the experience of sonic estrangement through *zaum*. From this, we understand the compilation of the poems into a more 'book-like' form assisting in both visibilising the attention mechanism of the speech recognition models, but also re-introduced the experiential body through *zaum-ing* the poems.

Multilingual Mediations

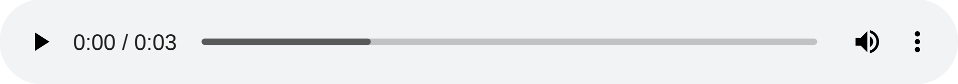
Within Kelsey's engagement with the Whisper ASR model, she uncovered another surprising AI-mediation of her voice. In the Whisper transcriptions, she observed that there were seemingly random instances of her audio files transcribed into different languages, mainly Japanese, Korean and Chinese. Although the potential for this to happen was not altogether unexpected[10], the way this multilingualism manifests across Kelsey's sound library was curious. She noted in the research logs that the original sound files were transcribed into Korean in 51 of the 1346 transcriptions, specifically in "monosyllabic" [p], [q] and [u] phonemes. Comparatively, Japanese transcriptions accounted for 46 of the 1346 transcriptions, occurring in "monosyllabic" [w], [y], [a], [e] and [i] phonemes.

An example of a Korean transcription is seen in Figure 5 below, paired with its original audio file (see Audio Excerpt 1). Here, we illustrate the phonemic similarities between Kelsey's original audio data and the phonemic events occurring within a reading of the transcribed audio, read in the language that Whisper identified from the original audio file. We understand this as a form of inverse *zaum*, in that Kelsey's vocal gesture is functionally codified and interpreted by Whisper.

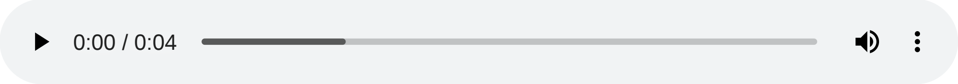


```
943   51-plosive consonants-190516_2043-glued-001.wav,en,Thank you.,Thank you.
944   51-plosive consonants-190516_2043-glued-002.wav,en,Thank you.,Thank you.
945   51-plosive consonants-190516_2043-glued-003.wav,en,Thank you.,Thank you.
946   51-plosive consonants-190516_2043-glued-004.wav,ko,끝까지도 끝까지도 끝까지도 끝까지도,끝까지도 끝까지도 끝까지도
947   51-plosive consonants-190516_2043-glued-005.wav,en,"Good job, good job, good job.","Good job, good job, good
948   51-plosive consonants-190516_2043-glued-006.wav,en,Thank you.,Thank you.
949   51-plosive consonants-190516_2043-glued-007.wav,en,Thank you.,Thank you.
```

Figure 5
An excerpt from the first author's research logs, depicting a Korean transcription.
Please see line 946 in the CSV file, featuring '51-plosive consonants-190516_2043-glued-004.wav' and a Korean transcription.

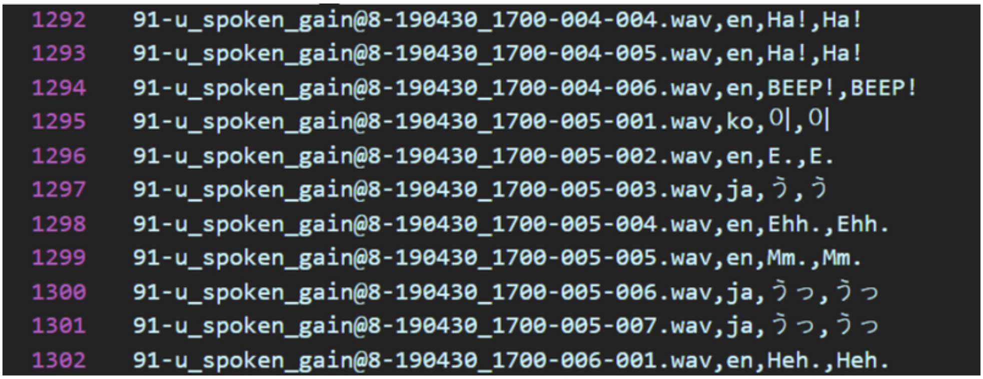


Audio Excerpt 1
The corresponding audio file for Figure 5.
Filename: '51-plosive consonants-190516_2043-glued-004.wav'



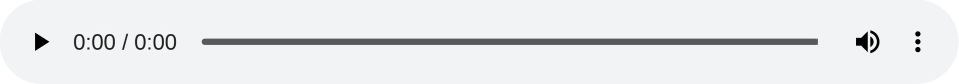
Audio Excerpt 2
A reading of the Korean normalized transcription, as seen in line 946 in Figure 5

In Figure 6, examples of several Japanese transcription are embedded from the research logs, paired with the original audio files. Here, we illustrate Whisper's inversion of *zaum*. The phonemes uttered by Kelsey are directly codified by Whisper according to phonemic similarity with Japanese vowels.

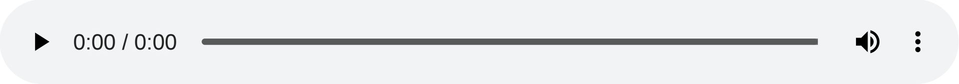


```
1292    91-u_spoken_gain@8-190430_1700-004-004.wav,en,Ha!,Ha!
1293    91-u_spoken_gain@8-190430_1700-004-005.wav,en,Ha!,Ha!
1294    91-u_spoken_gain@8-190430_1700-004-006.wav,en,BEEP!,BEEP!
1295    91-u_spoken_gain@8-190430_1700-005-001.wav,ko,이,이
1296    91-u_spoken_gain@8-190430_1700-005-002.wav,en,E.,E.
1297    91-u_spoken_gain@8-190430_1700-005-003.wav,ja,う,う
1298    91-u_spoken_gain@8-190430_1700-005-004.wav,en,Ehh.,Ehh.
1299    91-u_spoken_gain@8-190430_1700-005-005.wav,en,Mm.,Mm.
1300    91-u_spoken_gain@8-190430_1700-005-006.wav,ja,うっ,うっ
1301    91-u_spoken_gain@8-190430_1700-005-007.wav,ja,うっ,うっ
1302    91-u_spoken_gain@8-190430_1700-006-001.wav,en,Heh.,Heh.
```

Figure 6
An excerpt from Kelsey's research logs, depicting multiple instances of Japanese transcription.
Please see lines 1297, 1300 and 1201 in the CSV file, featuring '91-u_spoken_gain@8-190430_1700-005-003.wav'; '91-u_spoken_gain@8-190430_1700-005-006.wav'; and '91-u_spoken_gain@8-190430_1700-005-007.wav' and their respective Japanese transcriptions.



Audio Excerpt 3
The corresponding audio file for line 1297 in Figure 6.
Filename: 91-u_spoken_gain@8-190430_1700-005-003.wav'



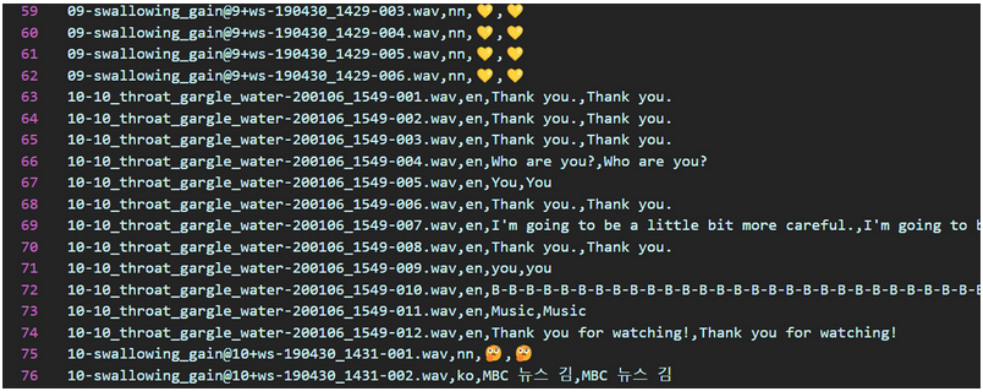
Audio Excerpt 4
The corresponding audio file for line 1300 in Figure 6.
Filename: '91-u_spoken_gain@8-190430_1700-005-006.wav

Audio Excerpt 5
The corresponding audio file for line 1301 in Figure 6.
Filename: 91-u_spoken_gain@8-190430_1700-005-007.wav

Audio Excerpt 6
A reading of the Japanese normalized transcription, as seen in lines 1297, 1300 and 1301 in Figure 6

Signs of Scraped Data and Attack of the Emojis

Another surprising outcome of utilising Whisper was that Kelsey found clear signs of scraped data in some select transcriptions of the audio files- predominantly amongst the multilingual transcriptions. We highlight line 76 in Figure 7 which features a transcription in Korean reading "MBC 뉴스 김".[11] She identified this as perhaps referring to the Munhwa Broadcasting Corporation, which is one of the leading South Korean television and radio broadcasters [76]. As previously noted, OpenAI does not disclose the exact audio sources of their dataset, but the connections made through Kelsey's very basic initial web search lead us to assume that data has been scraped from a broadcast from MBC.



```
59   09-swallowing_gain@9+ws-190430_1429-003.wav,nn,💛,💛
60   09-swallowing_gain@9+ws-190430_1429-004.wav,nn,💛,💛
61   09-swallowing_gain@9+ws-190430_1429-005.wav,nn,💛,💛
62   09-swallowing_gain@9+ws-190430_1429-006.wav,nn,💛,💛
63   10-10_throat_gargle_water-200106_1549-001.wav,en,Thank you.,Thank you.
64   10-10_throat_gargle_water-200106_1549-002.wav,en,Thank you.,Thank you.
65   10-10_throat_gargle_water-200106_1549-003.wav,en,Thank you.,Thank you.
66   10-10_throat_gargle_water-200106_1549-004.wav,en,Who are you?,Who are you?
67   10-10_throat_gargle_water-200106_1549-005.wav,en,You,You
68   10-10_throat_gargle_water-200106_1549-006.wav,en,Thank you.,Thank you.
69   10-10_throat_gargle_water-200106_1549-007.wav,en,I'm going to be a little bit more careful.,I'm going to b
70   10-10_throat_gargle_water-200106_1549-008.wav,en,Thank you.,Thank you.
71   10-10_throat_gargle_water-200106_1549-009.wav,en,you,you
72   10-10_throat_gargle_water-200106_1549-010.wav,en,B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-B-
73   10-10_throat_gargle_water-200106_1549-011.wav,en,Music,Music
74   10-10_throat_gargle_water-200106_1549-012.wav,en,Thank you for watching!,Thank you for watching!
75   10-swallowing_gain@10+ws-190430_1431-001.wav,nn,🫢,🫢
76   10-swallowing_gain@10+ws-190430_1431-002.wav,ko,MBC 뉴스 김,MBC 뉴스 김
```

Figure 7
An excerpt from the first author's research logs, depicting multiple instances of emojis in the non-normalized transcription.

We note in this same image in Figure 7, that some transcriptions also appeared as emojis (line 59-62) in the "nn" language code. Kelsey assumed this to be Norwegian Nynorsk, according to the ISO 639 language code protocol [77]. The appearance of emojis was completely shocking, and an explanation for why these appeared in "nn" transcriptions are still a mystery. Kelsey found that all transcriptions in "nn" yielded predominantly emoji transcriptions, with some instances of numerical transcriptions. Although we are yet to draw any firm conclusions, we speculate that the sonic profile of Kelsey's throat gargle must share similarities with a scraped audio clip in OpenAI's training ''nn" data subset. We note here, that this may indicate generative potential in non-normative usage of normative-ly trained models to forcibly "glitch" [78] [79] models into revealing more contextual information regarding the origin of their scraped or ambiguous datasets.

Having discussed the first author's engagement with SpeechBrain and Whisper, we now progress to our discussion of how the transcriptions emerging from the ASR phase were implemented within a text-to-speech synthesis phase using CoquiTTS's voice generation model. We outline the first author Kelsey's steps in utilising CoquiTTS's model for synthesis of the ASR transcriptions, cloned in her own voice.

After the transcriptions were generated using SpeechBrain and Whisper it became necessary to perform normalization of the transcriptions. This was due to the need to adapt the output text for subsequent re-synthesis through the CoquiTTS XTTS_V2[12] model. The XTTS model is optimized for short-form cloning, using audio clips of approximately 6 seconds duration and has a limit of 400 text tokens that it can successfully parse. As can be seen in the eBook Embedded Frame 1, there are a number of transcribed audio clips that clearly generated well over this 400 token limit due SpeechBrain's attention mechanism and the multilingual mappings triggered within Whisper.

Based on the models' limitations, the normalization of the transcriptions was functionally oriented: encompassing the editing of excessive punctuation marks[13] and Arabic numerals. There was also a limit to the tokens that the XTTS model could successfully synthesise, and so transcriptions exceeding a certain character amount had to be manually shortened to a maximum of 400 tokens. Kelsey shortened the transcriptions in a way that endeavored to retain a clear macro and micro structural organisation of the rhythms, phonemes and word structures that were evident in the transcribed audio files. This was informed by Kelsey's *zaum* of every poem aloud to determine the instinctive rhythms that they produced during the process re-performing her transcribed sounds. Here, the engagement with *zaum* assisted in determining the most appropriate places sonically and linguistically to constrain the input tokens.

Phase 3 - Live coding as digital zaum

After the normalized transcriptions and the paired audio from Kelsey's library were parsed through the XTTS model to yield a collection of synthesised and cloned audio files, she began the exploration of these files through a live coding musical practice. In this phase, she elected to work with strudelREPL[14]. strudel is JavaScript version of the live coding language tidalcycles[15], and is a browser-based environment for live coding music [80][81] [82] [83] [84] [85] [86] [87] [88]. strudel was chosen as the environment for exploring the 'original' and synthesised sounds for practical reasons, primarily it's ease-of-access as a platform. During this phase, Kelsey called her dataset and the cloned samples into the strudel editor to manipulate and transform them in real-time. This engagement took the form of a series of recorded improvisations using the original and TTS output. Kelsey recorded five improvisations, averaging 8-9 minutes per improvisation. In each improvisation, she worked with a maximum number of four samples: two original audio clips and their two correlating XTTS clones. In terms of the musical code, Kelsey would largely start with a "blank slate" and begin to build rhythmic patterns and include manipulations to the audio samples iteratively as the improvisations progressed. Working with strudel afforded the opportunity to efficiently work with and manipulate the original and cloned audio. Each improvisation session afforded the opportunity to explore the sonic similarities and tensions between Kelsey's original and cloned sounds. Similar to the engagement with *zaum* to '*make strange*' her sound library from its AI-mediated transcriptions, engaging and manipulating her original and cloned audio through strudel enabled a form of digital *zaum*.

Discussion

The rising attention concerning the cloning, synthesis and generation of voices with comparatively small amounts of original voice data has elicited much discourse about how we grapple with the role of human bodies implicated by AI technologies [89]. Attempting to "solve" this immensely complex issue would overlook the ripe opportunity we now find ourselves faced with: to re-form our understandings of what human- and AI- vocality might mean (and become). A critical and creative engagement with these tools affords understanding as to what this extra-normal AI vocality affords us, as we have demonstrated through our documented RtD process of Deep Learning AI voice technologies. This returns us to the earlier posed research question of *how does the engagement of AI tools for voice and speech in an auto-ethnographic voice practice contribute to novel understanding of human- and AI-vocality?*

When we contextualise auto-ethnographic research processes within our context of musical engagement with AI, we reach a point of collision with some perspectives from Research-through-Design. As Frayling tells us that *"[t]he artist, by definition, is someone who works in an* ***expressive*** *idiom, rather than a* ***cognitiv****e one, and for whom the great project is an extension of personal development: autobiography rather than understanding*" [23]. As we have demonstrated through our RtD engagement with AI models for ASR and TTS, this statement can no longer be assumed to be entirely true. In an age where the role of the artist is continually morphing and mutating to include and appropriate new technologies—especially AI— are we still ***truly and solely*** working within an expressive idiom? We argue and have demonstrated through this RtD with AI voice, that the expressive and cognitive idioms are interconnected—dependent—with one another. The inclusion of the body and bodily practices—such as singing—with novel technologies establishes that a musical AI RtD process is more than solely cognitive or expressive [90][91][92][93][94]. We have demonstrated that auto-ethnographic AI voice knowledge [33] is inherently cognitive ***and*** expressive in nature. Our research findings emerging from this engagement with AI voice technologies within an experimental voice practice reveal that "…the thinking is, so to speak, embodied in the artefact"[23]. We see this directly reflected as an outcome of our research question, that our engagement of AI tools for voice and speech in an auto-ethnographic voice practice contributes a novel understanding of human- and AI-vocality as both cognitively and expressively bound.

Whilst objective examination and assessment of how TTS models transcribe non-text-based vocalisations was **not** an aim of this auto-ethnography, we can anecdotally surmise they do not perform "successfully" in the conventional sense at transcribing non-textual vocalisations. We use the term "successfully" with regard to the intended benchmark standards within which these models have been built: to accurately transcribe human speech. In our case, the "failure" of the SpeechBrain ASR model is unsurprising, as we are intentionally forcing a glitch in our non-normative usage of an inherently functionally intended model. Anecdotally, Whisper performed significantly more “successfully”, which can be attributed to it’s multilingual capacities. As discussed above in our account of Kelsey’s engagement with SpeechBrain and Whisper, we noted this in instances of more monosyllabic or phonemic clips in her dataset in which the transcriptions revealed the multilingual capabilities of the XTTS model in surprising ways.

From our RtD, we establish an understanding of non-normative engagement with normative AI voice technologies as affording *extra-normal* mediations of the creative, expressive human voice. Our perspective on this pertains largely to how exploratory RtD assists in subverting normative AI voice, and utilising the outputs as novel sonic materials. We therefore understand that non-normative artistic research processes engage normative AI models to both re-form artistic understandings of how human vocality is re-formed by AI mediations, and constitutes a knowledge-creation mechanisms when working with these technologies.

Conclusion

In this paper, we constrained our discussion of experimental voice practice to the highly specific context of 20th and 21st century European traditions. We acknowledge this view as insular, and that contemporary and historical engagement with voice composition and performance is richly varied across geographies, timescales and is contextually informed by other cultural, social and technological developments. We note that there is important future work that may, and indeed should be done, in examining non-European-centrist experimental vocal practices.

In this paper we have presented an example of an auto-ethnographic study of Deep Learning models for speech recognition and synthesis. We have engaged with SpeechBrain, OpenAI and CoquiTTS voice and speech models within an experimental vocal practice in order to reveal their musical affordances as mediators of human vocality. We have demonstrated the generative potential of subverting normative models to provoke nonsensical AI-mediations of human voice, which has in turn been utilised as musical material within a series of live coding performances. We have contributed an example of Research-through-Design and auto-ethnography as generative methodologies for knowledge creation in this domain. We have further illustrated through our case study that the creative and critical engagement with AI voice and speech technologies may afford further consideration as to re-framing AI models as an additional technique of *extra-normal* vocality.

Ethics Statement

The paper has utilised open-sourced, pretrained AI models and the first author’s own personal dataset and auto-ethnographic research logs. No human participants (other than the first author) were recruited for this study, and no sensitive data were collected. The main methodology is based on an self-observational auto-ethnographic study, utilising video and audio materials and research diaries. This paper has the intention to contribute to musical AI research, and to support future research within this community. We predict the environmental impact of this work as minimal since the computation required to create this work was comparable to daily personal computer usage. Accessibility of the technology in this work is limited with the general accessibility to computers and computational development frameworks.

Acknowledgments

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

Footnotes

1. Kelsey Cotton ↩

2. Kelsey Cotton ↩

3. RtD hereafter ↩

4. XTTS hereafter ↩

5. strudel hereafter ↩

6. We note, with criticism, that OpenAI does not disclose the exact sources of their dataset. ↩

7. XTTS hereafter ↩

8. A term, word or phrase ↩

9. Specifically, '05-regular breathing_EE_mouth_gain @10 + windshield-190430_1415-01.wav' (on page 16); '06-regular breathing_OO_mouth_gain @10 + windshield-190430_1418.wav' (page 26) and '45-chest voice_eh-190516_2027-glued-001.wav' (page 96). ↩

10. As Whisper's ASR model is a multilingual one. ↩

11. Translated by X using Google Translate as: MBC News Kim ↩

12. XTTS hereafter ↩

13. Such as "?" and "," ↩

14. Referred to hereafter as strudel. ↩

15. Referred to hereafter as tidal. ↩

References

1. Sundberg, J. (1989). *Science of the Singing Voice*. Dekalb, Ill: Northern Illinois University Press. ↩

2. Sundberg, J. (1996). The Human Voice. In R. Greger & U. Windhorst (Eds.), *Comprehensive Human Physiology: From Cellular Mechanisms to Integration* (pp. 1095–1104). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-60946-6_54 ↩

3. Titze, I., & Alipour, F. (2006). *The Myoelastic-Aerodynamic Theory of Phonation*. Denver, Colorado, United States: National Center for Voice. ↩

4. Stevens, K. N. (2000). *Acoustic Phonetics*. The MIT Press. Retrieved from https://mitpress.mit.edu/9780262692502/acoustic-phonetics/ ↩

5. Eidsheim, N. S. (2015). *Sensing sound: singing & listening as vibrational practice*. Durham: Duke University Press. ↩

6. Eidsheim, N. S. (2011). Sensing Voice: Materiality and the Lived Body in Singing and Listening. *The Senses and Society*, *6*(2), 133–155. https://doi.org/10.2752/174589311X12961584845729 ↩

7. Coldewey, D. (2023). VALL-E's quickie voice deepfakes should worry you, if you weren't worried already. Retrieved from https://techcrunch.com/2023/01/12/vall-es-quickie-voice-deepfakes-should-worry-you-if-you-werent-worried-already/ ↩

8. David, E. (2023). RIAA wants AI voice cloning sites on government piracy watchlist. Retrieved from https://www.theverge.com/2023/10/11/23913405/riaa-ai-voice-cloning-threat-copyright-ustr ↩

9. Mannie, K. (2023). AI kidnapping scam copied teen girl's voice in $1M extortion attempt - National | Globalnews.ca. *Global News*. Retrieved from https://globalnews.ca/news/9629883/ai-kidnapping-scam-teen-girl-voice-cloned-extortion-arizona-jennifer-destefano/ ↩

10. Shah, H. (2023). Exploring the Pros and Cons of AI Voice Cloning. Retrieved from https://medium.com/@shahhardik2905/exploring-the-pros-and-cons-of-ai-voice-cloning-f4bb15514284 ↩

11. Kruspe, A. (2024). *More than words: Advancements and challenges in speech recognition for singing*. arXiv. https://doi.org/10.48550/arXiv.2403.09298 ↩

12. Seong, J., Lee, W., & Lee, S. (2021). Multilingual Speech Synthesis for Voice Cloning. *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 313–316. https://doi.org/10.1109/BigComp51126.2021.00067 ↩

13. Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., & McAuley, J. (2021). Expressive Neural Voice Cloning. *Proceedings of The 13th Asian Conference on Machine Learning*, 252–267. PMLR. Retrieved from https://proceedings.mlr.press/v157/neekhara21a.html ↩

14. Pecora, A. E. (2023). *Data Driven: AI Voice Cloning* (Laurea, Politecnico di Torino). Politecnico di Torino. Retrieved from https://webthesis.biblio.polito.it/27738/ ↩

15. Chen, W., & Jiang, X. (2023). *Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression*. Preprints. https://doi.org/10.20944/preprints202312.0807.v1 ↩

16. Hutiri, W., Papakyriakopoulos, O., & Xiang, A. (2024). *Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators*. arXiv. https://doi.org/10.48550/arXiv.2402.01708 ↩

17. Amezaga, N., & Hajek, J. (2022). Availability of Voice Deepfake Technology and its Impact for Good and Evil. *Proceedings of the 23rd Annual Conference on Information Technology Education*, 23–28. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3537674.3554742 ↩

18. Onuh Matthew Ijiga, Idoko Peter Idoko, Lawrence Anebi Enyejo, Omachile Akoh, Solomon Ileanaju Ugbane, & Akan Ime Ibokette. (2024). Harmonizing the voices of AI: Exploring generative music models, voice cloning, and voice transfer for creative expression. *World Journal of Advanced Engineering Technology and Sciences*, *11*(1), 372–394. https://doi.org/10.30574/wjaets.2024.11.1.0072 ↩

19. Cotton, K., De Vries, K., & Tatar, K. (2024). Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies. *Proceedings of the 9th International Conference on Movement and Computing*. Utrecht, Netherlands: Association for Computing Machinery. https://doi.org/10.1145/3658852.3659065 ↩

20. Napolitano, D. (2022). AI voice between anthropocentrism and posthumanism: Alexa and voice cloning. *Journal of Interdisciplinary Voice Studies*, *7*, 35–49. https://doi.org/10.1386/jivs_00053_1 ↩

21. Zellou, G., & Cohn, M. (2023). *Clear speech in the new digital era: Speaking and listening clearly to voice-AI systems* (Vol. 153). https://doi.org/10.1121/10.0018378 ↩

22. Allado-McDowell, K., & Bentivegna, F. (2022). Cybernetic animism: Voice and AI in conversation. *Journal of Interdisciplinary Voice Studies*, *7*, 107–118. https://doi.org/10.1386/jivs_00058_1 ↩

23. Frayling, C. (1993). Research in Art and Design. *Royal College of Art Research Papers*, *1*(1). Retrieved from https://researchonline.rca.ac.uk/384/ ↩

24. Gaver, W., Krogh, P. G., Boucher, A., & Chatting, D. (2022). Emergence as a Feature of Practice-based Design Research. *Designing Interactive Systems Conference*, 517–526. New York, NY, USA: Association

for Computing Machinery. https://doi.org/10.1145/3532106.3533524↩
25. Lynch, C. R. (2022). Glitch epistemology and the question of (artificial) intelligence: Perceptions, encounters, subjectivities. *Dialogues in Human Geography*, *12*(3), 379–383. https://doi.org/10.1177/20438206221102952 ↩
26. Olufemi, T. (2023). Blackness, Glitch Feminism and Evasion • Container Magazine. Retrieved from https://containermagazine.co.uk ↩
27. Russell, L. (2020). *Glitch Feminism*. Verso. Retrieved from https://www.penguinrandomhouse.com/books/646946/glitch-feminism-by-legacy-russell/ ↩
28. Chang, H. (2016). *Autoethnography as Method*. Routledge. ↩
29. Butz, D., & Besio, K. (2009). Autoethnography. *Geography Compass*, *3*(5), 1660–1674. https://doi.org/10.1111/j.1749-8198.2009.00279.x ↩
30. Adams, T. E., Jones, S. L. H., & Ellis, C. (2015). *Autoethnography*. Oxford University Press. ↩
31. Chang, H. (2016). *Autoethnography as Method*. Routledge. Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An Overview. *Historical Social Research / Historische Sozialforschung*, *36*(4 (138)), 273–290. Retrieved from https://www.jstor.org/stable/23032294 ↩
32. Wall, S. (2008). Easier Said than Done: Writing an Autoethnography. *International Journal of Qualitative Methods*, *7*(1), 38–53. https://doi.org/10.1177/160940690800700103 ↩
33. Polanyi, M. (2015). *Personal Knowledge: Towards a Post-Critical Philosophy* (M. J. Nye, Ed.). Chicago, IL: University of Chicago Press. Retrieved from https://press.uchicago.edu/ucp/books/book/chicago/P/bo19722848.html ↩
34. Schiphorst, T. (2011). Self-evidence: applying somatic connoisseurship to experience design. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*, 145. Vancouver, BC, Canada: ACM Press. https://doi.org/10.1145/1979742.1979640 ↩
35. Cotton, K., Afsar, O. K., Luft, Y., Syal, P., & Ben Abdesslem, F. (2021). SymbioSinging: Robotically transposing singing experience across singing and non-singing bodies. *Creativity and Cognition*, 1–5. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3450741.3466718 ↩
36. Kilic Afsar, O., Luft, Y., Cotton, K., Stepanova, E. R., Núñez-Pacheco, C., Kleinberger, R., … Höök, K. (2023). Corsetto: A Kinesthetic Garment for Designing, Composing for, and Experiencing an Intersubjective Haptic Voice. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. <conf-

loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581294↩

37. Edgerton, M. E. (2015). *The 21st-century voice: contemporary and traditional extra-normal voice* (Second edition). Lanham: Rowman & Littlefield. ↩

38. Noble, C. (2019). *Extended from What?: Tracing the construction, flexible meaning, and cultural discources of "Extended Vocal Techniques"* (Phdthesis, University of California). University of California, Santa Cruz. Retrieved from https://escholarship.org/content/qt6qn119zh/qt6qn119zh_noSplash_b604aff101759d9e818f6af5b5159091.pdf?t=ppqqs6 ↩

39. Edgerton, M. E. (2004). *The 21st-century voice: contemporary and traditional extra-normal voice*. Lanham, Md: Scarecrow Press. ↩

40. Esling, J. H., & Moisik, S. R. (2021). *Voice Quality*. https://doi.org/10.1017/9781108644198.010 ↩

41. Sundberg, J. (1989). *Science of the Singing Voice*. Dekalb, Ill: Northern Illinois University Press. ↩

42. Eidsheim, N., Meizel, K., Eidsheim, N., & Meizel, K. (Eds.). (2019). *The Oxford Handbook of Voice Studies*. Oxford, New York: Oxford University Press. ↩

43. Sundberg, J. (1977). The Acoustics of the Singing Voice. *Scientific American*, *236*(3), 82–91. Retrieved from https://www.jstor.org/stable/24953939 ↩

44. Eidsheim, N. S. (2008). *Voice as a technology of selfhood : towards an analysis of racialized timbre and vocal performance* (Phdthesis, UC San Diego). UC San Diego. Retrieved from https://escholarship.org/uc/item/0h8841kp ↩

45. Ihde, D. (2007). *Listening and voice: phenomenologies of sound* (2nd ed). Albany: State University of New York Press. ↩

46. Jones, A. (1998). *Body Art/Performing the Subject*. University of Minnesota Press. Retrieved from https://www.upress.umn.edu/book-division/books/body-art-performing-the-subject ↩

47. Tarvainen, A. (2018). Singing, Listening, Proprioceiving: Some Reflections on Vocal Somaesthetics. In *Aesthetic Experience and Somaesthetics* (pp. 120–142). Brill. https://doi.org/10.1163/9789004361928_010 ↩

48. Verstraete, P. (2011). *Vocal Extensions: Disembodied Voices in Contemporary Music Theatre and Performance*. Retrieved from https://www.academia.edu/3035586/Vocal_Extensions_Disembodied_Voices_in_Contemporary_Music_Theatre_and_Performance ↩

49. Warren, K. (2011). *Show More/Show Less: Extended Voice, Technology, and Presence*. Retrieved from https://www.academia.edu/33057645/Show_More_Show_Less_Extended_Voice_Technology_and_Presence ↩

50. Williams, R. (2023). *Voice in the Machine preprint*. https://doi.org/10.13140/RG.2.2.18134.01602 ↩

51. Young, M. (2015). *Singing the Body Electric: The Human Voice and Sound Technology*. Ashgate. Retrieved from https://libgen.pm/ads4648e3ef403a510369d410bd53e61e8eZEEH8AE3 ↩

52. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., … Bengio, Y. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. ↩

53. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (n.d.). *Robust Speech Recognition via Large-Scale Weak Supervision*. ↩

54. Tatar, K., Ericson, P., Cotton, K., Prado, P., Batlle-Roca, R., Cabrero-Daniel, B., … Hussain, J. (2023). *A Shift In Artistic Practices through Artificial Intelligence*. ↩

55. Newlands, G. (2021). Lifting the curtain: Strategic visibility of human labour in AI-as-a-Service. *Big Data & Society*, *8*(1), 20539517211016026. https://doi.org/10.1177/20539517211016026 ↩

56. DeBoer, A. (2012). Ingressive Phonation in Contemporary Vocal Music. *Doctor of Musical Arts Dissertations*. Retrieved from https://scholarworks.bgsu.edu/dma_diss/16 ↩

57. Fornhammar, L., Sundberg, J., Fuchs, M., & Pieper, L. (2022). Measuring Voice Effects of Vibrato-Free and Ingressive Singing: A Study of Phonation Threshold Pressures. *Journal of Voice*, *36*(4), 479–486. https://doi.org/10.1016/j.jvoice.2020.07.023 ↩

58. Nordgren, C. E. (2019). Phonetics of Vowels. In *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.405 ↩

59. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. South Brisbane, Queensland, Australia: IEEE. https://doi.org/10.1109/ICASSP.2015.7178964 ↩

60. Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, *89*(6), 2961–2977. https://doi.org/10.1121/1.400734 ↩

61. whisper/model-card.md at main · openai/whisper. (n.d.). Retrieved from https://github.com/openai/whisper/blob/main/model-card.md ↩

62. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. ↩

63. D'Ignazio, C., & Klein, L. F. (2023). *Data Feminism*. The MIT Press. Retrieved from https://mitpress.mit.edu/9780262547185/data-feminism/ ↩

64. Peña, P., & Varon, J. (2019). Consent to our Data Bodies: Lessons from feminist theories to enforce data protection. *Association for Progressive Communications*. Retrieved from https://codingrights.org/docs/ConsentToOurDataBodies.pdf ↩

65. Janecek, G. (1996). *Zaum: the transrational poetry of Russian futurism* (First published). San Diego, Calif: San Diego State University Press. ↩

66. Nilsson, N. Å. (1981). The Sound Poem: Russian Zaum' and German Dada. *Russian Literature*, *10*(4), 307–317. https://doi.org/10.1016/0304-3479(81)90008-9 ↩

67. Handbook of Russian Literature. (n.d.). Retrieved from https://yalebooks.yale.edu/9780300048681/handbook-of-russian-literature ↩

68. Perloff, N. (2009). Sound Poetry and the Musical Avant-Garde: A Musicologist's Perspective. In M. Perloff & C. Dworkin (Eds.), *The Sound of Poetry / The Poetry of Sound* (p. 0). University of Chicago Press. https://doi.org/10.7208/chicago/9780226657448.003.0009 ↩

69. Perloff, M., & Dworkin, C. (2009). *The Sound of Poetry / The Poetry of Sound* (Vol. 123). University of Chicago Press. Retrieved from https://www.jstor.org/stable/25501896 ↩

70. Perloff, N. (2009). Sound Poetry and the Musical Avant-Garde: A Musicologist's Perspective. In M. Perloff & C. Dworkin (Eds.), *The Sound of Poetry / The Poetry of Sound* (p. 0). University of Chicago Press. https://doi.org/10.7208/chicago/9780226657448.003.0009 ↩

71. Jakobovits, L. A. (1962). *Effects of repeated stimulation on cognitive aspects of behavior: some experiments on the phenomenon of semantic satiation.* (Phdthesis, McGill University). McGill University. Retrieved from https://escholarship.mcgill.ca/concern/theses/c821gp587 ↩

72. Das, J. P. (1969). *Verbal Conditioning and Behaviour* (1st Edition). Pergamon Press. Retrieved from https://shop.elsevier.com/books/verbal-conditioning-and-behaviour/das/978-0-08-012818-4 ↩

73. Withheld for anonymity. ↩

74. Bell, G., Blythe, M., & Sengers, P. (2005). Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Transactions on Computer-Human Interaction*, *12*(2), 149–173. https://doi.org/10.1145/1067860.1067862 ↩

75. Kumagai, A. K., & Wear, D. (2014). “Making Strange”: A Role for the Humanities in Medical Education. *Academic Medicine*, *89*(7), 973. https://doi.org/10.1097/ACM.0000000000000269 ↩

76. 만나면 좋은 친구 MBC [Official Webpage]. (n.d.). Retrieved from https://www.imbc.com/ ↩

77. Benhoff, M., Archibald, J., Ferrés Hernández, M., & Perou, N. (2023). *ISO 639:2023- Code for individual languages and language groups*. Retrieved from https://www.iso.org/obp/ui/en/#iso:std:iso:639:ed-2:v1:en ↩

78. Olufemi, T. (2023). Blackness, Glitch Feminism and Evasion • Container Magazine. Retrieved from https://containermagazine.co.uk ↩

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

80. Blackwell, A., Cocker, E., Cox, G., McLean, A., & Magnusson, T. (2022). *Live coding: a user's manual*. Cambridge, Massachusetts London: The MIT Press. ↩

81. Magnusson, T. (2011). Algorithms as Scores: Coding Live Music. *Leonardo Music Journal*, *21*, 19–23. https://doi.org/10.1162/LMJ_a_00056 ↩

82. Brown, A. R., & Sorensen, A. C. (2007). aa-cell in practice : an approach to musical live coding. *International Computer Music Conference*, 292–299. Retrieved from http://www.computermusic.org/ ↩

83. Collins, N., McLEAN, A., Rohrhuber, J., & Ward, A. (2003). Live coding in laptop performance. *Organised Sound*, *8*(3), 321–330. https://doi.org/10.1017/S135577180300030X ↩

84. Drymonitis, A. (2023). Live Coding Poetry: The narrative of code in a hybrid musical/poetic context. *Organised Sound*, *28*(2), 241–252. https://doi.org/10.1017/S1355771823000493 ↩

85. Diapoulis, G., & Dahlstedt, P. (2021). *The creative act of live coding practice in music performance*. ↩

86. Diapoulis, G. (2023). *Expression in Live Coding: Gestural Interaction for Machine Musicianship* (Phdthesis, Chalmers University of Technology). Chalmers University of Technology. Retrieved from https://research.chalmers.se/en/publication/537469 ↩

87. Diapoulis, G. (2022). Livecode me: Live coding practice and multimodal experience. *Proceedings of the 33rd Annual Workshop of the Psychology of Programming Interest Group (PPIG)*. Retrieved from https://research.chalmers.se/en/publication/533938 ↩

88. Diapoulis, G. (2021). *Primary and secondary aspects of musical gestures in live coding performance*. Retrieved from https://research.chalmers.se/en/publication/526966 ↩

89.

Cotton, K., De Vries, K., & Tatar, K. (2024). Singing for the Missing: Bringing the Body Back to AI Voice and Speech Technologies. Proceedings of the 9th International Conference on Movement and Computing. Utrecht, Netherlands: Association for Computing Machinery. https://doi.org/10.1145/3658852.3659065 ↩

90. Johnson, M. (2008). *The Meaning of the Body: Aesthetics of Human Understanding*. Chicago, IL: University of Chicago Press. Retrieved from https://press.uchicago.edu/ucp/books/book/chicago/M/bo5417890.html ↩

91. Svanæs, D. (2013). Interaction design for and with *the lived body*: Some implications of merleau-ponty's phenomenology. *ACM Transactions on Computer-Human Interaction*, *20*(1), 1–30. https://doi.org/10.1145/2442106.2442114 ↩

92. Sheets-Johnstone, M. (2015). *The Corporeal Turn: An Interdisciplinary Reader*. Andrews UK Limited. ↩

93. Shusterman, R. (2008). *Body Consciousness: A Philosophy of Mindfulness and Somaesthetics*. Cambridge University Press. ↩

94. Höök, K., Eriksson, S., Louise Juul Søndergaard, M., Ciolfi Felice, M., Campo Woytuk, N., Kilic Afsar, O., … Ståhl, A. (2019). Soma Design and Politics of the Body. *Proceedings of the Halfway to the Future Symposium 2019*, 1–8. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3363384.3363385 ↩

Paper V

glemöhnic

K.Cotton

AI Music Creativity (September 2024)

https://aimc2024.pubpub.org/pub/glemonic/release/1
This paper contains rich media, which is accessible in the linked Zenodo record.

glemöhnic

Project Description

glemöhnic is a ~20 minute performance piece that utilises SpeechBrain, OpenAI and CoquiTTS voice and speech models to transcribe, synthesise and clone historical and real-time improvised vocal gestures.

This piece explores how extra-normal [1] vocal sounds trigger and provoke nonsensical AI-mediations of human vocality, by utilising non-text audio as input material for text-expectant AI speech recognition and synthesis models. The mediation of non-textual human voice gestures by these ASR models yields eclectic, bizarre and poetic nonsense, which is further utilised as textual input for text-to-speech synthesis and voice cloning models. CoquiTTS' [2] XTTS_V2 model re-constructs the syllabic, phonemic and garbled poems into vocal clones that oscillate between their reference audio (the original audio dataset input) and the scraped audio data that the XTTS_V2 model has been trained on. The result of this is a collection of original and cloned audio samples that are utilised as sonic material in a live coded musical performance, using the strudelREPL platform.

To give a specific overview of how the AI model pipeline is implemented in ***glemöhnic,*** Image 1 below illustrates the overall flow for the models used, and will be systematically discussed.

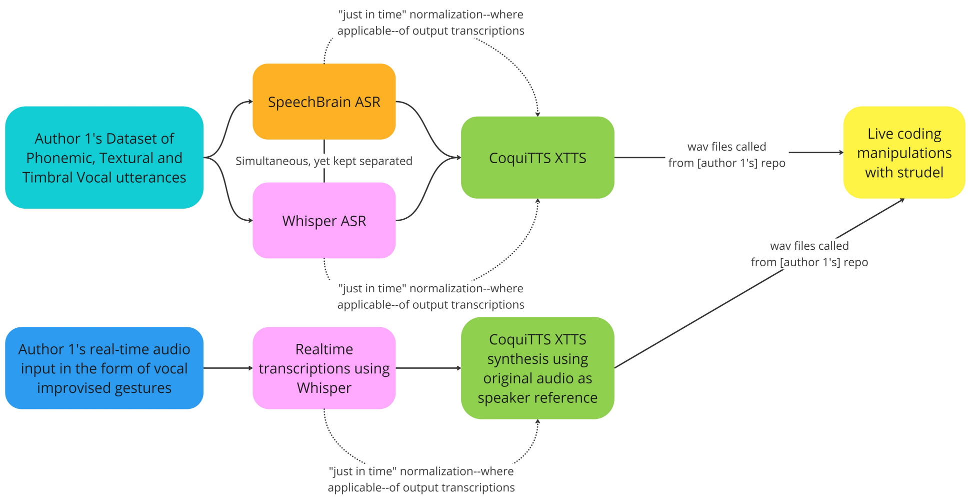
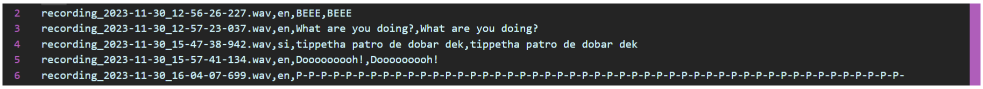

"just in time" normalization--where applicable--of output transcriptions
Author 1's Dataset of Phonemic, Textural and Timbral Vocal utterances
SpeechBrain ASR
Simultaneous, yet kept separated
Whisper ASR
CoquiTTS XTTS
wav files called from [author 1's] repo
Live coding manipulations with strudel
wav files called from [author 1's] repo
"just in time" normalization--where applicable--of output transcriptions
Author 1's real-time audio input in the form of vocal improvised gestures
Realtime transcriptions using Whisper
CoquiTTS XTTS synthesis using original audio as speaker reference
"just in time" normalization--where applicable--of output transcriptions

Image 1
Outline of the model pipeline for glemöhnic

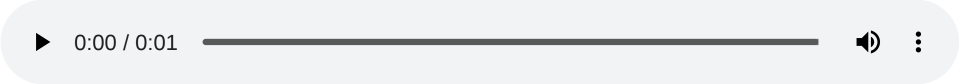
To explain the role of the SpeechBrain and Whisper ASR models pipeline, one branch (in orange in Image 1) utilises the SpeechBrain [3] wav2vec 2.0 ASR model. This model is an end-to-end system, comprised of a tokenizer block pretrained on the CommonVoice dataset and Facebook's acoustic model—Wav2Vec2-Large-LV60. The wav2vec 2.0 builds on Wav2Vec2 with the addition of 2 deep neural net layers and further finetuning on CommonVoice. The output representation is then parsed by a CTC decoder, which learns the alignment between input audio and output token sequences. The second branch (pink in Image 1) is OpenAI's [4] Whisper ASR model, which is a multilingual speech recognition model pretrained on 680,000 hours of audio and correlating transcripts scraped from the internet [5].

The purpose of using these two models in parallel is to expose the different quirks SpeechBrain and Whisper have as a result of the training of their respective datasets. The SpeechBrain model is pretrained on an English language corpus, which has difficulty parsing non-text-based speech and triggers the attention mechanism to enter phonemic "death-loops". Whisper, in comparison, is pretrained on a multilingual corpus of scraped audio and correlating transcripts, which yield unexpected mappings between more phonemic, textual, and timbral audio input and the original scraped audio dataset. This is illustrated below in Image 2.



```
2   recording_2023-11-30_12-56-26-227.wav,en,BEEE,BEEE
3   recording_2023-11-30_12-57-23-037.wav,en,What are you doing?,What are you doing?
4   recording_2023-11-30_15-47-38-942.wav,si,tippetha patro de dobar dek,tippetha patro de dobar dek
5   recording_2023-11-30_15-57-41-134.wav,en,Doooooooooh!,Doooooooooh!
6   recording_2023-11-30_16-04-07-699.wav,en,P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-P-
```

Image 2
Screenshot of transcriptions from Whisper, illustrating multilingual mappings



Audio 1
Audio excerpt referenced in line 2 of Image 2.

Audio 2
Audio excerpt referenced in line 4 of Image 2.

The next stage in the pipeline (green in Image 1) is CoquiTTS' XTTS_V2 cloning and text-to-speech synthesis model. This receives and loads individual input CSV transcriptions from the Whisper and SpeechBrain models and pairs them with their parent audio file (blue in Image 1). The transcriptions are used as prompt material for the XTTS_V2 model, and the parent audio is used as a speaker reference to synthesis the prompt material and clone it according to the parent audio. The resulting output is individual—or a bank[1] of—audio material that is then used as samples in the final pipeline stage (yellow in Image 1). In this final stage, the samples are then manipulated in real-time in the strudel platform to create BLAH. The overall intention of this real-time manipulation is to engage with a form of AI voice *to* construct sound collages inspired by Dada [6][7][8], grounded within rhythmic and gestural conventions of improvised voice.

Thematically, ***glemöhnic*** engages with the conference theme in its engagement of AI models that have been built, trained and deployed with certain intentionalities: namely to recognise, transcribe and synthesis human speech. It further ties into sound art exploration, historical art movements around language and text-as-sound (such as Dada and *zaum*).

Type of submission

glemöhnic would work best in Performance 2 at Oxford University. It could also work well at the Performance 3 club night at the Old Fire Station, as it utilises the strudel live coding platform to interact with the output XTTS cloned audio.

Technical/Stage Requirements

The performer will provide:

- own laptop
- microphones
- audio transmitter and receiver pack
- audio interface
- power plugs and UK power adaptors.

The performer can also bring all of her own audio cable, if required. The audio interface has two female XLR outputs, one for each stereo channel, which should go to the PA system.

The following is also required:

- a standing-height table that would suitably fit a laptop, audio interface, the transmitter and receiver
- stereo PA system (ideally with subwoofer) and basic theatrical lighting.

Program Notes

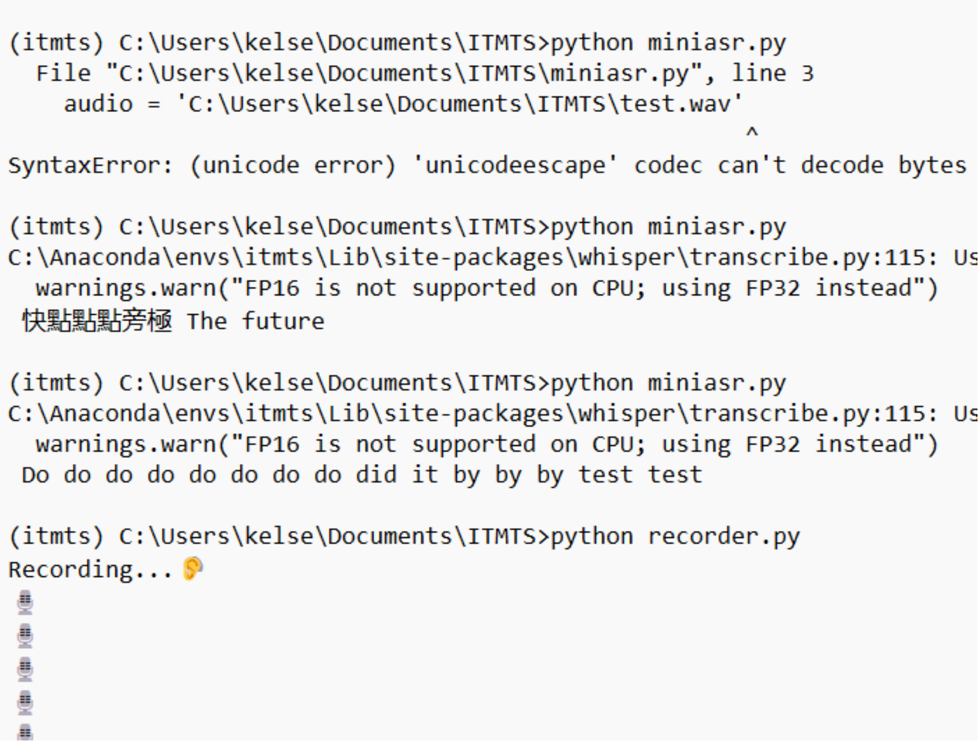
What secrets and poetry might be contained in our burps? Our giggles?

glemöhnic utilises explores how our vocal identities are pulled from us, and reformed into new vocal identities- through feeding a sigh or cough into text-oriented AI voice models and allowing them to unravel and warp.

This performance utilises open sourced speech recognition models from SpeechBrain, OpenAI and CoquiTTS to transcribe, synthesise, clone and mutate real-time vocal sounds and historical voice datasets. From this unraveling of non-text voice sounds by text-expectant AI models, scraped and forgotten voices are sung into being and rise from the dusty depths of scraped datasets. These new sounding bodies are reunited with their parent voice bodies, to create converging and diverging tapestries inspired by Dada and *zaum*.

glemöhnic catalyses the wayward mutations of human voice that occur when wordless voice is fed to word-dependent AI models.

Media



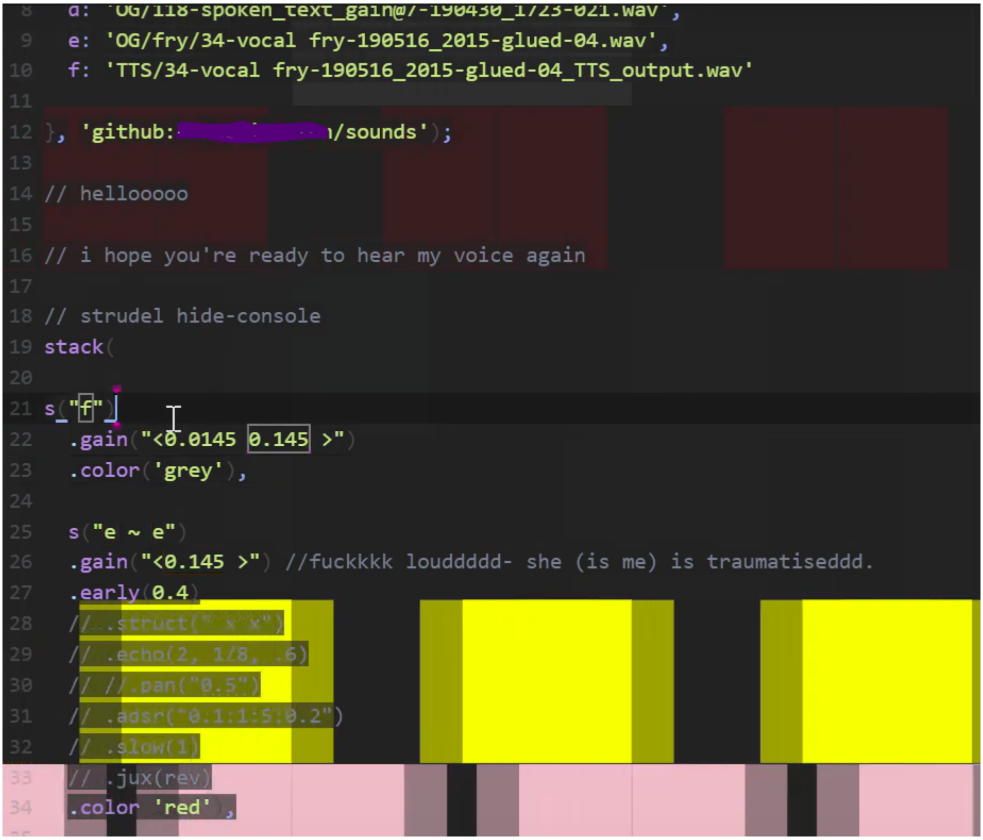
```
(itmts) C:\Users\kelse\Documents\ITMTS>python miniasr.py
  File "C:\Users\kelse\Documents\ITMTS\miniasr.py", line 3
    audio = 'C:\Users\kelse\Documents\ITMTS\test.wav'
                                                    ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes

(itmts) C:\Users\kelse\Documents\ITMTS>python miniasr.py
C:\Anaconda\envs\itmts\Lib\site-packages\whisper\transcribe.py:115: Us
  warnings.warn("FP16 is not supported on CPU; using FP32 instead")
 快點點點旁極 The future

(itmts) C:\Users\kelse\Documents\ITMTS>python miniasr.py
C:\Anaconda\envs\itmts\Lib\site-packages\whisper\transcribe.py:115: Us
  warnings.warn("FP16 is not supported on CPU; using FP32 instead")
 Do do do do do do do do did it by by by test test

(itmts) C:\Users\kelse\Documents\ITMTS>python recorder.py
Recording...👂
🎙
🎙
🎙
🎙
🎙
```

Image 3
Screenshot taken during a session gennerating glemöhnic samples



```
  d: 'OG/118-spoken_text_gain@7-190430_1723-021.wav',
  e: 'OG/fry/34-vocal fry-190516_2015-glued-04.wav',
  f: 'TTS/34-vocal fry-190516_2015-glued-04_TTS_output.wav'

}, 'github:[redacted]/sounds');

// hellooooo

// i hope you're ready to hear my voice again

// strudel hide-console
stack(

s("f")
  .gain("<0.0145 0.145 >")
  .color('grey'),

  s("e ~ e")
  .gain("<0.145 >") //fuckkkk louddddd- she (is me) is traumatiseddd.
  .early(0.4)
  //  .struct("  x x")
  //  .echo(2, 1/8, .6)
  // //.pan("0.5")
  //  .adsr("0.1:1:5:0.2")
  //  .slow(1)
  //  .jux(rev)
  .color 'red'),
```

Image 4
Screenshot taken during a live coding session with glemöhnic samples

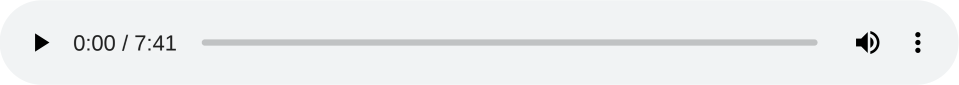
Some smaller audio excerpts are included below, as demonstration.

0:00 / 0:54

Audio 3
Smaller excerpt from glemöhnic

0:00 / 1:04

Audio 4
Smaller excerpt from glemöhnic



Audio 5
Smaller excerpt from glemöhnic

Artist Biography

Kelsey Cotton is a vocalist-artist-mover working with experimental music, Musical Artificial Intelligence, electronic textiles, soft-robotics, and Human-Computer Interaction. As a researcher, Kelsey is fascinated with pushing the limits of musical bodies, with her recent work delving deeper into designing artifacts which harness, augment and fuse different physiologies. She is passionate about somatic interaction, the potential for intersomatic experiences between fleshy and synthetic bodies, and first-person feminist perspectives of musical AI. Kelsey is currently undertaking PhD studies in Interactive Music and AI at Chalmers University of Technology in Gothenburg, Sweden.

Ethics Statement

The paper has utilised open-sourced, pretrained AI models and the first author’s own personal dataset. No human participants (other than the first author) were recruited for this study, and no sensitive data were collected. This paper has the intention to contribute to musical AI research, and to support future research within this community. The predicted environmental impact of this work as minimal since the computation required to create this work was comparable to daily personal computer usage. Accessibility of the technology in this work is limited with the general accessibility to computers and computational development frameworks.

Acknowledgements

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

Footnotes

1. If multiple audio files are called, or multiple live gestures are recorded in quick succession ↩

References

1. Edgerton, M. E. (2004). *The 21st-century voice: contemporary and traditional extra-normal voice*. Lanham, Md: Scarecrow Press. ↩
2. Kim, J., Kong, J., & Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. arXiv. Retrieved from http://arxiv.org/abs/2106.06103 ↩

3. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., … Bengio, Y. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. ↩

4. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. ↩

5. whisper/model-card.md at main · openai/whisper. (n.d.). Retrieved from https://github.com/openai/whisper/blob/main/model-card.md ↩

6. Beals, K. (2020). Decoding Dada: Avant-Garde Poetry in Its Cryptographic Context. *Dada/Surrealism*, *23*, 4. https://doi.org/10.17077/0084-9537.1359 ↩

7. Zurbrugg, N. (1979). Dada and the Poetry of the Contemporary Avant-Garde. *Journal of European Studies*, *9*(33–34), 121–143. https://doi.org/10.1177/004724417900903307 ↩

8. Mccaffery, S., Perloff, M., & Dworkin, C. (Eds.). (2009). Cacophony, Abstraction, and Potentiality: The Fate of the Dada Sound Poem. In *The Sound of Poetry / The Poetry of Sound* (p. 0). University of Chicago Press. https://doi.org/10.7208/chicago/9780226657448.003.0010 ↩

Part III

Appendix

Supplementary Materials for Paper IV and Paper V

All rich media materials included in Paper IV and Paper V are accessible via the following Zenodo record:

https://doi.org/10.5281/zenodo.13990776

- File name: P4-E1.pptx

  A collation of SpeechBrain ASR transcriptions into a volume of poetry, which is an artwork produced by Kelsey Cotton

- File names: P4-A1.wav; P4-A2.wav; P4-A3.wav; PA-A4.wav; P4-A5.wav; P4-A6.wav

  The above files are embedded within the publication "Sounding out extra-normal AI voice: Non-normative musical engagements with normative AI voice and speech technologies", published in the 2024 proceedings of the International Conference on AI and Musical Creativity in Oxford, United Kingdom. These files are excerpts of Kelsey's voice during various extra-normal vocal gestures; and translations of transcribed vocal gestures (from Kelsey) into different languages based on the deployment of an Automatic Speech Recognition model described in the paper. It is suggested to refer to the paper when referring to and listening to these examples.

- P5-A1.wav; P5-A2.wav; P5-A3.wav; P5-A4.wav; P5-A5.wav

  The above files are embedded within the publication "glemöhnic", published in the 2024 proceedings of the International Conference on AI and Musical Creativity in Oxford, United Kingdom. These files are excerpts from a performance pipeline described in the paper. The pipeline uses a combination of Automatic Speech Recognition models and Text to Speech synthesis models to produce a sample bank of warped clones of Kelsey's voice, which are then deployed as samples in the live coding platform strudelREPL. It is suggested to refer to the paper when referring to and listening to these examples.