Speaking of accent: A content analysis of accent misconceptions in ASR research

Kerri Prinos
k.prinos@wustl.edu
Washington University in St. Louis
Missouri, USA

Neal Patwari
neal.patwari@gmail.com
Washington University in St. Louis
Missouri, USA

Cathleen A. Power
cassie@relationalcommunities.com
Relational Communities
St. Louis, Missouri, USA

ABSTRACT

Automatic speech recognition (ASR) researchers are working to address the differing transcription performance of ASR by accent or dialect. However, research often has a limited view of accent in ways that reproduce discrimination and limit the scope of potential solutions. In this paper we present a content analysis of 22 papers published in 2022 in top conferences and journals on the topic of accent and ASR. We report on how accent is sometimes mistakenly viewed as something some people don't have; as having a default; and being an attribute only of the speaker, and not of the listener. We discuss the implications on research and provide recommendations to researchers who hope to reduce ASR biases by accent.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Speech recognition**.

KEYWORDS

accent, discrimination, speech recognition, AI fairness

ACM Reference Format:
Kerri Prinos, Neal Patwari, and Cathleen A. Power. 2024. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3658969

1 INTRODUCTION

An accent is, loosely, a *way of speaking* a language that varies by cultural or regional group. "Accents are loose bundles of prosodic and segmental features distributed over geographic and/or social space", where prosodic includes intonation, pitch contours, and cadence, and segmental includes the pronunciation of vowels and consonants [40]. Researchers in automatic speech recognition (ASR) have demonstrated multiple ways in which ASR algorithms are biased and discriminatory, i.e., have differing performance by the speaker's accent, providing better performance for speakers of historically favored accents [12, 16, 35, 41, 45, 61, 62]. Identifying, quantifying, and addressing ASR bias is an important and large subfield of speech recognition research.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658969

However, ASR research suffers from a limited view of what "accent" is. In this paper, we show how state-of-the-art ASR research often misrepresents accent as:

(1) something that some people don't have, or that only one accent is standard; and
(2) an attribute of a speaker but not also of a listener.

These represent two ways that accent is operationalized, i.e., modeled and then used, in current speech recognition research. We argue that these two characterizations of accent a) reproduce dominant, discriminatory narratives, and b) limit the scope of solutions developed or proposed to ameliorate ASR accent bias.

Accent is related to power, and has historically been used to control and oppress groups of people [8, 41]. A speaker's accent encodes information about class, caste, race, ethnicity, regional origin, sexual orientation, national origin, and age at immigration [43]. Colonial powers forced the use of colonial language instead of local languages as a means of usurping power, maintaining hierarchies, and aiding in capitalist goals [13, 48, 49]. In the US, the English language was forced on people who were enslaved to increase the surveillance power of enslavers [15], and forced on children to eradicate the culture of indigenous groups [40]. Xenophobic and nationalistic attitudes use the idea of a "standard" English language to denigrate the accents of Latine, Black, and indigenous speakers of English [43]. While speakers of English with European accents are valorized, those with accents associated with the Global South are seen as less intelligent, loyal, and influential [38]. Gloria Anzualdúa describes holding to her language identity in spite of the forces that wanted "for all Chicano students ... to get rid of our accents" [2]. Further, it is legal in the US to be fired because of one's accent if "the accent seriously interferes with the employee's job performance" [65]. While everyone's accent limits whom they can and cannot communicate with, people with disfavored accents are the ones fired for this reason [43]. This discrimination is driven by *standard language ideology* [49], the idea that one language variety is superior.

Systems that use ASR can reinforce such discrimination in a process Nina Sun Eidsheim calls *digital aural redlining* [19]. As the late Halcyon Lawrence reported in her influential article, "Siri Disciplines", from experience as a speaker of Caribbean English trying to use a voice-based navigation system that understood her only when she imitated a white American English accent, "to create conditions where accent choice is not negotiable by the speaker is hostile; to impose an accent upon another is violent" [38]. An Apple speech-recognition product manager said in 2015 that Apple was not working on improving performance for African American speakers because "Apple products are for the premium market" [6]. While some argue that companies will naturally address ASR

biases in order to increase their market, this is not necessarily true for brands. According to Lululemon founder Chip Wilson, "the definition of a brand is that you're not everything to everybody ... you've got to be clear that you don't want certain customers coming in" [10].

We note that age, disability, and gender can also noticeably impact characteristics of speech, but are not commonly referred to as "accent". However, ageism, ableism, and sexism are clearly dimensions of privilege and oppression, and we note that ASR performance is impacted by these characteristics.

Research to address differing ASR performance by accent is thus an important avenue for future language equity, and there are several papers published on this topic every year. In this paper, we study how research literature in ASR operationalizes accent, that is, how it defines what an accent is, what part it plays in ASR, and how ASR biases as a function of accent can be minimized. We perform a content analysis of ASR bias papers appearing in 2022 in top speech recognition conferences and journals in the field. We investigate the extent to which papers in the ASR literature misconceive accent in these two ways:

- **Question 1:** *Unaccented Default*: To what extent do papers assume there is one standard accent, and that speakers of this accent are unaccented?
- **Question 2:** *Speaker-based*: To what extent is accent understood to be an attribute of the speaker, without also being recognized as an attribute of the listener?

Regarding Question 1, it should go without saying, but: "Everyone who speaks a language, speaks it with an accent" [68]. Each person conveys information about their identity in how they speak. Moreover, dialects of the English language, including the dialects some call "standard American English", "African American Vernacular English", and "Southern American English", follow standard, consistent grammatical rules. Calling one dialect "standard" or "normal" is meant to privilege one identity over others. By normalizing one accent we make it invisible [44], while others are "perpetually accentuated" or made hyper-visible [19]. Additionally, with normalization comes a privilege hazard [17], in the case of accent, the problem that some cannot identify that they have an accent, and as such, cannot as accurately model how spoken language operates.

Regarding Question 2, it is important for designers of speech recognition systems to understand that communication is bidirectional. Verbal communication involves a speaker and a listener. Engineers know that a data communication system involves at least one transmitter and one receiver, and that a demodulator must be designed for the particular modulator to achieve efficient and reliable communication [58]. In terms of language, a listener can fluently understand speech if they are experienced in listening to the accent of the speaker [60]. A person's accent listening fluency might be more expansive than the accents they can fluently speak. We improve listening fluency with practice hearing from a person speaking an accent, and we may become fluent with frequent practice even if we never speak with that accent. An accent is primarily noticed by a listener when the speaker's accent is not matched to an accent for which the listener is fluent [40]. In some ways, when people speak with an accent a privileged person does not understand, the speaker is pathologized, often with ableist language [55]. But a privileged-group listener is never pathologized for lacking the fluency to understand the speaker [55]. This contradiction points in part to our lack of consideration of listening fluency when considering spoken communication. Transcribers, as all listeners, are most adept at understanding words spoken with an accent familiar to their own. A speaker-based operationalization of accent implicitly or explicitly denies this reality.

As ASR is intended to replace a human transcriber, when we ignore the accent of the human transcriber, we ignore a way in which accent bias can seep into ASR systems.

We believe that limiting our field's view of how accent operates also necessarily limits our view of how accent bias creeps into ASR algorithms, and how that bias might be addressed. Beyond this harm to the goal of equitable ASR performance, modeling accent in ways that reinforce the structural power of dominant groups is fundamentally exclusionary. As participants in a global conversation about AI-based language research, how we talk about language and accent should acknowledge and validate the skills and expertise of all in our community. Based on the results of our content analysis, we provide recommendations to ASR researchers working to reduce accent bias in Section 4.

The purpose of this paper is not to blame authors. Even authors who understand accent bias, perhaps from personal experience, may be forced to use phrasing that reiterates dominant norms of their research community in order to make a paper acceptable for reviewers from that community. For example, explicitly stating the norm is a violation of an unstated rule [29]. Instead, this paper is motivated by the idea that naming and debating the value of a norm can enable researchers to break the norm when they decide it is necessary [4].

Our Contribution: The contribution of this paper is to perform a content analysis of recent papers published on the topic of ASR that mention accent or dialect. We identify, read and code 22 papers published in top conferences and journals on the topic published in 2022. We use the results to elucidate how the speech recognition research community operationalizes accent in their work, focused on Question 1 and Question 2. We find that the vast majority of papers do not state explicitly what they consider to be the default accent, and more than a quarter of our sample represent accent as something that some people do not have. We find almost no discussion of the accent fluency of a listener in speech communications. Finally, we describe particular implications that these misconceptions about accent might have on ASR research, and make specific recommendations we hope will aid research that intends to reduce accent biases in ASR systems.

2 METHODS

To better understand how ASR papers operationalize accent, we conducted a content analysis of relevant papers published in English in the top venues (conferences and journal publications) in the area of ASR. Our methodology is informed both by content analysis as a research method [70] and specifically by recent use to analyze the operationalization of gender in automatic gender recognition research [32].

Our goal was to answer questions about the state-of-the-art research in ASR that mention some aspect of accent. Our initial

readingindicated that papers mentioned ASR either as an acronym or in words. When describing pronunciation or manner of speech, associated with first language (L1), class, or nation or region of origin, we found most papers used the word "accent", although some referred to "dialect". We included both terms. We chose not to limit our results to ASR research performed on English, both because searching for "English" eliminates many papers which don't meet the Bender rule [4], and that we did not see a reason to exclude other languages. In summary our search criteria became:

("ASR" OR "automatic speech recognition")
AND ("accent" OR "dialect") (1)

Searching for "speech" likely excludes papers on accent within automatic sign language recognition. We describe our separate search for sign language recognition papers that mention accent of the signer in Section 5.

We intended to study research in the mainstream of ASR that addresses accent, rather than every paper published on ASR. To do this, we selected the venues (conferences and journals) which consistently publish the most papers on this topic. Speech recognition crosses disciplinary boundaries of electrical engineering (the traditional home of signal processing) and computer science, and thus both journals and conferences and both IEEE and ACM societies are important avenues for publications.

To find these mainstream venues, we searched the IEEExplore and ACM Portal for publications that match the criteria in (1), from years 2018–2022 (inclusive). We excluded books and book chapters. There were 73 results on IEEExplore and 282 on ACM Portal.

Of the total 355 records, we counted the number in each unique conference series or journal title. Of the unique conference series or journals, we eliminated any with 3 or less papers, since these have fewer than one paper per year on the topic; it is not likely these are influential or major venues for researchers on this topic. There are 12 remaining venues. From these we dropped the venues with the lowest Scimego SJR ranking. Note there is almost no difference when ranking by the "citations per document" score. The top seven venues are listed in Table 1.

There are 135 papers in these 7 venues over the 5 year period. At the time of this research (May 2023), the most recent full year of published papers available was 2022. In this year, there are 31 papers that match the search criteria in these seven top venues, which we take as the set for our analysis.

Two people read and coded each paper in the set independently, and then each paper was discussed. Codes were determined by research questions 1 and 2, as described in Section 1.

Coders read each paper, and searched within each paper for the text that matched with the search terms in (1) along with other terms often related to language and accent (such as native, L2, L1, and standard). Nine papers were excluded because they did not include any writing about speech accent or ASR. For example, one paper used "accent" to refer to the property of a musical note, not of speech. The other papers referenced other papers on accent or ASR, but did not themselves have any discussion of the topic. Some referred to accent as a topic the paper does not address, e.g., when suggesting avenues for future work.

In order to ensure that the sample of included papers was robust enough to ensure qualitative rigor, we utilized the guiding qualitative principle of saturation. We chose to go beyond "code saturation," the point at which we were not deriving any new codes or themes from the papers, to "meaning saturation," in which we had adequate information to develop a textured understanding of the meaning of the codes [28]. By the end of our data set we were not adding any more new themes to our content analysis, nor were new papers extending our understanding of the meaning of those codes.

In the end, 22 papers were included in the content analysis: [1, 18, 20, 22, 23, 30, 31, 33, 34, 36, 37, 39, 42, 50, 52, 56, 57, 59, 67, 69, 71, 72].

3 RESULTS

To provide further context about the 22 papers selected for content analysis, we first describe the speech technology on which the paper focuses. Several (8/22) papers were focused directly on improving ASR algorithms. Two papers were on improving speech recognition but with the aid of additional sensors beyond speech, specifically, gaze [33] and jaw motion [59]. Five papers focused on a speech-based technology similar in many ways to ASR but with a different output (e.g., speech pathology diagnosis, acoustic-to-articulatory mapping, and text-to-speech), and seven papers were dedicated to applications that involve ASR as a component (e.g., conversational agents (CAs), dialog systems, mobile apps that use ASR, YouTube captioning). All of the papers discussed accent but to different extents. Six papers focused on the robustness of ASR to different accents.

For more context on the papers in the analysis, we describe their language(s) of study. The 22 papers in the analysis overwhelmingly studied or tested on the English language (16 papers). In addition several other languages were studied, including two papers each studying Mandarin and Cantonese, and one paper each on isiXhosa, Marathi, Mboshi, Javanese, and French. One paper [20] was a review article and thus indirectly covered research on multiple languages. We note that multiple papers did not clearly state the language of study, as will be covered in Section 3.2, but we could determine the language of most with some detective work. (For example, [67] discusses the ASR mistaking "male" for "mail" or "Mel", from which we infer its use of English.) For one paper [50] we were unable to infer the language used in the research. The total is higher than 22 because three papers study more than one language.

3.1 How is accent categorized?

It is informative to describe how the papers in our study have categorized accent, which we summarize in Table 2. No paper explicitly defined accent or dialect. Thus we first analyzed how each paper divides speakers by accent.

The largest group of papers (10/22) categorizes accent as a function of the nationality or region of the speaker. Another group of papers (4/22) categorizes accent by the first language (L1) of the speaker, which may be the same or different from the language of the ASR studied in the paper. For example, a speaker's accent is said to be due to their first language of Hindi [33]. Note that this is different from the nationality-based categorization (e.g., "Indian English" [37]) because a nation may contain speakers of many different L1 languages.

Top 7 venue names with more than 3 papers	Paper Count	SJR	Cites/Doc (2 year)
ACM Comput. Surv.	6	4.457	20.26
IEEE/ACM Trans. Audio, Speech and Lang. Proc.	80	1.348	5.98
Proc. of the CHI Conf. on Human Factors in Computing Systems	16	0.714	5.16
Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.	9	1.202	4.96
Proc. ACM Hum.-Comput. Interact.	5	0.715	3.74
ICASSP - IEEE Intl. Conf. Acoustics, Speech and Signal Proc. (ICASSP)	12	0.997	3.606
IEEE Automatic Speech Recognition and Understanding Wksp. (ASRU)	7	0.757	3.319

Table 1: Selected conference and journal venues on ASR and accent, based on number of papers from 2018-2022, and SJR rank (Scimego)

Categorization of Accent	Description	Literature Example	Paper Count
Geographical	Based on person's region or nationality of residence	"8 major dialect regions of the United States" [57]; "Indian English" and "US English" in [37]	10
Native binary	Either "native" or "non-native" speaker of the language	"it is therefore suggested that the voice prompts could be recorded slowly but clearly and preferably by native speakers" [67]	8
First language	Category for each possible first / L1 / native language of the person	"Participants in the dataset were diverse in terms of their native languages (3 Arabic, 3 Tamil, 2 Mandarin, 4 English, 2 Urdu, 2 Bengali, 7 Persian, 4 Sinhala, 2 Yoruba, 1 Bahasa speaker)" [33]	4
Race	Based on the race of the speaker	"improving the performance … for the understudied dialects of Southern American English and [African American English] AAE in children" [31]	2
None stated	No categorization given		2

Table 2: How analyzed papers categorized accent

More than one-third of the papers (8/22) provide a dichotomy of accent as being either "native" or "non-native". Few actually named the "non-native" category, but by naming one set of speakers as "native", and not naming the other speakers, the authors create a binary classification. One paper named "non-native" speech and did not name native speech. While a native/non-native binary might also be seen as a categorization by first language, it can be used to further specify speech corresponding to one accent within the language. For example, [57] describes "native American English speakers", which we assume refers to native speakers of American English, rather than Native American speakers of English.

Only two papers refer to speaker accent by racial group. Both name only African African English. Johnson et al. refers to "less standard dialects like Southern American English or African American English (AAE)" [31]. Garg et al. [20], referring to [53], states "They identified how CAs can help children who speak African American Vernacular English (AAVE) at home to code-switch between school-ratified English and AAVE in a school setting." Notably, in both of these cases, racial group name is only used to name an accent when the speakers are Black; other accents are not named as being used by speakers who are predominantly white. For example, the term "Southern American English" refers to an accent spoken primarily by white speakers in the Southern US. Although the name "Southern American English" implies it is distinguished purely by region, a history of discrimination, stigmatization, and resistance has led white and Black speakers in the Southern US to have distinct accents [25, 63]. Similarly, "school-ratified English", in the context used by [20], refers to the accent of white American English speakers. By not naming the race of the contrasting accent, it partially obscures the racial biases against speakers of African

American English in US education, including teacher biases and debunked language deficit models [26].

Two papers had no categorization of accent; that is, accent was referred to as a source of variation, or that there is a diversity of accents, but no further description of the accents in their work was given. Note that the total count in Table 2 is higher than the total number of papers (22) because some papers categorized accent in more than one way.

3.2 Standard or Non-existent Accent

For the papers in our set, is there, explicitly or implicitly, a *standard* accent? Is accent something that *some people don't have*? We address these questions by evaluating the context in which the accent of the speakers is named.

Some papers state specifically, in their categorization of accent, that some speakers do not have an accent. For example, participants are asked to rate the "accentedness" of speech samples on the scale, "1 = no accent/negligible accent, 2 = mild accent, 3 = strong accent, and 4 = very strong accent" [42]. In this case, it is explicit that there exists speech spoken with "no accent". Others state a dichotomy between "accented" speech and another category, which is named otherwise. For example, "data for new domains (e.g., data for accented English) is usually smaller than initial domains (e.g., data for native English)" [23]. In this case, one can presume that the "native English" category is not considered to be accented. Another example is in [50], which first classifies speech as "accent-unspecific" or "accented". In total, 6 of 22 papers (27%) describe one group of speakers as not having an accent.

Describing one group as being without an accent is a way to make the default accent invisible. Twenty of the papers in our set define, either implicitly or explicitly, one accent as the default or standard accent. However, only 3/20 state the standard explicitly. For example, [1] states that they use a dataset of Parisian French as considered the standard: "exhibiting a major Parisian accent for the BREF corpus (the closest to standardized French)", and they compare the features of a southwestern accent dataset in comparison to this standard. Similarly, Johnson et al. [31] states the standard accent in justifying which datasets to use in training and validation: "We use the training set containing California English because it is considered a widely-spoken American dialect. Adapting the California English training set to the Georgia English validation set then represents adapting from a more standard dialect to the less standard dialect as in low-resource scenarios."

In contrast, 17/20 papers that use a standard accent in their work do not state what the standard is, and the reader must infer it. In multiple cases, the paper uses one dataset for training the ASR model, and then tests performance when augmenting with, or simply testing on, another dataset with greater accent diversity [36, 50]. In this case, if the default accent is not stated, the accent of speech in the training set is the default. As another example, in their study on English-language speech, Khan et al. initially describes accent in terms of participants' "native languages", including English as one of the many native languages [33]. However, when explaining why the ASR error rate is so high (64%), the authors fault the diverse participants, stating that the ASR performance is "highly dependent on the accent of the speaker". This implies results would be better if participants spoke the accent for which the ASR algorithm performs best, but does not name this default accent [33]. Similarly, other papers describe accent as a problem, like noise, which causes ASR performance to decrease: "Furthermore, background noise, multi-talker speech, human accent, and disfluent speech may further downgrade the quality of automatic captions" [39]. Since "downgrade" must be with respect to another condition, this statement positions one accent as the default (along with noise-free, single-talker, and fluent speech). One paper explicitly calls some speech "normal speech", as in, "normal speech recorded from healthy, non-aged users" [22], but does not describe the accent of the speakers who produce "normal speech".

Leaving the accent implicit is reminiscent of papers in natural language processing (NLP) which leave the language of study implicit. Frustrated by the NLP research misconception that work on English is not language-specific, and that there is no need to name the language studied if it is English, Emily Bender proposed the #BenderRule, which can be succinctly summed up in a sentence: "Always name the language(s) you're working on" [4, 5]. Concerning the Bender Rule, of the papers in our analyses, 5/22 (23%) never named the language(s) they worked on. Four other papers don't state the language of the study but it can be inferred from text in the paper (e.g., listing "presented in a non-English language and did not have English caption" as an exclusion criteria [39]).

3.3 Is accent acknowledged to be an attribute of a listener?

As a reminder of the purpose of this question, most ASR papers use a speech corpus or dataset which contains at least partial human-generated transcriptions of the speech samples. The accent(s) for which the transcriber(s) are fluent impact how accurately they will be able to transcribe the speech.

3.3.1 Is the listener ever acknowledged at all? Thirteen of the 22 papers do not ever mention a listener to the speech in any way. In other words, although these papers involved words being spoken, there was no mention of it being listened to by a transcriber, participant, or anyone else.

Of the 9/22 papers which do mention a person listening to the speaker, many do so in the context of an application of speech recognition, for example, research on conversational agents (CAs). Papers on CAs in our set [18, 20] and one on automatic administration of voice surveys [67] describe a person both speaking and listening as part of the operation of the system of study. Two papers are focused on the detection or quantification of speech pathology [1] or articulation errors [42] and thus describe people (speech language pathologists or study participants) who listen to and rate speakers. A paper involving text-to-speech algorithms describes the study participants who rate the quality of the generated audio [36].

3.3.2 When the listener is acknowledged, is their accent fluency described? Six of the 9 papers described above as acknowledging a listener do not describe the accent of any listener. Some of these six do describe extensive details of a data set they use, or the participants they enroll to rate some aspect of speech. For example Kumar

et al. states about their participant-based judging of artificially generated speech, "We've recruited twenty judgers between the ages of 20 and 40. We selected English speakers who successfully completed a brief transcribing exam" [36]. While this acknowledges the language of the judgers, it does not mention the accent, even though the transcribing exam would presumably contain English speech from one or more particular accents. Presumably, a judger's fluent accent(s) would impact how they might rate the "naturalness" of artificially generated speech "of speakers with various accents" [36].

Of the three papers which do acknowledge the accent of a listener, two address the accent of people who listen, in general, but not any listener involved in their study. One of these papers discusses the accent of a listener in the context of sending interactive voice response (IVR) survey questions from a smart speaker to a participant [67]. It states "Previous studies on IVR surveys found that respondents tended to emulate the speaking styles of the voice, it is therefore suggested that the voice prompts could be recorded slowly but clearly and preferably by native speakers". In other words, since ASR is biased against second language (L2) speakers, people should listen to questions read by L1 speakers of the language, so that they are cued to use or imitate an L1 accent while responding. Another paper discusses how an ASR adaptation should work via analogy to human language learning: "Experiences change perception. For example, infants in different countries who are born with similar auditory organs can differentiate phoneme contrasts across languages; their perception is changed to bias their mother tongue after they have more listening experiences" [69]. In sum, neither of these two papers discuss the accent of the listeners who are used in evaluation of the proposed system.

Only one paper [42] describes the accent of any person who participates in labeling of speech data. The participants who rate the "accentedness" of speech samples on a scale from 1 to 4, are described as "thirteen native American speakers were recruited as annotators",[1] which does partially specify the accent fluency of the participants. However, in another part, the same paper uses speech pathologists to quantify articulation errors in US English speech samples, but the language or accent of the pathologists is not named.

In summary, it was extremely rare in our studied papers to describe or even acknowledge the accent fluency of any person transcribing or rating speech samples from the study.

4 DISCUSSION

As a field, how we model accent limits the research we value and the solutions we explore to improve the performance and robustness of ASR for all speakers of a language. Here, we share implications of the findings related to the research questions, and offer recommendations to encourage research that we believe would lead to improvement toward equitable performance across accents.

[1]Although a language is technically not named in this statement, we have assumed the authors intended to refer to native speakers of American English.

4.1 Implications of Question 1

Emily Bender describes, in developing what became the #BenderRule, how naming the language of study as being critical to understanding the specificity of a paper's research contribution. In particular, results for NLP research tested on English may or may not generalize to every other language. But the unstated assumption that results in English do generalize contributes to devaluing NLP research on other languages [4].

These concerns are also present in ASR research, when considering the accent that is studied. There is an unstated (and untested) assumption that ASR systems trained using "standard" American English accented speech will then generalize when trained using any speech accent. When we don't state the specific accent used in a study, we make this assumption invisible. And moreover, we devalue research on less studied accents, which is not necessary, given the unstated assumption.

> **Recommendation 1**: Emily Bender describes her rule as "the bare minimum" [4]. Given that researchers also need to consider variations within languages, in particular, accents, we extend the Bender rule: always name the accent(s) you're working on. Name the accent, even if it is white, non-immigrant, middle-to-upper class, Midwestern, US English. Or, as is common in our paper set, if using a data set with speakers of unknown cultural and geographic characteristics, explicitly state that. Only three papers in our analysis explicitly named the accent considered to be the standard; 17 others did not name the accent considered the norm. Naming the accent(s) being used in a paper's results is one step towards making assumptions about accent visible.
>
> Our recommendation to researchers extends the framework proposed by psychologist Elizabeth Cole, who outlined questions that human subjects researchers should answer in describing their research [14]. ASR researchers using or creating a speech corpus should consider intersectional identities and answer, who gets to be included in this corpus? Cole's question reminds us to identify intersectional identities while following the Bender accent rule. For example, African American English also has regional variation, just as white American English is described as having.
>
> This recommendation echoes the push from AI scholars to document datasets with datasheets, including details on what subset of the population is included, and how the data was labeled [21], so that downstream uses of the dataset are aware of its domain and limitations.

4.2 Implications of Question 2

Data from our analysis related to Question 2 does not give us confidence that ASR researchers are considering the accent fluency of transcribers and other people involved in listening to the speech samples in a dataset. As Joy Buolamwini says, "those with the power to build AI systems do not have a monopoly on truth" [11]. When researchers ignore that people listen with an accent, they ignore the fact that a transcriber with less familiarity with a person's accent will make more errors during transcription of their speech. If the transcribers are less likely than the people in the corpus to be fluent

in one accent, there will be more transcription errors for that group of people, and ASR algorithms will be trained to make more errors on their speech.

When researchers ignore this source of ASR accent bias, they may jump to a conclusion that accent bias is purely a function of under-representation of a group's speakers in the dataset. For example, in their highly cited work quantifying significant ASR bias against Black speakers of English, Koenecke et al. concludes "The likely cause of this shortcoming is insufficient audio data from black speakers when training the models" [35]. In a similar vein, after the Washington Post published an extensive evaluation of the accent biases of Amazon Alexa and Google Home [27], Amazon responded with a statement that said, "As more people speak to Alexa, and with various accents, Alexa's understanding will improve." [27].[2] But it is not at all clear that more data from Alexa users, on its own, will solve the problem. Beyond potential disparities in transcription performance discussed above, there is also a positive feedback mechanism. Deploying a product that works significantly worse on one accent will result in fewer speakers of that accent buying or using it. As a result, "Alexa's understanding" will not improve for that accent as fast as for the favored accent.

Three papers in our analysis involved the development of speech-based automatic systems to quantify speech or health pathologies, such as speech disorders caused by surgery for head or neck cancer [1], neurocognitive disorders [18], or dysarthritic speech [22]. None of these papers described the accent of the experts used to provide the ground truth diagnosis for each patient. We should consider whether the expert's accent fluency impacts the accuracy of their speech sample-based diagnosis, to know whether labelling is a mechanism which could bias performance by demographic group. This is another example of ASR research which should consider intersectional identities as it involves disability and accent.

> **Recommendation 2**: Researchers creating speech datasets should ensure that transcribers are fluent in the accent of the speakers whose speech they transcribe. Similarly, researchers developing speech-based health diagnostics should ensure that an expert providing a ground truth label from a speech sample is fluent in the speech accent so that they do not confuse accent features with pathology. Transcription projects like Mozilla's Common Voice[3] could simply survey participants (speakers, listeners) about their accent fluency, and match listener and speaker by fluency.

4.3 Broader Implications

The first two recommendations are stated without engaging with how they impact power and privilege. However, we should acknowledge that ignoring accent, provides power to the group with the default accent. Six papers in our analysis use an accent model that positions some people as not having an accent. When a person in a dominant group can ignore the fact that they have an accent, it makes their privilege invisible [44]. The result of invisible privilege is the feeling that their group's dominance is just natural. In the ASR case, the better performance of ASR on their group can be attributed to the idea that their speech is "unaccented". Further, the invisible privilege means that research on ASR systems focused on speakers with a different accent is in a completely different research category compared to research on ASR systems focused on speakers with the dominant accent. The former is "diversity work" that is then devalued [7].

Simultaneously, researchers who speak with an accent considered standard cannot as easily see the problems with their mental model of accent, a process referred to as privilege hazard [17], and thus can't as easily fix the problems with ASR accent bias. In our analysis, one paper's method to adapt an ASR model to accent assumed that some speech is "accent-unspecific" and other speech is "accented"; in the latter group, accented speech can either be an "unseen accent" (by the ASR model) or a "seen accent" [52]. The system classifies speech samples into one of these three categories; that may be unnecessary, given that all speech is accented. People who speak with an accent that is not considered standard are less subject to privilege hazard [17] and thus have considerable expertise to offer our field regarding accent-equitable ASR development. However, inequity, coupled with messaging that "the current status quo … is not only acceptable, but also unproblematic", negatively impacts recruitment and retention of minoritized scholars to the field [66].

Unfortunately, if research cannot resolve ASR biases by accent, then it will be more likely to recreate existing discrimination in new technologies in education [8], housing [3], health care [64], employment [6], and others.

> **Recommendation 3**: We extend a suggested question from [14] and suggest that accent, and the cultural and regional identities they represent, are not neutral categories devoid of privilege and power. Authors should answer the question: What role does inequity play in the performance disparities and in the application of the proposed system? For example, we should not give the impression that developing systems to teach children to switch dialects in order to do better in school is an apolitical goal [26]. As another example, we should discuss the racist history of blackface [47] and yellowface [46] when considering algorithms to alter the accent of recorded speech to a different nationality, gender, or race. Further, while we may want users to speak with the accent an ASR performs best on so that our system performs more reliably, forcing users to mimic the privileged group's accent is a form of violence [38]. By naming the problems in ASR, we can avoid giving the impression given that maintaining an oppressive status quo is acceptable [66].

Answering these questions do take time and thought. This reflection and metadata benefits transparency for the reader, but also benefits the researcher, in a similar vein to that described by Gebru et al. in [21].

5 LIMITATIONS AND CONCLUSIONS

The content analysis presented answers questions about how researchers model accent in their ASR research. However, we note several opportunities to extend our analysis that we believe would be fruitful for a broader understanding of ASR disparities. While

[2] Technically, the Washington Post evaluation and Amazon's response appeared simultaneously in the same article. Coincidentally Jeffrey P. Bezos owns them both.
[3] https://commonvoice.mozilla.org/

the papers from 2022 met the qualitative standard of meaning saturation, future work could look longitudinally for changes over time, and/or provide updates about the current state-of-the-art.

Our search terms include "accent" and "dialect", but as we mention, speech characteristics that are a function of age, disability, gender are not considered to be included in these terms. A broader set of terms could pull in other papers on ASR, and answer questions about how the field operationalizes the effect of age, disability, gender, sex, and sexual orientation on speech.

Our search terms excluded papers on sign language recognition that discuss accent. We performed an extensive search, but found only one matching paper in all of 2022 [9]. As our paper is limited to spoken languages, we have failed to include signed languages in the present study. Future work may need to widen the search terms, or include more years, to increase the number of papers in the sample.

The presented results from our content analysis reveal the extent to which researchers in ASR who published in top conferences and journals in the area align with misconceptions of accent: that there is a default accent; that some people don't even have an accent; and that accent is only a function of the speaker, not the listener. We offer several implications of these misconceptions on ASR research and development, and offer recommendations to researchers in the field. We are particularly motivated to change norms that we believe are helping to keep ASR inequitable by accent.

POSITIONALITY STATEMENT

How we are positioned socially and culturally shapes our understanding of the world around us [24]. This holds for engineers and computer scientists as it does for any person [11]. Our specific social locations have shaped our interest in, and our understanding of, the topic discussed in this paper. The authors have had experiences with marginalization that have sensitized us to how accent is connected to systems of power. However, our positionalities also create privilege hazards [17] that limit our understanding of this topic. In short, our identity and experiences shape how we write about accent, as well as how we read and analyze papers written by others.

Kerri Prinos is an electrical engineering graduate student with a liberal arts background in biology and applied math. She is a white American from New England. She speaks with an accent that is called white American English, and she is a fluent listener of English speakers whose native language is Ukrainian and Russian. Her fiancé is from Kyiv, Ukraine and is a native Ukrainian and Russian speaker and fluent English speaker. Her family's connection to the Italian, Greek, and Lithuanian languages has been lost through assimilation. Her great-grandparents, who spoke little English, were determined that their children would be "American" and speak only English even if it meant they would lose their heritage.

Neal Patwari is an electrical engineering and computer science professor. He is a U.S. Midwest-born second generation Indian immigrant. He speaks with an accent we describe as middle-class white American English, and is also a fluent listener, from experience, of English speakers whose first language is Hindi and Gujarati. His experience taught him that Indian English speech follows rules and is structured, not error-prone, noisy or abnormal (as some ASR papers imply). However, Neal was raised in an English-only household, as the most reputable American pediatricians of the time convinced his immigrant parents of the xenophobic myth that speaking their first language would confuse their children.

Cathleen A. Power holds a PhD in social psychology and gender studies. She is white American from the Mountain West. She speaks with an accent that is called white American English. Cathleen grew up straddling social class positions, and thus, learned early that valued speech norms are connected to systems of power, and thus are not neutral [51]. Her speech has, at times, been deemed "unprofessional" by middle-class American standards, though her accent is not othered because of her race and/or nationality.

The authors are monolingual, which limits our understanding of the experience of multilingual users of ASR services. Our paper's inclusion only of papers published in English is one negative impact. Further, our fluency in white middle-class American English, as reflected in our writing, provides an unearned and artificial advantage (i.e., privilege) in peer-review [54]. It is important for those of us whose accents are privileged to recognize that reading and hearing other varieties of language expands our fluency and thus should be valued, for example, in peer review. We urge other researchers to expand our understanding of speech accent and dialect in more expansive ways than we have here.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their constructive feedback. We also wish to acknowledge Os Keyes for their paper [32], which both provided an example and demonstrated to us the power of content analysis for advancing our research area.

No funding source supported this work.

REFERENCES

[1] Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. 2022. Interpreting Deep Representations of Phonetic Features via Neuro-Based Concept Detector: Application to Speech Disorders Due to Head and Neck Cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 200–214.

[2] Gloria Anzaldúa. 1987. *Borderlands/la frontera: The new mestiza.* Aunt Lute Books, Chapter How to Tame a Wild Tongue, 75–86.

[3] John Baugh. 2018. *Linguistics in pursuit of justice.* Cambridge University Press.

[4] Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient* 14 (2019).

[5] Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6 (2011), 1–26.

[6] Ruha Benjamin. 2023. *Race after technology* (1 ed.). Polity.

[7] Mary Blair-Loy and Erin A Cech. 2022. *Misconceiving merit: Paradoxes of excellence and devotion in academic science and engineering.* University of Chicago Press.

[8] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050 [cs.CL]

[9] Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring Collection of Sign Language Videos through Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[10] Curtis Bunn. 2024. Lululemon founder's remarks have some DEI experts calling for boycotts to combat 'regressive values'. *NBC News* (6 Jan. 2024).

[11] Joy Buolamwini. 2023. *Unmasking AI: my mission to protect what is human in a world of machines.* Random House.

[12] May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech 2022.* ISCA, 1273–1277. https://doi.org/10.21437/Interspeech.2022-10869

[13] Shefali Chandra. 2012. *The sexual life of English: Languages of caste and desire in colonial India.* Duke University Press.

[14] Elizabeth R Cole. 2009. Intersectionality and research in psychology. *American Psychologist* 64, 3 (2009), 170–180.

[15] Kumari Devarajan. 2018. Ready For A Linguistic Controversy? Say 'Mmhmm'. https://www.npr.org/sections/codeswitch/2018/08/17/606002607/ready-for-a-linguistic-controversy-say-mhmm. *Code Switch* (17 Aug. 2018).

[16] Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Performance disparities between accents in automatic speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. https://arxiv.org/abs/2208.01157

[17] Catherine D'Ignazio and Lauren F Klein. 2023. *Data feminism.* MIT Press.

[18] Zijian Ding, Jiawen Kang, Tinky Oi Ting Ho, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: a conversational agent using backchannels to engage older adults in neurocognitive disorders screening. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[19] Nina Sun Eidsheim. 2023. Rewriting algorithms for just recognition. In *Thinking With an Accent*, Pooja Rangan, Akshya Saxena, Ragini Tharoor Srinivasan, and Pavitra Sundar (Eds.). University of California Press, 134–150.

[20] Radhika Garg, Hua Cui, Spencer Seligson, Bo Zhang, Martin Porcheron, Leigh Clark, Benjamin R Cowan, and Erin Beneteau. 2022. The last decade of HCI research on children and voice-based conversational agents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[22] Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. 2022. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2597–2611.

[23] Shahram Ghorbani and John HL Hansen. 2022. Domain Expansion for End-to-End Speech Recognition: Applications for Accent/Dialect Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 762–774.

[24] Sandra Harding. 2001. Feminist standpoint epistemology. *The gender and science reader* (2001), 145–168.

[25] Michael Harriot. 2023. *Black AF History: The Un-Whitewashed Story of America.* Dey Street Books.

[26] Yvette R Harris and Valarie M Schroeder. 2013. Language deficits or differences: What we know about African American Vernacular English in the 21st century. *International Education Studies* 6, 4 (2013), 194–204.

[27] Drew Harwell. 2018. The accent gap: We tested Amazon's Alexa and Google's Home to see how people with accents are getting left behind in the smart speaker revolution. *The Washington Post* (18 July 2018).

[28] Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research* 27, 4 (2017), 591–608.

[29] Lauren N Irwin. 2022. White Normativity: Tracing Historical and Contemporary (Re)Productions of Whiteness in Higher Education. In *Critical Whiteness Praxis in Higher Education*. 48–69.

[30] Xiaofu Jin, Xiaozhu Hu, Xiaoying Wei, and Mingming Fan. 2022. Synapse: Interactive Guidance by Demonstration with Trial-and-Error Support for Older Adults to Use Smartphone Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–24.

[31] Alexander Johnson, Ruchao Fan, Robin Morris, and Abeer Alwan. 2022. LPC Augment: an LPC-based ASR Data Augmentation Algorithm for Low and Zero-Resource Children's Dialects. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8577–8581.

[32] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 1–22.

[33] Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[34] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. Mymove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[35] 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.

[36] Neeraj Kumar, Ankur Narang, and Brejesh Lall. 2022. Zero-Shot Normalization Driven Multi-Speaker Text to Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1679–1693.

[37] Pratik Kumar, Vrunda N Sukhadia, and S Umesh. 2022. Investigation of Robustness of Hubert Features from Different Layers to Domain, Accent and Language Variations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6887–6891.

[38] Halcyon M. Lawrence. 2021. Siri disciplines. In *Your Computer Is On Fire*, Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip (Eds.). MIT Press, 179–198. https://mitpress.mit.edu/books/your-computer-fire

[39] Franklin Mingzhe Li, Cheng Lu, Zhicong Lu, Patrick Carrington, and Khai N Truong. 2022. An exploration of captioning practices and challenges of individual content creators on YouTube for people with hearing impairments. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[40] Rosina Lippi-Green. 1997. *English with an accent* (1 ed.). Routledge.

[41] Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. 521–534.

[42] Vikram C Mathad, Julie M Liss, Kathy Chapman, Nancy Scherer, and Visar Berisha. 2022. Consonant-vowel transition models based on deep learning for objective evaluation of articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 86–95.

[43] Mari J Matsuda. 1991. Voices of America: Accent, antidiscrimination law, and a jurisprudence for the last reconstruction. *Yale Law Journal* (1991), 1329–1407.

[44] Peggy McIntosh. 1989. White privilege: Unpacking the invisible knapsack. *Peace and Freedom Magazine* (July/Aug 1989), 10–12.

[45] 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*. European Language Resources Association, Marseille, France, 6462–6468. https://aclanthology.org/2020.lrec-1.796

[46] Krystyn R Moon. 2005. *Yellowface: creating the Chinese in American popular music and performance, 1850s-1920s.* Rutgers University Press.

[47] Wesley Morris. 2021. Music. In *The 1619 Project*, Nikole Hannah-Jones, Caitlin Roper, Ilena Silverman, and Jake Silverstein (Eds.). One World, New York, Chapter 14, 358–379.

[48] Veena Naregal. 2001. *Language Politics, Elites and the Public Sphere: Western India under Colonialism.* Permanent Black, New Delhi, Delhi.

[49] Julia Nee, Genevieve Macfarlane Smith, Alicia Sheares, and Ishita Rustagi. 2022. Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society* 9, 1 (2022), 20539517221090930.

[50] Dino Oglic, Zoran Cvetkovic, Peter Sollich, Steve Renals, and Bin Yu. 2022. Towards Robust Waveform-Based Acoustic Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1977–1992.

[51] Cathleen A. Power. 2023. Just Hit Me Already: Obscured Workplace Abuse and Discrimination. *ADVANCE Journal* 4, 1 (2023).

[52] Yanmin Qian, Xun Gong, and Houjun Huang. 2022. Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2842–2853.

[53] Emilee Rader, Margaret Echelbarger, and Justine Cassell. 2011. Brick by brick: iterating interventions to bridge the achievement gap with virtual peers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2971–2974.

[54] Valeria Ramírez-Castañeda. 2020. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. *PloS one* 15, 9 (2020), e0238372.

[55] Pooja Rangan. 2023. From "Handicap" to Crip Curb Cut: Thinking Accent with Disability. In *Thinking With an Accent*, Pooja Rangan, Akshya Saxena, Ragini Tharoor Srinivasan, and Pavitra Sundar (Eds.). University of California Press, 54–72.

[56] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[57] Abdolreza Sabzi Shahrebabaki, Giampiero Salvi, Torbjørn Svendsen, and Sabato Marco Siniscalchi. 2021. Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 135–147.

[58] Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

[59] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[60] Pavitra Sundar. 2023. Listening with an Accent – or How to Loeribari. In *Thinking With an Accent*, Pooja Rangan, Akshya Saxena, Ragini Tharoor Srinivasan, and Pavitra Sundar (Eds.). University of California Press.

[61] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59. https://doi.org/10.18653/v1/W17-1606

[62] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*. 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[63] Erik R Thomas. 2004. *Rural white Southern accents*. Mouton de Gruyter Berlin), 300–324.

[64] Martin J. Tobin and Amal Jubran. 2022. Pulse oximetry, racial bias and statistical bias. *Annals of Intensive Care* 12, 1 (Jan 2022), 1–2.

[65] US Equal Employment Opportunity Commission (EEOC). [n. d.]. National Origin Discrimination. https://www.eeoc.gov/national-origin-discrimination. Accessed: 5 Dec. 2023.

[66] Alicia Nicki Washington. 2020. When twice as good isn't enough: The case for cultural competence in computing. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 213–219.

[67] Jing Wei, Weiwei Jiang, Chaofan Wang, Difeng Yu, Jorge Goncalves, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding How to Administer Voice Surveys through Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[68] Steven Weinberger. 2015. Speech accent archive. https://accent.gmu.edu/about.php. George Mason University.

[69] Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura. 2022. Modeling unsupervised empirical adaptation by DPGMM and DPGMM-RNN hybrid model to extract perceptual features for low-resource ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 901–916.

[70] Yan Zhang and Barbara M. Wildemuth. 2009. Qualitative Analysis of Content. In *Applications of Social Research Methods to Questions in Information and Library Science*, B. Wildemuth (Ed.). 308–319.

[71] Donghui Zhu and Ning Chen. 2022. Multi-Source Domain Adaptation and Fusion for Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2103–2116.

[72] Asier López Zorrilla, María Inés Torres, and Heriberto Cuayáhuitl. 2022. Audio Embedding-Aware Dialogue Policy Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 525–538.