EasyChair Preprint

№ 9678

Common Voice and Accent Choice: Data Contributors Self-Describe Their Spoken Accents in Diverse Ways

Kathy Reid and Elizabeth T. Williams

February 7, 2023

Common Voice and accent choice: data contributors self-describe their spoken accents in diverse ways

KATHY REID, School of Cybernetics, Australian National University, Australia
ELIZABETH T. WILLIAMS, School of Cybernetics, Australian National University, Australia

Fig. 1. Group of ethnically diverse but predominantly young and visually able-bodied people using smart phones, 2018. RawPixel Ltd via Flickr CC BY 2.0 - https://flic.kr/p/2cLeKBZ

The number of people using speech technologies, such as automatic speech recognition (ASR), powered by machine learning (ML), has increased exponentially in recent years [30, 41, 50]. Datasets used as inputs for training speech models often represent demographic features of the speaker – such as their gender, age, and accent. Often, those demographic axes are used to evaluate the training set and resultant model for bias and fairness [38]. Here, we first examine voice datasets to identify how accents are currently represented. We then analyse the speaker-described accent entries in Mozilla's Common Voice v11 dataset using a force-directed graph data visualisation. From this we formulate an emergent taxonomy of accent descriptors, of pragmatic use in accent bias detection. We find that accents are currently represented in ways that are geographically, and predominantly, nationally bound. More diverse representations are identified in the CV dataset. This work provides some early evidence for re-thinking how accents are represented in voice data, particularly where intended for use in building or evaluating ML-based speech technologies. Our tooling is open-sourced to aid in replication and impact.

Additional Key Words and Phrases: datasets, dataset documentation, voice data, speech data, accent recognition, accent data, metadata, bias, bias corpora, algorithmic audit, fairness, data visualization, force-directed graph

Manuscript submitted to ACM

ACM Reference Format:
Kathy Reid and Elizabeth T. Williams. 2023. Common Voice and accent choice: data contributors self-describe their spoken accents in diverse ways. In . ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND MOTIVATION

Accent may belie a speaker's geographical heritage [9, 33], socio-economic status [47] and educational attainment [18], and comprises features such as phonetics, intonation, emphasis and prosody. Dialect, by way of distinction, refers to the vocabulary and grammar that a speaker users [15, 39, 45].

We define *accent bias* as the systemic, real-world difference in treatment people experience due to the accent they speak with. Accent bias may occur without any technological intervention - for example if a listener perceives the speaker's accent and acts upon it in a manner distinct to that for other speakers [51]. However, accent bias is increasingly automated and scaled through speech technologies employing machine learning (ML). The harms of accent bias range from the inconvenient (such as mis-transcription [31]), to the egregious. To the latter, we cite examples such as modifying a person's accent in an act of cultural erasure [48], reduced professional opportunities [27], or the biasing of juries in a miscarriage of justice [11].

The real-world consequences of ML-enabled accent bias tend to stem from statistical or data bias in voice datasets, which are inherited by the resultant ML models. Accent descriptors may be included in voice datasets used for ML so that the trained model better reflects the characteristics of the real world setting into which it is deployed. For example, a model intended for deployment in Australia may be trained on speech with an Australian English accent. Accent groupings may also be used to detect bias in voice datasets, or in the models trained upon them [26, 38, 40, 55].

In our experience, accent bias in speech technology is not usually deliberate. However, even if a practitioner intentionally seeks to evaluate and fine-tune a model to prevent accent bias, they must use data containing labelled accents, and apply some form of accent taxonomy. As we shall see, few openly available datasets adopt a taxonomy for accents, and those which do use overly general labels. Moreover, these labels are inconsistent between datasets - making then difficult to combine commensurately. Thus, mechanisms to counter accent bias in speech technologies require researchers to attend to accent classification, description and representation.

Here, we pose the provocation: "What might we learn about accent representation for voice datasets if speakers were able to self-describe their accent rather than be ascribed an *a priori* categorisation?".

Using data visualisation approaches, we explore how data contributors self-identify their accents in the Mozilla Common Voice (CV) v11 dataset, attending to the descriptors used, their co-occurrence and emerging categorisations. We draw insights from this analysis to inform practices of accent representation in voice datasets for ML practitioners, thus contributing to the broader Fair ML movement. We propose an emerging taxonomy of accent descriptors in the English language from this analysis, against which ML practitioners may be better able to evaluate their datasets and models for accent bias. We believe the methodology outlined here could be readily applied to create taxa for other languages. Additionally, this work has implications for organisations who administer standards, supporting calls for accent description reform [24].

2 THE CURRENT STATE OF ACCENT REPRESENTATION IN VOICE DATASETS

To set the scene, we first explore how accents are *currently* represented in voice datasets intended for use in ML tasks such as speech recognition. Accent representation is a *boundary object* - an object which is malleable enough to meet the needs of multiple actors in a situation, yet rigid enough that it maintains a consistent identity [49]. ML engineers

use accent representation for dataset filtering, linguists use it to represent a person's speech, and researchers, like us, may use it to conduct inquiry. Accent representation - and its implications - thus links many communities of practice [54].

Dataset creators ascribe categories to which a speaker belongs in a process known as labelling or annotation. Accent categories ascribed to speech data items in a dataset allow practitioners to perform statistical bias measurements, and to evaluate models on data that is held back from training (e.g. [38]). The availability of accent data thus facilitates Fair ML practices.

There are many ways to describe a speaker's accent. It can be represented using a free text string - such as "American" or "Australian". Several standards also exist. For example, there are codes within the `ISO-639` [1] suite which may be used to represent languages, language groups and variants. Standards which allow for more nuance are also emerging. The `BCP-47` [2] standard facilitates labelling of spoken language data with additional granularity, such as "region" and "variant" [57]. A small group of scholars have advocated further extensions to `BCP-47` to allow for additional specificity, in the context of endangered languages [23].

It is important to note here how existing standards for representing spoken accent work to erase diversity. For example, two people could both speak English with an Australian accent - denoted `en-AU` in `BCP-47`. However, *within* this accent representation there could be significant variations in speech across characteristics such as formality, fluency or pitch. Standards for representing spoken language do not currently distinguish these variations, thus rendering invisible the diversity of accents. This is particularly the case for geographies with significant accent variation or for cohorts with accents of multiple heritage, such as migrants. The implications for ML are clear: the accent label may be the same, but the voices ascribed to it are qualitatively different.

Dataset documentation practices are therefore important to combating accent bias. Firstly, the accent of the speaker may not be recorded at the time of data capture. Secondly, if the accent *is* recorded, but uses an overly general label, then the variation in accents may not be represented. Without re-classifying the accents of the speakers in a dataset - which is at best probabilistic if this information is not stored at the time of capture – the ML practitioner has no basis on which to intentionally debias the training set or resultant model. The trained model is then deployed into speech technologies used by real people, who may then experience detrimental effects of accent bias (e.g. [37]).

How can we test if our ML-based speech technologies work well for people if we don't have ways to represent them?

Voice datasets are generally made available through one of three processes:

- Open source: Openly available datasets, such as those catalogued on the OpenSLR website [3], are often combined together into training sets for ML models, or used to evaluate those models [22, 44].
- Commercial data providers: Voice datasets can be purchased from commercial data providers. These may be off-the-shelf offerings, or created according to bespoke client specifications.
- Synthetic: Synthetic data is created using generative models, usually seeded by existing open source or commercial data captured from real people. This is a recent development which seeks to meet the increasing demand for voice data for ML.

To understand the current state of accent representation, we sought to examine several voice datasets from each group. Ten commercial providers were contacted. Three responded but were uncomfortable being directly quoted; their responses are synthesised. Open source voice datasets were identified through web searches. No synthesized

[1] https://www.iso.org/iso-639-language-codes.html
[2] https://www.ietf.org/rfc/bcp/bcp47.txt
[3] https://openslr.org

Table 1. How accent data is currently represented in voice datasets

Source	Dataset type	Representation	Standard used
Commercial organisations	Commercial	Per client specification, usually a country or subnational region descriptor	Unknown
Common Voice[1]	Open source	Country or supranational region descriptor	None
AusTalk[19]	Open source for research purposes	Provides granular speaker information including birthplace, native languages other languages and immersion information	None
Audio MNIST[3]	Open source	Country descriptor, as well as geographical origin of speaker down to level of town or city	None
George Washington University Speech Accent Archive[53]	Open source	Provides granular speaker information including birthplace, native language and language acquisition and immersion information	None
Corpus of Regional African American Language[29]	Open source	Provides granular speaker information including place of residence to city level, level of education, and occupation.	None
Multilingual Librispeech[43]	Open source	Data is categorised into separate corpora by language, e.g. "Dutch", but no accent data is provided.	None
Voxceleb[13]	Open source	Country descriptor, e.g. "USA"	None
English Accents in the British Isles[17]	Open source	Subnational regional descriptor, e.g. "Midlands English"	None
Librispeech[42]	Open source	None	None
TED-LIUM[46]	Open source	None	None
People's Speech[22]	Open source	None	None
Free ST American English Corpus[53]	Open source	None	None
None identified	Synthetic	–	–

datasets were available for this work, which is unsurprising, given that commercial research organisations are currently undertaking most of the work into synthetic speech data generation [20, 25].

Our results are captured in Table 1. In summary, where accent data *is* captured, it is often represented as a free text country or regional descriptor, however we note that some datasets contained much more granular geographic and demographic information, such as the speaker's city of residence and language immersion history. Moreover, standard formats for representing spoken language variants were not used at all in the datasets examined. Thus, it is difficult to make these datasets *commensurable* - accurately translatable and interchangeable [10]. This reduces their utility to ML practitioners who wish to reduce accent bias.

3 METHOD

Even though our analysis in 2 showed that spoken language standards, such as `ISO-639` and `BCP-47`, are rarely used in voice dataset documentation currently, we hold that it is a worthwhile inquiry to examine how people *self-describe* their own accents. In classification work, there is often a tension between "practical" and "formal" typologies . Contributors to Common Voice may not know how to formally classify their accent, but they will be able to describe their accent using everyday language - practical classification. This in turn has power to shape how formal classification evolves, hinting at our pragmatist research approach [8].

For this evaluation, we chose to examine self-styled accent data from the English corpus of Mozilla's CV dataset. It is the largest openly available voice dataset, and is comprised entirely of *elicited* speech, meaning data contributors read aloud from given text prompts. This removes the effects of *spontaneous* speech, where data contributors choose the words and phrases that are spoken.

Although CV now comprises over 100 languages, we chose English accents to explore because we are native English speakers, and because English had the largest variety of self-styled accents available to analyse [4]. We recognise that English is the highest-resource language globally and that choosing to focus on English here reinforces the Anglo-centrism apparent in ML practices. However, we believe our work is readily applicable to other languages, including low-resource languages where accent data is increasingly available. We open source all the analytical tools used here, in line with this axiological commitment.

Prior to 2022, data contributors to Mozilla's CV platform, when optionally providing demographic information in their profile, were *only* able to select from an *a priori* list of options to represent their accent. In mid-2022, Mozilla released an update to the CV contribution platform, allowing data contributors to self-specify an accent using free text. Subsequent releases of CV datasets included speaker-described accent data.

3.1 Data engineering

To extract the accent data, we used the tab-separated text file containing text transcriptions, audio file references and speaker demographic data for all transcript-validated utterances. This contained approximately 1.6 million rows. Using `pandas`, this list was filtered so that it contained one row per unique speaker, yielding 861,134 rows. This was further filtered to remove speakers who had not specified demographic data in their profile – meaning they did not supply any accent data to the CV platform. This left 14,822 unique accent entries. A speaker may specify multiple accents in an accent entry. These accent entries could be predetermined – that is, selected from an *a priori* drop-down list – or self-described by the speaker using a free text field – or a combination of both. The CV platform stores separate accents in an accent entry in a comma-delimited format, such as `United States English, Midwestern United States`. However, some accents also contained commas, and regular expressions were used to separate these accents. This allowed us to infer a co-reference relationship between the multiple accents expressed by a speaker, for example between `United States English` and `Midwestern United States`.

We then applied heuristics to reduce semantically identical accents into a single accent object for analysis. For example, there were around ten ways that `Midwestern United States` was expressed. We then separated distinct accents that were expressed in a single accent entry, while maintaining co-references. For example, `Indian with a tinge of an RP accent` was separated into three accents - the geographical descriptor `Indian`, the register marker and named accent `Received Pronunciation`, and the accent strength marker `Tinge`. Accents that were expressed in languages other than English were translated to

English using online machine translation tools[4]. For example, the Ukrainian descriptor Выраженный украинский акцент was translated to `pronounced Ukrainian accent` and split into the accent strength marker `pronounced` and the geographic descriptor `Ukrainian`. The heuristics used could be built upon in the future for processing self-described accents, and again are made openly available.

After applying heuristics, there were 164 distinct accents identified, with 297 co-reference relationships between them. These were represented as nodes (accents) and edges (co-reference relationships) for data visualisation. Of the 164 distinct accents, 16 were those available from the *a priori* drop down list, and 148 were self-described by speakers. Every accent was then ascribed one or more accent *descriptors*. The relationship between accents and accent descriptors was modelled using a one-to-many cardinality, so that one accent could have multiple descriptors. For example, the accent "Kiwi" is both a specifically named accent, and also refers to a geographical country. 171 accent descriptors were applied across the 164 accents. Seven accents had multiple descriptors applied. Accents which were available in the CV drop-down list were flagged as "predetermined" so they could be visually distinguished.

The accent descriptors were then coded and the taxonomy outlined in A emerged.

Accents were then exported in JSON format as *nodes*. Co-references for each speaker's accent entry, which could comprise multiple accents, were then calculated and exported as *edges*. For example, if a speaker specified their accents as `England English`, `Northern England` and `Northumbrian English`, this would be represented as nodes [1], [2] and [3], and edges [1,2], [2,3], and [1,3] respectively. Directionality of co-reference was not considered useful for analysis so bi-directional edges were removed.

The Python classes and Jupyter notebook that were used for data engineering are openly available[5], and could be re-used for similar future analyses.

3.2 Data visualisation

We then imported the JSON files into the Observable data visualization platform, and building on previous work by [7, 28, 56], we rendered a force-directed graph[6] to help explore the accents and their relationships. The force-directed graph was chosen because it allowed for analysis of *relationships* between accents, and is increasingly employed for similar analyses in the literature [5, 35].

In the visualisation itself, the nodes, representing accents, were colour-coded according to their taxonomic category. Semantically similar categories, such as geographical descriptors, were grouped within a similar colour range. Accents that were defined *a priori* in the CV platform were distinguished with a darker border. Interactive features were added to the visualisation so that mousing-over a node visually isolated its relationships, aiding analysis. Text labels were added to aid exploration. We decided to manually arrange the nodes, as clustering by accent descriptor obfuscated relationships. Insights were then drawn from the visualisation, detailed in 5.

4 VISUALISATION AND RESULTS

Table 2 provides a summary of the accent descriptors identified and the frequency with which they occurred. Descriptors that were available *a priori*, that is, before the ability for contributors to self-specify an accent, are noted. The resulting visualisation can be seen in Figure 2.

[4] https://translate.google.com
[5] https://bit.ly/facct23-commonvoice-accent-notebook
[6] https://bit.ly/facct23-commonvoice-accent-visualisation

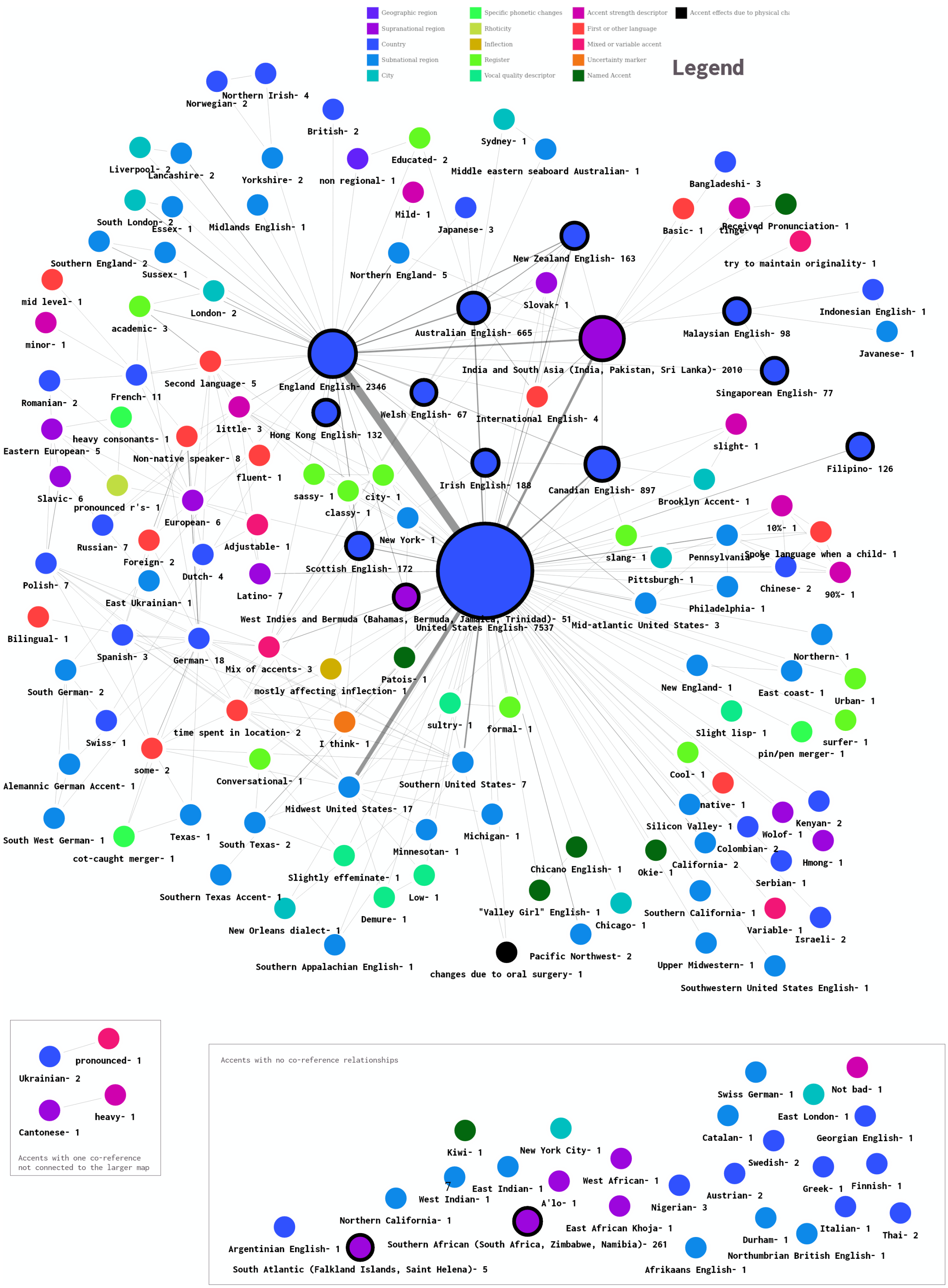

Fig. 2. Accent descriptors and their relationships visualised in a force directed graph. Labels indicate how many speakers in the CV dataset chose this descriptor. An interactive version of this map is available at https://bit.ly/facct23-commonvoice-accent-visualisation

Table 2. Accent descriptor count by taxonomic type in Mozilla Common Voice v.11 dataset

Taxonomic category	Count	No. of *a priori* descriptors	Percentage of total
Geographic descriptors	112	16	68.29%
Supranational region	15	4	9.15%
Country	42	12	25.61%
Subnational region	44	-	26.83%
City	10	-	6.10%
Other	1	-	0.61%
Register	12	-	7.32%
First or other language marker	13	-	7.93%
Accent strength descriptor	10	-	0.61%
Phonetic descriptors	5	-	3.05%
Specific phonetic changes	3	-	1.83%
Rhoticity	1	-	0.61%
Inflection	1	-	0.61%
Vocal quality descriptor	7	-	4.27%
Mixed or variable accent	4	-	2.44%
Uncertainty marker	1	-	0.61%
Accent effects due to physical change	1	-	0.61%
Named Accent	6	-	3.66%

5 ANALYSIS AND DISCUSSION

5.1 Geographical accent descriptors

Geographical accent descriptors were the most prevalent, accounting for around two-thirds of all descriptors. Within this category, descriptors predominantly fell into country- and subnational region-based categories. Where a country was specified, it tended to overlap with a distinct language - for example `German`, `French` and `Polish`. However, another interpretation of this result is that speakers are more likely to describe their accent using the *language* they speak rather than the *country* in which they reside. This interpretation does not resonate, however, with the frequent use of supranational accent descriptors - such as `European`.

Accent expression using subnational region descriptors tended to coincide with areas that had distinct spoken accents. For example, `Midwestern United States` was a frequent descriptor (expressed by 17 speakers), as was `Southern United States` (6). This pattern was also borne out at the city geographic level, with key examples being `London` (2) and `New York` (1), however, `Sydney` (1) was also expressed, which, as we understand, tends not to have an accent distinct from `Australian English`.

At the supranational region level, there were unexpected insights. `European`, `Slavic`, and `Eastern European` were frequently expressed descriptors (6, 6 and 5 speakers respectively). What we expected to find here were co-references to national or subnational regions - and there were some - to `Dutch` and `German` for `European`, to `Russian` for `Slavic` and to `Polish` for `Eastern European`, but not many. We can speculate, given the complicated history of these regions, and current conflicts, that speakers may not wish to identify with a particular national identity. Additionally, several supranational regions expressed as accents aligned to ethno-linguistic groups - such as `Latino` (7), `Wolof` (1), `Cantonese` (1), and `Hmong` (1). Wolof is a language and ethnic group of West Africa, while Hmong is one of South East Asia. This

again provides weak initial evidence for a stronger supranational than national language identity when self-specifying spoken accent.

5.2 Other descriptors

Speakers used several other categories to express their spoken accent. Accent strength descriptors were frequently used, and generally cast the speaker's accent as less pronounced, such as `tinge` (1), `mild` (1), `little bit` (3). Two notable exceptions were a speaker describing their `Cantonese` accent as "heavy" and one `Ukrainian`-accented speaker who expressed their accent as `pronounced`. First or other language markers were common. In particular, speakers frequently denoted if their accent was `non-native speaker` (8) or if they spoke English as a `second language` (5) - indicating that a native speaker is an "unmarked category" [52]. This category was also used for fluency markers, such as `Mid-level` (1) however we note that native speaker status and fluency are distinct. Future taxonomies may wish to separate this category. In linguistics, *register* refers to language used in a specific social situation [21], and this was used to provide a rubric for descriptors such as such as `academic` (3), `educated` (2) and `formal` (1). Occasionally, speakers expressed their accent using specific phonetic descriptors such as `heavy consonants` (1) or `prounounded 'r's` (1), or by employing vocal quality descriptors such as `sultry` (1) or `sassy` (1). We also identified a grouping expressing a mix or fluctuation of accents - such as `Variable` (1), `Mix of accents` (3) and `International English` (4), the latter two having, expectedly, several co-references with other accents.

6 IMPLICATIONS OF OUR WORK AND REFLECTIONS ON ITS ETHICAL CONSEQUENCES

Here, we discuss the implications of our work, and provide reflections on how we engaged with the ethical dilemmas the work presented.

The double-edged sword of granular data. Granular data presents a double-edged sword: on one hand, having more specific accent data assists ML practitioners to assess and correct fairness and bias using accent descriptors as an evaluative axis. On the other, the existence of this data *facilitates* discrimination by allowing the creation of accent classifiers (e.g. [34]). The possible uses - and abuses - are clear. Law enforcement could use accent classification for pre-emptive suspect identification. Indeed, this was the primary motivator of recent work in the Turkish language [32]. Call centers already make inferences about a client from their speech, such as emotional state [6]. It's a short stretch to imagine accent also being classified - with its attendant socio-demographic inferences. Alternatively, such classification could be used to *positively preference* speaker groups in particular contexts - such as migrants in service provision queues. In grappling with this challenge, we took a utilitarian view that the contribution of additional structures through which to undertake bias auditing outweighed the potential nefarious uses of this data.

Agency of data contributors. In choosing to analyse CV, we also took into consideration that contributors are clearly informed of the intended use of their data in voice applications[7] - unlike, for example, Librispeech, which was compiled, without speaker consent, from volunteer audio book recordings [42]. Moreover, data contributors here are *self-specifying* their accent descriptors. That is, they are applying a label to their own data, rather than having a label imposed on their data. This ascribes data contributors more agency in how accent descriptors are constructed and used in the world.

Reshaping language standards to better represent diversity of spoken language. Our work here has shown that the way voice data contributors express their accents differs markedly from existing industry standards that are available, but

[7] See the CV platforms privacy notice here

not often used, to represent them. These findings yield some early, tentative support for calls to reshape standards used to represent accents in spoken language. For example, data contributors to speech datasets may be more comfortable with selecting, or having ascribed, supranational accent or identity descriptors such as `European` or `Eastern European`. Existing standards, such as the `ISO-639` suite and `BCP-47`, do not currently have such groupings. However, `BCP-47` can be extended, and *could* be used to represent supranational descriptors. For example, the code `es-419` is used to represent `Spanish as spoken in Latin America`. It is conceivable that a code could be used to represent the accent `English as spoken in Europe`. We note this does not address the earlier problem identified of accent data frequently not being *captured* in the compilation of voice datasets. Retrospective classification of voice data using our taxonomy to provide accent representations, while ethically challenging, is a logical next research step - as would be training and evaluating supervised ML models to empirically test how accent classification contributes to performance. It is unknown whether the approach we have outlined here can be generalised to languages other than in English; applying a similar methodology may yield a useful accent taxonomy for other languages.

On classification work. In their seminal work, Bowker and Star [8] highlight how classification systems become woven into "working infrastructures". Classifications are imbued with power and with politics, and those of us who perform "classification work" must be mindful of our taxonomic sequelae. This, in part, motivates our work here. *Existing* accent taxonomies may render invisible the diversity of spoken language. This work is an early attempt to present an alternative data structure for consideration. It is yet to be shown conclusively whether this has utility for ML practitioners in addressing accent bias, and we hope to tackle this in future work.

The Common Voice platform. At a narrower scale, these findings have implications for the CV platform itself. Mozilla may consider whether to prompt data contributors to (optionally) provide a self-rating from *multiple accent categories* - such as geographical region, first or other language, strength of accent and so on. This would provide more granular accent information in the world's largest open source voice dataset, and help provide a rich resource for tools such as bias corpora that could be used to evaluate accent bias.

6.1 Limitations

We note several limitations with this work:

Small dataset with uncontrolled contributors. The dataset examined here is small, and represents accent entries of only 14,822 unique contributors to the CV voice dataset. Moreover, there is no control on the demographic contributions included in the CV dataset. The people who contribute are likely to have the free time, technical capability and computing resources *to* contribute, meaning many voices are likely to be unrepresented. Unsurprisingly, when we visualise the volume of data in the CV dataset along axes of gender and age[8], we see further evidence for unbalanced data contributions.

Accent dropdown list likely to influence data contributor accent description. Although a data contributor may specify their own accent, they are likely influenced by the *existing* listing. There is no way to control for this, other than by A/B testing the CV profile page, which is beyond the scope of this paper.

Common Voice accent data is not validated. Although CV speech data is validated for accuracy of *transcription*, there is no similar mechanism in place for assuring that accents specified are accurate.

[8] https://bit.ly/commonvoice-v11-metadata-coverage

Machine Translation (MT) tools not available for all languages in which accents may be expressed. Here, we used MT tools to translate accents expressed in French and Ukrainian, for which MT tools are readily available. Despite recent advances in MT (e.g. [2, 14]), these tools are not available, or accurate, for all of the other 7000 languages still spoken in the world. This places a constraint on replication.

7 CONCLUSION

"... moral questions arise when the categories of the powerful become the taken for granted; when policy decisions are layered into inaccessible technological structures; when one group's visibility comes at the expense of another's suffering." - *Geoff Bowker and Susan Leigh Star* [8]

In this paper, we reviewed voice datasets of three types to show that accent data is generally not represented using one of the existing standards for spoken language. We then filtered the Mozilla CV English corpus, which includes self-specified accent data from speakers, and applied a set of re-usable heuristics to separate and merge accents and descriptors into an emergent taxonomy. The accents and their co-references were visualised as nodes and edges interactively to aid analysis. We found speakers used a diverse range of descriptors to express their accent. In the geographic category, we identified a trend toward using supranational descriptors, specifically in Europe. Speakers also used categories such as accent strength, first or other language markers, specific phonetic changes and vocal quality descriptors to express their spoken accent.

This work contributes to broader efforts in bias detection and remediation in ML practice for speech technologies. Firstly, we introduced the concept of accent bias, and the need to address it as speech technologies scale. Next, we identified a gap in current voice ML practice - being that accent data is rarely captured, and when it is, it is not specified in a manner supporting commensurable data interchange. Further, we established that data contributors use a diverse range of categories to self-describe their accents. ML practitioners may be able to leverage the taxonomy we have developed as a structure for assessing voice datasets and models, thus helping to address the phenomenon of accent bias in rapidly-scaling speech technologies. Additionally, this work informs choices for annotating and labelling voice data in English, and suggests that the current standards available for codifying accents may be insufficient to represent speaker diversity. This is particularly relevant in an era of increasing globalisation and mobility, where a person's accent is dynamic rather than static over their lifetime [16, 36] and where new accents often emerge (e.g. [12]).

We are of the view that our methods here are applicable to other languages. To this end, we make a research software contribution to encourage replication by open sourcing our tool-sets.

ACKNOWLEDGEMENTS

Thanks are extended to Dr Jofish Kaye, Asst Prof Fran Tyers, Mr Ned Cooper and Mr Tom Chan for feedback on earlier iterations 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 and extends her thanks to the CV team, in particular EM Lewis-Jong and Asst Prof Fran Tyers.

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv preprint arXiv:1912.06670* (2019). arXiv:1912.06670

[2] Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building Machine Translation Systems for the next Thousand Languages. *arXiv preprint arXiv:2205.03983* (2022). arXiv:2205.03983

[3] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2018. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *arXiv preprint arXiv:1807.03418* (2018). arXiv:1807.03418

[4] Emily M Bender. 2019. The# Benderrule: On Naming the Languages We Study and Why It Matters. *The Gradient* 14 (2019). https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

[5] An Bing and Zhu Li-Gu. 2020. Film Big Data Visualization Based on D3. Js. In *2020 International Conference on Big Data and Social Sciences (ICBDSS)*. IEEE, 50–53.

[6] Milana Bojanić, Vlado Delić, and Alexey Karpov. 2020. Call Redistribution for a Call Center Based on Speech Emotion Recognition. *Applied Sciences* 10, 13 (2020), 4653.

[7] Mike Bostock. 2022. Force Directed Graph. https://observablehq.com/@d3/force-directed-graph

[8] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences*. MIT Press.

[9] Georgina Brown. 2014. *Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications*. Ph. D. Dissertation. University of York.

[10] Lawrence Busch. 2011. *Standards: Recipes for Reality*. MIT Press.

[11] Jason A Cantone, Leslie N Martinez, Cynthia Willis-Esqueda, and Taija Miller. 2019. Sounding Guilty: How Accent Bias Affects Juror Judgments of Culpability. *Journal of Ethnicity in Criminal Justice* 17, 3 (2019), 228–253.

[12] Jenny Cheshire, Paul Kerswill, Sue Fox, and Eivind Torgersen. 2011. Contact, the Feature Pool and the Speech Community: The Emergence of Multicultural London English. *Journal of Sociolinguistics* 15, 2 (2011), 151–196.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep Speaker Recognition. *arXiv preprint arXiv:1806.05622* (2018). arXiv:1806.05622 https://arxiv.org/abs/1806.05622

[14] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672* (2022). arXiv:2207.04672

[15] Ben Crystal and David Crystal. 2014. *You Say Potato: A Book about Accents*. Macmillan.

[16] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing* 3, 3 (2012), 349–365.

[17] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara E Rivera. 2020. Open-Source Multi-Speaker Corpora of the English Accents in the British Isles. (2020).

[18] John Edwards. 2009. *Language and Identity: An Introduction*. Cambridge University Press.

[19] Dominique Estival, Steve Cassidy, Felicity Cox, and Denis Burnham. 2014. AusTalk: An Audio-Visual Corpus of Australian English. (2014).

[20] Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking Synthetic Data for Speech Recognition. *arXiv preprint arXiv:2106.07803* (2021). arXiv:2106.07803

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

[22] Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. *arXiv preprint arXiv:2111.09344* (2021). arXiv:2111.09344

[23] 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.

[24] 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*. 3333–3339.

[25] Ting-Yao Hu, Mohammadreza Armandpour, Ashish Shrivastava, Jen-Hao Rick Chang, Hema Koppula, and Oncel Tuzel. 2022. SYNT++: Utilizing Imperfect Synthetic Data to Improve Speech Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7682–7686.

[26] Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 230–247.

[27] Kechi Iheduru-Anderson. 2020. Accent Bias: A Barrier to Black African-born Nurses Seeking Managerial and Faculty Positions in the United States. *Nursing Inquiry* 27, 4 (2020), e12355.

[28] John Alexis Guerra Gómez. 2021. Force-Directed Graph with Link Highlighting. https://observablehq.com/@john-guerra/force-directed-graph-with-link-highlighting

[29] Tyler Kendall and Charlie Farrington. 2018. CORAAL User Guide. The Corpus of Regional African American Language. Version 2018.04. 06. (2018).

[30] Bret Kinsella and Ava Mutchler. 2020. *Smart Speaker Consumer Adoption Report 2020*. Technical Report. Voicebot.AI.

[31] 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 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[32] Yunus Korkmaz and Aytuğ Boyacı. 2022. A Comprehensive Turkish Accent/Dialect Recognition System Using Acoustic Perceptual Formants. *Applied Acoustics* 193 (2022), 108761.

[33] Natalie J Lefkowitz. 1989. Verlan: Talking Backwards in French. *The French Review* 63, 2 (1989), 312–322.

[34] Jialu Li, Vimal Manohar, Pooja Chitkara, Andros Tjandra, Michael Picheny, Frank Zhang, Xiaohui Zhang, and Yatharth Saraf. 2021. Accent-Robust Automatic Speech Recognition Using Supervised and Unsupervised Wav2vec Embeddings. *arXiv preprint arXiv:2110.03520* (2021). arXiv:2110.03520

[35] Diaohan Luo, Chunfang Li, Chongyang Zhou, and Junshuai Xing. 2020. On the Knowledge Graphs of Postgraduate Entrance English Examination Based on WordNet and D3. Js. In *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, Vol. 1. IEEE, 991–996.

[36] Gretchen McCulloch. 2020. *Because Internet: Understanding the New Rules of Language.* Penguin.

[37] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I Don't Think These Devices Are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence* 4 (2021), 169.

[38] Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 6462–6468.

[39] Rosamund Moon. 2010. What Can a Corpus Tell Us about Lexis. *The Routledge handbook of corpus linguistics* (2010), 197–211.

[40] Maryam Najafian and Martin Russell. 2020. Automatic Accent Identification as an Analytical Tool for Accent Robust Automatic Speech Recognition. *Speech Communication* 122 (2020), 44–55.

[41] Kenneth Olmstead. 2017. *Nearly Half of Americans Use Digital Voice Assistants, Mostly on Their Smartphones.* Technical Report. Pew Research Center. https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

[42] 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)*. IEEE, 5206–5210.

[43] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. *arXiv preprint arXiv:2012.03411* (2020). arXiv:2012.03411

[44] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. *Robust Speech Recognition via Large-Scale Weak Supervision.* Technical Report. OpenAI. https://cdn.openai.com/papers/whisper.pdf

[45] Henry Rogers. 2014. *The Sounds of Language: An Introduction to Phonetics.* Routledge.

[46] Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: An Automatic Speech Recognition Dedicated Corpus.. In *LREC*. 125–129.

[47] Amee P Shah. 2019. Why Are Certain Accents Judged the Way They Are? Decoding Qualitative Patterns of Accent Bias. *Advances in Language and Literary Studies* 10, 3 (2019), 128–139.

[48] Catherine E. Shoichet. 2021. They're Building an App That Changes Accents | CNN. *CNN* (Dec. 2021). https://edition.cnn.com/2021/12/19/us/sanas-accent-translation-cec/index.html

[49] Susan Leigh Star and James R Griesemer. 1989. Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

[50] James Vlahos. 2019. *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think.* Eamon Dolan Books.

[51] Ze Wang, Aaron Arndt, Surendra Singh, and Monica Biernat. 2009. The Impact of Accent Stereotypes on Service Outcomes and Its Boundary Conditions. *ACR North American Advances* (2009).

[52] Linda R Waugh. 1982. Marked and Unmarked: A Choice between Unequals in Semiotic Structure. *Semiotica* 38, 3-4 (1982).

[53] S Weinberger. 2022. George Mason University Speech Accent Archive. https://accent.gmu.edu

[54] Etienne Wenger-Trayner and Beverly Wenger-Trayner. 2014. Learning in a Landscape of Practice: A Framework. In *Learning in Landscapes of Practice*. Routledge, 13–29.

[55] Cécile Woehrling and Philippe Boula de Mareüil. 2006. Identification of Regional Accents in French: Perception and Categorization. In *Ninth International Conference on Spoken Language Processing*.

[56] Stephen Wood. 2009. Creating Legends for Charts. https://observablehq.com/@wooduk/creating-legends-for-charts

[57] Sue Ellen Wright. 2019. Standards for the Language, Translation and Localization Industry. In *The Routledge Handbook of Translation and Technology*. Routledge London and New York, 21–44.

A EMERGENT TAXONOMY OF ACCENT DESCRIPTORS

Please refer 3.

Received Date; revised Date here; accepted Date here

Table 3. Emergent taxonomy of contributor-specified accent descriptors from Mozilla CV v11

Category	Definition applied to accent categorisation
Geographic region	
Country descriptor	Where the descriptor is a country or a nation-state.
Supranational region descriptor	Where the descriptor is a geographic region that overlaps multiple countries. An example would be `Slavic`, which refers to an ethno-linguistic group that covers several countries in Eastern Europe.
Subnational region descriptor	Where the descriptor is a geographic region that refers to a region within a country's national boundary. An example would be `Midwestern United States`.
City descriptor	Where the descriptor is a geographic region that refers to a city, town or municipality. An example would be `New York City` or `London`. Specific suburbs of cities have not been merged, for example `London` and `East London` are considered distinct accents.
First or other language descriptor	This descriptor was applied to accents where the data contributor expressed their accent with reference to whether they spoke English as a non-native or native speaker.
Accent strength descriptor	Where the data contributor expressed their accent using a marker of the strength of the accent.
Vocal quality descriptor	Where the data contributor expressed their accent using subjective stylistic qualities such as `sultry` or `sassy`.
Phonetic changes	This category refers to descriptors which express a particular phonetic change.
Specific phonetic change	Where the descriptor itself connotes the phonetic changes, such as `cot-caught merger` or `pin-pen merger`.
Rhoticity	Where the descriptor expresses how `/r/` and related phonemes are pronounced.
Inflection	Where the descriptor expresses an inflection change.
Register	Where the descriptor expresses a speech register - such as `formal`, `educated`, or `slang`.
Named Accent	Where the descriptor uses a popular name to describe an accent such as `Okie` or `Kiwi`.
Accent effects due to physical changes	Where the descriptor indicators physical changes to the speaker's vocal tract - for instance through surgery or disease.
Mixed or variable accent	Where the data contributor indicates that their accent is an amalgamation of accents but does not provide further information, which would allow for separation of descriptors.