Final Report: Preferential Accent Bias in Automatic Speech Recognition Models

Pavle Medvicović, Jingwen Qi, Ian Wu
{medvidov, qijingwe, ianwu}@usc.edu
Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA

(Dated: October 8, 2024)

Keywords: Machine Learning, Natural Language Processing, Fairness, Bias, Automatic Speech Recognition

I. ABSTRACT

As Automatic speech recognition (ASR) systems become more and more ubiquitous in our daily lives (smart home assistants, automatic closed captioning), the downstream effects of biased models becomes more and more pronounced. ASR models have already been shown to be biased in terms of accent, dialect, and gender, but these studies are generally conducted on overly simplified academic datasets, while most of the applied uses of these models occur in a much noisier setting, e.g., audio with multiple speakers/accents. To this end, our work investigates how the presence of multiple accents in a single audio clip can affect the transcription quality on different accent groups, potentially amplifying bias. We evaluate transcription quality on both single-accent and multi-accent audio using word error rate, character error rate, and Jaro-Winkler distance. Our findings affirm previous works showing clear and consistent bias in single-accent audio, but interestingly show that there is not a significant effect on transcription quality when these models are applied to multi-accent audio.

II. INTRODUCTION

Automatic speech recognition (ASR) is a rapidly advancing field that is widely used in applications in daily life. However, like most modern deep learning models, speech recognition systems can be subject to biased training data, often making mistakes as a result. The speech data used to create these systems is most often taken from white/Western men, resulting in models that can have lower performance for speech by women and minority groups [1]. Even among males, data has shown that different races are understood at different rates. For example, Indian English has about a 78 percent accuracy rate while Scottish English only has a 53 percent accuracy [2]. In another case, an Irish woman, although a highly educated native speaker, failed a spoken English proficiency test for oral fluency [3]. The ubiquity of ASR systems across the internet, television, and telephone means that these types of biases are highly problematic and frequently have impacts on daily life. In particular, ASR bias heavily impacts minorities and individuals from underrepresented groups by reinforcing existing social inequities, especially harming those with diverse accents, genders, ages, and non-native speakers [4, 5]. These biases limit these groups' access to technology, impede effective communication, and can negatively affect job opportunities and social activities, harming fairness and equity. Efforts to understand and mitigate these biases are crucial for ensuring that ASR technologies serve all communities equitably.

Existing work on understanding and reducing ASR bias spans a wide array of research areas, including focusing on fairness, inclusivity, evaluation metrics, and adaptation/few-shot learning techniques. This includes works on developing multi-domain training approaches to enhance recognition accuracy across various French accents [6], using neural style transfer for accent modification [7], and creating more inclusive ASR models through the Universal Speech Model, which aims to support a wide range of languages [8]. These initiatives represent a collective effort to address and reduce biases in the ASR machine-learning community. However, current research practices have simultaneously been shown to introduce bias into modern ASR systems, leading to systemic errors and harm [9]. Our work aims to quantify and advance the understanding of accent bias that current techniques create in modern ASR deep learning models through statistical analysis.

To this end, we have utilized the Common Voice 16 dataset [10], published by the Mozilla Foundation. This dataset contains over 500 hours of audio data, tagged by speaker age, sex, and accent. Accents are included from the British Commonwealth, various Oceanic and Asian countries, African countries, and America. We have evaluated Wav2Vec2, Canary, and OpenAI's Whisper, representing several popular ASR models. In addition, we evaluated the performance of these models on multi-accent audio samples, as most popular ASR models are transformer-based, which can cause problems with multi-accent audio [6]. This is because the model is forced to attend to multiple accents at once, making it more difficult to accurately embed the latent features of the sample. To do this, we constructed an original dataset of spliced audio samples by concatenating samples/transcripts and tagging each word with the accent in which it is spoken. We were able to study how different combinations of accents affect the accuracy of these models, and if certain accents (e.g., American) consistently outperformed others in multi-accent ASR. This issue has been previously studied with French language audio and different French accent variations [6], but not with a broad set of multi-cultural accents to our knowledge.

III. RELATED WORKS

Previous work has served to highlight biases present in ASR systems including gender, age, regional accents, and non-native accents [7, 11]. Much of this work has focused on quantifying different biases and employed various approaches in order to reduce bias in these systems. These methods include using larger datasets with more varied accents [12], including an accent embedding for use in training the model [13], and augmenting data to produce datasets more similar to those sourced from native speakers while retaining understanding for non-native speakers [6]. However, these works often take fairly simplistic approaches to measuring bias, which fail to capture the nuances of realistic day to day speech, such as multiple accents appearing in a single clip. In this work, we introduce new methods of measuring bias in audio containing multiple accents, with the goal of furthering efforts into bias reduction for ASR systems.

Radzikowski *et al.* [13] cite the fact that ASR systems can achieve human-like understanding, but only in cases where the system is optimized for that language and a native speaker is speaking. Their work then focuses on augmenting the speech data with autoencoder-based style transfer to improve comprehension in ASR models. They use various models to transcribe the resultant data and evaluate bias reduction based on character error rate (CER). Other approaches to reducing accent bias include the work of Wang *et al.* [14], which utilizes additional accent embeddings to improve model performance on accented speech, and that of Maison and Esteve [6], where they fine-tuned models with varied non-native French accent sample to reduce model bias. In contrast, our work augments our dataset to create multi-accent samples, increasing task difficulty for ASR models. Rather than attempting to improve transcription quality, we aim to better understand sources of bias and errors for these models. In addition, we utilize multiple metrics to achieve a more thorough analysis of results.

The performance of ASR systems is most often evaluated using word error rate (WER), with the goal of providing a standardized metric across multiple systems. However, WER alone is not a robust metric, particularly for CTC models such as Wav2Vec2. Other automatically calculable metrics such as CER, Levenshtein distance, Match Error Rates, and Word Recognition Rate are commonly used to supplement WER in many cases [15]. Jaro-Winkler Distance [16] is also an effective metric which penalizes unmatched words based on their edit distance to the most similar token in the target sentence. This approach focuses on token level accuracy, but reduces the penalty for small misspellings.

Although these alternatives avoid some of the drawbacks found in WER, some existing work advocates for deep learning based metrics. Roux *et al.* [17] specifically recommend the use of part-of-speech error rate (POSER) and embedding error rate (EmbER) in their work. While their work also evaluated other metrics, POSER and EmbER were notably robust. These metrics were able to both re-score the ASR systems tested and highlight the underlying linguistic shortcomings of those systems [17]. In particular, EmbER is a useful metric because it measures errors in context, which WER cannot achieve. However, context-based evaluation approaches are not universally well-suited. These approaches tend to work well in models with a purely language-focused component, such as transducer networks [18]. The models evaluated in our work use different model paradigms, thus we opted to use WER, CER, and Jaro-Winkler Distance in our evaluations.

IV. METHODS AND DATA

Dataset

We conducted our evaluations using the Common Voice 16 dataset [10], which is frequently used in low-resource language recognition settings and for tasks involving fairness/bias due to its relevant labels. This dataset consists of audio recordings of spoken words and corresponding text transcriptions. Tracks are also paired with labels for accent, age, gender, and locale. Currently, the dataset contains 19,673 validated hours of audio in 120 languages at over 8 million samples. However, in this work we limited our use to the approximately 411 thousand English language samples for simplicity of evaluation.

The Common Voice 16 dataset was created through the Mozilla Foundation's Common Voice initiative [10]. Participants were crowd-sourced and able to submit audio tracks either through the Common Voice website or mobile application. Participants were supplied with a simple text to read, which was paired with the recording in the dataset. The audio recordings were also validated through crowdsourcing with a simply voting system, where participants could "upvote" or "downvote" a track. Audio tracks with more downvotes than upvotes were marked as invalid and removed from the dataset.

Accent labels in Common Voice 16 are self reported and allow for open ended open ended responses, leading to many labels which only correspond to a single sample. As such, we grouped labels to allow for meaningful sample sizes and comparative analysis between groups.

Accent Group	Number of Samples	Number of Sub-lables
north american	13700	131
british/english	4172	83
south asia	3757	25
oceanian	1054	13
east asian	697	20
african	474	26
eastern/western european	237	64
atlantic and carribean	106	6
central and south american	73	16
middle eastern and mediterranean	22	12
northern european	21	7

TABLE I. Accent groupings generated for the Common Voice 16 Dataset. Groups with red labels were omitted from our analysis due to their limited sample size. The ”central and south american” group was reserved as an exception, as it represents a large part of the world's English speakers despite its underrepresentation in Common Voice 16

These groups are listed in Table I and were derived based on factors such as geographical distance for regions in the original labels and similar pronunciations/linguistic roots. Naturally, these groupings contain some bias, and a more empirically motivated approach to accent grouping would be an important avenue in future work. In addition, some groups have been omitted from our study due to low sample size.

One shortcoming of the Common Voice 16 dataset is that audio samples are only from single speakers, and are therefore also single-accent. We addressed this shortcoming through augmenting the dataset by creating multi-accent samples, constructed by combining samples already present in the dataset (Further discussed in IV). This did result in samples containing nonsensical conversation, which we noted could potentially affect the quality of transcriptions for some models, but ultimately was still effective in comparative evaluation of accent bias.

The Common Voice 16 dataset is available at `https://huggingface.co/datasets/mozilla-foundation/common_voice_16_1`

Models

Our evaluations primarily focus on OpenAI's Whisper model [19], which has 6 variants of differing sizes available through Hugging Face. Using this model, we aimed to evaluate how the size and multi-linguality of ASR models affects their accent bias. While the exact data used to train Whisper is not disclosed in the original paper, Common Voice 16 is used as an evaluation dataset, which is a good indication that the dataset was excluded from the training set. However, OpenAI's work is conducted with commercial interests in mind, so this is not necessarily guaranteed.

We also performed evaluation on NVIDIA's NeMo Canary model [20] and Facebook Research's XLSR-Wav2Vec2 (Wav2Vec2) model [21]. Canary is multilingual, being trained in English, Spanish, French, and German, and is currently the highest-scoring model on Hugging Face's Open ASR leaderboard. In contrast to this, Whisper large is ostensibly capable of transcription in 99 languages, although it is slightly larger with approximately 1.5 billion parameters, as opposed to 1 billion for Canary. Wav2Vec2 also performs well on the Open ASR leaderboard, although it achieves slightly lower score than Canary and Whisper and utilizes the connectionist temporal classification (CTC), which predicts characters rather than words/tokens. It is also a multilingual model, capable of processing audio in 53 languages. As the primary goal of this work is to furhter investigate existing biases in current state of the art ASR models we did not fine-tune or train any of the models before evaluation.

Model	Parameters (M)
whisper-large (.en)	1,540
whisper-medium (.en)	764
whisper-small (.en)	242
whisper-base (.en)	72.6
whisper-tiny (.en)	37.8
wav2vec2-large	317
wav2vec2-base	95
Canary	1000

TABLE II. Models used in our evaluations and their sizes

All models in our evaluations utilize a transformer based architecture, which is typical of current state-of-the-art ASR models. We hypothesized that this would prove to have a negative effect on transcription quality in multi-accent audio, as transformers utilize attention to calculate features for their inputs. This means that portions of the audio in one accent will have their features affected by other portions of the audio in another accent, which we hypothesized was less likely to occur in the training data than single accent clips. In particular, we were interested in whether this would affect some accents more than others. For example, in the case where models transcribe samples containing a North American accent and a South Asian accent, we expected them to perform more poorly on the portion with an South Asian accent than the portion with the North American accent. By evaluating with several different ASR model architectures, we aim to explore how these architectures can mitigate this issue. This analysis could also have value beyond accent bias by being applied to multi-lingual audio or audio from speakers with speech impediments.

Evaluation

After generating transcriptions for the Common Voice 16 dataset, we measured their quality using WER, CER, and Jaro-Winkler Distance, as discussed in III. As it balances the benefits and drawbacks for WER and CER, we opted to use Jaro-Winkler as our primary evaluation metric, although all three were still used.

Our evaluation was conducted 2 stages: single accent evaluation and multi-accent evaluation. Single accent evaluation was conducted first in order to inform multi-accent evaluation. Due to the high computational requirements for running many state-of-the-art ASR models, performing multi-accent evaluation for all models was infeasible. By taking a more fine-grained approach to evaluating single-accent bias, we were able to take a more direct approach to multi-accent evaluation.

In single-accent evaluation, we first applied the metrics listed above to transcriptions generated by different models. We then aggregated individual transcription scores to generate statistics for specific accent groups and models, which we then used for comparative analysis. Our experiments are further elaborated upon in Section V, but included observing the change in accent bias levels as model size increases for both multi-lingual and English only models.

In multi-accent evaluation, we performed statistical analysis to observe how transcript quality is affected for different accent groups when multiple accents are present in a single audio clip, as well as the change in transcription quality for different accents when they are in single vs. multi-accent audio samples. To achieve this, we first generated a set number of multi-accent samples for all possible combinations of accent groups. For example, given the combination (a, b), multi-accent samples are constructed by prepending the audio from a sample belonging to accent group a with the audio from a sample belonging to b. The corresponding texts are also concatenated together to construct the clip's ground truth text.

For these samples, we produced transcriptions using the multi-lingual and English-only versions of Whisper-tiny, Whisper-base, and Whisper-small. As mentioned above, this our evaluation was limited to these models due to the high computational requirements to run larger models, particularly on the longer multi-accent samples. Transcriptions were evaluated with respect to the subsamples of the audio belonging to each accent group. For a given sample, this yielded 2 scores for each metric: $score\_a$ and $score\_b$. The transcription quality for the subsequences belonging to each of the 2 accent-groups were then evaluated separately with WER, CER and Jaro-Winkler Distance. Finally, we utilized two-sample t-tests to evaluate if there were significant differences between the means for each metric for each accent group in single-accent audio transcriptions vs. for each multi-accent pairing at a confidence level of 95% ($\alpha = 0.05$). In doing this, we aimed to test if the model would "prioritize" quality of transcription for some accent groups over others when they occurred in the same sample. For example, in the case that audio with accent a yields better transcriptions than accent b in single accent evaluation, would $score\_b$ be lower than the average score for b in single accent evaluation? Would $score\_a$ also decrease by a proportional amount?

Due to the Whisper architecture, its outputs (transcription tokens) are not able to be directly paired with the portion of the input (audio array) they correspond to. Thus, we utilized sentence similarity calculated with the all-MiniLM-L6-v2 model from Hugging Face to roughly identify subsequences in the transcriptions which corresponded to the each of the accent-groups. By finding a split point in the output tokens which maximize the average cosine similarity between embeddings for the ground truth sentence for each segment ($gt\_sentence\_a$ and $gt\_sentence\_b$) of the multi-accent audio and the portion of the transcriptions to the right and left of the split point, we were able to automatically identify the portions of each transcription belonging to each subsample.

V. RESULTS AND DISCUSSION

Our results show clear and consistent bias amongst the models tested, particularly for Asian accents, with models consistently performing worst on East/South Asian Accented samples. Interestingly, the models consistently perform best for Oceanic accents, such as

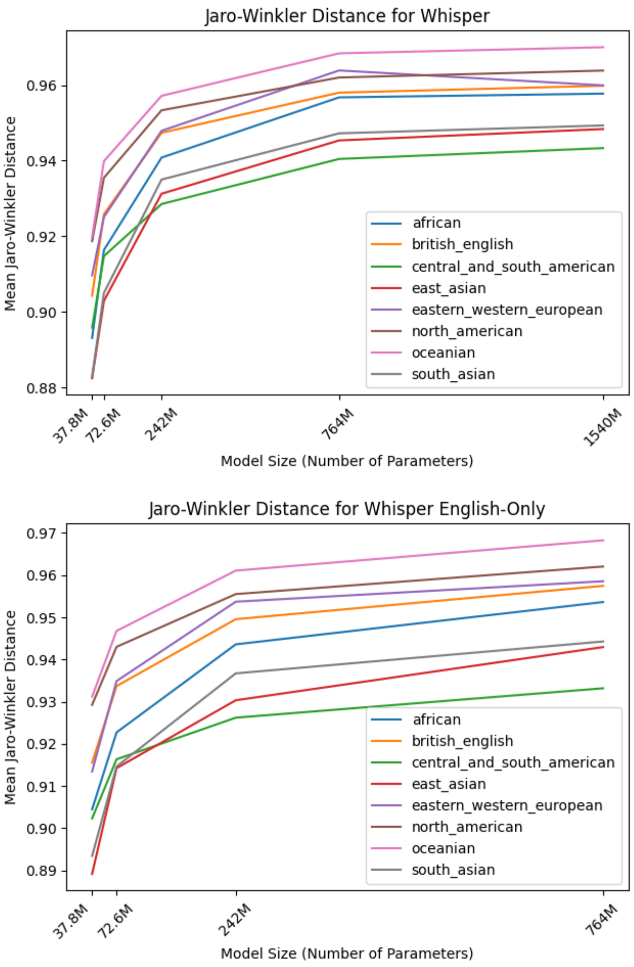

FIG. 1. Jaro-Winkler Distance for different accent-groups as Whisper model size changes.

Accent Group	Canary	w2v large	w2v base
oceanian	**0.97685**	**0.94355**	0.91175
north_american	0.97131	0.94327	**0.91417**
east/west_european	0.97031	0.93505	0.89884
british_english	0.96947	0.93648	0.901
south_asian	0.96406	0.91121	0.84404
african	0.96589	0.92452	0.88459
cent./south_american	0.96235	0.90825	0.86563
east_asian	0.96101	0.91551	0.87039

TABLE III. Jaro-Winkler Distance for non-Whisper model transcriptions across accent groups. The top score for each column is in **bold** and the lowest is underlined.

Australian/New Zealand accents, despite the fact that these populations make up a relatively small proportion of the world. Surprisingly however, we did not find that there was a significant difference in transcription quality for single-accent audio and multi-accent audio.

Model	WER	CER	JW-Dist
whisper_large	0.05048	0.03029	0.02671
whisper_medium	0.04853	0.02821	0.02797
whisper_small	0.06552	0.03585	0.02865
whisper_base	0.1029	0.0625	0.03697
whisper_tiny	0.12304	0.07396	0.03747
whisper_medium_en	0.06214	0.03555	0.03508
whisper_small_en	0.06963	0.03633	0.03486
whisper_base_en	0.09132	0.04872	0.03253
whisper_tiny_en	0.12096	0.0852	0.04202
wav2vec2_large	0.08963	0.04726	0.03531
wav2vec2_base	0.22615	0.12265	0.07013
canary	0.03232	0.01671	0.01584

TABLE IV. Difference between the best and worst average scores for different accent groups across models.

It is also notable that results for accent groups on which the models have lower performance are near-identical between the multi-lingual and English only versions of Whisper. Interestingly, marginal increases in performance on accents from English-speaking regions can be observed in the English-only models, as we would expect the increase in performance on accents from English speaking regions, but we would also expect a decrease in performance for other accent groups.

We also observed that as model size increases, accent bias actually decreases for both Whisper and Wav2Vec2 IV. This is counter to recent works in NLP which show that bias may actually increase with model size [22]. We hypothesize that while larger model sizes (and therefore higher variance) lead models to fit to stereotypes in NLP, in ASR tasks this higher level of variance allows models to better fit to phonetic differences between accents/speakers.

Our evaluation of multi-accent samples showed few statistically significant differences between single-accent and multi-accent evaluation for most accent groups at a confidence level of 95%. When observing differences in mean metric values for single-accent and multi-accent transcriptions, there are not any apparent trends for any accent groups 2. Across 56 possible unique accent combinations few accent combinations caused statistically significant changes in WER for any accent groups evaluated using the Whisper Tiny and Whisper Base models respectively. Whisper Small had no statistically significant differences in WER for any accent groups. While CER and Jaro-Winkler distance yielded more significant differences, this can be attributed to the longer sample length, as well as both the presence of homophones and punctuation errors arising due to the concatenation of samples.

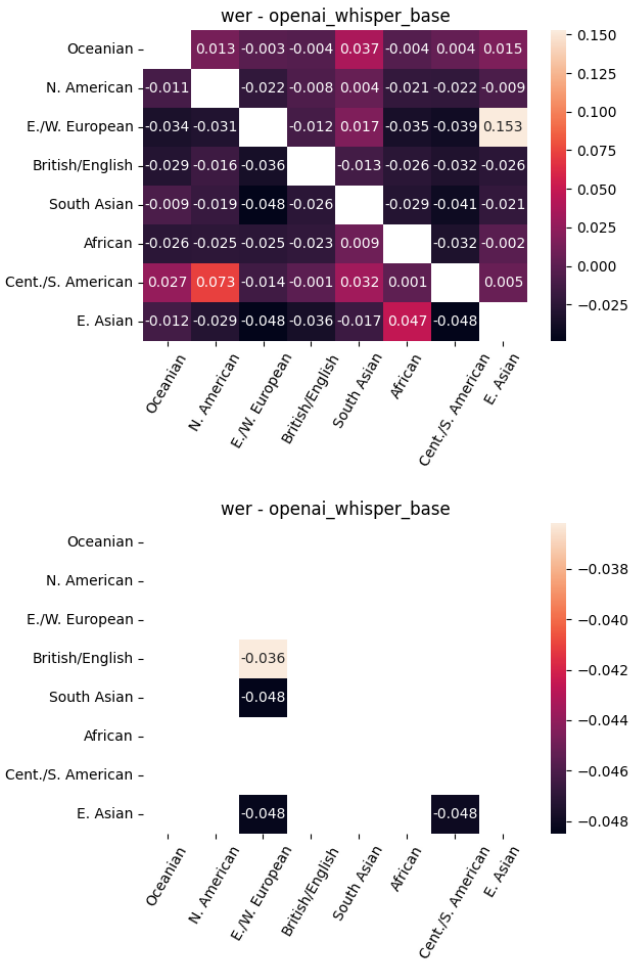

FIG. 2. Differences in average WER for transcriptions using Whisper-base before and after removing insignificant differences.

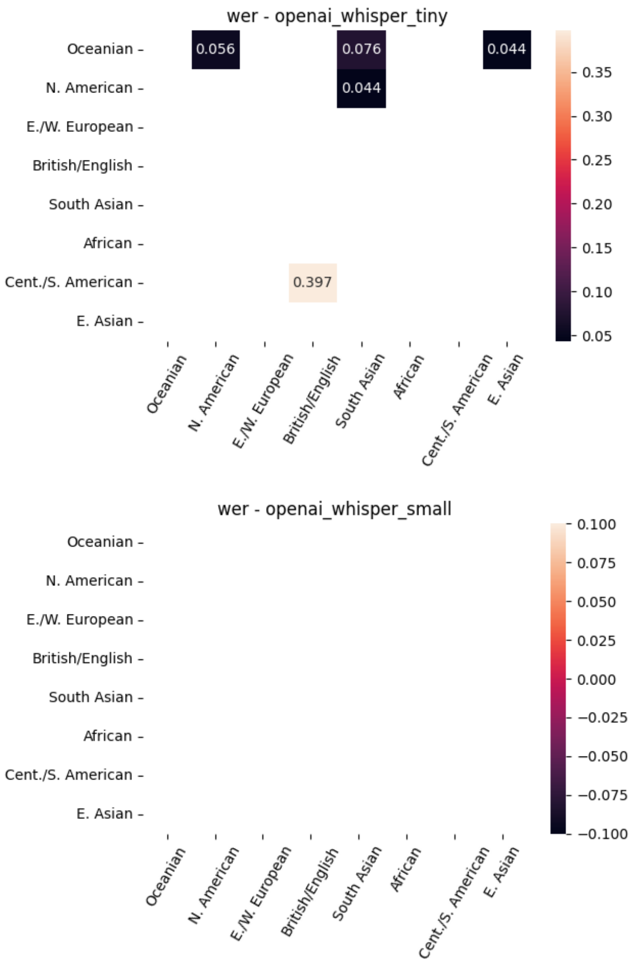

FIG. 3. Differences in average WER for transcriptions using Whisper-tiny and Whisper-small after removing insignificant differences.

VI. CONCLUSION AND FUTURE WORK

Our work was able to show a clear and consistent bias in state of the art ASR systems toward historically Western accents. While differences in Jaro-Winkler distance, WER, and CER were relatively small, our dataset also contained isolated samples and clear speech. When we forced the Whisper model to attend to multiple accents at once by evaluating mutli-accent samples there were few significant changes in WER for the accents that the model was shown to be biased against. Despite small differences, clear bias in such a controlled sample is cause for concern. In a real world situation, whether generating automatic captions, using an automated phone service, or going through an automated interview, ASR systems clearly will not serve all communities equitably.

Clear evidence of such bias provides ample direction for future work studying ASR systems. First, to further validate the results of this study, future work could focus on greater analysis of the variance of scores across model sizes and accent groups. Accent groups vary greatly in size, as seen in I, meaning that this analysis is needed to properly contextualize model performance. Future work could also work with more similarly sized accent groups to better understand bias. Alternatively, given the presence of bias, future work could focus on bias mitigation strategies for impacted accents. Maison and Esteve [6] provides a great example of strategies that could be taken, but training the models using a curated set of a targeted accent could also work well in a specific locale. Finally, more linguistically accurate accent grouping would provide more accurate insight into the bias present in these models. Whatever route future work takes, it is crucial to ensure that future ASR technologies serve all communities equitably.

CODE AVAILABILITY

Our codebase is available at `https://github.com/ianwu13/Accent-Bias-in-ASR-Models`.

Appendix A: Figures for Results

Accent Group	Canary	w2v large	w2v base
east/west_european	**0.05721**	0.1571	0.27372
oceanian	0.05963	**0.14145**	**0.23066**
cent./south_american	0.05996	0.19085	0.33581
north_american	0.06829	0.14545	0.23292
british_english	0.07153	0.15892	0.26332
african	0.07884	0.19198	0.31715
south_asian	0.08596	0.23108	0.45681
east_asian	0.08952	0.21948	0.37042

TABLE V. WER for non-Whisper model transcriptions across accent groups. The top score for each column is in **bold** and the lowest is underlined.

Complete results and graphics for all models is available at `https://drive.google.com/drive/folders/1hv7bWBM3qU7SNWl5A0kCVZkd4wELkI_h?usp=sharing`.

[1] W. T. Hutiri and A. Y. Ding, Bias in automated speaker recognition, in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (Association for Computing Machinery, New York, NY, USA, 2022) p. 230–247.

[2] R. Tatman, Gender and dialect bias in YouTube's automatic captions, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, edited by D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach (Association for Computational Linguistics, Valencia, Spain, 2017) pp. 53–59.

[3] A. A. Press, Computer says no: Irish vet fails oral english test needed to stay in australia — theguardian.com, https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia (2017), [Accessed 31-01-2024].

[4] C. Doty, Detecting and reducing bias in speech recognition (2022).

[5] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, Quantifying bias in automatic speech recognition (2021), arXiv:2103.15122 [eess.AS].

[6] L. Maison and Y. Esteve, Improving accented speech recognition with multi-domain training, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023) pp. 1–5.

[7] O. Y. Kacper Radzikowski, Le Wang and R. Nowak, Accent modification for speech recognition of non-native speakers using neural style transfer, Journal of Audio Speech and Music Processing **2021**, 10.1186/s13636-021-00199-3 (2021).

[8] Y. Zhang and J. Qin, Universal speech model: State-of-the-art speech ai for 100+ languages (2023).

[9] N. Markl and S. J. McNulty, Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (European Language Resources Association, Marseille, France, 2022) pp. 6328–6339.

[10] 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)* (2020) pp. 4211–4215.

[11] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, Quantifying bias in automatic speech recognition (2021), arXiv:2103.15122 [eess.AS].

[12] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, Training and typological bias in ASR performance for world Englishes, in *Proc. Interspeech 2022* (2022) pp. 1273–1277.

[13] K. Radzikowski, L. Wang, O. Yoshie, and R. Nowak, Accent modification for speech recognition of non-native speakers using neural style transfer, EURASIP J. Audio Speech Music Process. **2021**, 10.1186/s13636-021-00199-3 (2021).

[14] Y. Wang, H. Gu, R. Shen, Y. Li, W. Jiang, and J. Huang, Disentanglement of speaker identity for accented speech recognition, in *2023 8th International Conference on Communication, Image and Signal Processing (CCISP)* (2023) pp. 1–6.

[15] E. Draffan, M. Wald, C. Ding, and Y. Li, Exploring practical metrics to support automatic speech recognition evaluations, Assistive Technology: Shaping a Sustainable and Inclusive World 10.3233/shti230636 (2023).

[16] W. Winkler, String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, Proceedings of the Section on Survey Research Methods (1990).

[17] T. B. Roux, M. Rouvier, J. Wottawa, and R. Dufour, Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition, in *Interspeech* (Incheon, South Korea, 2022).

[18] A. Graves, Sequence transduction with recurrent neural networks, ArXiv **abs/1211.3711** (2012).

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust speech recognition via large-scale weak supervision (2022).

[20] K. C. Puvvada, P. Żelasko, H. Huang, O. Hrinchuk, N. R. Koluguri, S. Majumdar, E. Rastorgueva, K. Dhawan, Z. Chen, V. Larukhin, J. Balam, and B. Ginsburg, Nvidia nemo canary model pushes the frontier of speech recognition and translation (2024).

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, Unsupervised cross-lingual representation learning for speech recognition (2020), arXiv:2006.13979 [cs.CL].

[22] Y. Tal, I. Magar, and R. Schwartz, Fewer errors, but more stereotypes? the effect of model size on gender bias, in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, edited by C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, and H. Gonen (Association for Computational Linguistics, Seattle, Washington, 2022) pp. 112–120.