Fairness in Automatic Speech Recognition Isn't a One-Size-Fits-All

Hend ElGhazaly, Bahman Mirheidari, Heidi Christensen, Nafise Sadat Moosavi
School of Computer Science, The University of Sheffield
United Kingdom
{helghazaly1|b.mirheidari|heidi.christensen|n.s.moosavi}@sheffield.ac.uk

Abstract

Modern Automatic Speech Recognition (ASR) systems are increasingly deployed in high-stakes settings, including clinical interviews, public services, and educational tools, where equitable performance across speaker groups is essential. While pre-trained speech models like Whisper achieve strong overall accuracy, they often exhibit inconsistent group-level performance that varies across domains. These disparities are not fixed properties of the model, but emerge from the interaction between model, data, and task—posing challenges for fairness interventions designed in-domain. We frame fairness in ASR as a generalisation problem. We fine-tune a Whisper model on the Fair-Speech corpus using four strategies: basic fine-tuning, demographic rebalancing, gender-swapped data augmentation, and a novel contrastive learning objective that encourages gender-invariant representations. We evaluate performance across multiple aspects of fairness and utility, both in-domain and on three out-of-domain test sets: LibriSpeech, EdAcc, and CognoSpeak. Our findings show that the method with the best in-domain fairness performed worst out-of-domain, illustrating that fairness gains do not always generalise. Demographic balancing generalises more consistently, while our contrastive method offers a practical alternative: it achieves stable, cross-domain fairness improvements without requiring changes to the training data distribution, and with minimal accuracy trade-offs.

1 Introduction

Modern Automatic Speech Recognition (ASR) models are no longer trained from scratch for each task; instead, foundation models like Whisper (Radford et al., 2023) are deployed across domains with minimal fine-tuning. While this shift has led to impressive gains in average accuracy, it poses a new challenge for fairness. Interventions that previously relied on extensive task-specific training must now operate within a narrow fine-tuning window, and it remains unclear whether fairness improvements observed in one domain will hold under distributional shift.

In this work, we explicitly frame fairness as a generalisation problem. We fine-tune Whisper-small on Fair-Speech, a diverse dataset of 593 speakers and 26k utterances (Veliche et al., 2024), using four adaptation strategies: (i) basic fine-tuning (FT), (ii) demographic rebalancing (FT-Balanced), (iii) gender-swapped voice conversion augmentation (FT-Augmented), and (iv) a novel contrastive learning objective (CL) that we propose to explicitly encourage gender-invariant representations while preserving linguistic content. To our knowledge, this is the first use of contrastive learning for gender fairness in ASR. We evaluate both overall performance and fairness in-domain and across three out-of-domain test sets: LibriSpeech (read speech), EdAcc (Sanabria et al., 2023) (accented conversational English), and CognoSpeak (Pahar et al., 2025) (task-oriented clinical interviews).

Across the four adaptation strategies we evaluate, spanning both data-level and representation-level interventions, we find that fairness must be treated as both a multi-metric and multi-domain property. Reducing demographic gaps is not meaningful if it comes at the cost of degraded overall accuracy, or if improvements fail to persist under a distribution shift. This trade-off is most clearly illustrated by basic fine-tuning, which achieves the lowest gender gap in-domain but results in the worst outcomes in out-of-domain. This contrast demonstrates a central point of this work: methods that appear most fair in-domain can, under distribution shift, become the least fair overall. Our results highlight the risk of treating fairness as a local optimisation problem, rather than a generalisation challenge.

Other approaches show more promise. Demographic balancing generalises more consistently

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19169–19178
November 4-9, 2025 ©2025 Association for Computational Linguistics

across test sets, though its effectiveness may depend on the demographic diversity of the training data. Our proposed contrastive learning objective offers a complementary perspective: rather than training directly on synthetic data, we use gender-swapped utterances to construct contrastive pairs that encourage gender-invariant representations. This representation-level regularisation achieves competitive fairness improvements with minimal impact on performance, while avoiding modifications to the supervised training distribution.

To summarise, this work makes three **main contributions**: (1) we frame ASR fairness as a generalisation problem and provide the first evaluation across multiple domains; (2) we compare data-level and representation-level fairness interventions under a unified setup, including a novel contrastive objective that promotes gender-invariant representations; and (3) we demonstrate that in-domain fairness gains often fail to generalise, and argue for evaluating fairness interventions under distribution shift, using both performance and disparity metrics across domains.

1.1 Previous work

Recently, there have been increasing research efforts focused on addressing biases in speech recognition systems. Numerous works highlight disparities in Word Error Rates (WER) across different speaker genders[1] (Feng et al., 2021; Zanon Boito et al., 2022; Meng et al., 2022; Maison and Estève, 2023; Feng et al., 2024; ElGhazaly et al., 2025). Early studies found that ASR models exhibit higher WER for female speakers compared to male speakers, often attributed to training data imbalances and acoustic variations between genders (Tatman, 2017; Koenecke et al., 2020). To mitigate these disparities, various strategies have been explored, including data augmentation (Geng et al., 2020; Fucci et al., 2023; Zhang et al., 2023), adversarial learning (Zhang et al., 2018; Gorrostieta et al., 2019; Peri et al., 2023), and custom loss functions (Chang and Chen, 2022; Koudounas et al., 2024; Tang et al., 2024). Among these, enhancing training data, either through data augmentation or by constructing gender-balanced datasets, remains one of the most widely used techniques for improving ASR robustness and addressing demographic imbalances. Augmentation techniques have been employed to enhance model robustness by synthesising speech from under-represented speaker groups, thereby reducing bias in ASR predictions (Dheram et al., 2022; Peri et al., 2023; Zhang et al., 2023). While data augmentation can expand training diversity, it is not a perfect solution. Collecting balanced real-world data is often impractical due to resource constraints, while synthetic speech generation introduces artefacts that may not fully capture natural speaker variability. Directly training on synthetic data may also introduce biases inherited from the text-to-speech (TTS) system, potentially affecting ASR generalisation. These challenges highlight the need for alternative strategies beyond data manipulation to explicitly reduce demographic disparities in learned representations. Similarly, adversarial learning is a promising technique for mitigating ASR gender bias, but it has several limitations. Training these models can be unstable and require extensive hyperparameter tuning (Goodfellow et al., 2014). Furthermore, the adversarial component relies on effectively identifying and isolating speakers' attributes (Sun et al., 2018; Li et al., 2021), which can be difficult and can potentially harm overall accuracy (Tripathi et al., 2018) and reduce utility (Peri et al., 2023).

Prior work in NLP and computer vision showed promising results in achieving fairer representations and reducing biases using contrastive learning (Cheng et al., 2021; Shen et al., 2021). The use of contrastive learning in speech applications has been explored recently in improving robustness and representation learning. For example, Chang and Chen (2022) proposed a contrastive learning framework to align ASR outputs with manual transcripts, reducing error propagation in spoken language understanding. While prior work has leveraged contrastive learning for ASR robustness, its potential for bias mitigation remains unexplored. Our work introduces a novel application of contrastive learning to reduce gender-based disparities in ASR models. Instead of aligning ASR outputs with clean text representations, we minimise the embedding distance between male and female utterances of the same content, encouraging the model to learn gender-invariant speech representations. This approach aims to explicitly reduce demographic biases while preserving recognition performance, offering a promising direction for fairness-aware ASR optimisation.

[1] In this study, we analyse the performance differences between men and women, acknowledging that the gender spectrum is more diverse. Our focus on these two genders is driven by their representation in the dataset used for our investigation.

2 Methodology

2.1 Fairness-aware fine-tuning

In this work, we explore four fairness-oriented adaptation strategies: (i) basic fine-tuning on the original dataset (FT); (ii) fine-tuning on a demographically balanced subset (FT-Balanced); (iii) fine-tuning on more data using voice conversion augmentation (FT-Augmented) (Fucci et al., 2023); and (iv) adding to the standard cross-entropy loss a novel contrastive loss (CL) that we propose to enhance the model's fairness.

Our approach integrates voice conversion (VC)-based data augmentation with contrastive learning to encourage gender-invariant speech representations and hence mitigate gender bias in ASR models. The key idea is to expose the ASR model to the same textual content spoken by different genders, ensuring that it focuses on semantic information rather than speaker characteristics. To achieve this, we synthesise speech from an opposite-gender voice while preserving the original content using the XTTS-v2 model (Coqui, 2024). The XTTS voice generation model converts the specified speaker of the source audio file to the speaker of the target audio. The target audio contains different speech content, and the speaker's gender is the opposite of the source speaker. Thus, the generated augmented speech contains the content of the source with the voice of the target speaker. Both the source and target audio files are from the original dataset. This augmentation enables the model to encounter identical transcriptions from male and female speakers, reinforcing content-based rather than gender-dependent learning.

We fine-tune a pre-trained ASR model, where optimisation is guided by a combined loss function that includes cross-entropy loss for speech recognition accuracy and contrastive loss to enforce gender-invariant feature learning. Contrastive loss encourages the model to pull together embeddings of identical transcriptions spoken by different genders while maintaining discrimination between unrelated samples (Figure 1). The fine-tuned model is then evaluated for both performance (WER) and fairness (gender WER gap), ensuring that bias reduction does not degrade ASR accuracy, in- and out-of-domain.

2.2 Contrastive learning

Contrastive learning has been widely used to learn feature representations by bringing similar samples closer together while pushing dissimilar ones apart in the embedding space (Chopra et al., 2005; Hadsell et al., 2006). This technique has demonstrated effectiveness in various applications, including computer vision, natural language processing, and speech processing (Chen et al., 2020; Jaiswal et al., 2020; Han et al., 2021; Chang and Chen, 2022; Petrak et al., 2023; Koudounas et al., 2024). In this work, we leverage contrastive learning to mitigate gender-based disparities in speech recognition models, ensuring that utterances with the same linguistic content are represented similarly regardless of the speaker's gender. To achieve this, we introduce a novel approach to contrastive learning by strategically selecting positive and negative sample pairs within each training batch based on both gender and textual content. Our method aims to optimise model fairness, reducing performance gaps between genders while *maintaining good overall performance* (Wang and Liu, 2021; Chang and Chen, 2022).

In the experiments, positive pairs are constructed using gender-swapped utterances of the same content, i.e., male and female speakers saying the same sentence. The negative pairs are drawn from different-content utterances within each batch. We sample three negatives per anchor at random, as illustrated in Figure 1. Our choice of 3 negatives was empirically validated during development: increasing beyond 3 did not yield measurable gains in performance or fairness metrics, while reducing the number weakened the training signal. This formulation ensures that the model learns to differentiate based on speech content rather than speaker characteristics. The contrastive loss (CL) is added to the cross-entropy loss (CE) to form the total optimisation objective:

$$\mathcal{L}_{\text{total}} = (1-\alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{CL} \tag{1}$$

where α is a weight balancing the contributions of the two loss components. To determine the optimal α value, we analysed the trade-off between reducing the WER gap and maintaining a low overall WER on the validation set across epochs. We report the results for α values of 0.2 and 0.05 to evaluate the weight's effect on the efficacy of contrastive loss. The contrastive loss is defined as:

$$\mathcal{L}_{CL} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

where z_i and z_j are the embeddings of the positive pairs, male and female utterances of the same content, while z_k refers to the negative samples in each training batch, consisting of utterances with different content. The function $\mathrm{sim}(u, v)$ computes the cosine similarity between two embeddings, and (τ) is the temperature parameter, set to 0.1 as in Baevski et al. (2020).

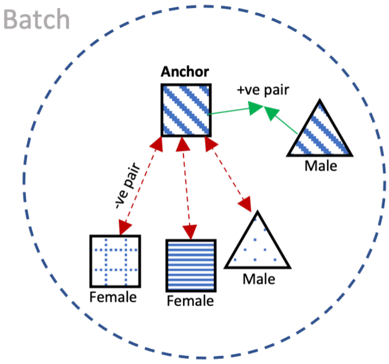

Figure 1: Proposed contrastive learning approach. Representations of the anchor and its gender-swapped augmentation (same text content) are pulled together (positive pair), while representations of other utterances in the batch are pushed apart (negative pairs)

2.3 Evaluation metrics

We assess the overall model performance with the Word Error Rate (WER) (Klakow and Peters, 2002); the lower the WER, the better the performance. To evaluate the model's fairness with respect to gender, we quantified the performance disparity by computing the absolute difference in WER between male and female speakers (WER gender gap) following the standard methodology outlined in Feng et al. (2024). We consider an ASR model to be fair if it recognises both genders equally well, i.e. if it has a small WER gap.

3 Experimental setup

3.1 Baseline model

We used the open-source pre-trained Whisper small model (Radford et al., 2023) as the baseline system for our experiments. Whisper small is a transformer-based model with 240M parameters and is trained on over 680,000 hours of weakly supervised labelled audio data from the web. It is widely adopted, as evidenced by its download frequency on Hugging Face's Model Hub.[2] Given this popularity and its computational efficiency, we selected Whisper small as our baseline model. We utilised OpenAI's Whisper Python package to load pre-trained checkpoints, ensuring consistency with the original model weights. The pre-trained checkpoint served as the starting point for fine-tuning. We optimised the learning rate and batch size for each model type based on the best WERs on the dev set. For fine-tuning the **FT** and **CL(α=0.05)** models, we used a batch size of 16 and a learning rate of 1.00E-05. For the **FT-Balanced**, **FT-Augmented** and **CL(α=0.2)** models, a batch size of 32 and a learning rate of 1.00E-07 were optimal. In all settings, training employed a cosine learning rate scheduler and weight decay to promote stable convergence and prevent overfitting.

3.2 Training data

We used the Fair-Speech dataset for both training and evaluation purposes. The Fair-Speech dataset is designed to support fairness-aware research in speech processing by providing labelled audio data with 6 speaker demographic attributes, such as gender and ethnicity (Veliche et al., 2024). It includes recordings from speakers with diverse linguistic and socio-demographic backgrounds, enabling the study of disparities in model performance.

We split the dataset randomly into 80% for training, 10% for development (dev), and 10% for testing (test), ensuring that speakers do not overlap across splits. Using this split, the train set had 21176 utterances in total, with 9639 male and 11537 female. This setup allowed us to monitor model performance and fairness metrics throughout the training process, while preserving a held-out test set for the in-domain evaluation. We used the whisper-normalizer Python library (Dettmers, 2023) to preprocess and normalise the original transcripts and model outputs.

We trained four model variants using different versions of the training set. The **FT** model was trained on the training set as it is. For the **FT-Balanced** model, we created a gender-balanced training set by sampling equal amounts of data from male and female speakers. For the **FT-Augmented** and **CL** models, we applied the data augmentation method described in Section2.1 to the original training split and used the generated augmented audio files as additional data in training. All models used the same dev set in validation for early stopping and hyperparameter tuning. The checkpoint that achieved the best performance on the dev set was then loaded to evaluate the fine-tuned models on the test set.

[2] Over the past month, Whisper small has been downloaded ~12,000 times, compared to 3,000 downloads for Whisper large.

3.3 Evaluation data

A key oversight in bias mitigation is ensuring that fairness gains are not merely artefacts of in-domain training but generalise to new and unseen data. If fairness interventions only reduce gender disparities on the in-domain dataset, they may not hold in real-world ASR scenarios. To address this, we evaluate the fine-tuned models on both **in-domain** and **out-of-domain** test sets. For in-domain evaluation, we use the Fair-Speech test set split.

To assess out-of-domain generalisation, we use three datasets from distinct contexts: LibriSpeech test-other (Panayotov et al., 2015), the Edinburgh International Accents of English Corpus (EdAcc) (Sanabria et al., 2023) and CognoSpeak (Pahar et al., 2025; Tao et al., 2025). The LibriSpeech test-other subset is a standard evaluation benchmark used to assess the robustness of speech models under more challenging acoustic and linguistic conditions (Panayotov et al., 2015). Unlike the test-clean subset, which features well-articulated speech with minimal background noise, test-other includes recordings with greater variability in speaker accents, pronunciation clarity, and recording quality. It is derived from audiobooks and represents more realistic and difficult test conditions. As such, test-other is commonly used to evaluate a model's ability to generalise beyond ideal scenarios, making it a valuable dataset for our out-of-domain evaluation.

The EdAcc is a publicly available speech dataset designed to support research in accent variation and robustness in speech processing systems (Sanabria et al., 2023). It contains recordings from virtual calls on Zoom between friends from a wide range of backgrounds, providing extensive coverage of non-native and regional accents. The corpus includes read and spontaneous speech in English, along with speaker metadata such as gender and linguistic background. EdAcc's diverse and controlled collection conditions make it particularly well-suited for evaluating model performance and fairness under domain shift.

CognoSpeak is an ongoing project aimed at remotely collecting audio and video recordings of individuals with cognitive decline through conversations with a computerised agent (Pahar et al., 2025). This agent prompts participants with a diverse set of clinically relevant questions and cognitive tasks. A subset of the data was used as a new challenge in ICASSP 2025, known as PROCESS (Tao et al., 2025). Over the course of 8 years, over 2000 recordings have been gathered from general practices and other clinical settings across the UK. For evaluation purposes, a carefully selected subset of the dataset—balanced for gender and matched for age range—has been curated as a fair test set. This includes 20 participants with healthy cognition, 20 with mild cognitive impairment, and 20 with dementia.[3]

We used the four datasets to evaluate our fine-tuned models on both in-domain and domain-shift settings. If the resulting ASR system is truly fair, gender fairness should persist even on out-of-domain evaluations rather than being confined to the training distribution.

4 Results and analysis

Table 1 presents the WERs per gender and absolute WER gender gap for each model evaluated on different test sets: the in-domain Fair-Speech test set, the test-other dataset from LibriSpeech, EdAcc and CognoSpeak. The baseline was the pre-trained Whisper-small model, which exhibits notable gender gaps, particularly in Fair-Speech and consistently worse performance (higher WER) with men speakers in all test sets. We evaluate and compare the models fine-tuned on the Fair-Speech dataset using our four different fine-tuning strategies, including (1) FT, (2) FT-Balanced, (3) FT-Augmented, and (4) CL with different α values (0.05 and 0.2). We highlight in green in Table 1 the improved performances; the more green a model has, the better and more generalisable it is. We discuss the main findings in the following sections.

4.1 Contrastive learning provides an effective generalisable solution

The results of our proposed models with contrastive learning (CL) demonstrate promising improvements in both performance and fairness across multiple evaluation datasets. The CL(α=0.05) model reduced the gender gap by half on the Fair-Speech test set, while also preserving reasonable performance. However, this model struggled with generalisation, particularly on the CognoSpeak dataset, where the gap and performance deteriorated substantially, indicating a potential over-fitting to the Fair-Speech domain. In contrast, contrastive loss

[3] Full details of the LibriSpeech, Fair-Speech, EdAcc and CognoSpeak corpora can be found in (Panayotov et al., 2015), (Veliche et al., 2024), (Sanabria et al., 2023), and (Pahar et al., 2025), respectively.

Model	Fair-Speech			LibriSpeech			EdAcc			CognoSpeak		
	W	M	G	W	M	G	W	M	G	W	M	G
Pre-trained	7.06	15.23	8.17	7.31	7.91	0.59	29.11	31.34	2.23	20.08	21.14	1.06
FT	5.22	8.39	↓3.17	12.57	13.09	↓0.52	39.06	40.83	↓1.77	43.59	42.23	↑1.36
FT-Balanced	5.71	10.14	↓4.43	7.95	8.78	↑0.83	26.33	29.13	↑2.80	16.58	17.54	↓0.96
FT-Augmented	5.45	9.90	↓4.45	8.41	9.31	↑0.90	26.80	29.48	↑2.68	17.60	18.65	↓1.05
CL(α=0.05)	5.24	9.31	↓4.07	11.18	12.40	↑1.22	36.82	37.67	↓0.85	46.26	43.11	↑3.15
CL(α=0.2)	5.61	10.08	↓4.47	8.16	8.98	↑0.82	26.33	29.08	↑2.75	19.27	20.80	↑1.53

Table 1: Evaluation results of the fine-tuned models across in-domain (Fair-Speech) and the three out-of-domain test sets. The fine-tuned models are: basic fine-tuning (FT), demographic balanced training data (FT-Balanced), more training data with augmentation (FT-Augmented), and our contrastive learning objective with different α values(CL). Gap (G) is the absolute difference between women's (W) and men's (M) WERs (values are in %). Green cells and downward arrows in (G) indicate better than the pre-trained model.

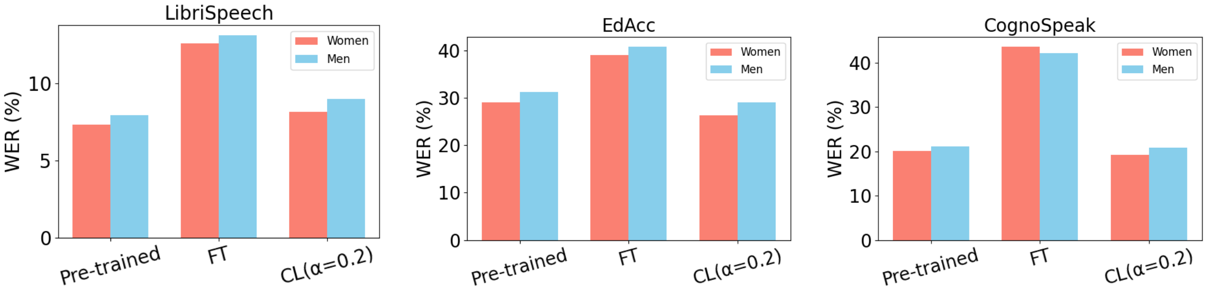

Figure 2: CL(α=0.2) model outperforms basic fine-tuning (FT) in out-of-domain evaluation.

with larger weight (CL(α=0.2)) maintained improved fairness on Fair-Speech (gap reduced by over 45% from 8.17 to 4.47) as well as EdAcc and CognoSpeak. In particular, the CL(α=0.2) model achieved the lowest WERs on the EdAcc dataset, and outperformed the basic fine-tuning (FT) on all the out-of-domain sets (Figure2). This indicates that contrastive learning is beneficial for enhancing the model's fairness while maintaining performance and improving generalisation.

4.2 Fairness interventions may not generalise under domain shifts

Although all fine-tuned models showed a substantial reduction in WER and gender gap on the Fair-Speech test set compared to the pre-trained model, these gains did not persist across different domains (Figure3). FT achieved the best WERs and over 60% reduced gender gap on the Fair-Speech dataset, indicating improved fairness within the training domain. However, these improvements disappeared when the data domain changed in LibriSpeech, EdAcc and CognoSpeak, suggesting that FT may overfit to the training data. On the other hand, FT-Balanced and FT-Augmented offered more balanced performance: both reduced gaps on CognoSpeak while maintaining or improving performance across domains. For example, FT-Balanced had the best results on CognoSpeak (lowest WERs and gap) while decreasing overall WERs in EdAcc by over 8%. The FT-Augmented model similarly showed strong fairness generalisation, achieving a smaller gap on Fair-Speech (4.45) and further gains on EdAcc and CognoSpeak, with gaps of 2.68 and 1.05, respectively. These results suggest that both FT-Balanced and FT-Augmented strategies lead to more robust and fair performance across domains, whereas FT primarily improves fairness in-domain at the expense of out-of-domain generalisability. The results of the FT, FT-Balanced, and FT-Augmented models suggest that the effectiveness of fine-tuning techniques on fairness is highly sensitive to domain shift.

4.3 Initial disparities impact fairness gains

A key insight from the results in Table1 is that fairness interventions are more effective when applied to a dataset with a large initial gender gap. This was particularly evident in the LibriSpeech dataset, which had worse WERs for both genders with the fine-tuned models. The pre-trained model

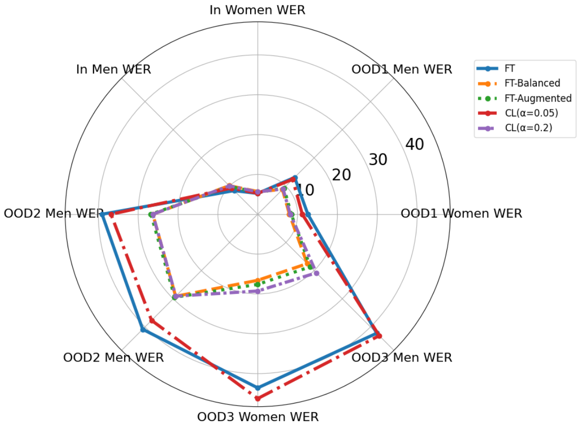

Figure 3: Inconsistent WERs on in-domain (In) and out-of-domain (OOD) test sets.

exhibited a low gender gap (0.59), which likely constrained the extent to which fine-tuning could further enhance fairness. Conversely, Fair-Speech started with a large WER gap of 8.17, allowing fine-tuning to have a greater impact. In addition, EdAcc and CognoSpeak had high WERs with the pre-trained model, which were further reduced by fine-tuning with FT-Balanced, FT-Augmented and CL(α=0.2). This suggests that fairness-aware fine-tuning is effective when applied to datasets with larger inherent gender disparities and WERs.

4.4 Fairness conclusions require reliable evaluation metrics

We reported the gap (G) in Table1 using the absolute difference that is frequently used as an indicator of disparity. While this is the most commonly used and intuitive metric, this measure does not necessarily reflect improved overall performance. The performance could be much worse for both genders, such as in the LibriSpeech evaluation. As shown in Table 1, the reduced gap values (downward arrow) do not always correspond with the improved performance (green cells). We therefore looked into other metrics that have been used in prior fairness research. The Word Error Rate Reduction (WERR) was suggested by Dheram et al. (2022) to quantify the relative improvement between systems on a cohort by comparing WERs before and after an intervention, normalised by the baseline WER. The more positive WERR indicates improvements on that cohort while negative WERRs indicate degradation. The WERR for the bottom cohort in our results (men) is defined as:

$$\mathrm{WERR_M} = \frac{\mathrm{WER\_Baseline_M} - \mathrm{WER\_Model_M}}{\mathrm{WER\_Baseline_M}} \tag{3}$$

Using this metric, the value is more reflective of a fair model with good performance. We hence computed the WERR on the men's WER (the bottom group) on all the test sets and compared them with the previously reported gap. Table2 summarises the results. We found that the WERRs clearly show all of the models deteriorated on the LibriSpeech test set. Although the absolute difference between genders' WERs (G) were small, the WERs were much higher than the pre-trained baseline model, therefore inaccurately showed improvements. Similarly, the CL(α=0.05) model showed almost equal WERs across both gender groups but suffered from high WER across all domains. For instance, in the EdAcc test set, the lowest gap was achieved by the CL(α=0.05) model (G=0.85) when in fact it worsened the performance on the disadvantaged group by over 20%. Whereas the CL(α=0.2) model achieved a balance between fairness and performance on the EdAcc test set relative to the other models, improving the baseline by 7.21%. The choice of metric can thus significantly affect the conclusions drawn about fairness and it is important to use a reliable fairness-specific measure.

4.5 Fairness is multifaceted

Fairness in speech recognition systems seems to be multidimensional and domain-dependent. It is influenced by a range of factors, including the domain of the evaluation dataset, the metric used to (e.g., WER improvements vs. group disparity measure), and the demographic subgroup being considered. Figure3 visualises the gender-based average WERs across the various test sets and fine-tuning approaches, illustrating efforts to improve fairness along one dimension might negatively impact performance along another dimension. None of the models consistently achieve good performance and fairness across all dimensions. For instance, FT significantly reduces the gender gap on the Fair-Speech dataset (from 8.17 to 3.17), demonstrating improved in-domain fairness. However, this improvement comes with a sharp increase in the WERs of EdAcc and CognoSpeak, suggesting poor generalisation of fairness gains. Moreover, the pre-trained model itself shows a very small gender gap on LibriSpeech (0.59), limiting further fairness improvement in that domain. This illustrates how

Model	Fair-Speech		LibriSpeech		EdAcc		CognoSpeak	
	G	WERR$_M$(%)	G	WERR$_M$(%)	G	WERR$_M$(%)	G	WERR$_M$(%)
Pre-trained	8.17	–	0.59	–	2.23	–	1.06	–
FT	**3.17**	**44.88**	**0.52**	-65.48	1.77	-30.29	1.36	-99.79
FT-Balanced	4.43	33.41	0.83	-11.06	2.80	7.04	**0.96**	**17.02**
FT-Augmented	4.45	34.96	0.90	-17.77	2.68	5.93	1.05	11.79
CL(α=0.05)	4.07	38.89	1.22	-56.78	**0.85**	-20.20	3.15	-103.96
CL(α=0.2)	4.47	33.82	0.82	-13.52	2.75	**7.21**	1.53	1.59

Table 2: WERR$_M$ offers more accurate evaluation than absolute difference. Best values per metric in bold.

initial model characteristics can constrain or shape the outcomes of fairness interventions. These results collectively emphasise that fairness cannot be meaningfully assessed using a single metric or dataset. A model that performs equitably in one domain may fail in another, and focusing solely on reducing group disparities may inadvertently sacrifice overall performance or fairness in other contexts. These findings underscore that no single approach universally outperforms the others; instead, the choice of method should be guided by the target domain and the desired outcome between fairness and generalisation. This underscores the critical need for multidimensional evaluation frameworks that account for domain, demographic subgroup, and performance trade-offs when developing and benchmarking fair speech recognition systems.

5 Conclusion

This work introduced fairness as a generalisation challenge in ASR, showing that interventions effective in one domain may not persist across others. We evaluated four strategies on the Whisper small model, comparing data-level and representation-level interventions across in-domain and out-of-domain benchmarks. Our findings show that in-domain fairness is not a reliable proxy for robustness, and that fairness must be assessed as a property of both model behaviour across groups and performance under distribution shift. Among the methods, demographic balancing showed the strongest cross-domain fairness but depends on access to balanced training data. Our proposed contrastive learning objective offers a complementary perspective: by encouraging gender-invariant representations using augmented data, it improves fairness without requiring changes to the supervised training distribution. Unlike data augmentation strategies that inject synthetic samples directly into the main loss, our contrastive objective isolates fairness regularisation, minimising the impact of potentially lower-quality generated data. Looking ahead, contrastive learning presents a promising direction for scalable, generalisable fairness in ASR. While this work focused on gender, the objective naturally extends to other protected attributes such as age, accent, or gender identity. By treating same-content utterances across these attributes as positive contrastive pairs, future models could be trained to align representations across multiple demographic axes simultaneously. In contrast, expanding data-level augmentation to cover all such dimensions would significantly increase training set size and compound quality concerns. Contrastive methods thus offer a principled path forward for multi-attribute fairness, especially in settings where demographic labels are available but rebalancing or augmenting each domain is infeasible.

6 Limitations

One limitation of our study is that the test sets used contain only binary gender labels, which do not capture the full spectrum of gender identities and may limit the generalisability of our fairness analysis. Additionally, while there exist numerous fairness metrics in the literature, we focused on a set of widely adopted metrics to maintain comparability with prior work. We acknowledge that other metrics might provide further insights and plan to explore them in future studies. Furthermore, our current analysis is restricted to gender-based fairness; evaluating disparities across other demographic factors such as age, race, and accents is an important direction for future research that we intend to pursue.

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.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive learning for improving asr robustness in spoken language understanding. In *Interspeech 2022*, pages 3458–3462.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 539–546.

Coqui. 2024. Xtts-v2. Accessed: August 25, 2024.

Tim Dettmers. 2023. whisper-normalizer: Text normalization for whisper transcripts. https://pypi.org/project/whisper-normalizer/. Version 1.0.7.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022*, pages 1268–1272.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring gender disparities in automatic speech recognition technology. *arXiv preprint arXiv:2502.18434*.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2023. No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. 2020. Investigation of data augmentation techniques for disordered speech recognition. In *Interspeech 2020*, pages 696–700.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. 2019. Gender de-biasing in speech emotion recognition. In *Interspeech 2019*, pages 2823–2827.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1735–1742.

Tao Han, Hantao Huang, Ziang Yang, and Wei Han. 2021. Supervised contrastive learning for accented speech recognition. *arXiv preprint arXiv:2107.00921*.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Allison Koenecke, Joon Sung Nam, Elizabeth Lake, Joseph Nudell, Michael Quartey, Zelalem Mengesha, Cody 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.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024. A contrastive learning approach to mitigate bias in speech models. In *Interspeech 2024*, pages 827–831.

Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. 2021. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*.

Lucas Maison and Yannick Estève. 2023. Some voices are too common: Building fair speech recognition systems using the commonvoice dataset. In *Interspeech 2023*, pages 4428–4432.

Yen Meng, Yi-Hui Chou, Andy T Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3258–3262. IEEE.

Madhurananda Pahar, Fuxiang Tao, Bahman Mirheidari, Nathan Pevy, Rebecca Bright, Swapnil Gadgil, Lise Sproson, Dorota Braun, Caitlin Illingworth, Daniel Blackburn, and 1 others. 2025. Cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*, pages 1–7. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. 2023. A study of bias mitigation strategies for speaker recognition. *Computer Speech & Language*, 79:101481.

Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. 2023. Arithmetic-based pretraining improving numeracy of pretrained language models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 477–493, Toronto, Canada. Association for Computational Linguistics.

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.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR. In *ICASSP 2023*.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. 2018. Domain adversarial training for accented speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4854–4858. IEEE.

Jiyang Tang, Kwangyoun Kim, Suwon Shon, Felix Wu, and Prashant Sridhar. 2024. Improving asr contextual biasing with guided attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12096–12100. IEEE.

Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Sophie Young, Yao Xiao, Hend Elghazaly, Fritz Peters, Caitlin Illingworth, Dorota Braun, Ronan O'Malley, and 1 others. 2025. Early dementia detection using multiple spontaneous speech prompts: The process challenge. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 53–59.

Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh. 2018. Adversarial learning of raw speech features for domain invariant speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5959–5963. IEEE.

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. In *Interspeech 2024*, pages 1385–1389.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. In *Interspeech 2022*, pages 1278–1282.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.