Sound Check: Auditing Recent Audio Dataset Practices

William Agnew[1], Julia Barnett[2], Annie Chu[2], Rachel Hong[3], Michael Feffer[1], Robin Netzorg[4], Harry H. Jiang[1], Ezra Awumey[1], Sauvik Das[1]
[1]Carnegie Mellon University
[2]Northwestern University
[3]University of Washington
[4]University of California Berkeley
wagnew@andrew.cmu.edu

Abstract

Audio AI models are increasingly used for a broad range of applications including music and sound generation, text-to-speech (TTS), voice cloning, emotion analysis, transcription, and audio classification. However, we have little understanding of the datasets used to create audio AI models, a gap that leaves the field without a powerful tool for understanding potential biases, toxicity, copyright violations, and other ethical and performance issues. We conduct a mapping literature review of hundreds of audio datasets used in recent music, sound, and speech AI papers. We first assess the sourcing, size, and usage of these datasets, finding that while there are hundreds of audio datasets, few are widely used. Next, we identify nine representative datasets and conduct several analyses to understand bias, toxicity, representation, and quality. We find that these datasets are often biased against women, have stereotypes about marginalized communities, and contain significant amounts of copyrighted work. We also find that audio datasets often come with scant documentation. To address this gap, we extend Gebru's datasheets for datasets to audio data, providing domain-specific documentation guidance. Finally, to facilitate public exploration of dataset contents and accountability, we developed an audio datasets exploration web tool which is available below in our links, along with our code and an extended version of our work including the appendix and augmented datasheets for datasets.
Content warning: this paper contains discussions of offensive language.

Web Tool — https://audio-audit.vercel.app/

Code — https://github.com/anniejchu/gen-audio-ethics

Extended version with appendix, including the augmented datasheets for datasets — https://arxiv.org/pdf/2410.13114

1 Introduction

Deep learning and ML-based techniques achieve state-of-the-art performance on a broad range of audio processing tasks, including speech transcription (Radford et al. 2023), pitch estimation (Kim et al. 2018; Liu et al. 2023b), and acoustic event classification (Wang et al. 2022). Beyond solving foundational problems in audio processing, these technologies support an increasing number of higher-level human-AI interactions. For example, virtual avatars, assistive technologies for the visually impaired, and novel UI paradigms use audio AI tools to improve user experiences (Danielescu et al. 2023; Upadhyay et al. 2023; Yu, Parde, and Chattopadhyay 2023). More recently, generative AI has led to the development of audio-based technologies for tasks including making reservations and generating music from text (Jacques et al. 2019; Wu et al. 2024).

This said, Audio AI has also been implicated and theorized in a range of ethical and societal concerns. Emotion analysis technologies have been deployed to make automated hiring decisions, with outcomes potentially worsened by bias (Akselrod and Venzke 2023). Voice actors and musicians have alleged their intellectual property (IP) is being improperly used in audio datasets, in particular by music generation startups Suno and Udio (Newton-Rex 2024a,b), the voice generation startup ElevenLabs (Soni 2024), and OpenAI, which allegedly copied Scarlet Johansson's voice without permission (Allyn 2024). Relatedly, voice AI, created using large and broad speech datasets, can accurately clone voices and have been used in financial scams (Granda 2024) and misinformation campaigns (Bond 2024; ANI 2024).

Our paper contributes a broad analysis of audio data currently used to train AI models spanning the domains of music, speech, and sound. This work is motivated by the fact that while the downstream ethical risks and harms of generative text and vision models have been the subject of significant prior work (Birhane et al. 2024a; Birhane, Prabhu, and Kahembwe 2021; Bianchi et al. 2023; Hong et al. 2024), there has been comparatively little focus and understanding of these issues in the context of generative audio, leading to a "documentation debt" (Bender et al. 2021) where widely used audio datasets are often poorly documented and understood. Inspired by audits of vision and text datasets that helped crystallize discussion of the ethical harms present in those generative AI modalities, in this paper we make the following contributions:

- We chart the recent landscape of audio datasets, systematizing their creation, size, usage, and submodality through a mapping literature review of hundreds of audio AI papers.
- We uncover issues with bias, representation, improper

use of intellectual property, and toxicity in nine widely used and representative audio datasets.

- We extend datasheets for datasets (Gebru et al. 2021) to audio datasets to help address their longstanding documentation debt.
- We create a web-based searchable index of these datasets to enable further exploration by the public.

2 Background

Before detailing the methods and results of our audio audits, we first briefly describe prior work in audio AI, research regarding ethical considerations of such developments, and work in dataset audits more generally.

2.1 Audio AI

To derive models from audio datasets, a number of methods have been introduced in the audio domain, mirroring developments in deep learning for text and images. This includes RNNs (Sturm et al. 2019), CNNs (Google Doodle Team 2019; Huang et al. 2017), combinations of the two (Donahue, Lipton, and McAuley 2017), and a recent turn to techniques leveraging transformer architectures (Agostinelli et al. 2023; Donahue et al. 2023; Garcia et al. 2023) and diffusion (Forsgren and Martiros 2022; Wang et al. 2023), especially as generative AI (GenAI) has captured public interest. The main architectures used in audio modeling are similar to those utilized in image and text, with adjustments made to handle the specifics of audio or speech data, such as the Audio Spectrogram Transformer (Gong, Chung, and Glass 2021), which utilize frequency representations of audio (e.g., spectrograms) to better model the different audio modalities. Large audio language models often will combine pre-trained large language models (LLMs) with audio-specific encoders to extend LLM capabilities to audio modalities (Tang et al. 2024; Gong et al. 2023; Chu et al. 2023). Generative audio approaches often borrow from LLMs (Borsos et al. 2023) or diffusion models that act on spectrograms or waveforms directly (Liu et al. 2023a).

While there are many semantically distinct audio modalities, they largely fall into three categories: music, speech, and (environmental) sound—largely referred to as audio by the community. While historically modeling in each of these modalities has primarily fallen under separate fields, recent advances in large audio language models have seen these distinct modalities being united under single frameworks (Ghosh et al. 2024; Gong et al. 2023). These models aim to perform classical tasks, such as Automatic Speech Recognition and Note Identification under a single modeling paradigm (Tang et al. 2024) and support applications spanning accessibility technologies to music generation. For more about audio AI, Civit et al. (2022) provide a review of music generation, Mehrish et al. (2023) review AI for speech processing, and Nogueira et al. (2022); Kelley and Dickerson (2020), and Palaniappan, Sundaraj, and Sundaraj (2014) review AI for sound processing.

2.2 Ethics and Societal Implications of Audio AI

In light of the recent turn to GenAI and improvements of other ML-based audio technologies, some researchers have started to grapple with corresponding ethical concerns and implications. However, as detailed by the literature review conducted by Barnett (2023) surveying generative audio research papers, few papers consider the potential negative impacts of their work. Even further, Morreale, Sharma, and Wei (2023) find that audio datasets are often created without permission of audio owners and creators. Some of these harms have started to be addressed, especially recently, such such as training data attribution of generative audio models (Barnett, Garcia, and Pardo 2024; Bralios et al. 2024).

Shelby et al. (2023) highlight the sociotechnical nature of AI harms and emphasize that harms cannot exist independent of societal norms and structures—they have to exist in a set of systems. Benjamin (2019) sheds light on how technological harms (not unlike most societal harms) have a disproportionate effect on people of color; technologies were built for the people in power and "often adopt the default norms and power structures of society." Audio harms are no exception; they can even be magnified when we do not understand the contents of the data. Modern voice cloning models boast of being able to capture diverse voices, but both being able and being unable to model voice archetypes like a "gay voice" comes with a host of both safety and representational harms (Sigurgeirsson and Ungless 2024). Being unable to model a particular voice archetype could result in harms of not representing all types of voices, while being able to could lead to harmful stereotyping and/or require data collection that could put participants at risk. Audio AI has also raised concerns about representation, culture, and data rights. For example, the Māori people have accused OpenAI's Whisper transcription AI of training on recordings of their language, *te reo* without permission, and then incorrectly transcribing *te reo*, potentially damaging its integrity (Mahelona et al. 2023a).

Audio deepfakes present a whole new set of harms separate from those already realized by visual and even video deepfakes. In 2023 alone, music featuring deepfake voices of popular artists went viral on social media (Coscarelli 2023; Feffer, Lipton, and Donahue 2023), prompting the music industry to start grappling with intellectual property concerns entailed by generative audio models (Hoover 2023; Johnson 2023; Patel 2023; Sisario 2024) and even take down online communities where deepfake audio was proliferating (Hook 2023). Audio deepfakes are particularly dangerous in the case of phishing and fraud, where bad actors can impersonate voices with high believability and deceive people or even bypass voice security systems (Habib et al. 2019; Sisman et al. 2020; Kim, Kim, and Yoon 2022). Many audio papers, especially text-to-speech papers (TTS), note the potential for misuse in form of audio deepfake (Wang et al. 2020; Kim et al. 2020; Kim, Kim, and Yoon 2022) and some even noted they had no plans to release their models due to the strong potential for misuse via deepfake (Kim, Kim, and Yoon 2022). Hutiri, Papakyriakopoulos, and Xiang (2024) detail the specific harms inherent to speech generators such as voice clones of voice actors, "bringing back the dead," and audio deepfakes of public figures. Batlle-Roca et al. (2023) focus on the specific aspect of transparency within generative music and highlight the link between transparency and creativity, originality, and ownership of AI-generated music, suggesting that we should move towards more transparent AI-based music generation. Within TTS, there are questions with regards to liability of harmful speech (Henderson, Hashimoto, and Lemley 2023) as well as harms of the reverse—a high potential for hallucination in speech-to-text (Koenecke et al. 2024) with an estimate of about 1% of transcriptions being entirely hallucinated.

Other ethical quandaries remain regarding the contribution of these models to climate change (Douwes, Esling, and Briot 2021; Holzapfel, Kaila, and Jääskeläinen 2024), speaker privacy and security (O'Reilly et al. 2024; Champion 2024), creativity (Khosrowi, Finn, and Clark 2023), GenAI's effect on music creators as a whole (Barnett 2023; Lee et al. 2022) and the ethics of using voice synthesis on deceased people (Feffer, Lipton, and Donahue 2023; Lee et al. 2022). Beyond risks for misinformation and economic harms to artists, recent high-profile instances of fraud (e.g., the transfer of millions of dollars to scammers leveraging GenAI to deceive targets (Lo 2024; Milmo 2024)), physiognomy (e.g., gender and sexual orientation classification (Lee et al. 2024)), and surveillance (e.g., gunshot detection for predictive policing (Crocco et al. 2016)) illustrate the real-world privacy, security, and ethical risks of these technologies.

2.3 Dataset Audits

Audits of datasets have proven vital for understanding the behavior and forecasting biases, toxicity, and other harms of downstream models. Prabhu and Birhane (2021) found that the 80 Million Tiny Images dataset contained racist and non-consensual intimate imagery (NCII), (Johnson 2020), and Birhane, Prabhu, and Kahembwe (2021) and Thiel (2023) uncovered evidence of child sexual abuse material (CSAM) in the LAION5B text-image dataset (Schuhmann et al. 2022), leading to removal of these datasets (Johnson 2020; Cole 2023). While dataset audits incorporate a variety of methods and aims—representation, toxicity, privacy, or copyright concerns—they all help determine how the targeted dataset's contents align with expectations in efforts to achieve accountability (Birhane et al. 2024b). Paullada et al. (2021) surveyed dataset audits and found they reveal representational harms and the presence of problematic content overlooked during data curation. Despite the impact of audits, Bender et al. (2021) argue that machine learning faces a dataset "documentation debt," with popular datasets having little if any documentation. Audio suffers acutely from documentation debt, with very few analyses of audio datasets outside of their suitability for increasing technical performance, with the notable exceptions of bias and representation audits of Mozilla Common Voice (Shuyo 2014) and Leschanowsky et al. (2024)'s audit of speaker recognition datasets used between 2012 and 2021. Our paper builds on these analyses to include recently used datasets across three major subgenres of audio–music, speech, and sounds–and concerns including bias, representation, stereotypes, data quality and quantity, data sourcing, and copyright.

3 Mapping Review of Current Audio Datasets and Models

To understand how many audio datasets exist, the distribution of their usage in the research community, and how these datasets were sourced, we conducted a mapping review enhanced with systematic elements (Ferrari 2015) utilizing the STAMP sampling method (Rogge et al. 2024) on audio modeling papers submitted to arXiv, a preprint platform previous studies have found to be an effective source for current and important audio AI papers (Barnett 2023). We searched for papers uploaded between May 1 2023 and May 1 2024 to capture one year of data and annotated the datasets included in these papers. We chose this time frame to capture contemporary usage of datasets in a field that has undergone rapid changes in recent years given fast growing academic and commercial interest in generative AI. We analyzed audio modeling papers until we approached dataset saturation (i.e., further analysis yielded few new datasets).

Our final corpus for the mapping review included 66 papers about music, 59 papers about speech, 19 papers about general audio (either environmental non-music, non-speech sounds, or general purpose audio), and five papers about music and speech (typically singing voice synthesis)—these categories are mutually exclusive. The authors then went through these papers and identified any audio datasets used for training or evaluation, yielding 175 unique datasets. Figure A5 in Appendix A.2 provides a visual representation of this process.

3.1 Analysis of Current Audio Datasets

We analyzed the 175 datasets found through our mapping review in order to understand practices, uses, and creation methods. For each dataset, we noted the number of times it was used by papers in our corpus, the number of times it had been cited overall (beyond our sample), its size in hours, the categorization of its contents (music/speech/general sound), how its corresponding data was collected, and concerns related to potential copyright infringement (see Table 1). For further detail on how we annotated and calculated these features, see Appendix A.2.

Distribution and Usage of Datasets Of the 175 datasets, the vast majority of them were only used in one ($n = 99$; 57%) or two papers ($n = 45$; 26%). The full distribution can be found in Figure A6. Only a handful of datasets were used more than 5 times. Speech datasets had the largest skew: most datasets were only used by one paper, while VCTK (Yamagishi et al. 2019) was used by 14. We find a wide variation in length even among the most popular datasets: the Mozilla Common Voice dataset is nearly two orders of magnitude larger than VCTK dataset, despite both being speech datasets. Speech datasets were also the largest by number of hours (see Appendix A.2) We documented 573,522 hours of speech data (median = 59 hours), the vast majority of which came from VoxPopuli (Wang et al. 2021), a 400,000 hour dataset consisting of European parliament

Overview of Datasets									
	Overall	Creation Method	Size (Hours)			Citations			Copyright
Category	Datasets	Scraped	Sum	Median	Mean	Sum	Median	Mean	Infringing
Music	61	36%	74,139	19	1,236	14,346	98	267	33%
Speech	80	24%	573,522	59	7,546	39,511	202	590	20%
Sounds	31	32%	37,178	64	1,377	11,998	159	461	35%
Music+Speech	3	0%	44	15	1	30	15	15	0%

Table 1: Descriptive statistics from 175 datasets identified in review. Split by music, speech and sound, we list the count of datasets, percent that were scraped, size in hours, total number of citations (beyond our corpus), and conservative estimate of the percent likely copyright infringing. This information was determined by two authors independently evaluating each dataset by reading the original papers proposing the datasets (when present), investigating all possible information provided online about the datasets, and lacking both of those downloading the dataset and assessing this information. For further detail on how we calculated hours and decided which datasets had potential copyright issues, see Appendix A.2.

event recordings, and the Spotify Podcast Dataset (Clifton et al. 2020), 100,000 hours of Spotify Podcasts that has been removed from public access. Music datasets totaled 74,139 hours, with a similar skew (median of 19 hours) driven by The Million Song Dataset (Bertin-Mahieux et al. 2011) (50,000 hours), Irish Massive ABC Notation Dataset (Wu et al. 2023) (7,200 hours), and Free Music Archive (Defferrard et al. 2017) (5,920 hours). These findings stand in contrast to text and image modalities, where there exist a smaller number of very large, widely used datasets that have significant source overlaps (Schuhmann et al. 2022; Gadre et al. 2024; Raffel et al. 2020; Soldaini et al. 2024).

We also calculated the total citations these datasets had received[1] to gauge popularity relative to usage. As seen in Figure 1, both usage and citations are quite fragmented, but actual citations have a heavier focus on a few important datasets. Similar to usage and hours, speech dominated the total citation count receiving 39,511 cumulative citations (median= 202), with LibriSpeech (Panayotov et al. 2015) in the lead with 6,136 citations. Music was second with 14,346 cumulative citations (median= 98); GTZAN led music datasets in citations with 4,345 citations. Datasets including sounds were least cited with 11,998 cumulative citations (median= 159), with AudioSet being most popularly cited at 3,204 citations.

In contrast to citing public datasets, many papers in our corpus did not release the data they used. Out of the 157 papers in our sample, 65 papers used at least one proprietary dataset, and there were 77 proprietary datasets in total. Of these 77 datasets, 79% ($n = 61$) were not released, often without a rationale.

Language Contents of Datasets When considering the linguistic diversity in speech datasets, the inclusion of underrepresented languages is frequently not prioritized, with many datasets predominantly featuring only one language. Out of the 77 speech datasets we examined, the majority (61) were monolingual, with 50 solely in English, followed by 7 in Mandarin. In contrast, 16 datasets encompassed between 2-30 languages, while only two datasets included more than 50 languages.

[1]Citation data accessed from Google Scholar in Fall of 2024 and may not be exhaustive. Moreover, oftentimes when someone cites a dataset, they are doing so in acknowledgment of the field (e.g., in the related literature section) as opposed to actually using that data for training or evaluation.

Sources of Datasets The two most prominent audio data sources were YouTube ($n = 25$ datasets) and LibriVox ($n = 13$). Other standouts were freesound.org ($n = 6$), Spotify ($n = 4$), and VCTK (Yamagishi et al. 2019) ($n = 3$). Other popular sources for audio content included podcasts ($n = 6$), marketplace websites ($n = 4$), TED talks ($n = 4$), TV shows ($n = 3$), and parliament/public speeches ($n = 3$). We found that LibriVox and its derivatives were referenced in 35 out of the 59 speech papers included in our literature review. LibriSpeech (Panayotov et al. 2015), a dataset comprised of public domain audiobooks read by volunteers across various languages, serves as a foundation for 13 derivative datasets, including both direct derivatives like the LibriSpeech dataset(Panayotov et al. 2015) and Musan (Snyder, Chen, and Povey 2015), as well as derivatives of derivatives like LibriLight (Kahn et al. 2020).

The most cited derivative datasets include LibriSpeech ($n = 12$) , LJSpeech ($n = 9$), LibriTTS ($n = 9$), and Libri-Light ($n = 6$). This trend aligns with the overall popularity of these datasets, gauged by citation numbers. It is crucial to emphasize that LibriVox predominantly comprises century-old texts, encompassing outdated and potentially problematic language, cultural perspectives, and social norms, which we corroborate in our audit in Section 4. Researchers utilizing LibriVox should be mindful of the potential introduction of bias and toxicity inherent in this dataset.

Takeaways from the Mapping Review The composition of audio datasets used by the research and commercial audio community is vastly different than that of text and vision. It is extremely fragmented—beyond a few notable datasets (e.g., VCTK (Veaux, Yamagishi, and MacDonald 2017) and AudioSet (Gemmeke et al. 2017)), researchers tended to use one-off, idiosyncratic datasets. These datasets also come with their own suite of problems; much of the contents in these datasets were likely sourced without creators' knowledge or consent, potentially infringing copyright or distributing content at scales beyond the original understanding of data providers. There is no standardized way to prepare, create, release, or even discuss audio datasets used in re-

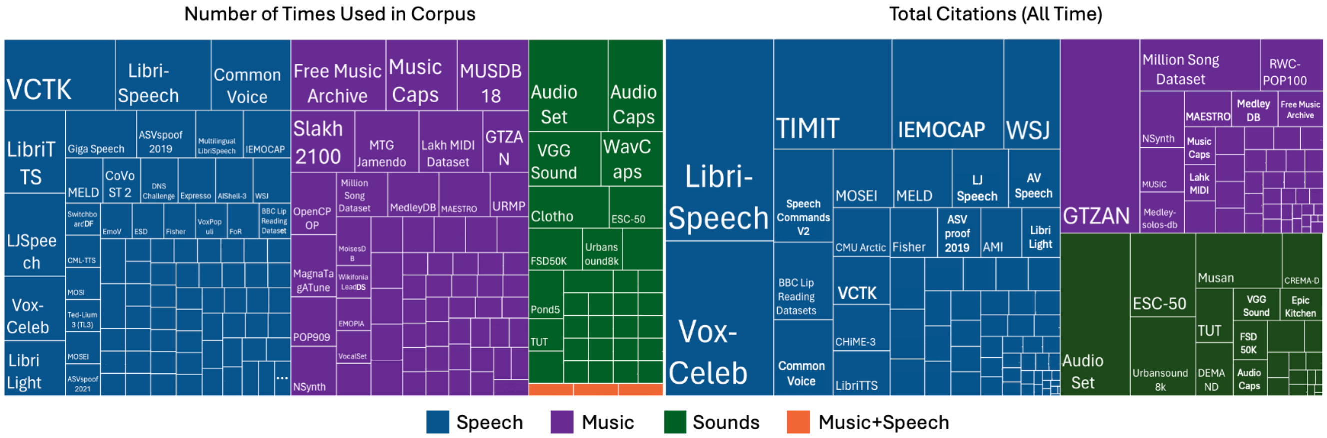

Figure 1: Area charts displaying the proportion of all 175 audited datasets by (1) number of times used in our paper corpus, (2) the cumulative total of citations received to date. Split into 4 categories: Speech, Music, Non-Music/Non-Speech sounds, Music and Speech combined. When datasets were so small their font was illegible, we instead display a blank box.

search. Much of the work done in this literature review was brute-force data diving to understand the composition of datasets—future datasets should be prepared and released in a more mindful and documented manner. We discuss a suggested approach at the end of our paper, adapting datasheets for datasets (Gebru et al. 2021) to the audio domain.

4 Audit of Audio Datasets

Given the lack of documentation and sheer number of audio datasets in recent use, we next conduct deeper inspections of nine audio datasets that we identified as representative. We choose these datasets by identifying the five most frequently used and the five largest datasets in each submodality: music, speech, and sound ($n = 24$ total after removing duplicates). We removed datasets that are not freely accessible ($n = 3$: Audiosparx, Million Song Dataset, and Spotify Podcast), or that were drawn from the same sources as larger datasets in our audit ($n = 10$: AudioCaps, Auto-ACD, Clotho, Kinnectics 70, Librilight, LibriTS, LJSpeech, MusicCaps, SLakh 2100, and VGG Sound). $n = 5$ datasets appeared multiple times in the largest and most used rankings. We exclude Voxpopulli ($n = 1$), as while it is very large, only a small fraction of it is transcribed, limiting its usefulness for audio AI. We exclude musdb18 ($n = 1$) due to its small size (150 tracks), and exclude the Irish ABC ($n = 1$) due to its specificity and lack of live recordings limiting its usefulness for training general purpose music or audio models. We are left with nine remaining datasets ($n = 9$), representative of the largest and most popular audio datasets in recent use across submodalities. We assess for bias, toxicity, representation, presence of potentially copyrighted content, and other important features for predicting behavior of downstream models.

1. *AudioSet* (Gemmeke et al. 2017), a dataset of 2 million 10-second YouTube clips, comprising a range of music, speech, and sounds;
2. *Mozilla Common Voice 17* (Ardila et al. 2020), a corpus of crowd-sourced sentences read by volunteers;
3. *VCTK* (Veaux, Yamagishi, and MacDonald 2017), a dataset of sentences read from the Herald, a Scottish newspaper, and several shorter accent elicitation texts;
4. *LibriVox* (LibriVox 2025), a dataset of volunteer recordings of public domain books;
5. *Free Music Archive* (Defferrard et al. 2017), a music-specific dataset scraped from an online repository;
6. *Jamendo* (Bogdanov et al. 2019), another scraped music-specific dataset,
7. *Wav-Caps* (Mei et al. 2024), a dataset of sounds sourced from AudioSet (Youtube), The BBC Sounds Effects library, FreeSound, and SoundBible;
8. *GigaSpeech* (Chen et al. 2021) a speech dataset sourced from YouTube, audiobooks, and podcasts; and lastly
9. the *Lakh MIDI Dataset* (Raffel 2016), a MIDI subset of the Million Song Dataset (Bertin-Mahieux et al. 2011), a currently unavailable dataset of music taken by The Echo Nest, a now defunct music analytics company.

4.1 Overview of Audio Datasets

Creation Dates Text-based audio datasets have two relevant creation dates: audio creation date and text creation date. VCTK was recorded in 2013 and features newspaper articles up to 2013 (Veaux, Yamagishi, and MacDonald 2017). LibriVox recordings were made between 2005 and present day (LibriVox 2025), but rely on public domain texts that typically enter the public domain 70 years after the death of their last living author (The University of California 2024), and are thus often over a century old. The Lakh MIDI dataset is derived from the Million Song Dataset, composed of songs released before 2012 (Raffel 2016). Similarly, AudioSet is composed of YouTube videos released before 2016 (Gemmeke et al. 2017). Of the audited datasets, only Mozilla Common Voice features both contemporary audio recordings and text, with creation dates between 2017 and present. Training cutoff date, or the date of the most recent training data, has emerged as a crucial concern of LLMs, determining which events they may possess knowledge (Cheng et al. 2024). The relative age of most audio datasets could lead to downstream models performing worse on new, or newly popular, words, sounds, and genres, reflecting a bias towards the past (Birhane et al. 2022).

Identity Representation Who is represented in the datasets is a central question of dataset audits with significant downstream impacts on bias. In Table 2 we summarize available information about audio creator demographics. One notable finding is that audio datasets often have limited demographic information, and several important demographic categories—including sexual orientation, disability, race, and ethnicity—are not documented in any of the datasets we audited. To assess representation and bias of different demographics, we instead investigated identity keywords in audio transcripts. In Figure 2, we present counts of identity keywords, using keywords from prior work (Dodge et al. 2021), for each dataset's transcripts, with plurals and common alternative spellings merged. With the exception of Mozilla Common Voice, we find "man" is used 2-10x more often than "woman" in these datasets. We also find that many identity keywords have a very low count in many datasets, with "Muslim" only appearing a thousand times (3.5x less than "Christian"), and "Nonbinary" appearing only 24 times. Gigaspeech and Mozilla Common Voice are the only datasets that contains at least one instance of every identity keyword studied. While this list of identity words is not comprehensive, and some words have alternate non-identity meanings, our results evidence low representation of marginalized groups in audio datasets.

Language Representation In Figure 3, we show the approximate number of hours of audio in each dataset for English and non-English languages, and in Table A5 we break down by language for non-English languages. We find that most audio datasets contain between 2x and 10x more data in English than in non-English, with the exception of Mozilla Common Voice, which contains 24,424 hours of non-English data and 3,507 hours of English data.

Sociodemographic bias We assess binary gender bias by comparing words with the highest pointwise mutual information (PMI) difference between "woman" and "man" keywords across all datasets (Appendix Figure A10). A higher PMI indicates that words are more strongly associated with each other. We find "woman"-related words are more associated with terms about families and childcare than "man"-related words, while "man"-related words are not correlated with typically gendered terms. In addition, we find that "woman"-related words have stronger associations with "baby," "beauty," and "b*tch." while "man"-related words have stronger associations with "dead" and "power." Overall, these associations provide evidence that women are commonly depicted in relation to families, childcare, and as subjects of the male gaze (Bloom 2017; Mulvey 2013). Non-binary genders were not represented well enough in these datasets to assess bias. However, given the prevalence of biased and toxic content towards queer people online (Queerinai et al. 2023), it is likely larger audio datasets will contain biased and toxic content towards these groups.

4.2 Toxicity

In this section we assess toxicity in the representative audio datasets. In Figure A8, we show the most common profane words in the datasets. We note that profanity is not the same as toxicity, and in some contexts these words are neither profane nor toxic. We find that, while all datasets contain at least some profane words, FMA, GigaSpeech and LibriVox contain by far the most, with many thousands of occurrences of racist and queerphobic terms. This finding may be due to LibriVox's sourcing from public domain texts, which are at a minimum 70 years old and generally much older, and represent times when toxic dialogues about marginalized populations were more overt.

In Figure 4, we present the number of hours of content in each dataset classified as toxic by the `pysentimento` toxicity classifier. This classifier considers not just profanity but additional textual cues to assess whether a text is toxic. Examples of and further discussion of sentences classified as toxic are available in Appendix A.4. While we find that each dataset has only approximately 1-3% of content flagged as toxic, this still amounts to hundreds of hours of toxic content. While levels of toxicity are low relative to the size of each dataset, LLMs have displayed an ability to recall text encountered only a few times during training (Carlini et al. 2022), raising the possibility that large audio models will exhibit similar behavior with even small amounts of profane or toxic content.

4.3 Audio Datasets Licensing

In Table A6, we summarize licenses of audio datasets. Mozilla Common Voice, VCTK, and LibriVox have permissive licenses that allow any use with minimal restrictions. The other six datasets all have licenses that potentially impact the ability of these datasets to be used for training or commercial applications. Lakh is derived from the Million Song Dataset, itself derived from Echo Nest, a music data service, which is subject to the Echo Nest License (The Echo Nest 2015), which prohibits commercial use. Lakh also contains many copyright tags, and we present the most frequent tags in Appendix Figure A11. AudioSet is derived from YouTube videos, which are licensed either under Creative Common Licenses (YouTube 2024a) with different levels of permissiveness, or the YouTube License (YouTube 2024b), where the creator retains all ownership. However, YouTube can use or modify videos in connection with YouTube's business, and users of YouTube can use or modify videos only as a feature of YouTube. We present the most common YouTube channel names in Figure A11 in our Appendix. Free Music Archive and Jamendo data are covered under several Creative Commons and other licenses, including many that prohibit commercial use and derivatives, require artist attribution, mandate that derivative works carry the same license, and restrict use to personal use only. Wav-Caps includes audio from YouTube under Creative Commons and Youtube licenses, and also content from the BBC Sound Effects Library under a BBC license that does not permit commercial use (BBC 2025). GigaSpeech

Dataset	Age	Gender	Sexual Orientation	Language	Locale/Country	Accent	Race/Ethnicity	Disability
Mozilla Common Voice	yes	yes	no	yes	yes	no	no	no
VCTK	yes	yes (binary)	no	yes	yes	yes	no	no
LibriVox	no	yes (binary)	no	yes	no	yes	no	no
Lakh	no	no	no	no	no	no	no	no
AudioSet	no	no	no	yes	yes	no	no	no
Free Music Archive	no	no	no	yes	yes	no	no	no
Jamendo	no	no	no	no	no	no	no	no
Wav-Caps	no	no	no	no	no	no	no	no
GigaSpeech	no	no	no	yes	no	no	no	no

Table 2: Documentation of demographics in audited datasets.

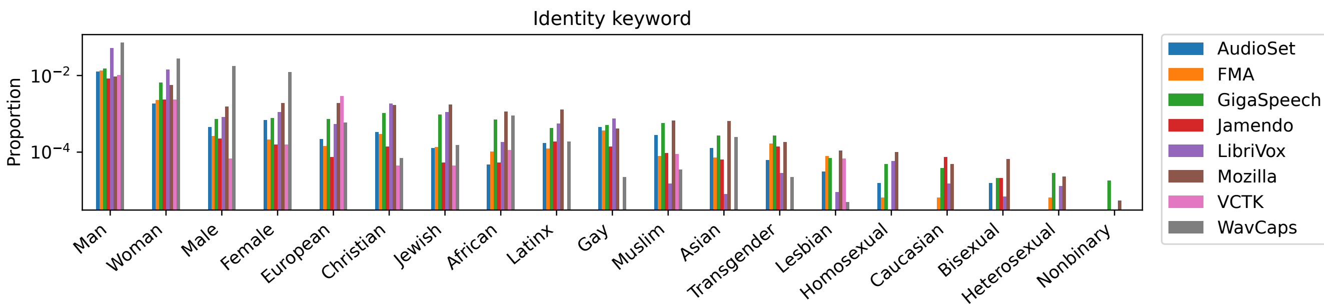

Figure 2: Proportion of identity keyword mentions for each dataset. Y-axis is in log-scale.

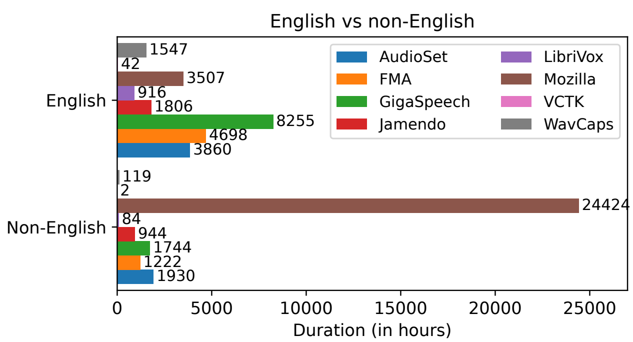

Figure 3: Estimated duration of each dataset by hours in English and not English.

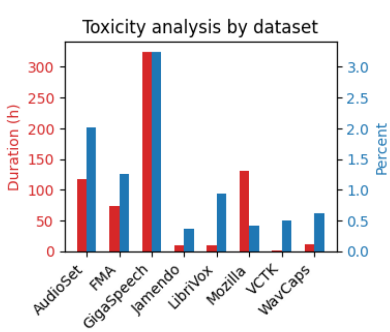

Figure 4: Estimated duration and percent of toxic-predicted English sentences split by dataset.

includes content from YouTube under Creative Commons and YouTube licenses, and also a variety of different podcasts, each with different licenses (including the 99% Invisible license (Visible 2025) and Australian Broadcasting Company license(ABC 2025)), many of which do not permit commercial use. In short, our analysis uncovers extensive presence of copyrighted material in these datasets from a broad range of artists and creators.

4.4 Audio Content: Listening Audit

Audio data is meant to be heard—so we conducted a "listening audit" of the selected datasets. Our descriptive review offers a qualitative assessment of audio quality and content. Three authors independently reviewed samples from the seven primary datasets analyzed in this paper. All three evaluated samples from AudioSet (Gemmeke et al. 2017), while at least two authors reviewed samples from each of the remaining six datasets. After listening, authors discussed their observations to reach consensus.

Quality Among the speech datasets considered, VCTK (Veaux, Yamagishi, and MacDonald 2017) stood out as very high quality, though exhibit minor issues, as some samples contain reverberation, muffling, or instances of speakers stumbling over words. In contrast, the majority of other datasets exhibit relatively low quality, often characterized by significant background noise and, in some cases, recordings that are nearly unintelligible. Mozilla Common Voice (Ardila et al. 2020) exhibited notably lower audio quality compared to VCTK and LibriVox. Frequent background noise often overshadowed the speech, and environmental or recording artifacts—such as audible keyboard clicks used to stop recordings—were common, increasing the likelihood of these artifacts being learned by models. Wav-Caps and GigaSpeech had varying levels of quality, with some clips apparently created in recording studio, and others with low quality microphones in noisy conditions.

In contrast, music datasets demonstrated considerably higher audio quality, featuring stereo channels and a sampling rate of 44.1 kHz. Both FMA and MTG-Jamendo datasets included a mix of vocal and non-vocal tracks, with a strong representation of samples having digital audio production elements such as sampling, synthesizers, and applied audio effects. Among these, MTG-Jamendo (Bogdanov et al. 2019) offered exceptionally high-quality audio, while FMA (Defferrard et al. 2017) was of noticeably lower quality, often exhibiting environmental noise and a grainy or distorted sound. Another notable observation was the lack of normalization, leading to significant variation in loudness levels across samples. The Lakh MIDI (Raffel 2016) dataset also demonstrated high quality, though it consists of MIDI files rather than raw audio, limiting its direct comparison with other datasets.

Content The source material for LibriVox—public domain books—primarily consists of older English texts, resulting in a vocabulary set that is not representative of contemporary English usage. Moreover, utterance lengths vary wildly—ranging from single-word utterances to fragmented sentences. This abrupt splicing of sentences before their natural conclusion led to orators delivering them as though they were complete, even when they ended mid-clause. These recordings may misrepresent the natural flow of human prosodic speech, reducing their suitability for tasks that rely on realistic speech patterns. In addition, we found that the Free Music Archive dataset (developed for music) contains many environmental sounds.

AudioSet (Gemmeke et al. 2017) differs from the other datasets in that it is a collection of URLs and 10 second time stamps links to Youtube videos, each with associated content tags. Many clips are over a decade old, have under 1,000 views, and would seldom surface to contemporary Youtube users. This practical privacy is disrupted by their inclusion in AudioSet. This concern is especially pertinent given that some of these videos had children or minors talking to camera (e.g., vlogs), raising ethical questions surrounding consent for using these data to train models. Categorizations (tags) were frequently incorrect or inconsistent—for example, a video tagged “bicycle” might include a bicycle in the footage but no accompanying “bicycle”-like sound. Inconsistent tagging is also evident in instances where two audio samples with similar content were assigned tags with different levels of detail, such as a produced song with vocals being tagged as “Music” in one case and “Music, Singing, female singing, musical instrument” in another. We found some categorizations to be potentially offensive, in particular the category “funny music” which we found applied to non-Western music. While “Speech” was a common tag, there was no distinction for language nor differentiation between background and foreground speech (e.g., speech meant to be intelligible vs. background conversation). Similarly, music is frequently tagged without differentiation between foreground music and incidental background music.

4.5 Privacy

Multiple datasets, including all datasets using YouTube, Free Music Archive, and Jamendo to source data, often contain metadata enabling determination of the names of speakers. Additional personally identifiable information is often available through metadata, audio, or–particularly on YouTube–video, including faces and locations. We found many instances of YouTube videos of children being included in datasets. Most critically, voiceprints are foundational to many of these datasets, and they contain voiceprints of tens of thousands of people. Voiceprints present with other PII, particularly names, are especially concerning, raising the possibility audio AI models could inadvertently learn to reproduce the voices of people in their training datasets, or malicious actors could use these datasets to impersonate the people within them.

5 Discussion

In this paper, we identified 175 audio datasets that have been recently used for generative audio AI. The distribution of their use is long-tailed, with a small number being used frequently, and a majority being used only once or twice. Many audio datasets were created specifically for research purposes, but approximately one third were scraped from online sources. These scraped datasets carry various licenses—often Creative Commons versions that bar commercial use, remixing, or require attribution—complicating their use for AI training. The corpora also contain many instances of profanity and toxicity. Sparse documentation and metadata obscure who is represented in audio datasets, highlighting a need for better documentation practices. Through a content analysis, however, we found that marginalized communities were less likely to be explicitly mentioned in audio datasets.

5.1 Impacts of Bias and Toxicity in Audio Datasets on Downstream Models

Our analysis indicates that the datasets we analyze in depth are biased both statistically (i.e., skewed inclusion of certain types of data) and socially (i.e., certain keywords and terms correlated with identity), in addition to containing non-trivial amounts of profane and potentially toxic content. Given that models in other modalities have displayed an ability to recall data encountered only a few times during training (Carlini et al. 2022), this raises the possibility that large audio models will exhibit similar behavior with even small amounts of profane or toxic content. These issues can also lead to disparate performance across socially salient categories. For instance, we observe heavy emphasis on the English language across datasets. Downstream of these datasets, Radford et al. (2023) note that while their Whisper transcription model obtains state-of-the-art performance on several tasks, they are limited in that “Whisper’s speech recognition performance is still quite poor on many languages” due to their “pre-training dataset [being] currently very English-heavy due to biases of [their] data collection pipeline.” Disparate performance of transcription via audio models across languages, and by extension cultures, is thus already documented as an impact of representation bias in audio datasets (Fuckner et al. 2023; Nacimiento-García, Díaz-Kaas-Nielsen, and González-González 2024).

Problems extend beyond transcription: translation, voice recognition, and audio generation face similar risks. Our PMI, identity keyword, and profanity analyses reveal gender associations with certain words, virtual omission of keywords pertaining to sexual orientation, and a high incidence of profanity related to sexual anatomy and racial slurs. If not properly addressed, these features of audio datasets may yield downstream audio generation models that perpetuate or even accelerate instances of social bias and toxicity that echo harmful stereotypes found in the wild (e.g., models that describe or portray women in a stereotyped way, models that are unable to create outputs relating to the LGBTQ experience, models that degenerate into profanity based on sexual anatomy or racial slurs without provocation, etc.).

Projects like the Common Voice corpus (Ardila et al. 2020), the Corpus of Regional African American Language (Kendall and Farrington 2023), and the Mid-Atlantic Gender Expansive Speech Corpus (Hope, Ward, and Lilley 2023) aim to improve the diversity of speech datasets. However, ethical dataset creation and use goes far beyond mere inclusion, especially in the context of AI and big tech—it can often be predatory, harming included communities by exposing them to data- and AI-intensified surveillance, poorly performing AI, and loss of control over data and culture (Mahelona et al. 2023b). Auditing datasets to uncover biases and a lack of representation is the start of the conversation, but effective solutions must be led by those who own the data (Mahelona et al. 2023b).

5.2 Scale of Audio Datasets

Audio datasets require considerably more storage per sentence than text datasets. AudioSet totals nearly 2.5TB in size and contains approximately 52 million spoken words. In contrast, C4 (Raffel et al. 2020), a text dataset approximately 30% of the (file) size of AudioSet, contains 153 billion words (Soldaini et al. 2024). Including the same breadth and depth of content in audio datasets as text datasets will require significantly more storage and compute which can both exacerbate known harms for large generative models, and create new ones. In particular, the requirement for more storage and compute limits which actors can train on such datasets. While most of the datasets of corporations training the largest audio models remain closed to external auditors, OpenAI Whisper (Radford et al. 2023) was trained on 680,000 hours of audio, and the BigSSL speech recognition model (Zhang et al. 2022) from Google was trained on one million hours of YouTube audio. Both of these datasets are over an order of magnitude larger than the largest freely accessible dataset in our survey. Thus, generative audio models may impose significantly higher barriers for participatory approaches to AI development than generative text models.

Finally, while the high carbon and water impacts (Criddle and Bryan 2024) of LLM training (Luccioni, Viguier, and Ligozat 2023) and inference (Everman et al. 2023; Luccioni, Jernite, and Strubell 2024) are well known, the scale of audio datasets raises concerns that audio and multimodal models will have even higher pollution and resource costs compared to purely text-based models (Douwes, Esling, and Briot 2021).

5.3 Source of Audio Datasets

The existence of proprietary datasets vastly larger than open datasets raises questions about the sources of these datasets. Google has explicitly sourced massive audio datasets from YouTube (Zhang et al. 2022), and Spotify released (and then removed) a 100,000 hour transcribed dataset from hosted podcasts (Clifton et al. 2020). While OpenAI has not indicated how the Whisper dataset was sourced, nor are they publicly known to host massive audio datasets, the few indications of the source of massive audio datasets we have point to existing commercial hosting and streaming corporations(Davis 2024).

Corporations once freely released massive audio datasets. Google released AudioSet, over 5,000 hours of YouTube audio, in 2017. However, AudioSet was released as links to YouTube videos that could be independently downloaded to obtain audio. When we tried to download this dataset in May 2024, we found YouTube had rate limiters and crawler IP blocking that would make downloading this dataset take several months. Similarly, Spotify released 100,000 hours of transcribed podcast audio in 2020, but then removed this dataset in December 2023, citing "shifting priorities" (Johnson 2023). Most explicitly, Universal removed its music from Tiktok in January 2024, partially citing concerns over AI generated music and AI covers of its songs (Hoskins 2024). In our analysis of current audio datasets (Section 3.1), there were 77 proprietary datasets utilized, of which 79% ($n = 61$) were not released. We argue these events constitute a pattern of increasing restrictiveness to proprietary audio datasets, mirroring recent trends in text data sources (Longpre et al. 2024). These restrictions seem in part motivated by the new value of massive audio datasets for training generative audio AI. As Ojewale et al. (2024) have noted, restrictions on dataset access is a major barrier auditors face, making the trend towards closed-source and proprietary datasets especially concerning.

5.4 Impacts to Rights and Intellectual Property

Any recording of a person's voice is a biometric identifier that can be misused for imitation, deception, or right-of-publicity violations. We found major audio datasets contained a wide array of content licensed under terms restricting commercial use, including audio from YouTube, the BBC, and a wide array of podcasts. A generated voice in an OpenAI product released in 2024 closely resembled that of actress Scarlett Johansson; she has since threatened legal action before the product, Sky, was quietly dropped (Allyn 2024; De Vynck 2024). The ability to mimic voice actors, musicians, and other artists raises concerns of economic and labor harms, where AI is used to undercut artists with their own data. With the threat to right of publicity greatly increased through GenAI, individuals in the US may seek recourse in the precedent of *Midler v. Ford Motor Co.* which ruled that voice may form part of an individual's identity, and that imitation of voice without approval is unlawful, though the specific boundaries of what consists of protected "identity" remains murky (Lapter 2007). Recently proposed US Senate legislation seeks to protect the public from generative AI; if passed, the bipartisan 2024 No AI Fakes Act would serve to protect individuals' voices and likenesses from AI deepfakes (Salazar 2024).

6 Recommendations

As initial steps towards ensuring harms that have plagued other modalities, including bias, toxicity, and intellectual property concerns, are not also present in future audio AI, we recommend (1) audio dataset developers adopt improved documentation to enable better assessment of bias and representation and (2) audio dataset developers only use data that permits remixing and commercial use at a minimum, but ideally seek active and informed consent for usage in AI. Both of these aims may be achieved by continuing existing practices of creating datasets in-lab specifically for AI.

6.1 Datasheets for Audio Datasets

To enable effective audio AI dataset documentation, we adapt and extend Gebru et al.'s "datasheets for datasets" (Gebru et al. 2021) and Papakyriakopoulos et al. (2023)'s augmented datasheets for speech datasets to the context of audio to guide ethical audio dataset development, documentation, and use. In particular, we add several questions that specifically speak to contents of and representation in audio data, in addition to questions assessing data provenance, consent, and copyright. This datasheet is intended to serve both as a reflexive practice for audio dataset creators, documentation standards for audio dataset publishers, and a guide for future audio dataset auditors. We provide the full datasheet in Appendix A1.

6.2 Educating and Mobilizing Data Workers

Documenting existing datasets is insufficient to improve dataset practices. As Birhane et al. (2024b) note in their review of audits, "Most of the academic work we reviewed focused on the process of evaluating AI systems for bias, fairness, or disparate impacts. Conversely, these studies rarely focused on other stages of auditing crucial to accountability in non-academic work, such as discovering harms, communicating audit results, or organizing non-technical interventions and collective action." Artists, creators, and other data workers learning about inclusion of their data in AI datasets have been a key catalyst for these communities discussing their wants and needs in regards to inclusion in AI datasets and subsequently organizing and advocating (Marx 2024). To support this collective action, we created https://audio-audit.vercel.app/, a website that enables anyone to search for their inclusion in prominent audio datasets. Modeled after dataset search tools used to assess inclusion of data in AI datasets (Willison 2023a, 2022b,a, 2023b), this website enables users to understand how their work is being used in audio AI.

7 Conclusion, Limitations, and Future Work

In this paper, we conducted a large-scale survey of audio datasets that are used in generative audio models: an audit broadly inclusive of speech, music, and sound datasets. We found that these datasets exhibit similar patterns of bias and toxicity as text and image datasets, raising the concern that audio models could, in turn, exhibit similar levels of bias and toxicity as LLMs and image models if these risks are not mitigated. We found that hundreds of audio datasets are in use, with several open datasets being significantly larger and more widely used than others. However, we found indications that datasets have recently started becoming more closed and commercial, with past sources of massive datasets, including YouTube and Spotify, taking down datasets or implementing new measures to block crawling of their audio repositories. The widespread presence of copyrighted material in many of these audio data sources, frequent use of proprietary audio datasets in research by corporations, and new commercial perceptions raise serious concerns for musicians, voice actors, and other audio workers releasing content online, especially on platforms like YouTube or Spotify.

Our audit has several limitations and only begins to address the dataset documentation debt (Bender et al. 2021) in audio AI. We were unable to assess the demographics of the people in audio datasets, and each of music, speech, and sound require specific and unique audits and analyses. We were only able to conduct high-level analyses of the contents of a small number of datasets, and each submodality has important distinctions that merit specialized treatment. We hope future audits will provide a deeper understanding of acoustic qualities of audio datasets, and that audio dataset curators will take steps to diversify their datasets, mitigate biases and toxicity, and remove copyrighted and other non-consensually sourced material.

Acknowledgments

William Agnew was supported by a CBI Bosch Fellowship. This work was funded, in part, by the National Science Foundation under grant CNS-2316287.

References

ABC. 2025. Terms of Use.

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325.

Akselrod, O.; and Venzke, C. 2023. How Artificial Intelligence Might Prevent You From Getting Hired.

Allyn, B. 2024. Scarlett Johansson says she is 'shocked, angered' over new ChatGPT voice.

ANI. 2024. Ranveer Singh's deepfake endorsing political party goes viral; actor flags alert on Instagram.

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech

Corpus. In *The Language Resources and Evaluation Conference (LREC)*, 4218–4222. Marseille, France: European Language Resources Association.

Barnett, J. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 146–161. Montreal, Canada: Association for Computing Machinery (ACM).

Barnett, J.; Garcia, H. F.; and Pardo, B. 2024. Exploring Musical Roots: Applying Audio Embeddings to Empower Influence Attribution for a Generative Music Model. arXiv:2401.14542.

Batlle-Roca, R.; Gómez, E.; Liao, W.; Serra, X.; and Mitsufuji, Y. 2023. Transparency in music-generative AI: A systematic literature review.

BBC. 2025. Licensing.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623. Virtual Event, Canada: Association for Computing Machinery (ACM).

Benjamin, R. 2019. *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Bertin-Mahieux, T.; Ellis, D. P.; Whitman, B.; and Lamere, P. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 591–596. Miami, US: International Society for Music Information Retrieval.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504.

Birhane, A.; Dehdashtian, S.; Prabhu, V.; and Boddeti, V. 2024a. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1229–1244.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024b. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. IEEE, Toronto, Canada: IEEE.

Bloom, J. D. 2017. *Reading the male gaze in literature and culture: Studies in erotic epistemology*. Cham, Switzerland: Springer.

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. Long Beach, CA, United States: PMLR.

Bond, S. 2024. A political consultant faces charges and fines for Biden deepfake robocalls.

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533.

Bralios, D.; Wichern, G.; Germain, F. G.; Pan, Z.; Khurana, S.; Hori, C.; and Le Roux, J. 2024. Generation or Replication: Auscultating Audio Latent Diffusion Models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1156–1160. IEEE, Seoul, South Korea: IEEE.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Champion, P. 2024. Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques. arXiv:2308.04455.

Chen, G.; Chai, S.; Wang, G.; Du, J.; Zhang, W.-Q.; Weng, C.; Su, D.; Povey, D.; Trmal, J.; Zhang, J.; et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Cheng, J.; Marone, M.; Weller, O.; Lawrie, D.; Khashabi, D.; and Van Durme, B. 2024. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. *arXiv preprint arXiv:2403.12958*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. arXiv:2311.07919.

Civit, M.; Civit-Masot, J.; Cuadrado, F.; and Escalona, M. J. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*, 209: 118190.

Clifton, A.; Pappu, A.; Reddy, S.; Yu, Y.; Karlgren, J.; Carterette, B.; and Jones, R. 2020. The Spotify Podcast Dataset. arXiv:2004.04270.

Cole, S. 2023. Largest Dataset Powering AI Images Removed After Discovery of Child Sexual Abuse Material.

Coscarelli, J. 2023. An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. *The New York Times*.

Criddle, C.; and Bryan, K. 2024. AI boom sparks concern over Big Tech's water consumption. *Financial Times*.

Crocco, M.; Cristani, M.; Trucco, A.; and Murino, V. 2016. Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4): 1–46.

Danielescu, A.; Horowit-Hendler, S. A.; Pabst, A.; Stewart, K. M.; Gallo, E. M.; and Aylett, M. P. 2023. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. In *Proceedings of the*

2023 CHI Conference on Human Factors in Computing Systems, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Davis, W. 2024. OpenAI transcribed over a million hours of YouTube videos to train GPT-4.

De Vynck, G. 2024. OpenAI rolls out voice mode after delaying it for safety reasons.

Defferrard, M.; Benzi, K.; Vandergheynst, P.; and Bresson, X. 2017. FMA: A Dataset For Music Analysis. In *18th International Society for Music Information Retrieval Conference*, 316–323. Suzhou, China: International Society for Music Information Retrieval.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Donahue, C.; Caillon, A.; Roberts, A.; Manilow, E.; Esling, P.; Agostinelli, A.; Verzetti, M.; Simon, I.; Pietquin, O.; Zeghidour, N.; and Engel, J. 2023. SingSong: Generating musical accompaniments from singing. arXiv:2301.12662.

Donahue, C.; Lipton, Z. C.; and McAuley, J. 2017. Dance dance convolution. In *International conference on machine learning*, 1039–1048. PMLR, Sydney, Australia: PMLR.

Douwes, C.; Esling, P.; and Briot, J.-P. 2021. Energy Consumption of Deep Generative Audio Models. arXiv:2107.02621.

Everman, B.; Villwock, T.; Chen, D.; Soto, N.; Zhang, O.; and Zong, Z. 2023. Evaluating the carbon impact of large language models at the inference stage. In *2023 IEEE international performance, computing, and communications conference (IPCCC)*, 150–157. IEEE, Anaheim, USA: IEEE.

Feffer, M.; Lipton, Z. C.; and Donahue, C. 2023. Deepdrake ft. bts-gan and taylorvc: an exploratory analysis of musical deepfakes and hosting platforms. In *Proceedings of the 2nd Workshop on Human-Centric Music Information Retrieval*. Milan, Italy: CEUR Workshop Proceedings.

Ferrari, R. 2015. Writing narrative style literature reviews. Medical Writing, 24 (4), 230-235.

Forsgren, S.; and Martiros, H. 2022. Riffusion: Stable diffusion for real-time music generation.

Fuckner, M.; Horsman, S.; Wiggers, P.; and Janssen, I. 2023. Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 146–151. IEEE.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Garcia, H. F. F.; Seetharaman, P.; Kumar, R.; and Pardo, B. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. In *Ismir 2023 Hybrid Conference*, 359–366. Milan, Italy: International Society for Music Information Retrieval.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, 776–780. New Orleans, LA: IEEE.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. arXiv:2406.11768.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778.

Gong, Y.; Liu, A. H.; Luo, H.; Karlinsky, L.; and Glass, J. 2023. Joint Audio and Speech Understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Google Doodle Team. 2019. Celebrating Johann Sebastian Bach Doodle - Google Doodles.

Granda, C. 2024. Fraudsters use voice-cloning AI to scam man out of 25,000.

Habib, R.; Mariooryad, S.; Shannon, M.; Battenberg, E.; Skerry-Ryan, R.; Stanton, D.; Kao, D.; and Bagby, T. 2019. Semi-supervised generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1910.01709*.

Henderson, P.; Hashimoto, T.; and Lemley, M. 2023. Where's the Liability in harmful AI Speech? *J. Free Speech L.*, 3: 589.

Holzapfel, A.; Kaila, A.-K.; and Jääskeläinen, P. 2024. Green MIR?: Investigating computational cost of recent music-Ai research in ISMIR. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Hong, R.; Agnew, W.; Kohno, T.; and Morgenstern, J. 2024. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. arXiv:2405.08209.

Hook, S. 2023. AI hub discord server hit with subpoena by RIAA.

Hoover, A. 2023. Spotify Has an AI Music Problem—but Bots Love It. *Wired*.

Hope, M.; Ward, C.; and Lilley, J. 2023. Nonbinary American English speakers encode gender in vowel acoustics. In *Proceedings of Interspeech*, 4713–4717. Dublin, Ireland: ISCA.

Hoskins, P. 2024. Universal Music to pull songs from TikTok. *BBC*.

Huang, C.-Z. A.; Cooijmans, T.; Roberts, A.; Courville, A.; and Eck, D. 2017. Counterpoint by Convolution. In *International Society for Music Information Retrieval (ISMIR)*, 211–218. Suzhou, China: International Society for Music Information Retrieval.

Hutiri, W.; Papakyriakopoulos, O.; and Xiang, A. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 359–376. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Jacques, R.; Følstad, A.; Gerber, E.; Grudin, J.; Luger, E.; Monroy-Hernández, A.; and Wang, D. 2019. Conversational Agents: Acting on the Wave of Research and Development. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, 1–8. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359719.

Johnson, A. 2023. Spotify Removes 'Tens Of Thousands' Of AI-Generated Songs: Here's Why. *Forbes*.

Johnson, K. 2020. MIT takes down 80 Million Tiny Images data set due to racist and offensive content. *VentureBeat*.

Kahn, J.; Rivière, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P. E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C.; Likhomanenko, T.; Synnaeve, G.; Joulin, A.; Mohamed, A.; and Dupoux, E. 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–7673. Barcelona, Spain: IEEE. https://github.com/facebookresearch/libri-light.

Kelley, T.; and Dickerson, K. 2020. A review of artificial intelligence (AI) algorithms for sound classification: Implications for human-robot interaction (hri). *Defense Technical Information Center (DTIC)*.

Kendall, T.; and Farrington, C. 2023. The Corpus of Regional African American Language. Version 2023.06. Eugene, Ore.: The Online Resources for African American Language Project.

Khosrowi, D.; Finn, F.; and Clark, E. 2023. Diffusing the Creator: Attributing Credit for Generative AI Outputs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 890–900. Montreal, Canada: Association for Computing Machinery (ACM).

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33: 8067–8077.

Kim, J. W.; Salamon, J.; Li, P.; and Bello, J. P. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 161–165. IEEE, Calgary, Canada: IEEE.

Kim, S.; Kim, H.; and Yoon, S. 2022. Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*.

Koenecke, A.; Choi, A. S. G.; Mei, K. X.; Schellmann, H.; and Sloane, M. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1672–1681.

Lapter, A. J. 2007. How the Other Half Lives (Revisited): Twenty Years Since Midler v. Ford. *Texas Intellectual Property Law Journal*.

Lee, H. P.; Yang, Y. J.; Von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. Honolulu, USA: Association for Computing Machinery (ACM).

Lee, K.; Hitt, G.; Terada, E.; and Lee, J. H. 2022. Ethics of Singing Voice Synthesis: Perceptions of Users and Developers. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 733–740. Bengaluru, India: International Society for Music Information Retrieval.

Leschanowsky, A.; Rusti, C.; Quinlan, C.; Pnacek, M.; Gorce, L.; and Hutiri, W. 2024. A Data Perspective on Ethical Challenges in Voice Biometrics Research. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

LibriVox. 2025. https://librivox.org/pages/about-librivox/. Accessed: 2025.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023a. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv:2301.12503.

Liu, Y.; Wu, P.; Black, A. W.; and Anumanchipalli, G. K. 2023b. A Fast and Accurate Pitch Estimation Algorithm Based on the Pseudo Wigner-Ville Distribution. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE, Rhodes Island, Greece: IEEE.

Lo, C. 2024. Hong Kong worker transfers HK$4 million after call from deepfake UK firm 'CFO'.

Longpre, S.; Mahari, R.; Lee, A.; Lund, C.; Oderinwale, H.; Brannon, W.; Saxena, N.; Obeng-Marnu, N.; South, T.; Hunter, C.; et al. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. *arXiv preprint arXiv:2407.14933*.

Luccioni, A. S.; Viguier, S.; and Ligozat, A.-L. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253): 1–15.

Luccioni, S.; Jernite, Y.; and Strubell, E. 2024. Power hungry processing: Watts driving the cost of AI deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 85–99. Rio de Janeiro, Brazil: Association for Computing Machinery (ACM).

Mahelona, K.; Leoni, G.; Duncan, S.; and Thompson, M. 2023a. OpenAI's Whisper is another case study in Colonisation.

Mahelona, K.; Leoni, G.; Duncan, S.; and Thompson, M. 2023b. OpenAI's Whisper is another case study in Colonisation.

Marx, P. 2024. How artists are fighting generative AI.

Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; and Poria, S. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99: 101869.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Milmo, D. 2024. UK engineering firm Arup falls victim to £20m deepfake scam. *The Guardian*.

Morreale, F.; Sharma, M.; and Wei, I.-C. 2023. Data Collection in Music Generation Training Sets: A Critical Analysis. In *Ismir 2023 Hybrid Conference*, 37–46. Milan, Italy: International Society for Music Information Retrieval.

Mulvey, L. 2013. Visual pleasure and narrative cinema. In *Feminism and film theory*, 57–68. London, UK: Routledge.

Nacimiento-García, E.; Díaz-Kaas-Nielsen, H. S.; and González-González, C. S. 2024. Gender and Accent Biases in AI-Based Tools for Spanish: A Comparative Study between Alexa and Whisper. *Applied Sciences*, 14(11): 4734.

Newton-Rex, E. 2024a. Suno is a music AI company aiming to generate $120 billion per year. But is it trained on copyrighted recordings?

Newton-Rex, E. 2024b. Yes… Udio's output resembles copyrighted music, too.

Nogueira, A. F. R.; Oliveira, H. S.; Machado, J. J.; and Tavares, J. M. R. 2022. Sound classification and processing of urban environments: A systematic literature review. *Sensors*, 22(22): 8608.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *arXiv preprint arXiv:2402.17861*.

O'Reilly, P.; Jin, Z.; Su, J.; and Pardo, B. 2024. Maskmark: Robust Neural watermarking for Real and Synthetic Speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4650–4654. IEEE, Seoul, South Korea: IEEE.

Palaniappan, R.; Sundaraj, K.; and Sundaraj, S. 2014. Artificial intelligence techniques used in respiratory sound analysis–a systematic review. *Biomedizinische Technik/Biomedical Engineering*, 59(1): 7–18.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE, Brisbane, Australia: IEEE.

Papakyriakopoulos, O.; Choi, A. S. G.; Thong, W.; Zhao, D.; Andrews, J.; Bourke, R.; Xiang, A.; and Koenecke, A. 2023. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 881–904.

Patel, N. 2023. Google and YouTube are trying to have it both ways with AI and copyright - The Verge.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Prabhu, V. U.; and Birhane, A. 2021. Large datasets: A pyrrhic win for computer vision. In *Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Applications of Computer Vision*. Virtual: IEEE.

Queerinai, O. O.; Ovalle, A.; Subramonian, A.; Singh, A.; Voelcker, C.; Sutherland, D. J.; Locatelli, D.; Breznik, E.; Klubicka, F.; Yuan, H.; et al. 2023. Queer in AI: a case study in community-led participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1882–1895. Chicago, USA: Association for Computing Machinery (ACM).

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR, Honolulu, USA: PMLR.

Raffel, C. 2016. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rogge, A.; Anter, L.; Kunze, D.; Pomsel, K.; and Willenbrock, G. 2024. Standardized sampling for systematic literature reviews (STAMP method): ensuring reproducibility and replicability. *Media and Communication*, 12.

Salazar, M. E. 2024. Nurture Originals, Foster Art, and Keep Entertainment Safe Act of 2024. Draft bill, 118th Congress, 2d Session, U.S. House of Representatives. Introduced 4 September 2024.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Shuyo, N. 2014. langdetect.

Sigurgeirsson, A.; and Ungless, E. L. 2024. Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices. *arXiv preprint arXiv:2406.07504*.

Sisario, B. 2024. Universal Music Group Pulls Songs From TikTok. *The New York Times*.

Sisman, B.; Yamagishi, J.; King, S.; and Li, H. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 132–157.

Snyder, D.; Chen, G.; and Povey, D. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; Hofmann, V.; Jha, A. H.; Kumar, S.; Lucy, L.; Lyu, X.; Lambert, N.; Magnusson, I.; Morrison, J.; Muennighoff, N.; Naik, A.; Nam, C.; Peters, M. E.; Ravichander, A.; Richardson, K.; Shen, Z.; Strubell, E.; Subramani, N.; Tafjord, O.; Walsh, P.; Zettlemoyer, L.; Smith, N. A.; Hajishirzi, H.; Beltagy, I.; Groeneveld, D.; Dodge, J.; and Lo, K. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv:2402.00159.

Soni, A. 2024. Actors Say AI Voice-Over Generator ElevenLabs Cloned Likenesses.

Sturm, B. L.; Ben-Tal, O.; Monaghan, Ú.; Collins, N.; Herremans, D.; Chew, E.; Hadjeres, G.; Deruty, E.; and Pachet, F. 2019. Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, 48(1): 36–55.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

The Echo Nest. 2015. Terms of Service.

The University of California. 2024. The public domain.

Thiel, D. 2023. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical report, Stanford University, Palo Alto, CA, 2023. URL https://purl . . . .

Upadhyay, P.; Heung, S.; Azenkot, S.; and Brewer, R. N. 2023. Studying Exploration & Long-Term Use of Voice Assistants by Older Adults. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Veaux, C.; Yamagishi, J.; and MacDonald, K. 2017. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.

Visible, . P. 2025. Terms of Service.

Wang, C.; Riviere, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; and Dupoux, E. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993–1003. Online: Association for Computational Linguistics.

Wang, R.; Juefei-Xu, F.; Huang, Y.; Guo, Q.; Xie, X.; Ma, L.; and Liu, Y. 2020. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*, 1207–1216.

Wang, Y.; Ju, Z.; Tan, X.; He, L.; Wu, Z.; Bian, J.; and Zhao, S. 2023. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. arXiv:2304.00830.

Wang, Y.; Wei-Kocsis, J.; Springer, J. A.; and Matson, E. T. 2022. Deep learning in audio classification. In *International Conference on Information and Software Technologies*, 64–77. Springer, Kaunas, Lithuania: Springer.

Willison, S. 2022a. Exploring 10m scraped Shutterstock videos used to train Meta's Make-A-Video text-to-video model.

Willison, S. 2022b. Exploring the training data behind Stable Diffusion.

Willison, S. 2023a. Exploring MusicCaps, the evaluation data released to accompany Google's MusicLM text-to-music model.

Willison, S. 2023b. What's in the RedPajama-Data-1T LLM training set.

Wu, S.; Li, X.; Yu, F.; and Sun, M. 2023. TunesFormer: Forming Irish Tunes with Control Codes by Bar Patching. arXiv:2301.02884.

Wu, S.-L.; Donahue, C.; Watanabe, S.; and Bryan, N. J. 2024. Music ControlNet: Multiple Time-Varying Controls for Music Generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32: 2692–2703.

Yamagishi, J.; Veaux, C.; MacDonald, K.; et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 271–350.

YouTube. 2024a. Creative Commons.

YouTube. 2024b. Terms of Service.

Yu, J. E.; Parde, N.; and Chattopadhyay, D. 2023. "Where is history": Toward Designing a Voice Assistant to help Older Adults locate Interface Features quickly. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Zhang, Y.; Park, D. S.; Han, W.; Qin, J.; Gulati, A.; Shor, J.; Jansen, A.; Xu, Y.; Huang, Y.; Wang, S.; et al. 2022. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1519–1532.