Right the docs: Characterising voice dataset documentation practices used in machine learning

Kathy Reid
School of Cybernetics
Australian National University
Canberra, Australia
kathy.reid@anu.edu.au

Elizabeth T. Williams
School of Cybernetics
Australian National University
Canberra, Australia
elizabeth.williams@anu.edu.au

Abstract

Voice-enabled technologies such as virtual assistants are quickly becoming ubiquitous. Their functionality relies on machine learning (ML) models that perform tasks such as automatic speech recognition (ASR). These models, in general, currently perform less accurately for some cohorts of speakers, across axes such as age, gender and accent; they are biased.

ML models are trained from large datasets. ML Practitioners (MLPs) are interested in addressing bias across the ML lifecycle, and they often use dataset documentation here to understand dataset characteristics. However, there is a lack of research centred on *voice* — spoken language — dataset documentation. Our work makes an empirical contribution to this gap, identifying shortcomings in voice dataset documents (VDD), and arguing for actions to improve them.

First, we undertake 13 interviews with MLPs who work with voice data, exploring how they use VDDs. We focus here on MLP roles and trade-offs made when working with VDDs. Drawing from the literature and from interview data, we create a rubric through which to analyse VDDs for nine voice datasets. Triangulating the two methods in our findings, we show that VDDs are inadequate for the needs of MLPs on several fronts. VDDs currently codify voice data characteristics in fragmented ways that make it difficult to compare and combine datasets, presenting a barrier to MLPs' bias reduction efforts.

We then seek to address these shortcomings and "right the docs" by proposing improvement actions aligned to our findings.

1 Introduction, motivation and previous work

Voice-enabled technologies, such as virtual assistants and smart speakers, are "going to scale" through axes such as volume (Kinsella and Mutchler, 2020; Bradley, 2020; Van der Meulen and Forni, 2016), geographies (Popović et al., 2015; Jones, 2020; Kendall et al., 2020), miniaturisation (Bouraoui et al., 2017), expanding use cases (Dale, 2020; Brewer et al., 2022; Jesús-Azabal et al., 2019) and use in multiple modalities (Baevski et al., 2022). Speech technology has become part of the fabric of modern *information infrastructures* — the technical capabilities, social norms, organisational practices and economic mechanisms (Bowker et al., 2009; Turow, 2021) — that collectively allow us to speak with machines and have them do our bidding. As voice technology becomes ubiquitous, so too does the potential societal impact of its bias. A person's poor voice interaction experience is no longer confined to a virtual assistant in the home, or to a mobile phone, but extends to the workplace, the car, healthcare, and customer service settings.

These systems use machine learning (ML)-enabled components like automatic speech recognition (ASR). However, they don't yet work well for everyone (Liu et al., 2022; Ngueajio and Washington, 2022; Feng et al., 2021). They exhibit bias — defined here as systematic and unfair discrimination against individuals or cohorts of individuals in favour of others (Friedman and Nissenbaum, 1996) [1] — across axes such as age (Vipperla et al., 2010; Gerosa et al., 2007), gender (Tatman, 2017; Tatman and Kasten, 2017; Garnerin et al., 2020), race (Koenecke et al., 2020), nationality (Hutiri and Ding, 2022), and accent (Hinsvark et al., 2021). Dataset documentation is a frequent tool used by MLPs to mitigate bias.

1.1 Dataset documentation and its use by MLPs

The ML-enabled components in voice-enabled technologies require large datasets to be effective.

[1] We recognise that bias manifests in many ways and has several interpretations, and suggest (Barocas et al., 2019) for a more complete treatment.

Dataset documentation — descriptive information characterising the nature, contents and provenance of a dataset — affords MLPs a clearer understanding of a dataset's characteristics before the dataset is used as an *input* to an ML model. This allows the detection of some forms of bias, such as underrepresentation of speakers having specific characteristics. In contrast, *model documentation* — descriptive information characterising the performance of a trained ML model against evaluative criteria — focuses on the performance *output* of ML processes. It provides MLPs the opportunity to detect and remediate bias issues such as poor inference accuracy for specific types of speech. Both types of documentation are well established in the literature as tools to detect and prevent bias in ML.

Bender and Friedman (2018) introduce *Data statements* for natural language processing (NLP), where they propose collecting information such as speaker demographics, annotator demographics, and the domain and context of the material as a way to address bias in written text corpora. Gebru et al. (2021) brings data provenance to the forefront of broader ML practice by outlining key areas MLPs should consider, such as the purpose and intended use of the dataset, the objects it stores, how they're represented, the relationships between them, sources of error and noise, sensitivity and identification considerations, how the data was collected and labelled, and how the datasets are distributed and maintained. Boyd (2021) seeks to empirically validate the utility of datasheets, and demonstrates their benefit by having MLPs ethically reflect on problematic datasets — directly connecting datasheets as an artifact with improved practice. From the field of computer vision, Miceli et al. (2021) also focus on praxis, emphasising the need for practitioner reflexivity in the production of ML datasets. Similarly, in an effort to make the ethical considerations and choices made during the production of datasets produced through crowd-sourced annotations more transparent, Díaz et al. (2022) develop the *CrowdWorkSheets* framework. McMillan-Major et al. (2023) focus on adoption of dataset documentation, working with NLP practitioners to increase uptake.

In Costa-jussà et al. (2020), we see the adaptation of data statements and datasheets for datasets from NLP to other written language technologies — in this case — machine translation. Bandy and Vincent (2021) tie dataset documentation to the concept of technical debt, and retrospectively produce a datasheet for a text corpus. Pushkarna et al. (2022), based on their work with text corpora at Google, then introduce the concept of *data cards*, concentrating on descriptive information that cannot be inferred from the dataset itself. Building on this work, and drawing from an extensive literature review, Papakyriakopoulos et al. (2023) propose augmented datasheets specifically for *spoken language* datasets — the only one of its kind to date.

Similarly, there has been increasing research attention toward model documentation. *Model cards* were first introduced by Mitchell et al. (2019) and built on by Shen et al. (2022), who produced a practitioner toolkit to aid in generic model card development. Crisan et al. (2022), recognising that many laypeople also use model documentation, develop an interactive approach to aid in model exploration. McMillan-Major et al. (2021) seek to join both datasheets and model cards, proposing a standard format for datasets in NLP.

However, data and model documentation in itself is not sufficient for tackling bias. An MLP creates or consumes that documentation, providing a *feedback loop* which motivates MLP *action*: re-balancing a training set, gathering more diverse data, or fine-tuning a model.

Accordingly, recent work from Microsoft Research shifts the focus of inquiry to practitioners' use of dataset and model documentation and approaches to fairness more broadly. Heger et al. (2022) find that dataset documentation practices are "largely ad-hoc and myopic in nature", with many practitioner needs unaddressed. Similarly, Holstein et al. (2019) find, in a set of interviews with MLPs in industry, that while they saw the datasets as "the most important place to intervene to improve fairness in their products", the teams did not have in place processes — such as dataset documentation — "to help support the collection and curation of balanced or representative datasets".

1.2 The research gap

People are increasingly using speech to interface with services and sources of support in the real world. ML-enabled voice technology systems continue to have pronounced biases; they work better for some people than others. If we wish to make the socio-technical systems of our world fairer, then we need to generate effective approaches for tackling bias in these systems. The approaches, motivations

and actions of MLPs around dataset documentation have been shown to assist in this regard. However, there is a lack of research here covering *spoken* language data — the kind of data used to build voice technology systems.

We therefore pose the following research provocations: How may we characterise current VDD artefacts and practices? And what work is needed to make VDDs more useful in addressing bias in voice technologies?

2 Methodology

We devise an exploratory study that combines two methods, one focusing on ML practitioners and their *experiences* creating or consuming voice dataset documentation and the other on dataset documentation *artefacts*.

Firstly, we undertake 13 semi-structured interviews with MLPs who work with voice or closely adjacent data. We explore their voice dataset document (VDD) approaches across the ML lifecycle. Secondly, we turn our attention to existing VDDs. VDDs represent how MLPs generate datasets and release them to the world — they encode practices, beliefs and assumptions (Birhane et al., 2022). We select nine VDDs for their varied purposes, collection methods and source data.

Drawing both from our literature review in Section 1.1 and from participant data, we develop a rubric for analysis, and assess the VDD artefacts across seven categories. We then triangulate the two methods, showing how VDD practices differ by MLP role, and how VDDs may help or hinder MLPs in making trade-off decisions.

2.1 Semi-structured interviews

Semi-structured interviews are established as an appropriate exploratory method for inquiring about phenomena, particularly in ML practice (Baier et al., 2019; Jöhnk et al., 2021; Følstad et al., 2018).

2.1.1 Participant selection

Potential participants were identified using professional networks, snowball sampling, and via collaborative code sites. Inclusion criteria were (i) that the participant must work with voice or closely adjacent data, and (ii) be currently practicing in industry, academia or open source fields. Purposive sampling was used to ensure representation of perspectives from diverse genders, professional disciplines, and geographic locations, and to help establish trustworthiness of findings (Campbell et al., 2020; Lincoln and Guba, 1985; Groves et al., 2011; Ezzy, 2013). A summary of participants by characteristic is shown in Appendix A.

Interviews were conducted via videoconferencing, and participants were able to make corrections and redactions to the resulting transcript. We concluded our interviews at 13 participants as themes were becoming repetitive, and we had sufficient data to inform our document analysis method.

2.1.2 Semi-structured interview design

We adopted an inductive approach, seeking to accumulate many perspectives around how VDDs are produced and consumed, whilst varying their contexts, applications and geographic sites of practice (Creswell and Creswell, 2018). Drawing from both Spradley (1979) and Minichiello et al. (1990), we structured our interview questions around "the lifecycle of creating a voice dataset" — a "grand tour" approach.

2.1.3 Coding approach

Based on our literature review, we identified several *a priori* categories (Saldaña, 2021) and used them to code the 13 interviews. We combined this with open coding — a way to capture new categories as they emerge in the data (Williams and Moser, 2019). Axial coding — a way to frame the contextual conditions of the phenomena being studied (Ezzy, 2013) — was then used to categorise how VDDs were produced and consumed. Selective coding — a way to collapse and combine several codes into core categories for analysis (Corbin and Strauss, 1990) — was then applied, yielding 14 broad categorisations across a total of 1889 codes. Here, we focus on only two of those broad categories; different MLP *roles* involved in VDDs, and how VDDs are used in the *trade-offs* MLPs make.

2.2 Document analysis

As a complementary method to our semi-structured interviews, we then undertook document analysis — "a systematic procedure for reviewing or evaluating documents" (Bowen, 2009).

2.2.1 Selection of documents

Datasets used for ML are often released with accompanying documentation in the form of a dedicated web site, code repository or online catalogue entry. Additionally, some datasets contain a metadata file *within* the dataset. We considered all of these in scope for analysis.

To identify VDDs for analysis, we performed a web search, using the terms "voice dataset" or "speech dataset". We purposively sampled nine datasets that varied by intended task; by whether the speech was elicited or spontaneous; the domain of speech; the curation rationale; funding source; license; and vocabulary size [2]. A summary is provided in Appendix B.

2.2.2 Document analysis rubric

To create the rubric used to analyse the VDD artefacts, we drew from previous work in dataset documentation (see Section 1), broader reading in metadata and research infrastructure, and participant data, arriving at 41 elements across seven categories. Here, we outline the contents of each category and justify their inclusion in the rubric.

Dataset identification Here we included persistent identifier — a uniquely identifying string, separate from the location of the dataset itself, which provides a referral to the current storage location of the dataset (Zeng and Qin, 2016) — and version as a way to distinguish dataset releases over time (Bhattacherjee et al., 2015). Efforts have been long underway to ensure datasets have persistent identifiers (Klump and Huber, 2017), and they tie closely to work on making research datasets more findable (Wilkinson et al., 2016).

Intent, purpose and curation rationale Here, we draw on the definition given by Schlangen (2021); a language task is a mapping between an input and an output, and a dataset provides examples of this mapping. Clear descriptions of intent and purpose are therefore important so the MLP can identify if the dataset is task-appropriate. We adapt "curation rationale" as given in Bender and Friedman (2018) to *spoken* language, and define it as determining which speech utterances are included in the dataset, and why.

Dataset creation process, sources and actors Here, we draw again from Bender and Friedman (2018), who place emphasis on understanding the social standpoint of annotators. For many speech tasks, written transcriptions are also required as inputs. Noting the work of Bucholtz (2007, 2000) — that transcription has both variation and politics in its production — we also identified whether the transcription method was provided. Referencing Barbiers et al. (2007) work on spoken language variation from corpus linguistics, we also included the source of elicited speech prompts as an element.

Characteristics of the dataset itself Here, we drew on from material on research data infrastructure. Working with "big data" presents many challenges to MLPs (Kitchin, 2014); and so it is beneficial to provide an overview of the size, shape and constituency of the dataset.

Constitution of the dataset by speaker, recording environment and spoken language attributes In our exploratory interviews (see 3), comprehending contents was a key consideration for many participants. Speech recognition requires a wide variety of voice samples, while speech synthesis needs many samples from a single speaker. It is therefore important that characteristics of the speech utterances captured in the data are clearly represented:

"...Sometimes you really need to dig deeply into the corpus to find it. Sometimes you just don't find it. And sometimes this is well documented. ... This is important ... because we need to have a balanced corpus for training your system. And then also to be able to evaluate, gender wise, the performance of your system." — *SB*

We drew both from the literature and from exploratory interviews to identify specific attributes to assess. Bender (2019) makes the case for clearly identifying the languages we work with in, and Bender and Friedman (2018) advocate both for representing the languages in a dataset in `BCP-47` format *and* providing a "prose description" of the language's "axes of variation".

Participant *TS* highlighted additional areas of spoken language variance to scrutinise when evaluating trained models: "... We have a lot of folks who have code-switched data ... it's also domain variation or register variation, or all your training data is super formal ..." —*TS*.

Code-switching is where the speaker alternates between two or more "codes" — usually languages — within a conversation (Auer, 2013). The domain of spoken language is usually taken to be the subject matter of the conversation, while register is how spoken language varies by social situation; we speak differently in formal and informal settings (Finegan, 2014).

Models, benchmarks and academic papers We adapt this category from Gebru et al. (2021), who

[2] We note here that the AusTalk dataset in the ALVEO repository is currently offline; had it been available we would have also included it due to its focus on Australian speech.

recommend documenting where a dataset should *not* be used, as also echoed by an interview participant: "When you think about kind of building a dataset, it's easy to think about, 'Okay, I'm building a dataset, it's going to be used for this. This is what I want it to be used for.' Unfortunately, people are going to use it for things you didn't intend." —*CG*.

Similarly, noting increasing calls for benchmarks to be tightly linked to the intended task of a dataset (Raji et al., 2021), we included these as an attribute in the analysis.

Privacy, bias, limitations and social impact Here, we drew from Bender and Friedman (2018); Gebru et al. (2021); Papakyriakopoulos et al. (2023), who all underscore the importance of documenting privacy and sensitivity considerations of a dataset, and their potential social consequences, and we use this category to assess whether biases and limits of the dataset are considered in VDDs.

2.2.3 Performing the analysis

To perform the analysis, we reviewed each dataset's documents against the criteria in each category of the rubric. If fulfilment of a criterion was implied but not explicit in the document(s), then we made a finding of "Implied" and provided a rationale. If a criterion was not applicable to a dataset, we made a finding of "N/A" — for example, in speech synthesis datasets like "LJSpeech", speech samples are usually taken from only one speaker and so the number of unique speakers in the dataset is not applicable. Our analysis is summarised in Appendix C.

3 Findings

Here, we triangulate our two methods. We characterise the experience of MLPs with VDDs through the frames of MLP roles and trade-offs the MLP makes, quoting from interview transcripts to highlight key points. At the same time, we corroborate the interview findings by referencing results from the document analysis. This layered approach provides a richer characterisation of VDDs.

3.1 Characterising practices by role

Our interview data showed that MLPs could be categorised into four distinct roles, depending on how they discovered, commissioned, produced or consumed voice datasets. We use a "food" analogy to label the roles — which seems odd at first glance — but which we believe accurately characterises a role's relationship with voice datasets. The results of the document analysis had different implications for each role, which we unpack below.

Chefs We characterise as *Chefs* those MLPs who are provided with a dataset specification against which to create a voice dataset: "... we would have a data collection spec, [with a] percentage of different accents or gender or whatever." —*BP*. *Chefs* are mostly likely to be *producers* of VDDs.

Diners *Diners* form a complement to *Chefs*, being the MLPs who are in a position to order voice datasets from commercial companies. These companies offer both bespoke options — à la carte — as well as subscriptions to regular dataset updates — a grocery box. There are many such providers: "So there are many companies that offer services in terms of annotating data, transcribing data. There are many companies that collect some data and sell data." —*SS*.

Scavengers Alternatively, an MLP may be a *Scavenger* — where they must discover freely available voice datasets to meet their needs due to cost constraints. "... us open source folks we're scavengers, right? ... The ordering options are there ... and I've looked at them and they want tens of thousands of dollars, for access. And I'm like, '"I don't have that."'" —*PS*.

Importantly, it was this remark that helped us arrive at our role categorisation.

Hoarders *Hoarders*, in contrast to Scavengers, Chefs and Diners, do not have a clear intent in mind for the voice data they accumulate; they store it for some future, unspecified purpose in the hope that it will be of use. Voice data accumulated this way is usually a byproduct of business operations: "We know that often companies, they have a plan to extract and collect as much data as possible before they even know what it's potentially useful for."—*PP*.

3.2 The focus of VDD practices differs by role

Discovery For the *Scavenger*, dataset documentation is important to their discovery efforts — and their ability to comprehend the contents of a dataset when found. Based on our document analysis, their needs are currently poorly served. While eight of the nine datasets represented speaker gender, only two represented accent, and only one represented

speaker nationality or age. Speaker occupation, language heritage or education attainment were absent, save for an overview of speaker occupation in the *African languages* VDD. There was very little information provided on the recording environments used, and the only representation of variance of spoken language tended to be the way in which the dataset language(s) were specified — with five of the nine VDDs representing language using a `BCP-47` or `ISO-639` code.

Representation *Chefs* may produce documentation as part of their creation efforts, and in doing so, must make choices about how to represent that data. With both an absence of agreed or *de facto* standards for documenting voice and speech data[3], as well as multiple standards for language representation (Wright, 2019), some participants faced challenges in determining *how* some data items should be reported: "There is no unified format. Everybody has their own `JSON` [4] that might have similar information."—*BP*.

Another *Chef* practitioner faced similar data representation dilemmas in regard to dialect, grappling with what level of granularity to represent in the VDD: "...what if we label what dialect they are speaking in? Or what if they self label what dialect they think they are speaking in? Then we do things like how about we review this? Meaning let's write whether we think this is pronounced correctly. It's either yes or no. Okay. Wait, what if we can label every single character in the sentence and say whether the character was pronounced correctly?" —*EG*

VDDs are still relevant for the *Hoarder* role, even though they may not yet know what tasks their datasets will be used to perform. Hoarders still wrestle with how to represent the data they are collecting. Here, there was a desire to create VDDs that allowed the broadest scope of future use for the data:

"...it's always good to document, to label your data to the maximum extent that you can in terms of fidelity." —*RW*.

The desire to chronicle datasets with high fidelity places additional onus on the MLP to define *how* the data is represented. We see here a tendency to reproduce that which has come before: "...we didn't put a lot of thought into the choosing of the structure of the [Dataset] dataset, because we just used it as it was. And the reason that we chose the [Dataset] as an example dataset was because it was a fairly common, well-known speech recognition dataset"—*RW*.

This effect serves as a reinforcing loop, anchoring practice to the status quo.

Diversity of data Both *Scavengers* and *Diners* need to know whether the data within a voice dataset is useful for their intended purpose: "...the dataset documentation would give me an idea, does this dataset work for my application? ... Is this dataset going to be useful?" —*CG*.

Drawing from our document analysis, it appears *Scavengers* and *Diners* are well served by current VDDs — all nine datasets examined provided an executive summary or description, and eight of the nine provided both intended tasks or use cases, as well as a curation rationale.

However, even if a dataset appears to meet an MLP's need based on the contents of the VDD, variation in how the dataset is transcribed can be problematic, requiring that the MLP spend time "listening to the data": "...All transcription is subjective. And so each of these databases will have been transcribed by different people, maybe following different conventions, and those conventions are especially important with semi words, ums and uhs and mm-mms, and stuff like that." —*BP*.

Cross-referencing our participant's statement with our document analysis, we note that only one of the three datasets that had transcribed spontaneous speech provided a description of the transcription process.

Another salient example here deals with the lack of variation of accents in the dataset not being apparent from the VDD, a realisation the practitioner makes only *after* listening to the data, and having to cross-check with the dataset's related academic paper:

"...I had worked with it for a while, I thought I knew the data. It was a very popular dataset. And it wasn't until I started listening to it, that I realized that these are only North American voices. It wasn't obvious to me until then. And then I went back and I read the paper, the actual paper ... and it was explicit like, yes, they chose voices that were North American. And it's something simple as that, you don't know until you start listening to the data." —*CD*

[3] We note here the work of Papakyriakopoulos et al. (2023), however this was not available at the time the interviews were conducted.

[4] `JSON` is a data structure format commonly used for voice data

Again, cross-referencing, only two of the nine datasets provided a representation of the speaker's accent.

3.3 Characterising practices through trade-offs the practitioner must make

Changing our analytical lens, we now explore practices by exploring the trade-offs a practitioner must make. Drawing from the field of social learning theory, Wenger-Trayner and Wenger-Trayner (2014) hold that practitioners operate across multiple disciplinary communities in a "landscape of practice". An MLP may need to span disciplines such as data engineering, machine learning, metadata specification and linguistics; each with their own accepted practices (e.g. Deng et al. (2022); Balayn et al. (2021)). These practices may be in tension, requiring the practitioner to make trade-offs. While our interview data uncovered many trade-offs, we focus here on the most frequently recurring.

3.3.1 Big data vs storage

"The problem is, data gets big. And then you have a problem, right?" —*AG*

Speech technologies may require thousands of hours of data, in turn requiring large volumes of disk storage capacity. For example, one dataset we analysed, Mozilla Common Voice, is nearly 80GB in size. This scale causes practical problems for MLPs, such as one *Chef* who created voice datasets, and needed to store them on a server. His frustration at having to frequently move datasets was palpable: "Yeah. We would find somewhere on [University web server] we'd be, 'Oh yeah. No, we'll serve it off our little file server here and it'll be no worries.' And we'd put it up there and we'd create a website for it. And we'd point people at the website. And then the IT guys would go, 'Oh yeah, no. We don't want to do that [...] We're going to shut that down. You're going to have to find somewhere else to put that.'" —*RW*.

One mechanism that exists to overcome this limitation is the use of a *persistent identifier*. In our document analysis, only three of the nine datasets were found to have persistent identifiers applied (see C *Dataset identification*), and these were verified using Crossref [5]. More positively, all datasets bar one indicated storage requirements, and all provided the number of hours of overall speech in the dataset (see C *Characteristics of the dataset itself*).

[5] https://search.crossref.org/

3.3.2 Big data vs understanding data contents

We also identified trade-offs that the MLP had to make in comprehending the contents of a voice dataset. Earlier, in 3.2, we showed that an MLP compensated for lack of variation description in VDD by "listening to the data". The size of voice datasets makes this practice more onerous, as highlighted by one interview participant: "We ended up with 12,000 recordings, which was humanly transcribed and those 12,000 recordings equated to 20 hours of speech. So we literally had a team of people listening to recordings and typing the recordings out verbatim." —*SS*.

This again points to the need for more focus on capturing data related to recording environment in particular: "And with a hundred thousand hours of data, how are you going to listen to all that as one person especially? You can't. You can randomly sample and hope for the best that you catch something. But if you precisely knew exactly the conditions of the recordings and all that stuff, if you could control all that then I think you could do a much better job." —*PS*.

Triangulating this with our document analysis (see C *How the dataset represents the recording environment*), we find that only the CHIME-5 dataset provided explicit information on the recording environment. This is likely due to its relevance in the dataset's purpose of speech separation. Other datasets implied some recording information — such as the HUB5 dataset being of recorded telephone conversations.

Again, we find that VDDs are inadequate for MLPs' needs.

4 Righting the docs: Towards VDD that help MLPs mitigate bias in speech technologies

Drawing from the gaps in VDD practice uncovered from our exploratory study above, we now propose a program of work to begin to address them.

4.1 A unified description format for spoken language datasets

The VDDs we analysed contained a patchwork of information in varying formats. This presents hurdles for dataset consumers, such as *Scavengers* and *Hoarders*, in understanding dataset contents, as corroborated in 3.2. This is a necessary step before datasets can be effectively combined for training ML models. A unified datasheet format for spo-

ken language datasets is likely to go some way to addressing this weakness. Here, we welcome the work of Papakyriakopoulos et al. (2023) in formulating *Augmented datasheets for speech data*. This work provides both a minimal description structure, and tools to enable the dataset producer to create it. However, this alone is insufficient to address the challenges we uncovered.

4.2 Automating the creation of descriptive information for voice data

Augmented datasheets for speech data assumes that dataset producers act reflexively *before or during* the dataset creation process. Indeed, reflexivity has been shown to improve dataset practice (Boyd, 2021). We found some evidence of reflexivity in our interviews, with *Chefs* considering how to represent data items (see 3.2). However, given the lack of descriptive information found in many of the nine datasets analysed, it is reasonable to claim that much VDD work happens *after the fact*, if at all.

Here, classification models, such as for gender, age and accent, are needed to help provide better descriptive information for speech datasets, reducing the need for the MLP to "listen to the data" (see 3.2). This would be particularly helpful for datasets where granular VDD was not captured at the point of creation, providing the ability to create parts of VDD retrospectively — although we acknowledge that inferred VDD are likely to represent dataset contents less accurately.

There is some emerging work in this space, such as Sánchez-Hevia et al. (2022), who use a range of neural models to accurately predict gender and age on the Common Voice dataset, and Najafian and Russell (2020), who use automatic accent identification to make a model more robust to accented speech. We note, however, that such classification can be used for ethically dubious purposes, such as pre-emptive policing (e.g. such as that recently done in the Türkçe language (Korkmaz and Boyacı, 2022)). We also note that Gebru et al. (2021) caution *against* automating the creation of dataset documentation, championing instead the use of reflexive processes. We hold that there is a practical middle ground here; to be reflexive during dataset creation, but to have tools available when VDDs of existing datasets are insufficient.

4.3 Common representation taxonomies for voice data

In section 3.2, several participants highlighted the lack of consistency in formats used for representing variance in speaker characteristics, context of speech and the spoken language itself. Here, common taxonomies would assist MLPs in combining datasets in ways that aid in addressing bias. For example, MLPs may wish to compile spoken language data of a particular accent to assess if a neural model performs well on that accent. However, if different datasets represent accents in different ways, combining datasets becomes much harder. Indeed, the need to capture speaker demographics in particular more systematically was highlighted in our interviews:

"I would say that each time a new speaker is registered to the system, is going to start making a recording, we should have a nice interface, an easy to use interface, to quickly fill all the information that we need." —*SB*.

Although there is some recent work in the accent space, such as calls to extend the `BCP-47` format to better represent low-resource languages (Gillis-Webber and Tittel, 2019, 2020), and work to represent gender bias more accurately in text corpora (Havens et al., 2022), we still lack accepted taxonomies for representing the linguistic heritage of a speaker (language acquisition, L1 and L2 status etc), domains of speech (such as medical, quick service restaurant ordering, industrial automation) and the recording environment (such as cafe, quiet office, family home, studio). Having such reusable and inter-operable taxonomies would also align with efforts to make research data, and speech archives specifically, more "FAIR" (Wilkinson et al., 2016; Calamai and Frontini, 2018).

4.4 Incentivising adoption of unified formats

Even if unified description formats and common taxonomies for VDDs are available, a mechanism is needed to incentivise their *adoption*, particularly given the practice identified in our interview data of replicating existing dataset formats (see 3.2). Bender and Friedman (2018) outline several incentives which would be useful here, such as requiring adherence to dataset documentation formats for publication in key journals.

With increasing usage of collaborative coding platforms in ML practice (Berman, 2023), another available incentive is to require complete VDDs

before datasets are uploaded. For example, while Hugging Face displays dataset datasheets on the platform, there is no requirement for them to be completed, and they are often blank [6].

5 Limitations

Additional methods to triangulate findings We recognise the small, although purposive, sample of participants and datasets in our exploratory study. We now intend to administer a questionnaire to a broader group of MLPs, to validate or invalidate these initial findings.

Only publicly knowable datasets were analysed In identifying and selecting datasets for analysis, we recognise that our approach was limited to only publicly knowable datasets; private and/or proprietary datasets used internally by organisations may exhibit very different dataset documentation practices, although this is unlikely based on the work of Heger et al. (2022) and Holstein et al. (2019).

6 Conclusion

Here, we have situated voice dataset documentation (VDD) practices conducted by machine learning practitioners (MLPs) within broader efforts to reduce bias in ML-enabled speech technologies as they go to scale. We first provided a brief literature review of ML-related dataset documentation work, identifying that VDD practices are understudied. We presented an exploratory study that combined two methods — semi-structured interviews and document analysis — to provide a rich characterisation of practices surrounding VDDs.

We find that VDDs are currently inadequate to meet the needs of MLPs who create and consume voice datasets. In particular, they often fail to describe voice dataset contents accurately, if at all, and the range of representation formats used makes it difficult for MLPs to combine datasets effectively — as is often required in bias reduction efforts.

Drawing from these findings, we propose actions that seek to "right the docs", focusing on unified formats for dataset documentation, as well as the need for common taxonomies for data items common to voice datasets.

[6] For example, the datasheet for Common Voice on Hugging Face omits large sections, such as curation rationale and limitations

7 Ethics statement

The interviews conducted as part of this research have received human ethics approval from Australian National University's Human Research Ethics Committee, protocol number `417/2021`. This protocol requires using pseudonyms to refer to research participants, which has been done here.

Acknowledgements

Thanks are extended to Glen Berman for feedback on earlier versions of this paper. Kathy Reid's PhD research is funded by an Australian Research Training Scholarship and via the Florence Violet MacKenzie Scholarship. Kathy Reid holds a Research Partnership with Mozilla Foundation however this Partnership does not provide research funding. Kathy is grateful to the interview participants who gave generously of their time, knowledge and experience, and to the School of Cybernetics PhD cohort for their support. We thank the reviewers for their feedback.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, Marseille, France. European Language Resources Association (ELRA).

Peter Auer. 2013. *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Lucas Baier, Fabian Jöhren, and Stefan Seebacher. 2019. Challenges in the deployment and operation of machine learning in practice. In *ECIS*, volume 1.

Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. 2021. Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768.

John Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in machine learning: A retrospective datasheet for BookCorpus. In *Proceedings*

of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1. Curran.

Sjef Barbiers, Leonie Cornips, and Jan Pieter Kunst. 2007. The Syntactic Atlas of the Dutch Dialects (SAND): A corpus of elicited speech and text as an online dynamic atlas. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 54–90.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'CHiME' Speech separation and recognition challenge: Dataset, task and baselines. In *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Emily M Bender. 2019. The# benderrule: On naming the languages we study and why it matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Glen Berman. 2023. Machine Learning practices and infrastructures. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 466–481.

Souvik Bhattacherjee, Amit Chavan, Silu Huang, Amol Deshpande, and Aditya Parameswaran. 2015. Principles of dataset versioning: Exploring the recreation/storage tradeoff. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 8, page 1346. NIH Public Access.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Hasna Bouraoui, Chadlia Jerad, Anupam Chattopadhyay, and Nejib Ben Hadj-Alouane. 2017. Hardware architectures for embedded speaker recognition applications: A survey. *ACM Transactions on Embedded Computing Systems (TECS)*, 16(3):1–28.

Glenn A Bowen. 2009. Document analysis as a qualitative research method. *Qualitative research journal*, 9(2):27–40.

Geoffrey C Bowker, Karen Baker, Florence Millerand, and David Ribes. 2009. Toward information infrastructure studies: Ways of knowing in a networked environment. In *International Handbook of Internet Research*, pages 97–117. Springer.

Karen L Boyd. 2021. Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27.

Anthony J. Bradley. 2020. Brace Yourself for an Explosion of Virtual Assistants. https://blogs.gartner.com/anthony_bradley/2020/08/10/brace-yourself-for-an-explosion-of-virtual-assistants/.

Robin Brewer, Casey Pierce, Pooja Upadhyay, and Leeseul Park. 2022. An empirical study of older adult's voice assistant use for health information seeking. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(2):1–32.

Mary Bucholtz. 2000. The politics of transcription. *Journal of Pragmatics*, 32(10):1439–1465.

Mary Bucholtz. 2007. Variation in transcription. *Discourse studies*, 9(6):784–808.

Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, 47(4):339–354.

Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. 2020. Purposive sampling: Complex or simple? Research case examples. *Journal of research in Nursing*, 25(8):652–661.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, pages 1086–1090.

Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21.

Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. MT-adapted datasheets for datasets: Template and repository. *arXiv preprint arXiv:2005.13156*.

John W Creswell and J David Creswell. 2018. *Research Design*, 5th edition. Sage publications Thousand Oaks, CA.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439.

Robert Dale. 2020. Voice assistance in 2019. *Natural Language Engineering*, 26(1):129–136.

Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 473–484.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351.

Douglas Ezzy. 2013. Coding data and interpreting text: Methods of analysis. In *Qualitative Analysis*, pages 80–110. Routledge.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Edward Finegan. 2014. *Language: Its Structure and Use*. Cengage Learning.

Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. In *International Conference on Internet Science*, pages 194–208. Springer.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2020. Gender Representation in Open Source Speech Resources. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6599–6605, Marseille, France. European Language Resources Association (ELRA).

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-Saharan African languages for automatic speech recognition: A case study of wolof. *LREC*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2007. Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10-11):847–860.

Frances Gillis-Webber and Sabine Tittel. 2019. The shortcomings of language tags for linked data when modeling lesser-known languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Frances Gillis-Webber and Sabine Tittel. 2020. A framework for shared agreement of language tags beyond ISO 639. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3333–3339.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey Methodology*, volume 561. John Wiley & Sons.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of British English text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, Seattle, Washington. Association for Computational Linguistics.

Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in automated speaker recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 230–247.

Keith Ito. 2017. LJ Speech.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. 2018. Jakobovski/free-spoken-digit-dataset: V1.0.8. Zenodo.

Manuel Jesús-Azabal, Javier Rojo, Enrique Moguel, Daniel Flores-Martin, Javier Berrocal, José García-Alonso, and Juan M Murillo. 2019. Voice assistant to remind pharmacologic treatment in elders. In *International Workshop on Gerontechnology*, pages 123–133. Springer.

Jan Jöhnk, Malte Weißert, and Katrin Wyrtki. 2021. Ready or not, AI comes—an interview study of organizational AI readiness factors. *Business & Information Systems Engineering*, 63(1):5–20.

Dewi Jones. 2020. Macsen: A voice assistant for speakers of a lesser resourced language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for under-Resourced Languages (CCURL)*, pages 194–201.

Linus Kendall, Bidisha Chaudhuri, and Apoorva Bhalla. 2020. Understanding technology as situated practice: Everyday use of voice user interfaces among diverse groups of users in urban India. *Information Systems Frontiers*, 22:585–605.

Bret Kinsella and Ava Mutchler. 2020. Smart Speaker Consumer Adoption Report 2020. Technical report, Voicebot.AI.

Rob Kitchin. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Jens Klump and Robert Huber. 2017. 20 Years of persistent identifiers–Which systems are here to stay? *Data Science Journal*, 16:9–9.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Yunus Korkmaz and Aytuğ Boyacı. 2022. A comprehensive Turkish accent/dialect recognition system using acoustic perceptual formants. *Applied Acoustics*, 193:108761.

Librivox. 2021. Librivox - Acoustical liberation of books in the public domain.

Yvonna S Lincoln and Egon G Guba. 1985. *Naturalistic Inquiry*. Sage Publications, Inc, Beverly Hills, California, United States of America.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166. IEEE.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.*

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135.

Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 161–172, New York, NY, USA. Association for Computing Machinery.

Victor Minichiello, Rosalie Aroni, Eric Timewell, and Loris Alexander. 1990. *In-Depth Interviewing: Researching People*. Longman Cheshire, Melbourne, Australia.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Maryam Najafian and Martin Russell. 2020. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 122:44–55.

Mikel K Ngueajio and Gloria Washington. 2022. Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. In *HCI International 2022–Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*, pages 421–440. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 881–904.

Branislav Popović, Edvin Pakoci, Nikša Jakovljević, Goran Kočiš, and Darko Pekar. 2015. Voice assistant application for the Serbian language. In *2015 23rd Telecommunications Forum Telfor (TELFOR)*, pages 858–861. IEEE.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible AI. In

Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, pages 1776–1826, New York, NY, USA. Association for Computing Machinery.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: An Automatic Speech Recognition dedicated corpus. In *LREC*, pages 125–129.

Johnny Saldaña. 2021. *The Coding Manual for Qualitative Researchers*. Sage.

Héctor A Sánchez-Hevia, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera. 2022. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81(3):3535–3552.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674.

Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 440–451.

James P. Spradley. 1979. *The Ethnographic Interview*. Harcourt Brace Jovanovich College Publishers, Fort Worth, Texas, United States.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and YouTube automatic captions. In *INTERSPEECH*, pages 934–938.

Joseph Turow. 2021. *The Voice Catchers*. Yale University Press.

Rob Van der Meulen and Amy Ann Forni. 2016. Gartner Says Worldwide Spending on VPA-Enabled Wireless Speakers Will Top $2 Billion by 2020. *Gartner*.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2010. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–10.

Etienne Wenger-Trayner and Beverly Wenger-Trayner. 2014. Learning in a landscape of practice: A framework. In *Learning in Landscapes of Practice*, pages 13–29. Routledge.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Michael Williams and Tami Moser. 2019. The art of coding and thematic exploration in qualitative research. *International Management Review*, 15(1):45–55.

Sue Ellen Wright. 2019. Standards for the language, translation and localization industry. In *The Routledge Handbook of Translation and Technology*, pages 21–44. Routledge London and New York.

Marcia Lei Zeng and Jian Qin. 2016. *Metadata*. Facet.

A Interview participant summary

Table 1: Interview participant summary (n=13)

Characteristic		Total
Gender	Female	3
	Male	10
	Other gender expressions	0
Occupational field	Research scientist or academic	5
	ML or NLP Engineer	2
	Software Engineer	2
	Data annotator	1
	Developer Relations Advocate	2
	UX Designer / researcher	1
Country of residence	United States	5
	Australia	3
	South Africa	1
	Aotearoa New Zealand	1
	Nigeria	1
	France	1
	Canada	1

B Summary of voice dataset documents analysed

Table 2: Summary of voice dataset documents analysed

Characteristic of the dataset or voice dataset document (VDD)	Mozilla Common Voice(Ardila et al., 2020)	Librispeech (Panayotov et al., 2015)	African languages in the field (Gauthier et al., 2016)	Voxceleb (Chung et al., 2018; Nagrani et al., 2020)	LDC 2000 HUB5 English Evaluation Speech	TED-LIUM corpus (Rousseau et al., 2012; Hernandez et al., 2018)	Free spoken digit dataset (Jackson et al., 2018)	CHIME 5 Speech separation challenge dataset (Barker et al., 2018)	LJSpeech Speech dataset (Ito, 2017)
Type of document(s) analysed	CommonVoice website, GitHub repository, related paper	Entry on OpenSLR website, related paper	Entry on OpenSLR website, `README` file in dataset, related paper	VoxCeleb website, Metadata file archived on archive.org, related papers	LDC Catalogue entry	TED-LIUM website, `README` file in dataset, related paper	GitHub repository, Zenodo dataset record, `metadata.py` file in dataset	Data page on CHIME website, `JSON` file in dataset	LJ Speech website
Year of initial release & latest version	2018; 2023 (version 13)	2015; no newer version	2005; no newer version	2017; 2018 (version 2)	2005; no newer version	2012; 2018 (version 3)	2018; no newer version	2018	2017; 2017 (version 1.1)
Intended language task	Speech recognition	Speech recognition, multilingual	Speech recognition, monolingual	Speaker identification, speech separation, monolingual	Speech recognition, monolingual	Speech recognition, monolingual	Speech recognition, monolingual	Speech separation, monolingual	Speech synthesis, monolingual
Nature of speech in dataset	Elicited, large vocabulary, multiple domains	Elicited, large vocabulary, out of copyright works	Elicited, large vocabulary, multiple domains	Spontaneous, large vocabulary, multiple domains	Spontaneous, large vocabulary, multiple domains	Spontaneous, large vocabulary, multiple domains	Elicited, constrained vocabulary, spoken digits	Spontaneous, large vocabulary, multiple domains	Elicited, large vocabulary, non-fiction books publishes between 1884 and 1964
Motivation and funding source	Ecosystem development; Grant-based for particular languages; additional funding from NVIDIA	Research; funding unknown.	Research; ALFFA Research Project, funded by agence nationale de la recherche.	Research by Oxford University, funded through EPSRC programme grant	Commercial; Sponsored by National Institute of Standards and Technology.	Research, funding not specified.	Research, funding not specified.	Research challenge sponsored by Google and Microsoft Research.	Research, funding not specified, independent researcher.
Method of collection of dataset	Volunteer speakers recorded on web-based platform.	Secondary use dataset from Librivox volunteer audio book project (Librivox, 2021)	Original dataset, volunteer speakers recorded in field.	Secondary use dataset from YouTube; speakers' consent not provided.	Original dataset, recruited speakers recorded via telephone.	Secondary use dataset from TED videos; speaker consent unknown.	Original dataset, speaker recruitment and recording unknown.	Original dataset, speaker recruitment unknown, recorded in speakers' homes.	Tertiary dataset, subset of Librispeech containing single speaker. Speaker consent not stated.

C Summary of dataset documentation analysis

Table 3: Descriptions and data items included in current voice dataset documentation

Data item	Mozilla Common Voice	Librispeech	African languages in the field	Voxceleb	LDC 2000 HUB5 English Evaluation Speech	TED-LIUM corpus	Free spoken digit dataset	CHIME 5 Speech separation challenge dataset	LJSpeech Speech dataset
Dataset identification									
Persistent identifier for the dataset	No	No	No	No	Yes	No	Yes	Yes	No
Dataset versioning	Yes	Yes	No	Yes	Yes	Yes	Yes	Implied via yearly competition	Yes
Dataset release date	Yes	Implied through related paper	Yes	Implied through related paper	Yes	Implied through related paper	Yes	Yes	Yes
Intent, purpose and curation rationale									
Executive summary or description	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intended tasks or use cases	Yes	Yes	Yes	Yes	Yes	Yes	Implied through GitHub repository tags	Yes	No
Curation rationale	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Dataset creation process, sources and actors									
Dataset collection method	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
For elicited speech, the source of prompts	Implied through GitHub history	Yes	Yes	Yes	No	Yes	No	Yes	Yes
For spontaneous speech, description of the annotation/transcription process	N/A	N/A	N/A	N/A	No	Yes	N/A	No	N/A
For spontaneous speech, description of the annotators	N/A	N/A	N/A	No	N/A	No	N/A	N/A	N/A
Characteristics of the dataset itself									
Structure of dataset, such as field mapping, described	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dataset storage size provided	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Overall hours of speech in dataset specified	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
License specified	CC0	CC-BY-4.0	MIT	CC-BY-SA-4.0	LDC User Agreement	CC-BY-NC-ND-3.0	CC-BY-SA-4.0	Dataset specific	Public domain
# of distinct voices in dataset specified	Yes	Yes	Yes	Yes	Implied via # of conversations	Yes, in paper	Yes	Yes	Yes
# of utterances in dataset specified	Yes	Yes	Yes	Yes	No, only # of conversations given	Yes, in paper	Yes	Yes	Yes
Length of utterances given	Yes	No	Yes, averaged	Implied via each utterance having same length	No	No	Implied via each utterance being a single digit	Inferred via JSON file	Yes, averaged
Split information (test, train, dev etc) provided	Yes	Yes	Yes, in data structure	No	No	Yes, in data structure	Yes	Yes	N/A, splits not used in speech synthesis
Audio file type specified	Yes, in data structure	Yes, in data structure	Yes, in data structure	No	File type implied by sample file	Yes	Yes	Yes	Yes
Audio file format details (resolution etc) provided	No	Yes (some)	No	No	Yes	Yes	Yes	Yes	Yes

Table 4: Descriptions and data items included in current voice dataset documentation (continued)

Data item	Mozilla Common Voice	Librispeech	African languages in the field	Voxceleb	LDC 2000 HUB5 English Evaluation Speech	TED-LIUM corpus	Free spoken digit dataset	CHIME 5 Speech separation challenge dataset	LJSpeech Speech dataset
How the dataset represents characteristics of the speaker(s)									
Representation or distribution of speaker accent	Yes	No	No	No	No	No	Yes	No	No
Representation or distribution of speaker nationality	No	No	No	No	No	No	Yes	No	No
Representation or distribution of speaker age	Yes	No	No	No	No	No	No	No	No
Representation or distribution of speaker gender	Yes	Yes	Not in dataset, but distribution specified in paper	Yes	No	Yes, in paper	Yes	Yes	Implied, single speaker, gender specified
Representation of speaker occupation	No	No	Not in dataset, but overview given in paper	No	No	No	No	No	No
Representation of speaker language acquisition or heritage	No	No	No	No	No	No	No	No	No
Representation of speaker educational attainment	No	No	No	No	No	No	No	No	No
How the dataset represents the recording environment									
Constitution by recording hardware	No	No	No	No	Implied (telephone	No	No	No	No
Constitution by recording environment	No	No	No	Implied (interview)	Implied (phone conversations)	Implied (TED talks)	No	Yes	No
How the dataset represents characteristics of spoken language									
Dataset language(s) represented using a standard such as BCP-47 or ISO-639	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Multilingual flag	Implied through dataset structure	Yes	Implied through dataset structure	No	No	No	No	No	No
Representation of phonetic distribution or variation	No	No	No	No	No	No	No	No	No
Representation of dialect, lexical or non-phonetic variation	No	No	No	No	No	No	No	No	No
Representation of domain of speech	No	Implied due to Librivox source	No	No	No	No	Implied - digits	No	Implied due to Librivox source
Constitution by formality or register of spoken language	Varies with prompt	Varies with prompt	No	No	No	Implied - TED talks	No	No	Varies with prompt
For spontaneous speech, whether code-switching is indicated	N/A	N/A	N/A	No	No	No	N/A	No	N/A
Models, benchmarks and academic papers									
Benchmarks specified or linked to	No, uses CER for eval'n but no benchmark	Yes, WSJ	No	Yes, previous speaker recog'n datasets	No	No	No, uses WER and CER for eval'n but no benchmark	No	No
Models trained from dataset specified or linked to	Yes, specified in paper	Yes, specified in paper	Yes, specified in paper	Yes, specified in paper	No	No	No	Yes, in results page	No
Papers based on dataset specified or linked to	Yes, on website	Yes, on website	Yes, on website	Yes, on website	No	Yes, on website	Yes	Yes, in results page	No
Privacy, limitations and social impact									
Privacy or sensitivity statement of the dataset	Some info on website	No	No	Has a privacy statement	No	No	No	No	No
Social impact statement of the dataset	Some info on website	No	No	No	No	No	No	No	No
Statement of biases in dataset	No	No	No	No	No	No	No	No	No
Statement of limitations of dataset	No	No	No	No	No	No	No	No	No