Connecting the Dots: A Longitudinal Study of Performance Disparities in Automatic Speech Recognition

ALEXANDER METZGER*, ARUNA SRIVASTAVA*, and RUSLAN MUKHAMEDVALEEV, Koel Labs LLC, USA
EUNJUNG YEO, University of Texas at Austin, USA
SYED ISHTIAQUE AHMED, University of Toronto, Canada
NINA MARKL, University of Essex, England
SACHIN KUMAR, Ohio State University, USA
FARHAN SAMIR, University of Toronto, Canada

Automatic Speech Recognition (ASR) system evaluations have consistently revealed performance disparities along lines of race, gender, socioeconomic status, and language variety. While these disparities are well-documented, they are often (implicitly or explicitly) treated as issues that subsequent, larger, and more universal models will correct. We challenge this assumption of steady progress through the first longitudinal study of ASR performance disparities, evaluating 11 prominent systems released between 2021-2025 across 5 datasets representing diverse English accent varieties. Our findings reveal that while speakers of standard varieties maintain stable and improving performance, speakers of minority accents face substantial performance instability across model generations, with degradations of up to 65% absolute Word Error Rate between successive releases. These patterns demonstrate that performance disparities are not temporary anomalies but rather systemic failures of language technology development infrastructure that disproportionately affect underrepresented speaker populations. We argue that standard ASR benchmarks enact an implicit language policy privileging Mainstream US English varieties. Given that ASR systems are increasingly deployed in high-stakes contexts where speakers of minority varieties already face linguistic discrimination (immigration systems, workplace surveillance, educational assessment), continuous auditing for disparate impacts is critical.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Natural language processing**; Language resources.

Additional Key Words and Phrases: ASR, fairness, evaluation, longitudinal, sociolinguistics

ACM Reference Format:
Alexander Metzger, Aruna Srivastava, Ruslan Mukhamedvaleev, Eunjung Yeo, Syed Ishtiaque Ahmed, Nina Markl, Sachin Kumar, and Farhan Samir. 2025. Connecting the Dots: A Longitudinal Study of Performance Disparities in Automatic Speech Recognition. 1, 1 (January 2025), 19 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this work.

Authors' Contact Information: Alexander Metzger; Aruna Srivastava; Ruslan Mukhamedvaleev, Koel Labs LLC, Seattle, Washington, USA, {alex,aruna, ruslan}@koellabs.com; Eunjung Yeo, University of Texas at Austin, Austin, USA, eunjung.yeo@utexas.edu; Syed Ishtiaque Ahmed, University of Toronto, Toronto, Canada, ishtiaque@dgp.toronto.edu; Nina Markl, University of Essex, Colchester, England, nina.markl@essex.ac.uk; Sachin Kumar, Ohio State University, Columbus, USA, kumar.1145@osu.edu; Farhan Samir, University of Toronto, Toronto, Canada, fsamir@dgp.toronto.edu.

Manuscript submitted to ACM

1 Introduction

Automatic Speech Recognition (ASR) systems are among one of the oldest, formative pursuits in computational linguistics research [53]. Concerns about the limited robustness of these systems, owing to the relative homogeneity of their training datasets, have been prominent for decades [9]. A wide array of studies have reported on performance disparities among important demographic constructs. At different points in time, English ASR systems have been shown to be worse for children and the elderly [9, 89]; for Black speakers compared to white speakers in the US [47, 50]; for speakers from lower socioeconomic backgrounds [15]; and for speakers of stigmatized varieties of English [37, 56], among others. While these studies are situated in particular temporal contexts, relying on prominent models and organizational APIs at the time, the same types of biases are found across linguistic and technical contexts.

One common response to these reported shortcomings is the use of different strategies to reshape the training data distribution: targeted training data acquisition, corpus resampling, data augmentation, among other methods; for a review see Table 4 in [63]. However, these ad-hoc patches risk concealing broader systemic challenges in the benchmark-driven infrastructure that has been the dominant paradigm in Machine Learning (ML) system development research [28]. That is, this ad-hoc approach treats performance disparities as temporary shortcomings, or glitches [7], rather than enduring aspects of norms and ideologies pertaining to building machine learning systems for speech [22]. While biases furthermore can be targeted and optimized, this does not guarantee that they are taken up.

In this work, we evaluate 11 prominent models on datasets that are designed to capture a wide range of varieties of English (Section 4.1). These datasets then provide an opportunity to assess performance of ASR models at the margins of the contexts the models were designed for. We find that speakers of minority accent varieties frequently observe major performance degradations between successive model releases, while speakers of more standard varieties benefit from stable performance over new generations of models. Thus, our results complicate commonplace narratives of monotonic progress in machine learning research and show empirically that improvements in models are distributed unevenly across speaker groups. Moreover, our findings are aligned with previous positions that ML benchmarks ought to be treated like physical infrastructure, requiring continuous and documented updating [42, 59]. Finally, these findings call for greater transparency of speech processing systems that are deployed in downstream applications where speakers of marginalized English varieties would come into sensitive contact with them, especially in contexts characterized by hierarchy, including but not limited to schools, workplaces, and incarceration facilities where language practice is highly scrutinized [58].

2 Global Englishes, Language Ideologies and ASR evaluation

To contextualize and motivate our analysis, we briefly discuss Global Englishes and linguistic discrimination, We then highlight common ideologies about language in language technology research, before exploring how these different issues (linguistic discrimination and language variation, computationalist approaches to language) come together to shape current benchmarking practices in automatic speech recognition.

2.1 English Varieties and Linguistic Hierarchies

English has long served as the global lingua franca, the result of complex intertwined histories of colonialism, globalization, and US cultural hegemony. More specifically, applied linguistics scholars have argued that the dominance of the standard prestige varieties Mainstream US English (MUSE) and Standard Southern British English (SSBE) is inextricable from the central role that the British empire, and, later, the United States played in implementing capitalist

world-economy [67]. Close adherence to prestige varieties is robustly rewarded in the labor market, and deviance is penalized [67]. This has been a longstanding effect, with recent studies showing that standard varieties are rewarded by greater perceived competency, and therefore greater access to prestigious employment [43, 84]. While connections between language varieties and ostensible relationships with professional and intellectual competencies have been dispelled in sociolinguistics and linguistic anthropology circles, perceptions rooted in all too familiar bigotries, they still remain deeply and broadly entrenched among the public [21].

Global Englishes exhibit vast diversity among both standard and non-standard speakers, varying across geographical regions, cultural contexts, age groups, educational backgrounds, and linguistic histories. Varieties spoken by both first-language and second-language speakers of English beyond and within hegemonic centers of English (especially the US and the UK) differ from hegemonic varieties on multiple linguistic axes (phonetics, phonology, syntax, lexicon, among others). These varieties and their speakers represent the vast majority of English speakers globally, but remain marginalized. In addition to the social and economic discrimination discussed above, they remain under-represented (and often entirely forgotten) in the development of language technologies. This exclusion can have dramatic consequences as larger parts of our professional and social lives involve interacting with or through such technologies.

2.2 Ideologies about language and computation

The long prevailing understanding in science and technology studies is that by default new technological innovations, while often framed commercially as disruptive and libratory, will tend towards reflecting and amplifying existing social orders [8, 35, 88]. We should expect this to be the case for speech technologies all the same, and indeed by now many reports of hierarchical performance in ASR systems trend in the way we would expect. But here we will place these studies in a broader cultural context, to understand why the disparate effectiveness of these systems are systemic problems rather than episodic and independent glitches. To see this, we should first observe that the conceptualization of language in computational culture tends to be one that is isolated from people, speakers and listeners, communities of practice [10]. Instead, languages tend to be treated as discrete, bounded objects, in stark opposition to the study of Global Englishes that in its very name emphasizes its pluricentricity [67, Chapter 7]. In language technology scholarship, the taxonomization of language into discretized, separable, bounded, and countable codes— practice critiqued in linguistic anthropology [32, 68, 83]—is remarkably prevalent.

Consider, for example, papers titled "How to adapt your pretrained multilingual model to 1600 languages" [30], "Scaling speech technology to 1000+ languages" [72], and "Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data", just to name a few. From these publications at major conferences, we can start to gleam that this formation of languages as natural, distinctive, and countable objects is largely taken for granted – "an objectualization, of language and its use, in which language acquires a thinginess" [86]. Nor is this conceptualization a recent development; in his famous and formative memorandum on machine translation, Natural Language Processing pioneer Warren Weaver understood languages as being independent codes (or as he originally wrote them, "routes") that had a "univocal interpretation" [35, pp. 86]. Thus, the computational treatment of language lends itself towards objective, standard definitions of language. While there is of course flexibility in how this standard for English might be defined, in practice, it has been shaped by people and institutions in the Global North.

2.3 ASR Evaluation Infrastructure

ML benchmark datasets have been theorized as a form of *infrastructure* [25, 42], in many senses of the concept [12]. Most importantly for our purposes, they undergird the development of new speech-processing models; and the human

judgments that went into the data curation process [82] become more invisible as they become naturalized as standard benchmark datasets [25]. For ASR benchmarks, these are socio-linguistic judgments about the type of speech that is considered "typical" and thus important to model, and the types that can be made invisible and subsequently ignored. In effect, this process of formalization into a benchmark dataset can be interpreted as a type of language policy [59], specifying the spoken English varieties and contexts that are most appropriate. Many Global Englishes are foregone in the curation of standard ASR benchmarks. While some recent work draws attention to these concealed positionalities [34, 51, 57], established benchmarks are used in model evaluation without much discussion as part of conventional evaluation practices. What remains marginal in technical research, despite the inclusion of more diverse benchmarks [e.g., 46], is deeper engagement with the limitations of benchmarks and approaches which incorporate social constructionist view of language [cf. 80].

There is clearly value in building out such shared infrastructure, enabling large-scale community collaboration across varied institutional contexts. Simultaneously, however, it raises major concerns around construct validity, especially as the culture of AI research has been criticized for over-claiming the generality of the tasks defined by these benchmarks [77]. As originally recounted by Raji et al., the ImageNet dataset was described as "an attempt to map the entire world of objects"; the GLUE benchmark as frameworks for developing "general-purpose language understanding technologies"; both served as infrastructure by way of grounding development efforts into a single defined performance number [77]. Through such generalistic framings, these benchmarks obfuscated the normative decisions that went into standardizing visual and textual information in specific ways [77].

Similarly, construct invalidity in speech processing benchmarks is also a major concern. Because mainstream English varieties are "aggressively hegemonic" [86], they are made to feel neutral, unmarked, or objective. This aligns well with the political values in constructing machine-learning datasets, described by [82], in particular, the high appraisal of impartiality, or objectivity, in data collection. That geographic, temporal, and social considerations inform this objective standard is rarely acknowledged [82]; instead, MUSE and SSBE speech datasets are typically labeled as simply "English". The failure to contend with the positionalities in data collection then necessarily leads to concerns of construct validity [77]; namely, that the understanding of English that is constructed is a rather narrow sampling of Global Englishes. Benchmarking infrastructure is understood to not only conceal normative biases but amplify them through the "reputation of neutrality and fair judgment" that is given to numerical knowledge [64], in the form of a "single performance number."

3 Prominent Evaluation Benchmarks

Here we argue that prominent evaluation benchmarks effectuate an Anglocentric language policy. While there are no firm prescriptions (e.g., by publishers, regulators or professional organisations) regarding benchmark selection, there is nevertheless a clearly observable norm within the field. As observed in [59], "the absence of an official policy...often serves only to reinforce the power and hegemony of prestige varieties, and marginalize others." We can still analyze regularities in benchmark selection, and these historically entrenched regularities can be constituted as a de facto language policy [59].

3.1 Benchmark Selection

We study these regularities by first selecting a range of 11 prominent ASR models over the last 4 years, shown in Table 1. We selected models from 2021 onwards, as this period is when large-scale pre-training emerged as the dominant

paradigm in machine learning [38].[1] We selected organization that are notable for prominent speech-processing technologies: Mozilla, Microsoft, OpenAI, Meta, CMU, Alibaba, NVIDIA, and Google. We list the benchmarks that were employed in assessing the English ASR capabilities of these models, gathered from the models' corresponding technical reports that are referenced in the table. In the case of composite benchmarks like SUPERB [92], OpenASR [87], and, SpeechStew [17], we list all of the constituent datasets individually rather than the aggregated benchmark.

We then bisect the set of evaluation benchmarks to highlight ones that are conducive to evaluating robustness to accent variation (Section 3.1.1), especially outside of prestige varieties like Mainstream US English, and those that are centered on prestige English varieties (Section 3.1.2). We explain this categorization in the following section.

3.1.1 *Categorizing benchmarks – Diversity of English Varieties.*

VoxForge	"Our source for accented speech is the publicly available VoxForge (http://www.voxforge.org) dataset, which has clean speech read from speakers with many different accents." [2]
Artie	"The Artie Bias Corpus contains information on 3 gender classes, 17 English accents, and 8 age ranges." [60]
VoxPopuli-En	"VoxPopuli provides 29 hours of transcribed speech data of non-native English intended for research in ASR for accented speech" [33]
Earnings22	"Our attention focused on aggregation of accented public English-language earnings calls from global companies." [24]
SPGISpeech	"There are roughly 50,000 speakers in SPGISpeech, drawn from corporate officers and analysts appearing in English earnings calls and spanning a broad cross-section of L1 and L2 accents" [66]
CommonVoice-EN	"The 2021 release of Common Voice English (7.0) contains 2,015 hours of (validated) speech submitted by over 75,000 speakers some of whom opted to provide some information about their gender and accent" [57]

There are notable multi-accent benchmarks, as we list above. The concern is not whether such benchmarks exist at all, but rather that they are peripheral – they are sometimes used in evaluations, and often not.

3.1.2 *Categorizing benchmarks – Homogeneous English Variety Representation.*

Other representational axes – multilinguality. Other benchmarks also sought to increase representational diversity, but along other axes rather than L2-English speaker representation. One prominent axis was on multilinguality; for example, *Common Voice*, *Fleurs*. This is a reasonable, well-motivated axis to pursue, as a prominent critique of NLP as a discipline has been its focus on English datasets emerging from central Anglophone countries [55], a critique that applies for audio datasets just as well [1]. In collecting datasets that contain a larger number of distinct languages, however, these works typically adopt the normative stance that languages are geographically and culturally homogenous artifacts [83], as we argued in Section 2.2. Often this stance is implicit, though can be ascertained from the categorization of the speech data into different language bins, with no discussion of the particular varieties that the speakers are employing as well limited sociodemographic information [1], e.g., in Multilingual Librispeech [73]. Sometimes, this stance is explicit, for example the Fleurs dataset explicitly recruits "native" speakers in constructing its multilingual speech evaluation datasets: "We collected three recordings by three different three recordings by three different native speakers" [20].

[1]Though, the general computational philosophy to take a data-centric approach to ASR has been around since the 1960s [53].

Inspection of the English speech in this dataset indicates a Mainstream US English variety. Thus, the construction of "linguistic diversity" is complex and value-laden, and the representation of World Englishes may even be better in a monolingual dataset than one that contends to serve as evaluation infrastructure for "universal" [20] speech models.

There are multilingual datasets that have more pluralistic representations of English, for example, CommonVoice [3] and VoxPopuli [90]. However, the most common accent representations in CommonVoice is by far Mainstream US English, based on audits of v7 and V13 [57, 78]. VoxPopuli contains 15 European accent varieties, but is a minority of the total transcribed corpus (29/1,800 hours). How these corpora are evaluated matters. For example, the MMS model evaluates on CommonVoice over the whole corpus [72, Table 5]. Since minority accent varieties are a small proportion of the English subset of the corpus, and necessarily an even smaller subset of the entire corpus, the average word error rate would be ineffective in tracking robustness to accent variation, particularly variation outside of from Mainstream US English. As recognized in [11], operationalizing performance as "correctness across individual predictions" is a value-laden modus operandi, and reflects Anglocentric language policy due to the skewed constructions of these datasets. By comparison, the Whisper model presents their evaluation targeted on the English partitions of both VoxPopuli and Whisper, thus more effectively tracking accent robustness [75, Table 2]. For this reason, we distinguish between CommonVoice and CommonVoice-EN, depending on whether the subset was evaluated specifically or whether it was the entire set. We draw the same contrast for VoxPopuli and Voxpopuli-EN.

Other representational axes — scale. Other evaluation benchmarks have sought to increase through scale, implementing the aphorism "there is no data like more data", attributed to Robert Mercer during his tenure at IBM's Continuous Speech Recognition team [53], as a data collection policy. The *CALLHOME* [54], *SwitchBoard* [36], and WSJ [70] corpora were funded by the US Department of Defense to advance speech recognition research from "specific database inquiry tasks, characterized by medium vocabularies" [70] towards open-ended, multi-speaker dictation. These were characterized by increases in number of speakers, number of words in transcripts, and recording durations – in a word, *scale*. For example, CALLHOME emphasized "18.3 hours of transcribed spontaneous speech, comprising about 230,000 words" [14]. In a similar vein, we have "WSJ corpus will provide DARPA its first general-purpose English, large vocabulary, natural language, high perplexity, corpus containing significant quantities of both speech data (400 hrs.) and text data (47M words)." The focus on quantification and scale as a stand-in for diversity continues to be pervasive in ML data collection [6, 82], like TedLium [79] and Librispeech [69], both describing their datasets first and foremost in terms of scale ("a total of 774 talks, representing 118 hours of speech"; "contains 1000 hours of speech"), rather than other potentially salient characteristics, like speakers' demographic attributes.

As previously observed by Scheuerman et al., a foundational assumption in targeting large volumes of data is that such "unconstrained" data collection methods will naturally represent diversity on sociodemographic axes [82]. Optimizing for scale first and foremost is thus not unlike an attempt to build up the spoken version of the hypothetical *Everything in the Whole Wide World Benchmark* [77]. In these works, the disparities between the availabilities of different sociodemographic speaker groups is rarely acknowledged let alone interrogated. This inevitably leads to the proliferation of speech benchmarks that have skewed representations of Mainstream English varieties [57], mirroring the broad trend surrounding skewed representations of WEIRD populations in other machine learning evaluation datasets [81] and in experimental sciences broadly construed [39].

3.2 Observations

Descriptive statistics. Having explained our classification of these individual evaluation benchmarks into pluricentric ones as opposed to univocal ones, we now study their aggregate distribution in Table 1. First, we observe that out of the 43 evaluations conducted across the technical reports, only 10 of them include some consideration of accent diverisity. Out of the 18 unique evaluation benchmarks, 6 of them have pluricentric considerations of English varieties. Four of the systems (WavLM, MMS, Chirp-3, Omnilingual) include no explicit considerations of accent pluralism.

It almost goes without saying that Mainstream US English is never excluded from the evaluation infrastructure. Moreover, ASR performance on this variety is often assessed in multiple contexts, from spontaneous vs. read speech (e.g., Switchboard vs. Librispeech); to individual speakers vs. multiple (TedLIUM vs. AMI Meeting Corpus). By comparison, if accent pluralism is considered in the evaluation, it tends to be in a more limited range of contexts and registers. For example, the OWSM and Qwen models model evaluates performance on CommonVoice, only assessing for read-speech of neutral-register Wikipedia sentences. Whisper [Table 2 75] and Canary [Table 3 74] diverge from this, evaluating on a variety of registers and contexts.

Discussion. Benchmarks serve a dual purpose in ML. Most commonly, they serve as evidence of improved performance on a target task, understood as the most common goal in ML [11]. But there are other commonplace values, including parameter and data efficiency, transparency, and theoretical analyzability [11]. In pursuing primary goals that are not standard benchmark dominance, benchmarks serve a role that is analogous to regression tests in software development – preventing against performance degradations.

Indeed, we observe benchmarks assuming this regression prevention role in some of these systems, that have primary goals other than achieving state of the art ASR performance on notable benchmarks. For example, Qwen-2 expands the set of speech-processing tasks that the model can complete [19]. The OWSM system was meant to serve as an open-data, open-weights (near-)replication of OpenAI's Whisper [71], the latter being trained on a dataset of 680,000 hours of speech recordings of suspicious origin [75]. The MMS system was aimed at expanding the set of languages that ASR models can transcribe [72]. The models are often not the state-of-the art on some benchmarks, yet this is not understood as a failure, but rather evidence that the model did not regress too much along other performance axes of their primary one; for example, both the MMS and OWSM reports are transparent that Whisper outperforms them on some standard benchmarks. However, we see that performance degradations for standard varieties of English are thoroughly tested, while minority varieties are not. We thus test the following hypothesis next.

H1: Minority English varieties are likely to degrade in performance relative to prior system generations.

4 Experiments and Results

Here we perform a longitudinal analysis of the prominent ASR models (listed in Table 1) on multi-accent datasets (Section 4.1), assessing whether minority accents are more prone to performance degradations between successive releases of speech processing models. We describe the setup for our longitudinal analysis in Section 4.2 and present results in Section 4.3.

4.1 Datasets

To select datasets for the evaluation, we compiled a list of multi-accent datasets, specifically those with speakers for whom English is a second language (L2-English speakers). We aim for datasets that are not considered standard benchmarks in the ASR community; none of the ones in our evaluation were used to by the development teams for the

systems listed in Table 1. We built up our evaluation corpora list through a combination of scholarly search engines and consulting with sociophoneticians. This aggregation phase left us with 39 potential datasets. However, many of these were unsuitable for the purposes of our analysis. Several datasets lacked accent annotations for their recordings, making it unsuitable for a comparative analysis between performance on standard and minority accents. In the same vein, other datasets did not have recordings from L1-speakers, precluding a comparative analysis. Other datasets required requests for access that are still pending. For the remaining datasets, we downloaded and listened to random samples, excluding ones that had an overrepresentation of noisy, unintelligible recordings. Ultimately, we used 4 multi-accent corpora for our evaluation; we describe these in further detail below. In addition, we provide preprocessing details in Appendix A.

4.1.1 Speech Accent Archive. Speech Accent Archive [91] is an English speech corpus containing both non-standard and standard speakers reading the same passage: "*Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*"

It has previously been used for assessing performance disparities in ASR systems across different Global English varieties [for example, 16, 27, 37, 91]. The dataset controls for the noise level and the distance to the microphone. Each subject is given one minute to read over the passage and ask questions about unfamiliar words before recording. However, this was done by many different researchers, so the quality and procedure does differ between samples. As of writing, the corpus contains 3038 samples across 392 standard languages. The corpus contains a smaller subset of high quality expert-annotated samples. This constitutes approximately 10 hours of speech from 1238 speakers across 831 birthplaces and 190 standard languages. Of these speakers, 181 are listed as standard English speakers. We only kept the accents where 20 or more speakers had contributed a recording. Each sample has median duration of 26 seconds with 8 second standard deviation, varying by speaking rate.

4.1.2 L2 ARCTIC. L2 ARCTIC [95] contains about 1 hour of scripted English from each of 24 non-standard speakers, utilizing source texts originally derived from Project Gutenberg [48]. The speakers are evenly divided, 2 male and 2 female, between standard Arabic, Hindi, Vietnamese, Spanish, Korean, and Mandarin. There is also an unscripted subset, the "suitcase" subset from 22 of the speakers where they were prompted with the suitcase story [26] and asked to explain what was happening in the pictures. About 1 hour of the scripted subset is expert-annotated (3,599 samples, 150 utterances for each speaker). The entire suitcase subset is expert-annotated. Additionally, we record 1 sample of L1 North American English speech by one of the authors using the suitcase prompt, given that L2 Arctic does not contain an unscripted sample spoken by a standard English speaker.

4.1.3 OpenSLR83. OpenSLR83 is a dataset of transcribed speech featuring English speakers from various dialectal regions across the UK and Ireland. The original dataset comprises 17,877 utterances across nine dialect-gender combinations: Irish English, Midlands English, Northern English, Scottish English, Southern English, and Welsh English.

4.1.4 ALLSSTAR. ALLSSTAR Corpus [13] is a multilingual speech dataset containing paired recordings from each participant in both their first language (L1), and in English as a second language (L2). The dataset features 102 L2 English speakers representing 21 standard language backgrounds, along with 26 standard English speakers.

The corpus comprises both read and spontaneous speech (e.g., picture descriptions and question–answer sessions), while we focused on the read-speech portion for this paper. The read materials include short sentences and paragraphs, with each speaker producing 120 unique sentences and 3 unique paragraphs. Sentence recordings were collected in two

	Model	Release date	Organization	Paper	Evaluation Datasets (ASR)
m_1	Deepspeech	2021-10	Mozilla	[2]	WSJ, Librispeech, VoxForge, CHiME, CommonVoice-EN, Multilingual Librispeech
m_2	WavLM	2021-12	Microsoft	[18]	Librispeech
m_3	whisper	2022-09	OpenAI	[75]	Librispeech, WSJ, Switchboard, CORAAL, TedLium, CallHome, VoxPopuli En, Fleurs, AMI Meeting corpus, CommonVoice, CHiME, Artie Bias Corpus, Earnings22
m_4	whisper-v2	2022-12	OpenAI	[75]	Same as V1
m_5	MMS	2023-05	Meta	[72]	Fleurs, CommonVoice, VoxPopuli, Multilingual Librispeech, MMS-lab (bible)
m_6	whisper-v3	2023-11	OpenAI	[75]	Same as V1
m_7	OWSM	2024-01	CMU	[71]	CommonVoice-EN, Fleurs, Librispeech, TedLium, VoxPopuli, WSJ, Switchboard, Multilingual Librispeech
m_8	Qwen2-Audio	2024-08	Alibaba	[19]	Librispeech, CommonVoice-EN, Fleurs
m_9	Canary	2025-07	NVIDIA	[74]	AMI corpus, Earnings22, Gigaspeech, SPGIspeech, Tedlium, Voxpopuli, MUSAN, Casual Conversations Dataset, Multilingual Librispeech
m_{10}	Chirp-3	2025-10	Google	[94]	AMI Meeting Corpus, Broadcast News, Common Voice, LibriSpeech, Switchboard, Tedlium, and WSJ, Fleurs, Youtube captions
m_{11}	Omnilingual	2025-11	Meta	[65]	MMS-Lab (bible), Fleurs, MLS, CommonVoice, AllASR

Table 1. Models evaluated in our diachronic analysis.

sessions, each captured as a single continuous audio file. The paragraph topics were: Little Prince (LPP), Declaration of Human Rights (DHR), and The North Wind and the Sun (NWS). In total, the dataset includes 129 recordings from L1 English speakers and 570 from L2 English speakers. The final analyzed subset comprised 10,753 short sentences (2,862 L1; 7891 L2), 2,548 LPP recordings (727 L1; 1,821 L2), and 1,869 DHR paragraphs (514 L1; 1,355 L2). Recording duration varied by material type, with short sentences averaging 1.68 seconds, LPP averaging 4.16 seconds, and DHR paragraphs averaging 6.04 seconds.

4.2 Setup

Consider two models m_i and m_j, where $j > i$ indicates that m_j was released after m_i. We define m_j to exhibit a performance degradation over m_i if the Word Error Rate for the later model is higher than the one for the earlier model: $WER(m_j) > WER(m_i)$. We measure this as $d(m_i \rightarrow m_j) = max\left(0, WER(m_j) - WER(m_i)\right)$, where the WER is computed on a held-out test set. When we have not only two models but a series of them ordered by release date $[m_1, \ldots, m_n]$, we can consider the set of degradations

$$\mathbb{D} = \{d(m_i \rightarrow m_j) | i \in [n], j \in [i]\}$$

We then obtain the *Mean Degradation* (MD):

$$MD = \frac{\texttt{sum}(\mathbb{D})}{n(n-1)/2}$$

We can compute MD for every accent variety in every dataset from Section 4.1. We use the set of models listed in Table 1 for our longitudinal analysis. We follow the ordering in the table, setting $m_1 = \texttt{Deepspeech}, \ldots, m_n = \texttt{Omnilingual}$.

Dataset	Background	Mean Degradation (*MD*)	*MD* rank
L2-Arctic	Vietnamese En.	**3.7**%*	3/7
	Mainstream US English (MUSE)	2.1%	2/7
L2-Arctic-Spontaneous	Korean En.	**3.2**%*	3/7
	MUSE	1.0%	1/7
Speech Accent Archive	Korean En.	**4.6**%	12/12
	MUSE	4.3%	2/12
OpenSLR-83	Irish En.	**5.1**%	6/6
	Southern En. (UK)	3.3%	5/6
ALLSTAR-LPP	Vietnamese En.	**1.7**%*	8/23
	MUSE	1.0%	2/23
ALLSTAR-DHR	Cantonese En.	**1.6**%*	17/23
	MUSE	0.0%	1/23
ALLSTAR-HT1	Nkore En.	**4.3**%*	23/23
	MUSE	1.0%	3/23
ALLSTAR-HT2	Nkore En.	**5.6**%*	23/23
	MUSE	1.1%	2/23

Table 2. Mean Degradation (MD) across 8 datasets

4.3 Results

We present all of our results in Figure 1. We highlight two language backgrounds for each dataset: (1) what we interpret as the standard variety in that dataset, (2) we rank all accent varieties by their average word error rate across all models in descending order, then highlight the highest-ranked accent variety in this list.

4.3.1 Mean Degradation Results. In Table 2, we show the Mean Degradation (*MD*) for both the standard and the selected accent varieties. In all cases, we find that the minority accent variety has a higher MD score, indicating that performance degradations will be more impactful for these varieties.

We test for statistical signficance using a bootstrap percentile test [31]: re-sampling the WER distributions for both varieties, re-computing their MD scores, and checking whether the minority accent variety has a higher MD than the standard variety. We repeat this procedure with $B = 1000$ bootstrap samples. We find that the procedure is statistically significant for all datasets but the OpenSLR-83 dataset and the Speech Accent (SAA) archive dataset. For OpenSLR-83, we find that Irish English (minority) has similar regression patterns to Southern British English (standard). For the SAA dataset, the similar MD scores can be explained by a large increase in error rates with the release of m_3 = `whisper-v1` for the Mainstream US English variety (MUSE). `whisper-v1` also performed poorly on the Korean English variety, but since the prior models (m_1, m_2) also struggled with Korean English, the regression ($d(m_3, m_2)$ or $d(m_3, m_1)$) was not nearly as pronounced.

Ordinal rankings of accent varieties. We also find that the attenuated risk of degradations for standard Mainstream US English (MUSE) is robust across all accent varieties in the datasets we tested. In each dataset, we ranked the varieties by their MD scores, finding that MUSE is ranked 2nd out of 7 for L2-Arctic, then 1/7 (L2-Arctic Spontaneous), 2/12 (Speech Accent Archive), 2/23 (ALLSTAR-LPP), 1/23 (ALLSTAR-DHR), 3/23 (ALLSTAR-HT1), and 2/23 (ALLSTAR-HT2).

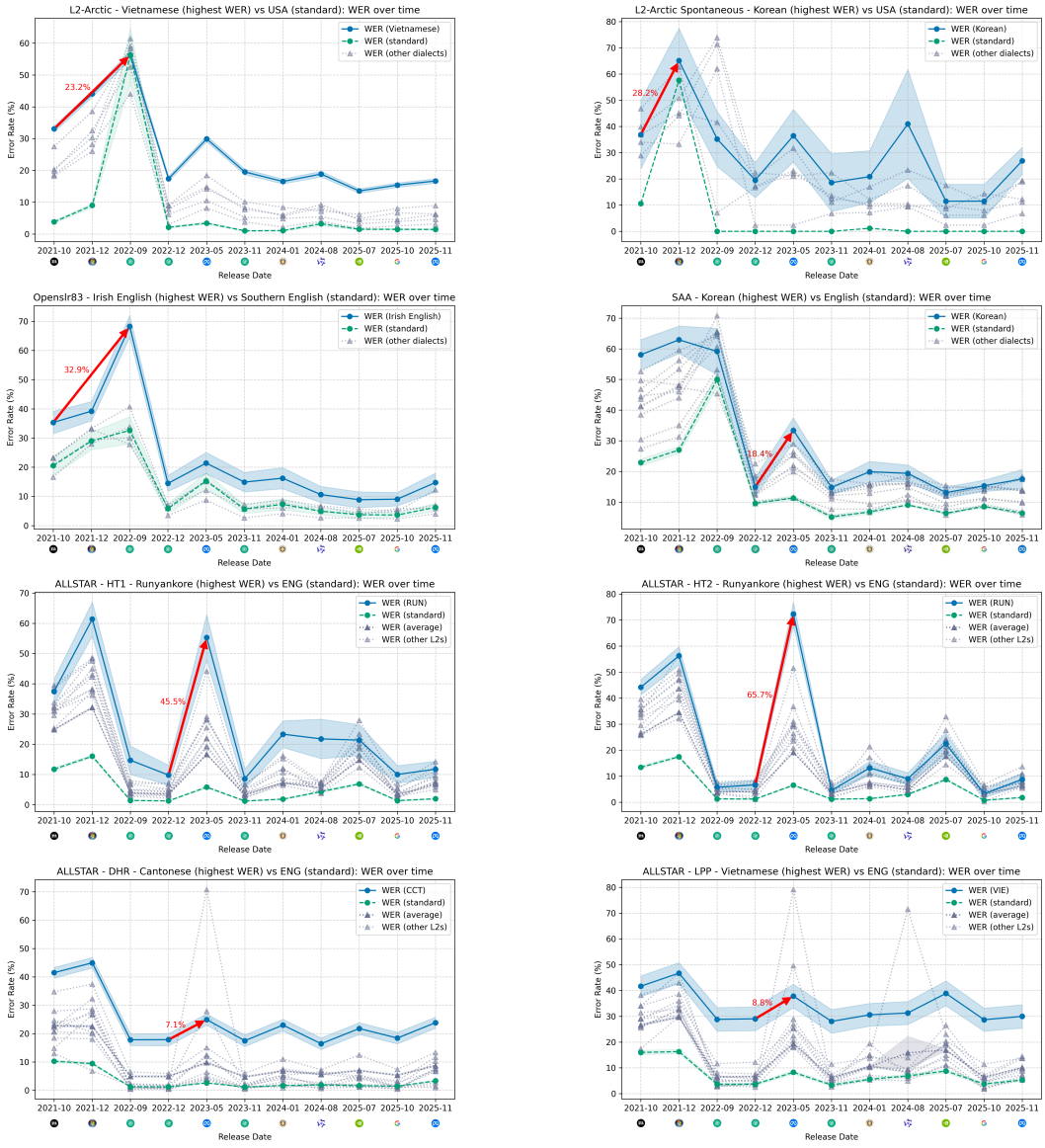

Fig. 1. Longitudinal system performance on the 8 datasets from Section 4.1. For L2-Arctic (the top row), WER for Mainstream US English (MUSE) stays mostly constant with 1-2 minor degradations that immediately revert. Meanwhile on the scripted portion, the dialect with the highest average WER, Vietnamese, achieves a maximum degradation of $d(m_1 = \texttt{Deepspeech} \rightarrow m_3 = \texttt{whisper}) = 23.2\%$ absolute WER. For L2-Arctic-Spontaneous, Korean has the highest WER and degrades frequently with a maximum of $d(m_1 = \texttt{Deepspeech} \rightarrow m_2 = \texttt{WavLM}) = 28.2\%$. In the second row, the OpenSLR83 dialects degrade in tandem following the same trends, but the degradations for Irish English are much larger than Southern UK English, peaking at $d(m_1 = \texttt{Deepspeech} \rightarrow m_3 = \texttt{whisper}) = 32.9\%$ while the Southern degradation never exceeds $\max(\mathbb{D}) = 12.1\%$. For Speech accent Archive, Korean peaks at $d(m_4 = \texttt{whisper-v2} \rightarrow m_5 = \texttt{MMS}) = 18.4\%$. In the bottom two rows (ALLSTAR), we observe very modest degradations for MUSE but some of the largest degradations for other dialects. For the HINT sentences, the Runyankore dialect peaks at $d(m_4 = \texttt{whisper-v2} \rightarrow m_5 = \texttt{MMS}) = 45.5, 65.7\%$ for the first and second half respectively. For the DHR and LPP passages, Runyankore has similar peaks but the max WER dialects, Cantonese and Vietnamese, have more modest $d(m_4 = \texttt{whisper-v2} \rightarrow m_5 = \texttt{MMS}) = 7.1, 8.8\%$ respectively.

Dataset	Avg. d($m_i \rightarrow m_j$)	Prediction (m_i)	Prediction (m_j)	Transcript
L2-Arctic	12.5%	Whisper Large v2 (2022-12): Without a doubt, some of them have dinner engagements.	MMS 1B All (2023-05): wit dou a doubt some of them hapdiner engagements	without a doubt some of them have dinner engagements
L2-Arctic Spontaneous	15.4%	Canary (2025-07): Okay, um I can see skyscrapers tall buildings and I think	Omnilingual 7B (2025-11): i can see skyscripers tall buildings and i think	ok uh i can see skyscrapers tall buildings and i think
OpenSLR-83	5.7%	Google Chirp 3 (2025-10): Your app facilitates or promotes content that includes gratuitous violence or dangerous activities.	Omnilingual 7B (2025-11): your off facilitates or promotes content that includes gratuitous flylence or dangerous activities	Your app facilitates or promotes content that includes gratuitous violence or dangerous activities
SAA	18.3%	Whisper Large v2 (2022-12): Please call Stella. Ask her to bring these things with her from the store.	MMS 1B All (2023-05): please costella escr to bring these things with her from the store	please call stella Ask her to bring these things with her from the store.
ALLSTAR-HT1	6.7%	Whisper Large v1 (2022-09): the salt shaker is empty	Canary (2025-07): the sword sheikah is empty	the salt shaker is empty
ALLSTAR-HT2	3.1%	Whisper Large v1 (2022-09): he is washing his face with soap	Omnilingual (2025-11): he is washing his face with sore	he is washing his face with soap
ALLSTAR-LPP	10.1%	Whisper Large v2 (2022-12): draw me a sheep	Canary (2025-07): turn me a say	draw me a sheep
ALLSTAR-DHR	6.1%	Whisper Large v3 (2023-11): everyone has the right to a nationality	OWSM 3.1 (2024-01): every one has to wai t t t snr t	everyone has the right to a nationality

Table 3. Illustration of degradation from earlier model m_i to later model m_j across all datasets listed in Section 4.1.

These results support our argument in Section 3, namely that evaluation benchmarks serve as robust regression testing infrastructure for standard accent varieties, but not minoritized ones.

4.3.2 Large degradations across successive systems. The degradations between successive iterations of systems for minority accent varieties can be dramatic. We highlight the largest degradation for the selected minority accent variety in Figure 1. This is especially notable with m_5 = `MMS`, where $d(m_4 \rightarrow m_5) = 18\%$ (`Speech Accent Archive`), as well as $d(m_4 \rightarrow m_5) = 45\%$ (`ALLSTAR-HT1`), $d(m_4 \rightarrow m_5) = 65\%$ (`ALLSTAR-HT2`), $d(m_4 \rightarrow m_5) = 7\%$ (`ALLSTAR-LPP`), and $d(m_4 \rightarrow m_5) = 9\%$ (`ALLSTAR-DHR`). The degradations for the MUSE variety are considerably smaller. To their credit, the MMS authors acknowledge in their technical report titled "Scaling Speech Technology to 1,000+ Languages" that their system degrades on English: "*Improvements at low resource languages result in a small degradation in some of the high-resource languages such as English...*" [72]. Yet this statement also reflects the normative treatment of English as a straightforwardly measurable construct that is commonplace in computational text and speech processing (Section 2), resulting in employing evaluation practices that obscures the starkly uneven degradations across English varieties (Section 3).

We demonstrate some of these degradations for `MMS` in the first and fourth rows of Table 3, showing that `MMS` predicts a considerable amount of nonce words (*hapdiner* instead of *have dinner*, `L2-Arctic`; *costella* instead of *call Stella*, `Speech Accent Archive`), specifically transcriptions that resemble eye-dialect and pronunciation spellings [44]. These are not harmless idiosyncracies, prior sociolinguistic work had found that "non-standard orthographies almost invariably index sociolinguistic stigma." [44]. Our results demonstrate that ASR systems can reproduce these stigmatizing non-standard orthographies in a predictable fashion, in that they are generated for non-standard English varieties but not for MUSE.

MMS is not the only model that exhibits major degradations compared to its predecessors. In Table 3, we list examples of other pairs of models, one pair for each dataset we evaluated on the minority accent we highlighted in Figure 1. For the selected pair of models m_i and m_j, we compute the average degradation across the entire dataset, listed in the second column of Table 3, and select a random example from the top quartile of the degradation distribution across the dataset. We find that the m_{11} = `Omnilingual` model, the most recent model we assessed, also degraded on some varieties: $d(m_9 \rightarrow m_{11}) = 15.4\%$ on `L2-Arctic Spontaneous`; $d(m_{10} \rightarrow m_{11}) = 5.7\%$ on `OpenSLR-83`. Similar to MMS, it produces non-standard respellings (*skyscrapers* → *skyscripers*; *violence* → *flylence*). m_9 = is recognized as the top-performing multilingual ASR model on the OpenASR leaderboard, boasting a 5% error rate. However, these error rates nearly double for Cantonese and Nkore-influenced accents in the ALLSTAR corpus ($d(m_4 \rightarrow m_9) = 6.7\%$ and $d(m_6 \rightarrow m_9) = 10.1\%$). The top 10 models on the OpenASR leaderboard do not deviate by more than a single percentage point, further emphasizing the significance of a performance degredations over 6 percentage points.

5 Discussion

Speech processing systems are fundamentally embedded in social contexts, aiming to "to supplement, reproduce, or replace human actions" [93]. As such they constitute [52] called E-class systems; socially embedded software systems where "the pressure for change is built in" [52]. Even as new improvements and capabilities are sought after, model performance may also degrade. In fact, they are more likely to degrade than traditional open-source software; as Raffel writes, "there is currently no effective approach for updating models... Instead, after being released, they are typically used as-is until a better pretrained model comes along." [76]. Wholesale replacement, rather than gradual updating and patching carries the risk of major unanticipated degradations [23], especially as major releases are deployed by "small resource-rich teams", rather than broad-participation "community-led model development". This heightened risk of major degradation is attenuated by benchmarking infrastructure, ensuring that visible dimensions of success in speech processing are measured against, more or less serving as a regression-testing suite. Yet, as we argued in this work, this infrastructure is socially positioned and deeply value-laden [42, 82]. While presented as objective and complete [77], it conceals other important axes of variation and thus, other definitions of success. Here, we focused on one such dimension, the variation among World Englishes that is well studied in sociolinguistics. But there are practically countless other such axes, including physical impairments and disfluencies [62], gender differences [47], and socioeconomic class differences [15, 49], and of course the intersections of all of these. Without changes to the infrastructure that facilitates the development of future speech processing systems, this risk will continue to be borne by marginalized speakers, in the high-stakes contexts where speech-processing systems are deployed.

We review some of these deployments in Section 5.1. Then, we discuss how organizations developing these systems should adjust their evaluation practices to reduce the risk of major performance degradations for marginalized speakers (Section 5.2), and how procurers of Speech AI services should evaluate vendors systems (Section 5.3.

5.1 Known deployments of ASR systems

ASR-based systems are already widely deployed in contexts marked by stark power imbalances, notably in workplaces [45], incarceration facilities [4, 5], and immigration bureaucracies [61]. A large fraction of companies worldwide employ speech AI systems like HireVue in their labor recruitment processes [85]. Video conferencing platforms like Zoom and Google Meet are widely integrated into workplaces, including performance management systems that are hugely consequential for workers [45]. All of these systems were reported to have significant accent-based performance

disparities [29, 85]. As our study demonstrates, these disparities should not be expected to straightforwardly decline as a function of time; speakers of some accent varieties may well be afflicted by major system degradations.

Speech AI systems are also pervasive in state bureaucracies. Several US states purchased automated conversation transcription and monitoring software from the California-based LEO Technologies firm, surveilling 300 million minutes (at the time of reporting) of conversations between inmates and their lawyer and social contacts [4]. The predictions made by these transcription systems can be life-altering, providing pretext for harsh punishments like solitary confinement [5]. Immigration offices are also known to rely on speech AI systems, assessing for proficiency in a standardized form of English. The Pearson Test of English (PTE) Core test that is accepted for some immigration programs in Canada explicitly make use of automated scoring systems [61]. Crucially, in all of these contexts, the capacity for speakers subject to these systems to contest the automatically generated transcriptions is limited. The longitudinal performance instability that we demonstrated in publicized ASR systems shows that there is reason to believe that minority English speakers are at risk of being unfairly represented by these systems.

5.2 Disaggregated Evaluations

Speech AI system developers can take steps in their evaluation practices to mitigate this risk.[2] The most straightforward amendment to current evaluation practices would be to follow practices set by the teams that trained `Canary` and the `Whisper` series of models, specifically in measuring performance on dis-aggregated multi-accent datasets (see Table 1, and Tables 2 and 3 respectively in [74, 75]). This approach is not completely foolproof – see `Canary` degradations and `whisper-v1` degradations in Figure 1. Still, this will provide greater robustness than summarizing English performance with a "single defined performance number" [28], that implicitly replicates the status-quo hierarchy of English accent varieties. In general, more extensive robustness testing should be pursued for developing speech AI systems for major institutional languages, otherwise they burden speakers of minority accent varieties with the risk of being penalized by algorithmic language management systems [58].

5.3 Procuring Speech AI-based Services

In a similar vein, procurers should require disaggregated evaluation results from speech AI system vendors. They should be skeptical of an aggregated "single defined performance number," as the computation of such individual error rates obscure the role of normative language ideologies (Section 2.2), in no small part due to the perceived objectivity of numerical knowledge [64]. Even when vendors provide some disaggregated results, it is unlikely that their evaluation will be as targeted and exhaustive as necessary for procurers. Consider, for example, HireVue's 2022 disclosure of error rates of their speech transcription system (which itself is procured from Rev.AI) [40, 85], that mentioned how there was a "12% WER for Canadian accent, and 22% WER for participants from China)". [3]

While this provides some insight into performance disparities, it is likely to be insufficiently detailed for procurers of HireVue's services, unless firms are only assessing candidates from Canada and China. Instead procurers should have their own in-house evaluation benchmarks, applying domain knowledge of the types of accent varieties that are present in the populations they work with. These bespoke evaluation benchmarks should be re-executed whenever vendors upgrade their service, to test for possible unanticipated performance degradations in accordance with our results.

[2]While it is unclear whether vendors of Speech AI services like HireVue or LEO Technologies leverage open-source pretrained models like those we listed in Table 1, the high cost of pretraining such models [76] gives us good reason to believe they would.
[3]Unlike their 2022 statement, their 2025 statement does not mention the specific error rates for other L2-English varieties [41].

6 Conclusion

A common fixture in the development process of novel ASR systems is evaluation on standard benchmarks. Among these standard benchmarks, Mainstream US English accents are always represented, often disproportionately so. This conventionalized cultural practice ensures that novel ASR systems exhibit stable performance on this prestige variety. But other World Englishes are often marginalized in the computational operationalization of English, especially those in the Global Majority [57]. We study the effects of this systemic exclusion on a longitudinal basis, investigating performance on marginalized English accent varieties on a number of notable models across a critical four-year period (2021-2025), using sociolinguistically varied speech datasets that are outside of the standard benchmarking milieu. We find that while performance can be improved on these varieties, substantial performance degradations are more likely to re-emerge for minority accent varieties than standard Mainstream US English ones. Given the widespread deployment of ASR systems in sensitive environments like immigration, diagnostics, and hiring, we call for dis-aggregated and continuous evaluations by developers and domain-specific monitoring by service providers to ensure these systems serve all speakers equitably.

7 Generative AI Usage Statement

We did not use Generative AI to prepare this manuscript.

References

[1] William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H Jiang, Ezra Awumey, and Sauvik Das. Sound check: Auditing recent audio dataset practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 26–40, 2025.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, and others. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[3] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[4] Avi Asher-Schapiro and David Sherfinski. Crack down on u.s. prison surveillance tech, rights groups urge. *Thomson Reuters Foundation*, February 2022. Updated February 10, 2022.

[5] Chelsea Barabas. Care as (re) capture: Data colonialism and race during times of crisis. *New Media & Society*, 26(12):7351–7370, 2024.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[7] Ruha Benjamin. *Race after technology: abolitionist tools for the New Jim Code*. Polity Press.

[8] Ruha Benjamin. Race after technology. In *Social Theory Re-Wired*, pages 405–415. Routledge, 2023.

[9] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, and others. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007. Publisher: Elsevier.

[10] Steven Bird. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 173–184, Seoul Republic of Korea, June 2022. ACM.

[12] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 2000.

[13] A. R. Bradlow. Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts and recordings. https://speechbox.linguistics.northwestern.edu/allsstar, n.d. Accessed: 2025-?

[14] Alexandra Canavan, David Graff, and George Zipperlen. CALLHOME American English Speech (LDC97S42), 1997. ISBN: 1-58563-111-6 Place: Philadelphia Published: Web Download.

[15] Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. Classist Tools: Social Class Correlates with Performance in NLP. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[16] May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole R Holliday. Training and typological bias in asr performance for world englishes. In *INTERSPEECH*, pages 1273–1277, 2022.

[17] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021.

[18] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. 2021. Publisher: arXiv Version Number: 5.

[19] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report, 2024. Version Number: 1.

[20] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805, Doha, Qatar, January 2023. IEEE.

[21] Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. 6(1):389–407.

[22] Jay L Cunningham, Adinawa Adjagbodjou, Jeffrey Basoah, Jainaba Jawara, Kowe Kadoma, and Aaleyah Lewis. Toward responsible asr for african american english speakers: A scoping review of bias and equity in speech technology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 665–678, 2025.

[23] Jessica Dai, Inioluwa Deborah Raji, Benjamin Recht, and Irene Y Chen. Aggregated individual reporting for post-deployment evaluation. *arXiv preprint arXiv:2506.18133*, 2025.

[24] Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jetté. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*, 2021.

[25] Remi Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

[26] Tracey M. Derwing, Murray J. Munro, Ronald I. Thomson, and Marian J. Rossiter. The relationship between l1 fluency and l2 fluency development. *Studies in Second Language Acquisition*, 31(4):533–557, 2009.

[27] Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. Global performance disparities between english-language accents in automatic speech recognition. *arXiv preprint arXiv:2208.01157*, 2022.

[28] David Donoho. Data science at the singularity. *arXiv preprint arXiv:2310.00865*, 2023.

[29] Daniel J Dubois, Nicole Holliday, Kaveh Waddell, and David Choffnes. Fair or fare? understanding automated transcription error bias in social media and videoconferencing platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 367–380, 2024.

[30] Abteen Ebrahimi and Katharina Kann. How to adapt your pretrained multilingual model to 1600 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online, August 2021. Association for Computational Linguistics.

[31] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[32] Iker Erdocia, Bettina Migge, and Britta Schneider. Language is not a data set—why overcoming ideologies of dataism is more important than ever in the age of ai. 28(5):20–25.

[33] Facebook Research. facebookresearch/voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. https://github.com/facebookresearch/voxpopuli, 2021. GitHub repository (archived).

[34] Maria Goldshtein, Jaclyn Ocumpaugh, Andrew Potter, and Rod D. Roscoe. The social consequences of language technologies and their underlying language ideologies. In *Universal Access in Human-Computer Interaction: 18th International Conference, UAHCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, Proceedings, Part I*, page 271–290, Berlin, Heidelberg, 2024. Springer-Verlag.

[35] David Golumbia. *The cultural logic of computation*. Harvard University Press, 2009.

[36] David Graff, Alexandra Canavan, and George Zipperlen. Switchboard-2 Phase I (LDC98S75), 1998. Place: Philadelphia Published: Web Download.

[37] Calbert Graham and Nathan Roll. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2024. Publisher: AIP Publishing.

[38] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics.

[39] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[40] HireVue Science Team. Ai explainability statement, April 2022. Explainability statement providing a high-level overview of AI-based assessments by HireVue, reviewed by UK ICO and published as an industry first in 2022.

[41] HireVue Science Team. Ai explainability statement, 2025. HireVue 2025 Explainability Statement (PDF) accessed via Google Drive.

[42] Ben Hutchinson, Andrew Smart, Alex Hanna, Remi Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA, 2021. Association for Computing Machinery.

[43] Samantha Jackson and Derek Denis. Same script, different sway: ethnolinguistic accent hierarchies in hiring evaluations in southern ontario. *Journal of Multilingual and Multicultural Development*, pages 1–21, 2025.

[44] Alexandra Jaffe. Introduction: Non-standard orthography and non-standard speech. *Journal of sociolinguistics*, 4(4):497–513, 2000.

[45] Mohammad Hossein Jarrahi, Gemma Newlands, Min Kyung Lee, Christine T Wolf, Eliscia Kinder, and Will Sutherland. Algorithmic management in a work context. *Big data & society*, 8(2):20539517211020332, 2021.

[46] Tyler Kendall and Charlie Farrington. The corpus of regional african american language. *Version*, 6:1, 2018.

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

[48] John Kominek and Alan Black. The cmu arctic speech databases. *SSW5-2004*, 01 2004.

[49] William Labov. *Dialect diversity in America: The politics of language change*. University of Virginia Press, 2012.

[50] Li-Fang Lai and Nicole R Holliday. Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation. In *INTERSPEECH*, volume 2023, pages 1284–1288, 2023.

[51] Halcyon M. Lawrence. Siri Disciplines. In Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip, editors, *Your Computer Is on Fire*, pages 179–198. The MIT Press.

[52] Manny M Lehman. Program evolution. *Information Processing & Management*, 20(1-2):19–36, 1984.

[53] Xiaochang Li. *Divination engines: A media history of text prediction*. PhD thesis, New York University, 2017.

[54] Linguistic Data Consortium. CABank English CallHome Corpus, 2013.

[55] Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. AboutMe: Using self-descriptions in webpages to document the effects of English pretraining data filters. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7393–7420, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[56] Nina Markl. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 521–534, Seoul Republic of Korea, June 2022. ACM.

[57] Nina Markl. Mind the data gap(s): Investigating power in speech and language datasets. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–12, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[58] Nina Markl. Algorithmic language management: How do language technologies affect linguistic practices and beliefs? *Available at SSRN 5042410*, 2024.

[59] Nina Markl and Stephen Joseph McNulty. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France, June 2022. European Language Resources Association.

[60] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France, May 2020. European Language Resources Association.

[61] Asheesh Moosapeta. Ircc now accepting pearson test of english for canadian immigration. *Pearson*, January 2024.

[62] Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[63] Joyce Nakatumba-Nabende, Sulaiman Kagumire, Caroline Kantono, and Peter Nabende. A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource african languages. *ACM Computing Surveys*, 58(4):1–24, 2025.

[64] Silvia Naydenova. Global benchmarking: contestations from a localized everyday perspective. *Globalizations*, pages 1–18, 2025.

[65] Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages, 2025. Version Number: 1.

[66] Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*, 2021.

[67] John P O'Regan. *Global English and political economy*. Routledge, 2021.

[68] Ricardo Otheguy, Ofelia García, and Wallis Reid. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. 6(3):281–307.

[69] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, April 2015. ISSN: 2379-190X.

[70] Douglas B. Paul and Janet M. Baker. The Design for the Wall Street Journal-based CSR Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[71] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-weon Jung, Soumi Maiti, and Shinji Watanabe. Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data, 2023. Version Number: 3.

[72] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling Speech Technology to 1,000+ Languages, 2023. Version Number: 1.

[73] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020.

[74] Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. Less is More: Accurate Speech Recognition & Translation without Web-Scale Data, 2024. Version Number: 1.

[75] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022. Version Number: 1.

[76] Colin Raffel. Building machine learning models like open source software. *Communications of the ACM*, 66(2):38–40, 2023.

[77] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[78] Kathy Reid and Elizabeth T. Williams. Common voice and accent choice: data contributors self-describe their spoken accents in diverse ways. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery.

[79] Anthony Rousseau, Paul Deléglise, and Yannick Estève. TED-LIUM: an automatic speech recognition dedicated corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[80] Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The edinburgh international accents of english corpus: Towards the democratization of english asr, 2023.

[81] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics.

[82] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

[83] Britta Schneider. Methodological nationalism in Linguistics. *Language Sciences*, 76:101169, November 2019.

[84] Devyani Sharma, Erez Levon, and Yang Ye. 50 years of british accent bias: Stability and lifespan change in attitudes to accents.

[85] Natalie Sheard. Algorithm-facilitated discrimination: a socio-legal study of the use by employers of artificial intelligence hiring systems. *Journal of Law and Society*, 52(2):269–291, 2025.

[86] Michael Silverstein. Monoglot "standard" in america: Standardization and metaphors of linguistic hegemony. In *The matrix of language*, pages 284–306. Routledge, 2018.

[87] Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2025.

[88] Kentaro Toyama. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs, 2015.

[89] Ravichander Vipperla, Steve Renals, and Joe Frankel. Longitudinal study of ASR performance on ageing voices. In *Interspeech 2008*, pages 2550–2553, 2008.

[90] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics.

[91] Steven Weinberger. Speech accent archive. *George Mason University*, 2015.

[92] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. Superb: Speech processing universal performance benchmark, 2021.

[93] Rua M Williams. *Disabling Intelligences: Legacies of Eugenics and How We are Wrong about AI*. Springer Nature, 2025.

[94] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, 2023. Version Number: 3.

[95] Guanlong Zhao, Evgeny Chukharev-Hudilainen, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Ricardo Gutierrez-Osuna, and John Levis. L2-arctic: A non-native english speech corpus. 2018.

A Appendix – Dataset Filtering Details

Here we provide additional details for filtering out some samples from the datasets we listed in Section 4.1, used for identifying performance degradations in Section 4.1.

Speech Accent Archive. Each recording is published as a single wave file. The expert annotated subset comes with phonetic and text transcriptions in a Praat TextGrid file. We keep only the expert annotated subset. We remove untranscribed sections of the audio labeled as “chatter”, “self-introduction”, “self-talk”, “experimenter comments”, “background speech”, “noise”, “bird song”, “cough” and any variant mispellings thereof. Furthermore, we standardize numbers to text form (“2” → “two”), fix typos, and clip audio into consistent duration segments of 50 phonemes each.

L2-Arctic. Each utterance is published as a single wave file with a corresponding TextGrid file containing phonetic and text transcriptions. Except for the expert-annotated subset, these are machine generated. We keep only the high quality, expert-annotated subsets. In this subset, each scripted utterance is about 3.6 seconds of speech. We clip the suitcase utterances to be approximately the same length. As part of the clipping process, we remove audio that hasn’t been annotated, e.g., the interviewer speaking or the subject mumbling.

OpenSLR83. We remove utterances containing fewer than 4 words. To maintain balanced representation across dialects and genders, we select 100 samples per dialect with equal gender distribution (50 female, 50 male samples per dialect). For Irish English, which contains only male speakers, we select 50 samples. This yields a balanced subset for analysis.

ALLSTAR.. Each recording is accompanied by Praat TextGrid annotations at the sentence, word, and phoneme levels. For the present analysis, full-session audio files were segmented into individual sentence-level utterances based on these annotations, and NWS paragraph recordings were excluded due to the absence of utterance-level time alignments. Instances in which the recordings did not correspond to the reference text were automatically filtered out by removing utterances for which all ASR models produced a word error rate greater than 100.