Addressing Accent Disparities in Automatic Speech Recognition: A Comparative Study of Single and Two-Step Adaptation

Mykhailo Danilevskyi, Fernando Perez-Tellez, Jelena Vasic
Technological University Dublin
Blessington Rd, Dublin, D24 FKT9, Ireland
D22126578@mytudublin.ie, {Fernando.PerezTellez,Jelena.Vasic}@TUDublin.ie

Abstract

Automatic speech recognition (ASR) systems often exhibit uneven performance across accents, raising concerns about fairness and bias. This study investigates the impact of model fine-tuning strategies on ASR performance and accent-related disparities. We conduct a controlled empirical evaluation of two adaptation approaches—single-step and two-step fine-tuning—using pretrained Whisper (small) and Wav2Vec2-XLSR-53 models on African-accented English speech from the AfriSpeech-200 dataset, covering Yoruba, Igbo, Swahili, and Hausa accents. Both fine-tuning strategies substantially reduced mean word error rate (WER) for all models. However, these improvements did not translate into consistent reductions in accent-related performance gaps. When analysed separately across general and clinical subsets, WER gaps often increased due to uneven gains across accents. Although two-step fine-tuning provided modest improvements over single-step adaptation, its impact on reducing disparities remained limited. These findings indicate that fine-tuning primarily optimises performance without effectively addressing systematic bias across speaker groups, even when models are specialised for individual accents. This highlights the limitations of per-accent specialisation as a practical bias mitigation strategy.

Keywords: automatic speech recognition, accent disparities, fairness in AI; bias in speech recognition, fine-tuning

1. Introduction

Automatic Speech Recognition (ASR) performance has improved significantly in recent years, largely driven by self-supervised acoustic models such as Wav2Vec2 (Baevski et al., 2020; Conneau et al., 2020) and large-scale multilingual transformer models such as Whisper (Radford et al., 2022). The current level of ASR performance has led to its widespread acceptance and deployment across industries, accelerating the adoption of transcription technologies in healthcare, customer service, education, and other domains.

Despite strong performance on general English speech, ASR systems remain sensitive to non-native accents and domain-specific terminology, resulting in persistent performance disparities across speaker groups (Veliche et al., 2024; Liu et al., 2021). These disparities raise important ethical concerns related to fairness, equal access, and equitable distribution of technological benefits.

Unequal ASR performance across accents can result in systematically higher word error rates (WER) for speakers of non-dominant language varieties. Such discrepancies risk marginalising already underrepresented groups and may undermine trust in AI-driven systems. As ASR increasingly replaces or supplements human transcription, performance gaps can have tangible social and professional implications. Therefore, addressing accent-based disparities is not only a technical challenge but also a matter of linguistic equity. Reflecting this growing concern, research on accent bias in ASR continues to expand, with numerous studies published each year (Prinos et al., 2024).

The contributions of this study are:

- Analysed accent-related model bias by comparing WER and WER gaps under single-step and two-step fine-tuning strategies.
- Investigated whether additional fine-tuning and accent-specific models mitigate performance disparities across accent pairs.
- Provided a systematic comparison of how Whisper and Wav2Vec2-XLSR-53 models exhibit disparities across accents.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the methodology, including the dataset, evaluation metrics, models, and fine-tuning procedures. Section 4 presents the experimental results. Section 5 discusses the findings, and Section 6 concludes the paper.

2. Related work

Prior studies have proposed a range of strategies for mitigating accent-related bias in ASR. These can be broadly classified into accent-agnostic and accent-specific approaches. Accent-agnostic methods typically rely on large-scale training with

Proceedings of SPEAKABLE @ LREC 2026, pages 79–85

adversarial bias mitigation to encourage accent-invariant representations. In contrast, accent-specific approaches incorporate explicit accent labels or leverage architectures such as mixture-of-experts (MoE), in which specialised submodules handle specific accents (Bagat et al., 2025; Lee et al., 2026). However, MoE-based solutions require accurate routing between experts, which can be overly complex in scenarios with a limited number of accents and ASR system users.

In these constrained settings, simpler adaptation strategies may prove to be more practical. Following the ASR model adaptation presented in (Meyer et al., 2020; Pillai et al., 2024), we explore a two-step fine-tuning procedure to improve ASR performance in accent-aware contexts without requiring routing or explicit accent classification at inference time. In this approach, the pretrained model is first adapted to the entire dataset and then further fine-tuned on each target accent individually. In (Meyer et al., 2020), this strategy was applied with the DeepSpeech model (Hannun et al., 2014), fine-tuning first on the full Common Voice dataset and then on each demographic group, yielding five models. The authors (Pillai et al., 2024) used a similar two-step strategy to adapt Whisper from Tamil to Malasar, leveraging the linguistic similarities between resource-rich and low-resource languages.

While recent work has primarily focused on improving model performance for specific accents or languages, this paper instead examined accent-related bias and investigated whether fine-tuning on accent-specific data can reduce WER disparities. In contrast to (Torgbi et al., 2025), which evaluated Whisper on spontaneous telephone speech from native English speakers across two Scottish accents, our work analysed both Whisper and Wav2Vec2-XLSR-53 models on non-native English speakers reading scripted text, providing a different linguistic and acoustic setting. Furthermore, we investigated whether additional fine-tuning on a specific accent improves performance for that accent while also reducing performance disparities across accents. Although (Özyilmaz et al., 2025) adopted a similar fine-tuning strategy, its primary focus remained on improving performance for Arabic dialects rather than measuring bias across groups. In addition to the original Afrispeech-200 paper (Olatunji et al., 2023), which introduced the dataset and evaluated several state-of-the-art models, including Whisper and Wav2Vec2-XLSR-53, we extended this work by examining the impact of further model fine-tuning on both performance and bias. Specifically, we assessed overall model bias and evaluated whether observed differences between accents were statistically significant.

In this paper, we adopted a two-step fine-tuning approach on the Afrispeech dataset, first fine-tuning pretrained Whisper and Wav2Vec2-XSLR-53 models on the full dataset, followed by accent-specific fine-tuning. We compared this with single-step fine-tuning on the full dataset to evaluate overall performance and accent disparities.

3. Methodology

3.1. Data

Our experiments were conducted on the Afrispeech-200 dataset (Olatunji et al., 2023), which contains audio recordings, transcripts, and accent labels covering 120 African accents from 13 countries and 2,463 unique speakers, covering both general and clinical domains. The data were originally partitioned into training, development, and test splits with no speaker overlap across the splits[1]. Its composition is summarised in Table 1, with the proportion of clinical data indicated in brackets.

Table 1: Speaker counts and total duration (minutes) per accent across train, development, and test splits. Clinical-domain data proportions are indicated in brackets.

Accent	Speakers			Duration (min)		
	Train	Dev	Test	Train	Dev	Test
Yoruba	454 (51%)	57 (53%)	172 (47%)	2527 (62%)	55 (53%)	107 (36%)
Igbo	246 (64%)	34 (50%)	94 (56%)	1457 (70%)	35 (50%)	55 (53%)
Swahili	46 (59%)	12 (58%)	61 (44%)	805 (71%)	47 (53%)	80 (44%)
Hausa	176 (80%)	19 (74%)	53 (74%)	1125 (85%)	15 (66%)	38 (78%)
Others	544 (51%)	125 (60%)	370 (60%)	4459 (55%)	371 (55%)	841 (55%)
Total	**1466** (57%)	**247** (58%)	**750** (56%)	**10373** (63%)	**523** (54%)	**1121** (56%)

For performance and bias analysis, we selected the four most represented accents in the dataset: Yoruba, Hausa, Swahili, and Igbo, which we refer to as the target accents. Model fine-tuning was performed using the training and development splits of the dataset, whereas evaluation was conducted on the test split of the target accents.

The duration of speech per speaker varied across splits and accents. On average, speakers in the training split contributed 7.1 min of audio, compared to 2.1 and 1.5 min in the development and test splits, respectively. For the target accents, the average audio duration in the training split was approximately 6 min per speaker; however, Swahili

[1] https://huggingface.co/datasets/intronhealth/afrispeech-200

exhibited substantially longer recordings, with an average of 17.5 min per speaker. The data were unevenly distributed across the general and clinical domains, with Hausa exhibiting the highest proportion of clinical audio recordings (~80%) and Yoruba the lowest (36% in the test split).

All audio files were resampled from 44.1 kHz to 16 kHz to be consistent with the model requirements. To ensure correct comparison across models, both reference transcripts and model predictions were normalised using the Hugging Face Whisper basic normaliser, followed by punctuation removal.

3.2. Evaluation Metrics

The transcription accuracy in all experiments was measured using the word error rate (WER) metric.

To estimate the overall model bias in the ASR system, we used the relative *WERgap* metric, defined as the difference between the maximum and minimum WER across groups (e.g. accents), normalised by the maximum WER:

$$\text{WER}_{\text{gap}} = \frac{\text{WER}_{\text{max}} - \text{WER}_{\text{min}}}{\text{WER}_{\text{max}}} \quad (1)$$

For pairwise disparity evaluation between accents, we employed the methodology proposed by Liu et al. (2021) using the following test statistic:

$$\theta_{i,j} = \frac{\text{WER}_i}{\text{WER}_j} - 1 \quad (2)$$

where $\theta_{i,j}$ represents the bias between the two groups, and WER_i and WER_j are the word error rates for groups i and j, respectively. Statistical significance was assessed using the bootstrap method (Marriott et al., 1995; Bisani and Ney, 2004) with 95% percentile confidence intervals (CI). If the CI $[\text{ln}, \text{un}]$ does not include 0, the WER difference between the two groups is considered statistically significant. The bootstrap method was also used to estimate whether the difference in WER between fine-tuning methods is statistically significant.

3.3. Models

The experiments were conducted using the pretrained Whisper small model with 244M parameters (Radford et al., 2022) and the Wav2Vec2-XLSR-53 model with 317M parameters (Conneau et al., 2020). Whisper is a transformer-based encoder–decoder architecture, commonly described as a sequence-to-sequence model. It was trained on approximately 680,000h of labelled speech data annotated through large-scale weak supervision. Wav2Vec2-XLSR-53 model builds on the Wav2Vec 2.0 architecture (Baevski et al., 2020) by extending it to a cross-lingual setting. Wav2Vec2-XLSR-53 is an encoder-only architecture that predicts output tokens using connectionist temporal classification (CTC) (Graves et al., 2006).

For this study, we used pretrained models available on Hugging Face: wav2vec2-large-xlsr-53-english[2] and whisper-small multilingual model [3]. Both models are end-to-end architectures and were pretrained on multilingual data, which have been shown to be beneficial for improving ASR performance on accented speech (Vu et al., 2014).

3.4. Experimental method

We evaluated two ASR model adaptation approaches using Whisper (small) and Wav2Vec2-XLSR-53: (i) single-step fine-tuning on the full dataset and (ii) a two-step fine-tuning consisting of accent adaptation on the entire dataset followed by accent-specific fine-tuning yielding specialised models for each accent (Figure1). The two-step approach separates general accent adaptation and accent-specific specialisation. This allows the model to better capture acoustic and linguistic characteristics of each accent and may help reduce performance disparities across accents.

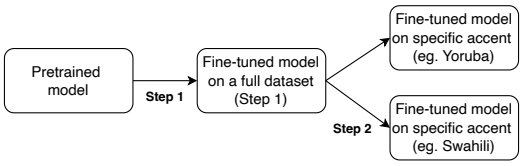

Figure 1: Two-step fine-tuning

For each architecture, we fine-tuned a pretrained baseline model on the entire dataset to implement the single-step approach, selecting the best model based on the lowest word error rate (WER). This approach yields a single model specialised for all accents in the dataset. For the two-step approach, we first fine-tuned the model on the entire dataset, selecting the best checkpoint based on minimum validation loss. In the second step, we continued the adaptation separately for each target accent, selecting the best accent-specific model using WER as the evaluation criterion. This strategy results in one specialised model per accent.

Whisper models were trained for 10 epochs using a linear learning rate scheduler with a 0.025 warm-up ratio, weight decay of 0.03 and effective batch size of 32 for both adaptation strategies. The learning rate was set to 1e-5 for adaptation on the entire dataset and reduced to 1e-7 for accent-specific fine-tuning. Wav2Vec2-XLSR-53 models were trained for 20 epochs with a linear scheduler and a 0.1

[2]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

[3]https://huggingface.co/openai/whisper-small

warm-up ratio for both strategies, a learning rate of 1e-5 and effective batch size of 32. All models were fine-tuned with the encoder unfrozen to allow maximal adaptation to the acoustic and linguistic characteristics of each accent. Early stopping was applied to terminate training once performance plateaued, preventing overfitting. These hyperparameters were selected based on prior literature (Olatunji et al., 2023; Baevski et al., 2020; Bagat et al., 2025) and preliminary experiments, providing stable and reasonable performance within our computational constraints.

4. Results

Table 2 reports WER results for the pretrained baselines and their single-step (FT1) and two-step fine-tuned (FT2) variants of Whisper (small) and Wav2Vec2-XLSR-53 across four African accents: Yoruba, Igbo, Swahili, and Hausa. We assess whether additional fine-tuning and accent-specific models reduce WER and cross-accent differences, analysing (i) within-model WER changes and (ii) pairwise variation across accents.

The pretrained Whisper (small) model achieved a mean WER of 0.446 across the four target accents, with an absolute WER gap of 0.420. Both fine-tuning strategies led to substantial reductions in mean WER, reaching 0.189 and 0.183 for FT1 and FT2, respectively. Across all configurations, Swahili consistently exhibited the lowest overall WER. In contrast, the highest WER was observed for Hausa in the clinical subset, while the lowest within Hausa occurred in the general subset. The overall WER gaps remained relatively stable across all model variants (0.404–0.425). However, when analysed separately by domain, the gaps increased after fine-tuning. In the general subset, the gaps were primarily driven by differences between Yoruba and Hausa, whereas in the clinical subset they were driven by Hausa and Swahili. Although fine-tuning increased the WER gaps in both subsets compared to the baseline, FT2 resulted in a lower gaps than FT1. Overall, FT2 achieved both lower WER and smaller WER gaps than FT1 across domains.

Wav2Vec2-XLSR-53 baseline model achieved a mean WER of 0.631 across the target accents with an absolute WER gap of 0.384. Both fine-tuning strategies substantially reduced the mean WER, to 0.270 (FT1) and 0.259 (FT2). Swahili consistently exhibited the lowest WER across all model variants, while Hausa showed the highest WER overall, except in the general subset where its WER was lower than that of Swahili. FT2 achieved lower WER than FT1 across all accents and domains, with statistically significant differences observed for Yoruba and Swahili. The overall WER gaps ranged from 0.318 to 0.384 and was primarily driven by differences between Hausa and Swahili. FT1 reduced the absolute WER gap by approximately 0.06 (from 0.384 to 0.318), whereas FT2 reduced it to 0.346. This reduction was primarily driven by improvements in the clinical domain, whereas the gap increased in the general domain due to lower WER for Hausa. Overall, FT2 achieved lower WER across accents and a smaller WER gap in the general subset.

The results indicate that fine-tuning primarily improved overall recognition accuracy, but it exacerbated performance disparities due to uneven gains across accents and data domains. Although

Table 2: WER comparison across accents (statistically significant difference between FT2 and FT1 within accents and models are in bold).

Accent	Base	FT1	FT2	
			Step 1	Step 2
Whisper (small)				
Yoruba	0.523	0.226	0.231	0.223
general	0.496	0.229	0.235	0.221
clinical	0.562	0.221	0.226	0.227
Igbo	0.472	0.174	0.198	0.178
general	0.389	0.137	0.137	0.138
clinical	0.541	0.205	0.249	0.211
Swahili	0.303	0.141	0.133	**0.132**
general	0.289	0.138	0.132	0.131
clinical	0.323	0.147	0.134	**0.135**
Hausa	0.520	0.239	0.224	0.219
general	0.351	0.089	0.097	0.096
clinical	0.582	0.293	0.269	0.263
WER mean	0.446	0.189	0.191	0.183
general	0.399	0.169	0.169	0.164
clinical	0.496	0.212	0.216	0.205
WER gap	0.420	0.407	0.425	0.408
general	0.417	0.611	0.587	0.566
clinical	0.445	0.498	0.502	0.487
Wav2Vec2-XLSR-53				
Yoruba	0.676	0.298	0.296	**0.281**
general	0.563	0.280	0.274	**0.260**
clinical	0.835	0.323	0.327	0.309
Igbo	0.640	0.260	0.250	0.256
general	0.535	0.215	0.209	0.212
clinical	0.728	0.298	0.285	0.293
Swahili	0.501	0.228	0.226	**0.216**
general	0.433	0.204	0.205	**0.193**
clinical	0.590	0.261	0.253	**0.246**
Hausa	0.813	0.335	0.333	0.331
general	0.426	0.175	0.183	0.181
clinical	0.954	0.392	0.386	0.385
WER mean	0.631	0.270	0.267	0.259
general	0.507	0.234	0.231	0.221
clinical	0.765	0.312	0.309	0.302
WER gap	0.384	0.318	0.321	0.346
general	0.243	0.375	0.332	0.304
clinical	0.382	0.334	0.345	0.361

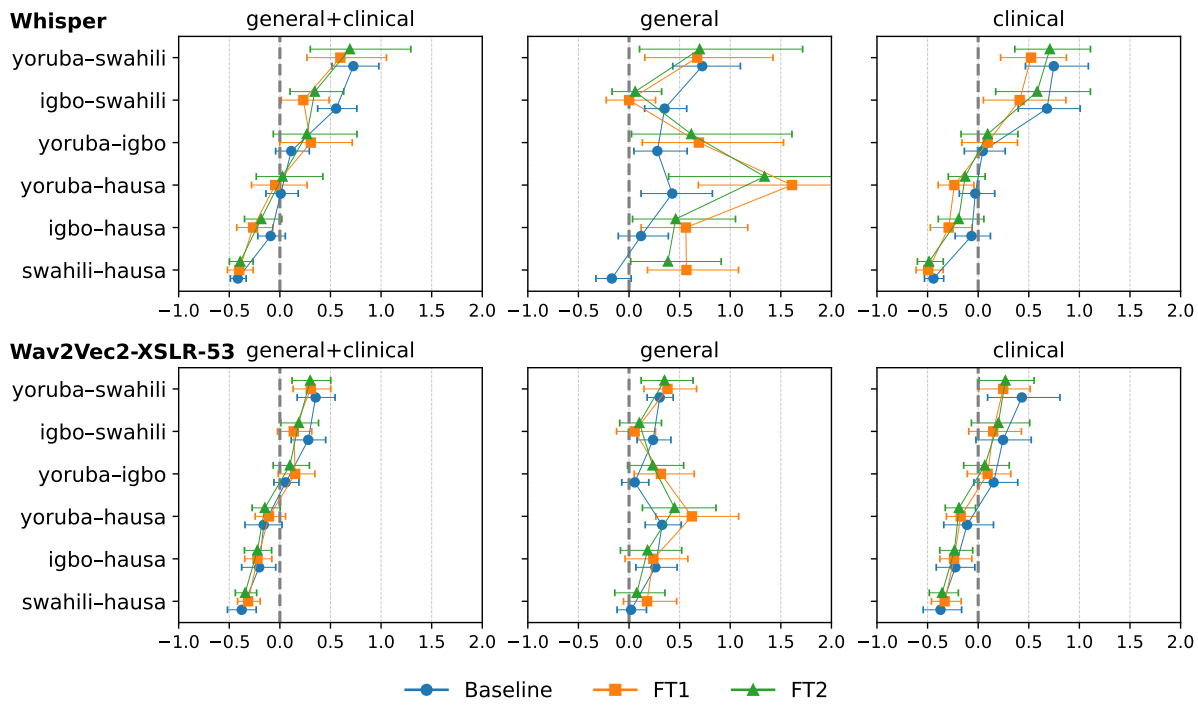

Figure 2: Pairwise WER comparison of relative difference using test statistic θ (Eq. 2). The x-axis was limited to [-1, 2] to enhonce results visibility, truncating the upper confidence intervals for the Yoruba-Hausa pair, which reach 2.98 and 2.96 for FT1 and FT2, respectively.

Whisper achieved higher overall performance than Wav2Vec2-XSLR-53, WER gaps remained larger across all model configurations, and both fine-tuning approaches even amplified these gaps in the general and clinical data subsets. The performance of Wav2Vec2-XSLR-53 is less varying across accents and data domains, which reflects on further pairwise analysis.

Pairwise WER measurements between accents using the test statistic θ (Eq.2) presented on Figure 2. For the Whisper model, bias patterns differed across domains. In the combined (general + clinical) setting, all pairs involving Swahili exhibited statistically significant differences at the baseline; these were reduced by both FT1 and FT2. In the general domain subset, baseline differences were observed for all pairs involving Yoruba, as well as for Igbo–Swahili. Both fine-tuning approaches resolved only the Igbo–Swahili difference, while increasing gaps for pairs involving Hausa; this increase was slightly smaller for FT2. In the clinical domain, all three pairs involving Swahili showed statistically significant differences at the baseline. These were not resolved by either fine-tuning approach. FT1 increased gaps for all pairs involving Hausa to statistically significant levels, while both methods slightly reduced differences for pairs involving Swahili. Overall, improvements in Swahili-related disparities appear to come at the cost of increased disparities for Hausa, highlighting a trade-off in bias mitigation across accents.

For the Wav2Vec2-XLSR-53 model, bias patterns also varied across domains. In the combined (general + clinical) setting, four pairs showed significant differences at baseline, including three involving Swahili and the Igbo–Hausa pair. Both FT1 and FT2 produced similar outcomes, primarily reducing differences for pairs involving Swahili. In the general domain, four pairs involving Swahili and Hausa showed significant differences at baseline. FT1 resolved differences for pairs involving Igbo (with Swahili and Hausa), but introduced a significant difference for the Yoruba–Igbo pair. In contrast, FT2 resolved two differences, both involving Igbo, without introducing new effects; for most pairs, bias gaps moved closer to zero. In the clinical domain, three pairs showed significant differences at baseline, including two involving Hausa (paired with Igbo and Swahili) and one Yoruba–Swahili pair. Both fine-tuning approaches slightly reduced bias gaps; however, these differences remained for all three pairs.

In conclusion, Whisper exhibited larger absolute performance gaps between certain accents, particularly involving Swahili, yet fewer statistically significant pairwise disparities overall. In contrast, Wav2Vec2-XLSR-53 demonstrated a smaller overall WER gaps and more consistently distributed disparities across accent pairs. These results indicate that Whisper's bias is more concentrated around

specific accent, whereas Wav2Vec2-XLSR-53 exhibits more distributed differences across groups.

5. Discussion

This study investigates single-step (FT1) and two-step (FT2) fine-tuning approaches, focusing on their impact on reducing accent-related performance disparities in ASR models. The results show that for Whisper and Wav2Vec2-XSLR-53 models both fine-tuning approaches significantly outperform baseline and FT2 demonstrated lower WER in most cases. In contrast, WER gaps increased after fine-tuning, though after FT2 they were lower than after FT1. The main driver of WER gaps was Swahili accent with the lowest WER. Notably, this occurs despite Swahili having the smallest number of speakers in the training split. However, Swahili recordings contain substantially longer speech segments per speaker (approximately 17 min per speaker), whereas the other accents average around 6 min per speaker. Hausa is another accent with lowest WER for general subset that drove high gaps in the general subset, but this can be driven by the limited test subset containing only 8 min of audio.

Whisper model achieved lower WER but exhibited larger WER gaps compared to Wav2Vec2-XLSR-53. Pairwise analysis highlights distinct bias patterns between the models, likely arising from differences in architecture and training data. These findings suggest that Whisper's sequence-to-sequence architecture, together with its autoregressive decoding, may contribute to accent-specific biases. By contrast, Wav2Vec2's cross-lingual pretraining and CTC-based design appears to yield more consistent performance across accents.

The study is constrained by the limited size and uneven distribution of accent-specific data, particularly for Hausa, where some subsets contained as little as 8min of recordings compared to 17–68min for other accents. This imbalance likely contributed to amplified gaps in general domains after fine-tuning which requires further investigation. Additionally, our analysis focused on only four African accents and two ASR models; conclusions may not generalize to other languages and accents. Finally, while the two-step fine-tuning approach improves WER, it may not fully address systematic bias arising from pretraining data composition or domain-specific acoustic variability. Future work should explore balanced data augmentation, multi-task objectives, or bias-aware loss functions to improve both accuracy and fairness. Further analysis of domain-specific acoustic characteristics may also help in designing more equitable ASR systems for clinical and general speech.

6. Conclusion

In this paper, we investigated the effectiveness of single-step and two-step fine-tuning strategies for reducing accent-related performance differences in automatic speech recognition models across four African accents using Whisper (small) and Wav2Vec2-XLSR-53. The results demonstrated that fine-tuning significantly improved overall WER for both models; however, its impact on reducing accent-dependent disparities remained limited. Performance gaps across accents persisted under all examined fine-tuning approaches. Although two-step fine-tuning yielded incremental improvements over single-step adaptation, these gains were modest and unevenly distributed across accents - Hausa consistently exhibited the highest WER, whereas Swahili achieved the lowest.

More broadly, this study demonstrates that optimisation of overall performance does not necessarily lead to improved parity across groups. Therefore, mitigating accent-related disparities may require reweighting strategies, adversarial invariance training, or fairness-regularised optimisation. However, a two-step fine-tuning approach that yields one model per speaker group may be beneficial in settings with a limited number of users, such as medical laboratories. In such constrained environments, more complex approaches, such as mixture-of-experts architectures or adversarial debiasing, may be impractical. Consequently, two-step fine-tuning can serve as a method for improving performance in accent-specific deployments, although its bias mitigation capabilities are limited.

7. References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv:2006.11477 [cs].

Raphaël Bagat, Irina Illina, and Emmanuel Vincent. 2025. Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition. In *Interspeech 2025*, pages 1143–1147. ArXiv:2505.20006 [cs].

M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409–12, Montreal, Que., Canada. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael

Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. ArXiv:2006.13979 [cs].

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Wonjun Lee, Hyounghun Kim, and Gary Geunbae Lee. 2026. Mixture-of-Experts with Intermediate CTC Supervision for Accented Speech Recognition. ArXiv:2602.01967 [cs].

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions. ArXiv:2111.09983 [eess].

Paul Marriott, B. Efron, and R. J. Tibshirani. 1995. An Introduction to the Bootstrap. In *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, volume 158, page 347. ISSN: 09641998 Issue: 2 Journal Abbreviation: Journal of the Royal Statistical Society. Series A (Statistics in Society).

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 Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne I. Osuchukwu, Salomey Osei, A. Tonja, Naome A. Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Leena G. Pillai, Kavya Manohar, Basil K. Raju, and Elizabeth Sherly. 2024. Multistage Fine-tuning Strategies for Automatic Speech Recognition in Low-resource Languages. ArXiv:2411.04573 [cs].

Kerri Prinos, Neal Patwari, and Cathleen A. Power. 2024. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1245–1254, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv:2212.04356.

Melissa Torgbi, Andrew Clayman, Jordan J. Speight, and Harish Tayyar Madabushi. 2025. Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the united kingdom.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset. ArXiv:2408.12734 [cs].

Ngoc Thang Vu, Yuanfan Wang, Marten Klose, Zlatka Mihaylova, and Tanja Schultz. 2014. Improving ASR performance on non-native speech using multilingual and crosslingual information. In *15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, Singapore, September 14-18, 2014*, pages 11–15. ISCA.

Ömer Özyilmaz, Matt Coler, and Matias Valdenegro. 2025. Overcoming data scarcity in multi-dialectal arabic asr via whisper fine-tuning. pages 1158–1162.