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

AIMC 2024 (09/09 - 11/09)

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

Published on: Aug 29, 2024

URL: https://aimc2024.pubpub.org/pub/extranormal-aivoice

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.

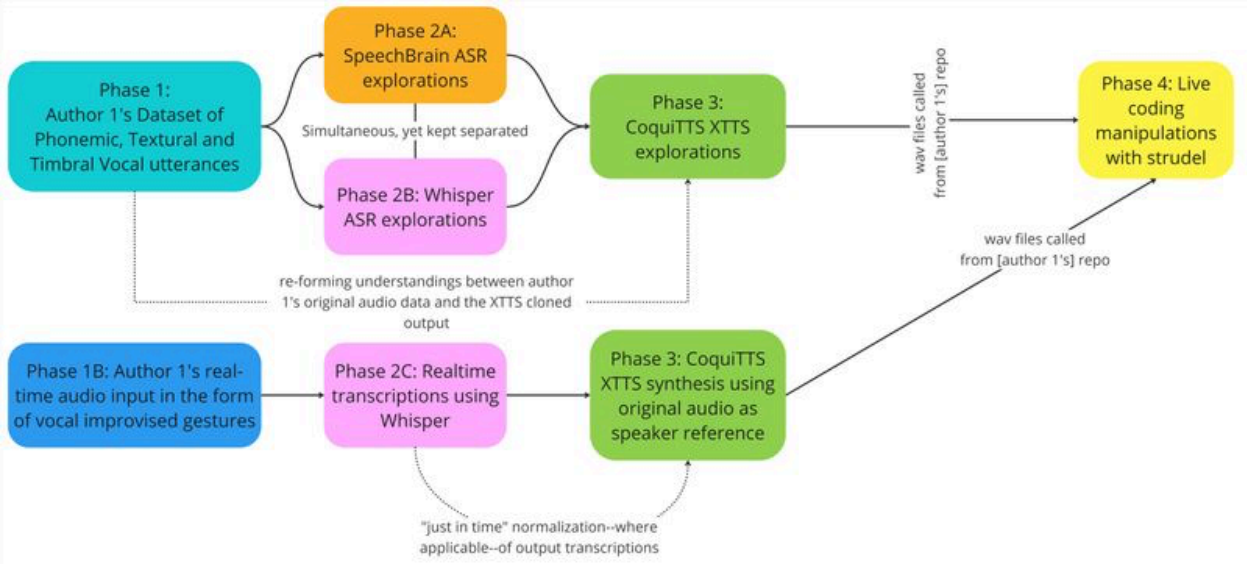


Figure 1
Chronology of Kelsey's working phases

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:

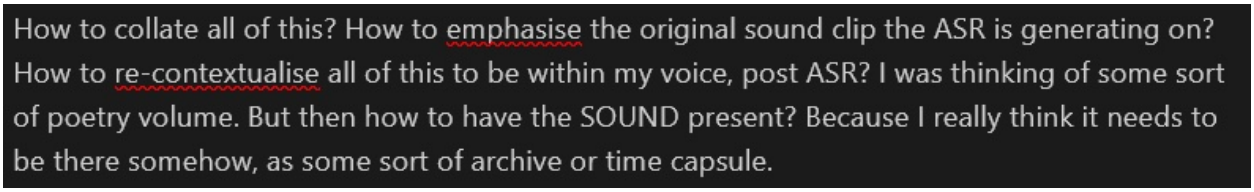


Figure 2
A screenshot from the first author's research diary

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.

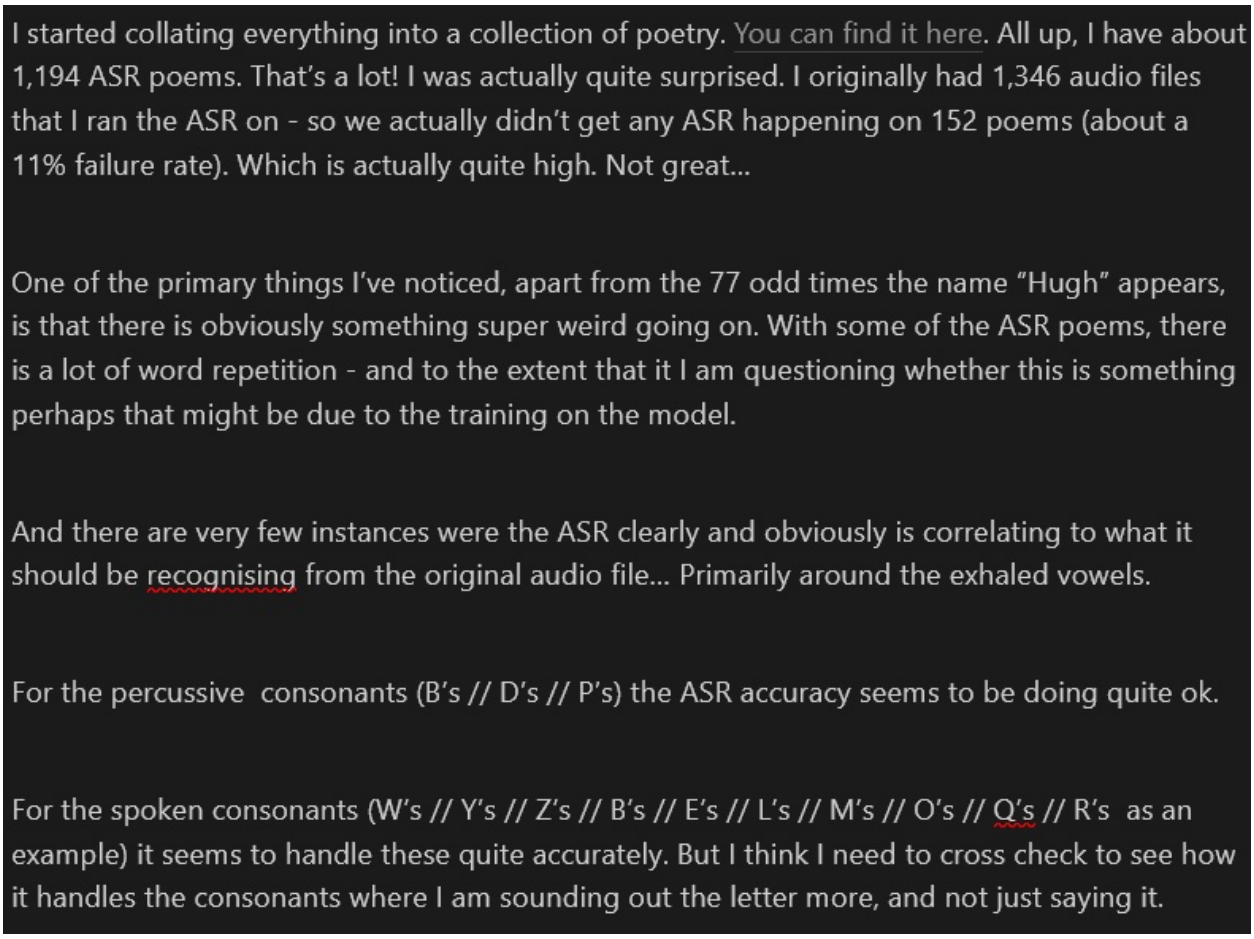


Figure 3
A screenshot from the first author's research diary

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:

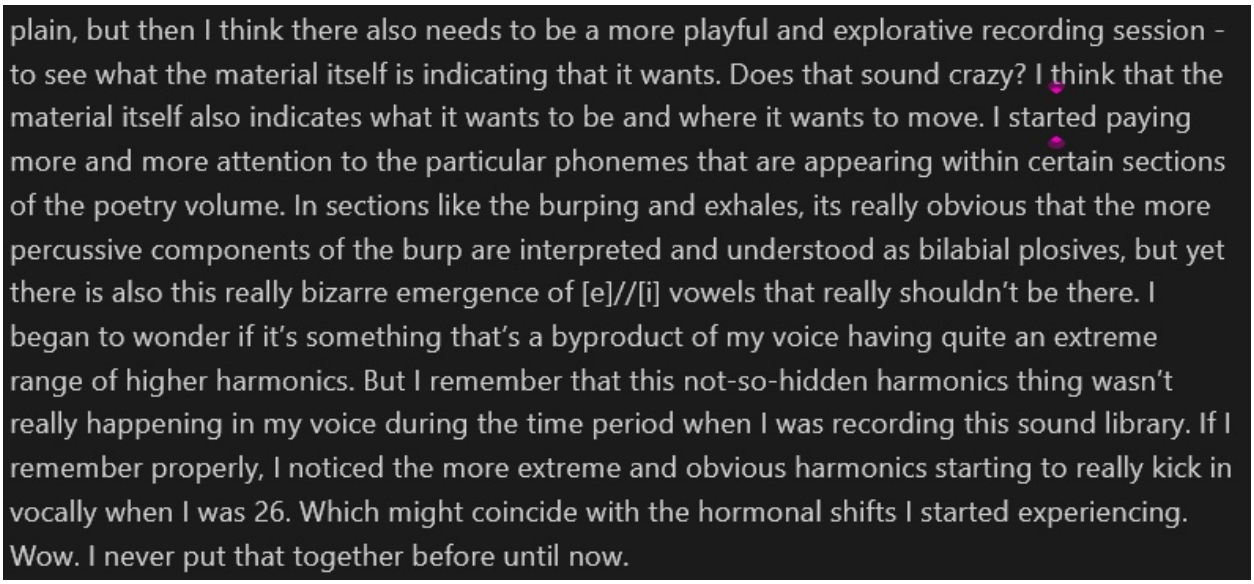


Figure 4
A screenshot from the first author's research diary

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.

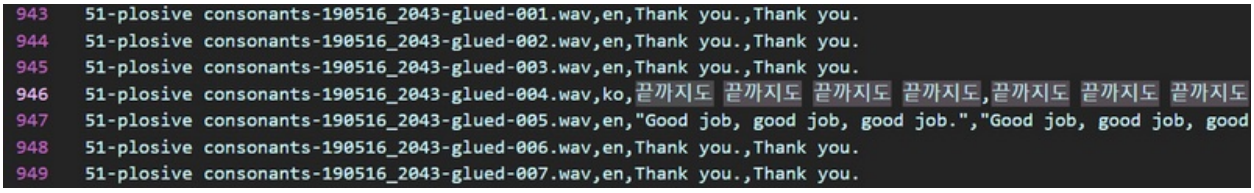


Figure 5
An excerpt from the first author's research logs, depicting a Korean transcription. Please see line 946 in the CSV file, featuring '51-plosive consonants-190516_2043-glued-004.wav' and a Korean transcription.

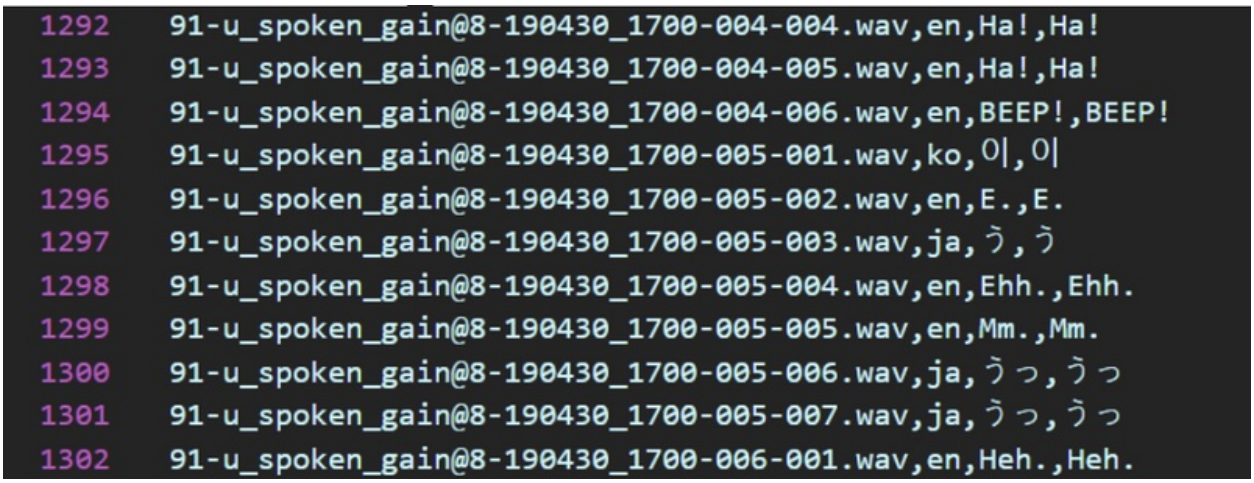
0:00 / 0:03

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

0:00 / 0:04

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,ㅇ|,ㅇ|
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

0:00 / 0:00

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.

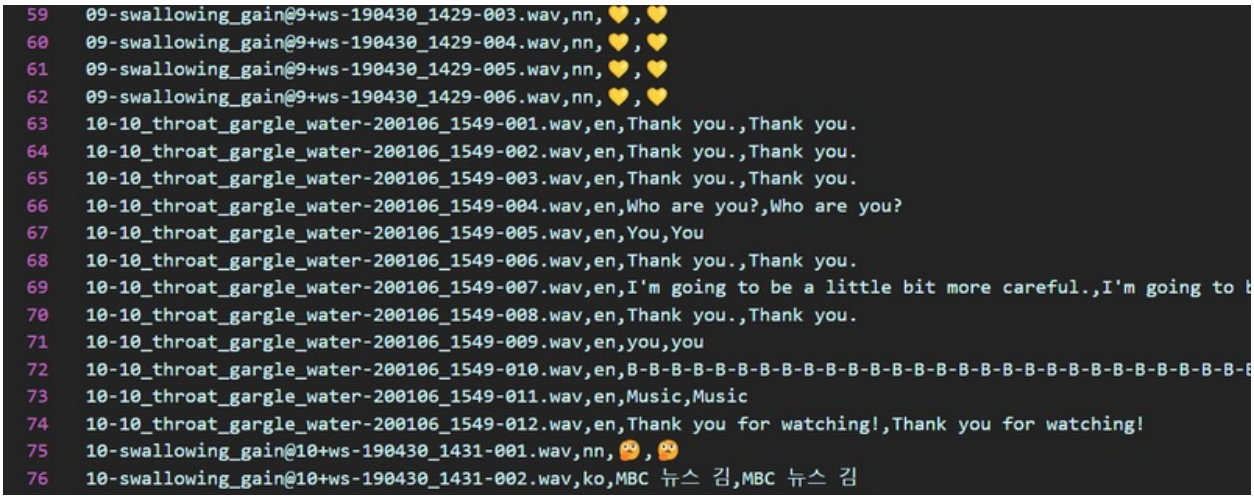


Figure 7
An excerpt from the first author's research logs, depicting multiple instances of emojis in the non-normalized transcription.

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 ↩