Modeling Bias in Automatic Speech Recognition

Camille Harris
Georgia Institute of Technology
charris320@gatech.edu

Chijioke Chinaza Mgbahurike
Stanford University
cmgbahur@stanford.edu

Diyi Yang
Stanford University
diyiy@stanford.edu

Abstract

Dialect and gender-based biases have become an area of concern in automatic speech recognition (ASR). In this work, we aim to benchmark the performance of ASR systems across different genders, and across U.S. based English language variations: African American English, Spanglish, Chicano English, and Standard American English. We build and analyze a novel audio dataset labeled for gender and dialect, and use the dataset to better understand disparities in state-of-the-art models and speech. Our initial results show a clear disparity between minority dialect speakers across gender and women standard English speakers.

1 Introduction

Recent work in natural language processing has identified dialect and gender bias in several applications (Sun et al., 2019), including automatic speech recognition (Koenecke et al., 2020; Tatman, 2017; Tatman and Kasten, 2017; Wassink et al., 2022). As a result, minority dialect speakers and women across dialects struggle to have their speech captioned accurately. Human-Computer Interaction studies have found minority groups have negative experiences with downstream ASR applications, such as captions on video based social media platforms (Harris et al., 2023) and voice assistants (Cunningham, 2023; Harrington et al., 2022). Mitigating these discrepancies is an important step towards developing equitable technologies that work well regardless of a user's identity. In this work, we investigate dialect and gender biases with our novel dataset specifically aimed at assessing dialect and gender bias. We frame our work around two research questions: (1) How do state-of-the-art ASR models perform across dialects, across genders, and within categories? (2) How do various finetuning approaches impact performance on these groups?

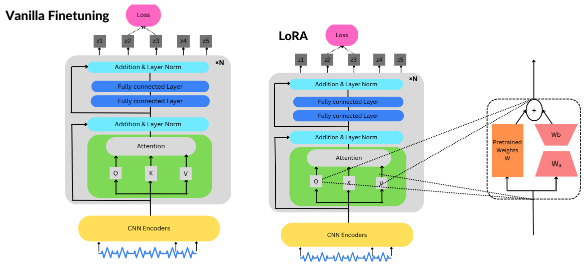
2 Background

Prior work from Koenecke et al. (2020) found industry ASR systems from IBM, Apple, Microsoft, Google, and Amazon have significantly worse performance for Black speakers. Another analysis of Client Libraries Oxford captioning system found disparities for Chicanx and African American speakers (Wassink et al., 2022). In the social media context, one evaluation of YouTube captions shows higher error rates for women than men (Tatman, 2017), while another analysis of YouTube across ethnic groups found the highest error rates for African Americans (Tatman and Kasten, 2017). Radford et al. (2023) studied the performance of wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021) and whisper (Radford et al., 2023) other models on several datasets, including the corpus of regional African American language (CORAAL) (Kendall and Farrington, 2023) which represents African American speech. This study identifies the Word Error Rate (WER) of English transcription on several datasets, giving some understanding of how models perform on underrepresented dialects, but doesn't explicitly explore racial or gender disparities. Prior studies of bias in ASR do not explore potential discrepancies within marginalized groups, further, most studies use the same dataset, CORAAL to represent African American speech, with minimal analysis of Spanglish. We fill these gap in the research by exploring bias with our novel dataset, labeled for minority dialect speech and gender.

3 Methods

3.1 Dialect-Centered Data Collection

We take an approach of data annotation centered on representing the minority dialects and demographic groups among annotators that are represented in our data. We collect data starting with the Spotify podcast dataset (Clifton et al., 2020). We collect



(a) Vanilla finetuning (b) Lora Adapter

Figure 1: Finetuning approaches

data for specific demographic groups of focus by using demographic related keyword searches. We start with and expand upon keywords used similarly in prior work (Richard and Kafai, 2016), the full list of keywords is found in the Appendix. We identify an audio sample as a potential match for a demographic group if it contains a related keyword in the podcast title or podcast description which annotators confirm. We recruit data annotators who are speakers of non-standard English dialects to annotate our data. Annotators listen to audio and transcribe the audio samples, using automatically generated transcripts from whisper as a base. Annotators are instructed to pay special attention to properly transcribing words, grammar patterns, and phrases that are unique to dialects of interest. These linguistic differentiation are often the source of automatic speech recognition errors. This process resulted in 14 hours of audio data.

3.2 Benchmarking

We benchmark baseline model performance across gender, dialect, and gender-dialect combinations with wav2vec2, HuBERT, and Whisper, using Word Error Rate (WER) as the evaluation metric

3.2.1 Vanilla Fine-tuning

Next we finetune models on our dataset with representation from each demographic group to understand how finetuning can impact performance (shown in Figure 1A).

3.2.2 Low-Rank Adaptation (LoRA)

Low Rank Adaptation or LoRA (Hu et al., 2021) is a parameter efficient training technique that freezes pre-trained model weights and injects a small amount of new weights into a model. We use this method (shown in Figure 1B) to understand the impact of a parameter efficient method on performance across marginalized groups.

.

	Men	Women
Whisper	0.280	0.471
HuBERT	0.314	0.478
wav2vec2	0.415	0.548

Table 1: Word Error Rate of Whisper, HuBERT, and wav2vec2 on our full dataset with respect to gender.

	AAVE	Chicano English	Spanglish	SAE
Whisper	0.670	0.453	0.363	0.279
Hubert	0.760	0.255	0.445	0.273
wav2vec2	0.820	0.318	0.528	0.356

Table 2: Word Error Rate of Whisper, HuBERT, and wav2vec2 on our full dataset with respect to dialect.

4 Baseline Benchmarking Results

Our initial results performance benchmark Word Error Rate of models with respect to gender, dialect, and gender dialect combinations. Results with gender are shown in Table 1. Results with respect to dialect are shown in Table 2. Results with respect to gender-dialect combinations are shown in the appendix. Results show disparities between men and women and between minority dialects and Standard American English. Results at the gender-dialect level are show Standard American English speaking men having better performance than all other sub-groups.

5 Conclusion

We present a novel dataset to explore fairness of automatic speech recognition across English language dialects and gender. Initial results of our analysis show clear performance disparities between Standard English Speaking men and all other groups, with the worst performance with African American English speaking women. Future results will show how different fine-tuning approaches on SoTA models using our dataset impacts the performance on these groups. Further, the size of the dataset will be increased to improve representation of some gender-dialect subgroups with low data representation.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Advances in Neural Information Processing Systems, 33:12449–12460.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jay L. Cunningham. 2023. Collaboratively mitigating racial disparities in automated speech recognition and language technologies with african american english speakers: Community-collaborative and equity-centered approaches toward designing inclusive natural language systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. "honestly, i think tiktok has a vendetta against black creators": Understanding black content creator experiences on tiktok. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Tyler Kendall and Charlie Farrington. 2023. The corpus of regional african american language. version 2023.06. eugene, ore.: The online resources for african american language project.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Gabriela T Richard and Yasmin B Kafai. 2016. Blind spots in youth diy programming: Examining diversity in creators, content, and comments within the scratch online community. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*, pages 1473–1485.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

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

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

A Appendix

Table 3 displays keywords used in creating the dataset. Table 4 displays full results on gender-dialect combinations.

	Keyword list
Women	women, girls woman, ladies
Men	men, man boys, boy guys, male
Latino	hispanic, hispanic american, boricua, mexican american, latino, latina, lantinx, chicano, chicana, chicanx
Black	african american, black women, black woman, black men, black man, black people

Table 3: Keywords used to identify podcasts of demographic groups.

	African American Vernacular English		**Chicano English**		**Spanglish**		**Standard American English/ White Mainstream English**	
	Men	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
Whisper	0.355	0.709	0.444	0.459	0.317	0.386	0.244	0.294
Hubert	0.538	0.767	0.242	0.265	0.531	0.369	0.224	0.280
Wav2Vec2	0.632	0.827	0.333	0.307	0.634	0.439	0.327	0.355

Table 4: WER on our full dataset of models without fine-tuning on gender-dialect combined categories. Results across some sub-categories are statistically insignificant, however all results with respect to Standard American English men are significant.