Towards Fair Speech Recognition: Mitigating Demographic Bias in End-to-End ASR Systems

Maliha Jahan[1], **Thomas Thebaud**[1], **Zsuzsanna Fagyal**[2], **Jesus Villalba**[1], **Mark Hasegawa-Johnson**[2], **Laureano Moro-Velazquez**[1], **Najim Dehak**[1]
[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]University of Illinois Urbana-Champaign, Champaign, IL, USA

Abstract

Demographic bias in the performance of speech and language technology has been an active area of recent research. A lot of studies have shown the existence of demographic biases in Automatic Speech Recognition (ASR) systems. In this work, we propose a novel model-agnostic and demographic label-agnostic approach, called DARe, to mitigate any existing bias in an ASR system towards certain speaker groups. We built a debiasing module that goes between the feature extractor of an ASR and the rest of that ASR. The module includes content-group disentanglers to separate content and group, a demographic classifier, and adversarial reweighting. To eliminate the need for demographic labels, we generated pseudo-group labels by extracting speaker embeddings and clustering them. We worked with three ASR systems–Wav2Vec2 base, SEW tiny, and Whisper small. We used the FAI dataset, which contains naturalistic conversations with speakers who self-identify their demographic attributes. We used Word Error Rate (WER) as a metric of ASR performance and a Poisson regression-based approach to evaluate the racial fairness of the models. We compared the racial bias of the models before and after applying our proposed approach and observed a significant improvement in fairness.

Keywords: fairness, demographic bias, automatic speech recognition (ASR), poisson regression, drop-in-deviance testing, disentanglement, reweighting, clustering; model-agnostic mitigation, word error rate (WER).

1. Introduction

Speech and language technology have seen notable progress in recent years. However, they continue to struggle with fairness across speaker demographics. These demographics include individuals who share attributes such as gender identity (e.g., male, female, non-binary), native language, and other relevant characteristics. A growing body of research highlights the presence of demographic biases in ASR systems, spanning gender, racial, social, and cultural dimensions. For example, Koenecke et al. (Koenecke et al., 2020) found that five major ASR systems—Amazon, Apple, Google, Microsoft, and IBM—performed disproportionately poorly on Black speakers who use African American Vernacular English (AAVE) (Thomas, 2007; Bailey and Thomas, 2021). They attributed these disparities primarily to acoustic features rather than language model components. Similarly, Feng et al. (Feng et al., 2021) observed biases related to gender, age, and accent in Dutch ASR systems. Other studies (Tatman, 2017; Tatman and Kasten, 2017; Martin and Tang, 2020; Wu et al., 2020) have further documented accent and racial biases. Regarding gender bias specifically, some systems show improved performance for female speakers (Adda-Decker and Lamel, 2005; Sawalha and Shariah, 2013; Koenecke et al., 2020), while others demonstrate a preference for male voices (Garnerin et al., 2019; Tatman, 2017; Garnerin et al., 2021), indicating that such biases can vary depending on the system and context.

We propose a novel model-agnostic demographic label-agnostic bias mitigation approach to address the bias present in ASR systems. Our approach starts with extracting speaker embeddings from the speech and clustering them to attain cluster labels that obviate the need for demographic labels. We experimented with several speaker encoders to choose the best one. Then, we propose a debiasing module that goes between the feature extractor and the rest of the ASR system. We experimented with three ASR systems, Wav2Vec2 base, SEW tiny, and Whisper small, on the FAI dataset (Jahan et al., 2024). We used a Poisson regression-based approach (Liu et al., 2022b) to evaluate the fairness of the ASR systems. We performed the drop-in-deviance test and used the p-value as a fairness indicator. The experiments show that the bias mitigation approach increases the p-values, indicating improved fairness, while maintaining a similar or lower WER. The following are the contributions of this paper.

- We introduce DARe, a bias mitigation method that:
  - Is **Model-agnostic**, which allows the use of this approach to mitigate any existing demographic bias independently from the ASR model at hand.
  - Is **Demographic label-agnostic**, which makes the fairness application to an ASR independent from the dataset's labels; it

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4115–4125
11–16 May 2026.

can be done using our method, regardless of whether the dataset has any demographic information available.

  - Correct bias **in multiple attributes and their intersectionality**, since we do not focus on any particular attribute, and rather use cluster labels.

- We experiment with several systems for speaker encoders and show their performance in capturing demographic information and mitigating biases.

2. Previous Works

In recent years, numerous efforts have been undertaken to address the various forms of bias present in ASR systems. These approaches span model-based, data-based, and hybrid strategies.

Data Augmentation: One line of work focuses on various kinds of data augmentation. For example, (Sarı et al., 2021) proposed a method that mitigates bias by generating counterfactual data latent representations disentangled from protected attributes (e.g., gender, age). The phonetic content is preserved while the protected attribute is modified using group-average spectrograms. The model is trained to minimize the difference between logits from factual and counterfactual pairs. Their method reduces the Male-Female WER gap by 1.5 when augmenting gender data. However, the use of average spectrograms to represent groups can result in embeddings that do not sound realistic, often perceived as synthetic or "dummies" by listeners.

Similarly, (Zhang et al., 2022) used autoencoder-based cross-lingual voice conversion (Chen et al., 2021) to simulate non-native accents in Dutch. They explored two training strategies: fine-tuning on augmented data and domain adversarial training (DAT) (Sun et al., 2018). DAT involved adversarial learning to distinguish native from non-native accents. Their best model reduced the non-native bias, measured by WER difference, for read data from 28.83% to 13.59%, and for human-machine interaction data from 29.49% to 12.62%. However, voice conversion systems may fail to fully capture the diversity of accents and may not preserve all linguistic nuances. They used the cosine speaker similarity method to find source speakers and target speakers (for voice conversion) that are relatively similar. This may lead to limited diversity in terms of the intelligibility of the accented speech.

Another approach leveraging signal-level augmentation is presented by (Fucci et al., 2023), who simulate female speech by manipulating F_0 and formants, features known to differentiate male and female voices (Coleman, 1976; Hillenbrand and Clark, 2009). Their method yielded up to a 9.87% relative WER reduction for female speakers. While potentially effective, this approach risks reinforcing gender stereotypes by treating certain acoustic changes as universally representative of gender, and may overlook within-group variability (Merritt and Bent, 2022).

Speaker Group Based Approaches: Other works have explored the use of clustering and embeddings to implicitly represent speaker characteristics. Authors of (Veliche and Fung, 2023) perform unsupervised clustering on acoustic data and use the resulting cluster IDs as features in an RNN-T model with an Emformer encoder (Shi et al., 2021). During training, they mask cluster IDs in randomly selected samples to teach the model to infer this information, while during inference, they assign all samples to an 'unknown' class to preserve speaker privacy. Using their approach, the performance of male speakers (under-represented) in the Casual Conversation(Liu et al., 2022a) dataset saw a statistically significant performance gain of 11.17%. They also showed improvements of 5% relative WER on average across demographic groups in the Voice Command dataset, which they collected. This method risks penalizing underrepresented clusters if the model prioritizes overall performance.

Relatedly, some methods embed explicit group or cohort information into the model. Authors of (Dheram et al., 2022) introduced one-hot cohort embeddings (e.g., "bottom" vs. "top" performance groups) concatenated to acoustic features, improving equity across cohorts. This method reduced the relative WER gap between the cohorts by 31.6%. Still, this technique is limited in scalability, as it only handles binary groupings and would need redesign for continuous or multi-class attributes.

Data Balancing Techniques: A straightforward mitigation strategy is balancing training data across demographic groups. For instance, authors of the study (Ezema et al., 2025) fine-tuned OpenAI's Whisper ASR on speech data balanced across Black and White speakers, reducing WER for both groups and narrowing the performance gap from 37% to 26% relative difference. Likewise, (Dheram et al., 2022) simulated WER-based cohorts and applied oversampling of underperforming groups in a semi-supervised framework, achieving a 17.9% relative WER gap reduction between cohorts. While effective, these strategies require group labels and are dependent on known demographic distinctions.

Fairness-Aware Regularization: Adversarial training has also been applied to reduce bias in ASR. (Na and Park, 2021) employed Domain-Adversarial Neural Networks to build accent-invariant ASR systems, reducing disparities between standard and accented speech. Their approach decreased the WER for British and Indian

accents by 5% and 3%. However, tuning the adversarial strength (λ) is crucial— excessive regularization can degrade performance. Moreover, this approach typically relies on known domain labels (e.g., accent, gender), limiting generalizability to unseen or intersectional groups.

Data-Centric Approaches: Some bias mitigation efforts focus on data curation and evaluation. (Meyer et al., 2020) proposed an expert-validated subset of the Common Voice dataset (Mozilla) containing demographic information. They used this dataset for finetuning, which did not improve WER difference for gender (0.08 increase in gap) but did improve for accent (5.67 decrease in gap). While designed to evaluate ASR bias, the dataset includes sparse and often missing demographic labels, especially for accents, and lacks controls for intra-group speaker diversity. Similarly, (Liu et al., 2022a) released the Casual Conversations dataset—an 846-hour corpus labeled by age, gender, and skin tone. Although they achieved lower overall WER after fine-tuning models on this dataset, the subgroup disparities did not decrease. These data-driven methods require significant resources for data collection, validation, and annotation.

Beyond data quantity, diversity within groups is also critical. The studies (Feng et al., 2024, 2021) differentiate between direct (balancing dataset size) and indirect (increasing diversity within groups) methods of bias reduction. However, both approaches demand extensive data engineering.

Finally, there is growing recognition that bias is not purely a data issue. Authors of (Markl, 2022) emphasize the importance of treating datasets as infrastructure. Drawing from (Paullada et al., 2021; Bender and Friedman, 2018; Hutchinson et al., 2021), they argue for documentation practices that anticipate and mitigate bias, including annotator diversity and awareness of differing values. Furthermore, (Hooker, 2021) stresses that model architecture itself can introduce or amplify bias, highlighting the need for holistic evaluation beyond just data collection.

In our method, we try to tackle the bias introduced by ASR model architectures. The existing model-based approaches are often tailored for a particular model, our proposed approach will work for any end-to-end model. Our method also does not require specific labels. Also, several of the discussed works only show improvement in WER of each demographic group compared to the baseline, overlooking the relative improvement and the statistical significance. Through our experiments, we make sure the demographic bias is reduced and the reduction is statistically significant.

3. Materials: Data and Pre-trained ASR Models

We used the following three ASR systems from Hugging Face[1] for our experiments.

- **Wav2Vec 2.0 Base (Baevski et al., 2020) (w2v2):**
  Model Name: facebook/wav2vec2-base-960h
  Number of Parameters: ~95 million
  Pretraining Dataset: LibriSpeech (960 hours of labeled speech for fine-tuning) and Libri-Light (53K hours of unlabeled audio for pretraining)
  Description: Developed by Facebook AI, this model is based on the Wav2Vec 2.0 architecture, which learns speech representations in a self-supervised fashion from raw audio. It achieves strong performance in ASR tasks with minimal labeled data.

- **Whisper Small (Radford et al., 2023) (whisper):**
  Model Name: openai/whisper-small
  Number of Parameters: ~244 million
  Pretraining Dataset: 680,000 hours of multilingual and multitask supervised data collected from the web
  Description: Whisper is OpenAI's sequence-to-sequence ASR model trained on a large and diverse dataset. It supports multilingual transcription, translation, and language identification. The "small" version offers a good balance between accuracy and efficiency.

- **SEW-Tiny (Wu et al., 2022) (sew):**
  Model Name: asapp/sew-tiny-100k-ft-ls100h
  Number of Parameters: ~6 million
  Pretraining Dataset: Pretrained on 100k hours of unlabeled audio from VoxPopuli; fine-tuned on LibriSpeech 100h
  Description: SEW stands for Speech-Equivalent Wav2Vec. This "Tiny" version is highly compact and efficient, targeting resource-constrained environments. Despite its small size, it delivers solid ASR performance due to effective self-supervised pretraining.

Dataset: to adapt and evaluate the ASRs, we used the FAI dataset described in (Jahan et al., 2024). This is an English speech corpus featuring 526 speakers (145.2 hours of audio) who self-identify or are identified by others in real-life conversation as belonging to various racial, ethnic, and national origin groups. The available labels include race, ethnicity, and national origin of the identified speaker, the type of identification (by speaker or

[1] https://huggingface.co/

by someone else), and the sentence or segment where the identification was made. For this study, we focused on the Asian, White, and Black racial categories. We used 75% of the data for training, 12.5% for the development, and 12.5% for testing. The development and test speakers are not present in the training split.

4. Proposed Method

The proposed approach involves attaching a debiasing module, DARe, between the feature extractor and the rest of the ASR. The input can be raw wave or spectrograms, depending on the ASR system. The method overview is shown in Figure 1. The code for the debiasing module is available on GitHub[2].

4.1. Demographic clustering

We extract speaker embeddings (Z_S) from the speech using a pretrained speaker encoder system. Then we perform K-means clustering to cluster the embeddings. The value of K is determined using the elbow method. To ensure that the speakers from the same demographic group end up in the same cluster, we experimented with the following seven speaker embedding systems and chose the one that captures the most demographic information.

1. Microsoft WavLM-Base for Speaker Verification, pretrained on 960 hours of LibriSpeech; fine-tuned on VoxCeleb1.

2. SpeechBrain X-Vector Speaker Recognition, pretrained on VoxCeleb1 and VoxCeleb2.

3. SpeechBrain ECAPA-TDNN Speaker Recognition, pretrained on combined VoxCeleb1 and VoxCeleb2 datasets.

4. SpeechBrain ResNet TDNN Speaker Recognition, pretrained on combined VoxCeleb1 and VoxCeleb2 datasets.

5. NVIDIA TitaNet-Large for Speaker Verification, pretrained on combined VoxCeleb1, VoxCeleb2, Fisher, Switchboard, LibriSpeech, and NIST SRE (2004–2010).

6. Facebook Wav2Vec2-Large-XLSR-53, pretrained on 53,000 hours of multilingual data from the CommonVoice dataset.

7. Facebook HuBERT-Large (LL-60k), pretrained on 60,000 hours of Libri-Light.

[2]https://github.com/Maliha-Jahan/DARe

Then we used the cluster labels as demographic labels. This way, we can obviate the use of specific demographic labels like gender and race and target multiple demographic attributes and their intersectionality for bias mitigation.

4.2. Encoders and disentangler

We extracted two embeddings- content embedding and group embedding, from the speech, and disentangled them. The content embedding (Z_C) represents what is being said, and the group embedding (Z_G) represents the demographic information in the signal. To extract the embeddings, we have a content and a group encoder. The ASR model's feature extractor is used as the content encoder, which is optimized through disentangling to provide only content and remove group information. The group encoder has the same architecture but different (non-shared) weights, followed by a pooling layer. We used disentanglers to separate content from group information. The disentangler $G2C$ tries to predict the content embedding (Z'_C) from the group embedding (Z_G), and the disentangler $C2G$ does the opposite. The goal is to make Z_C uninformative about Z_G, and vice versa. We achieved this via a min-max optimization: disentanglers are trained to improve prediction performance, while the encoders are trained to make the embeddings less predictable to the disentanglers, thus enforcing disentanglement in the encoders. We took the Mean Squared Error (MSE) loss between the predicted and actual embeddings and added it to the total loss, which is explained in Section 4.5.

4.3. Cluster classifier

The content embedding ends up with content information because the content encoder is trained with the ASR loss. To ensure that the group embedding ends up with group information, we used a cluster classifier. The classifier predicts cluster labels from group embeddings, forcing the group embeddings to learn to distinguish between the clusters. During training, we add the cross-entropy loss from this classifier to the total loss. The reason why we need to ensure that group embeddings capture group information is so that when the group and content embeddings are disentangled, the group information captured by the group embedding will be removed from the content embedding.

4.4. Adversarial reweighting

Taking inspiration from (Jin et al., 2022), a work on speaker verification fairness, we used adversarial reweighting for ASR fairness. We concatenated the cluster labels with the content embeddings and predicted weights for each sample using

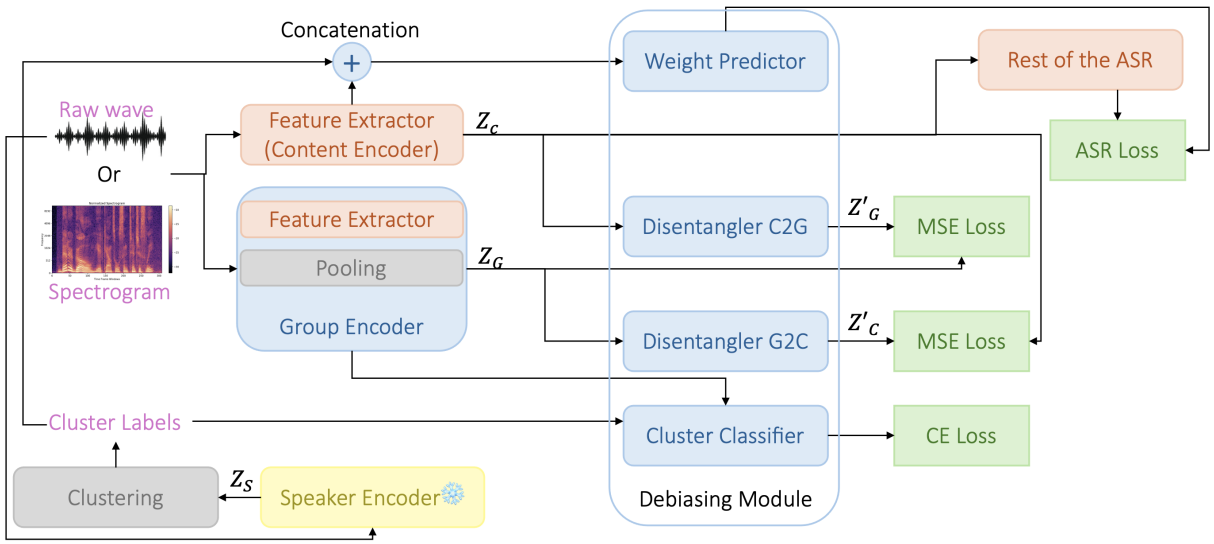

Figure 1: Overview of the proposed approach. Inputs are shown in pink, debiasing-related submodules in blue, ASR components in orange, speaker encoder in yellow, and losses in green.

a weight predictor. Then we multiplied the ASR loss of each sample by the weights before aggregating the loss. The purpose of the weights is to penalize the ASR more for getting the underrepresented samples wrong. We use min-max optimization to minimize total loss with respect to the ASR parameters and maximize the total loss with respect to the weight predictor parameters. So, the weight predictors will produce weights that maximize the loss, generating larger weights for samples that are harder to recognize and smaller weights for easier samples. The ASR will focus more on the harder samples since those samples have higher weights.

4.5. Loss function

The loss function has three components:

1. **ASR term:** The ASR loss can be CTC loss or sequence-to-sequence loss, based on the ASR system.

$$\min_{\theta_{ASR},\phi_c} \max_{\psi} \mathcal{L}_{ASR}(f_c(x;\phi_c), y; W(\psi)) \quad (1)$$

2. **Classifier term:** The classifier term is the cross-entropy loss of the cluster classifier.

$$\min_{\theta_{Clf},\phi_g} \mathcal{L}_{Clf}(f_g(x;\phi_g), d) \quad (2)$$

3. **Disentangler term:** The disentangler loss is the sum of the two Mean Squared Error (MSE) losses between the two embeddings and their predictions.

$$\min_{\theta_{D\_C2G},\theta_{D\_G2C}} \max_{\phi_c,\phi_g} \mathcal{L}_{Dis}(f_c(x;\phi_c), f_g(x;\phi_g)) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_{ASR} + \lambda_{Clf}\mathcal{L}_{Clf} + \lambda_{Dis}\mathcal{L}_{Dis} \quad (4)$$

In the above equations, θ_{ASR}, θ_{Cls}, θ_{D_C2G}, and θ_{D_G2C} are the parameters of the ASR, the classifier, and the disentanglers of content from the group and group from the content. ϕ_c and ϕ_g are the parameters of the content and group encoders. ψ indicates the parameters of the weight predictor. $\mathcal{L}_{ASR}$, $\mathcal{L}_{Clf}$, and $\mathcal{L}_{Dis}$ are the ASR, the classifier and the disentangler loss. x, y, and d are the inputs, the transcriptions, and the demographic labels. W, f_c, and f_g are the weight predictor, content encoder, and group encoder. λ_{Clf} and λ_{Dis} are the contributions of the classifier loss term and the disentangler loss term to the total loss. These weights start very small and are gradually increased using the following formula to slowly introduce the fairness components into the loss.

$$\lambda_{\text{effective}} = \lambda * \frac{1}{1 + \exp(-(\text{epoch} - \text{mid\_point}))} \quad (5)$$

Here, λ is either of the two weights. $\lambda_{effective}$ is the effective value of the weight being used in the current epoch. $epoch$ is the current epoch, and mid_point is the epoch where we want to apply half of λ as the effective weight.

5. Experiments

5.1. Speaker encoder selection experiment

To extract speaker embeddings that are rich in demographic information, we performed experiments on several speaker encoder systems to choose the best one. We extracted speaker embeddings from the FAI train split using each of the systems

described in section 4.1, then we calculated the silhouette score, the within-group and between-group distances, and performed PLDA classification. We used the race labels to calculate these scores. The silhouette score is a metric that evaluates the quality of clustering. It indicates how the samples in each cluster, or race group in our case, are similar to their own cluster than to the other clusters. Silhouette scores can range from -1 to 1, where higher values indicate better-defined and more distinct clusters. Within-group and between-group distances are calculated across the race groups. We calculate the ratio (ρ) of the between-group distance to the within-group distance. The higher this ratio, the better the system is as our speaker encoder. Finally, we trained a PLDA (Prince and Elder, 2007) to classify the races and measure the prediction accuracy. For classification using PLDA, we used an 80-20 train-test split of the FAI training data. We used these scores to determine which embeddings contain the most race information.

5.2. ASR fairness evaluation metric

To evaluate the fairness of the ASR systems, we used the approach proposed in (Liu et al., 2022b). They proposed that the number of word errors in the j^{th} utterance by speaker i, E_{ij}, is a Poisson random variable with rate parameter λ_{ij}:

$$E_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{6}$$

$$\log(\lambda_{ij}) = \log(N_{ij}) + \mu_{f(i)} + \theta^\top x_{ij} + r_i. \tag{7}$$

N_{ij} is the number of words in the utterance. $f(i)$ is the group label of the speaker: in our experiments, $f(i) \in \{\text{Asian}, \text{Black}, \text{White}\}$. $\mu_{f(i)}$ is the effect of demographic group on WER. x_{ij} is a vector that contains any known potential confounding effects, including any available demographic labels of speaker i other than the label being tested. θ has the learnable coefficients of each confounding effect. r_i is the speaker-level random effect: in this model, inter-speaker variability is a random variable that is normally distributed with zero mean and with a variance, σ^2, that is assumed to be independent of group identity.

By explicitly modeling confounding effects, this model avoids confounds that may have affected the WER. Previously published studies of ASR fairness, e.g., studies of CORAAL (Kendall and Farrington, 2023) report a higher WER for male than for female speakers, but when confounding variables like profession is explicitly modeled, the apparent dependence on gender becomes non-significant (Jahan et al., 2025). In our case, as confounding effects, we took the gender logits and predicted age from a pretrained gender and age classifier (Burkhardt et al., 2023) since FAI did not have these labels. In (Burkhardt et al., 2023), the classifier achieves a Mean Average Error between 7.1 years and 10.8 years for age, and a 91.1% accuracy for gender. We also used signal-to-noise ratio (SNR) as a confounding variable. To calculate SNR, we used Silero-VAD (Team, 2021) to get the noise segments and then squared the noise segment values and took the mean as the noise power. We squared the signal and took the mean as the signal power, then calculated the SNR.

Spk. Emb.	Silhouette ↑	ρ ↑	PLDA Classi. ↑
1	-0.069	*1.05*	0.71
2	-0.012	**1.18**	0.81
3	**0.008**	1.03	0.80
4	*0.005*	*1.05*	**0.83**
5	0.004	1.03	0.82
6	-0.069	0.7	**0.83**
7	-0.058	1.09	0.79

Table 1: The results of the speaker encoder selection experiments in Section 5.1. From left to right-speaker embedding system, silhouette score, ratio ρ as described in Section 5.1, PLDA classification accuracy. Top score is in **bold**, second best is in *italic*.

We performed the drop-in-deviance test to evaluate fairness concerning the attribute of interest, which is race in our experiments. In the simpler model, we took the speaker effect and confounding variables (male, female logits, and SNR) as predictors. In the larger model, we added race as an additional predictor. Then we used the drop-in-deviance test to test whether adding the race variable significantly improves the fit of the Poisson model to the data. If racial bias does exist, the fit will improve significantly. The drop-in-deviance is calculated by taking -2 times the difference in the simpler and larger models' log likelihood. If the fit does not improve after adding race, then the drop-in-deviance will follow a χ^2 distribution with degrees of freedom equal to the difference (d_{diff}) between the larger and simpler models. We get the p-value by taking the probability of a χ^2 random variable with d_{diff} degrees of freedom being as extreme as the observed drop-in-deviance. If race is significant in predicting word error, then we will get a low p-value. So, a higher p-value indicates better fairness.

5.3. ASR experiment

The empirical study contrasts five system configurations—three unaltered baselines and two variants of the DARE debiasing framework, across three backbone recognizers (WAV2VEC2, SEW, and WHISPER). A concise classification of the

model variants is presented below.

Baseline Systems:

- **BL** (*Frozen Baseline*). The off-the-shelf pretrained ASR model is evaluated *as is*, with all parameters frozen and no debiasing module attached. This configuration provides a reference point for native performance and demographic disparity.
- **FT** (*Feature-Tuned Baseline*). Only the acoustic feature extractor is unfrozen and fine-tuned on FAISST; decoder and other layers remain fixed. The setting gauges the benefit of limited adaptation without fairness intervention.
- **Full** (*Fully-Tuned Baseline*). All network parameters, including the encoder, decoder, and any attention modules, are fine-tuned end-to-end. This variant represents the upper bound of conventional task-specific adaptation.

Proposed Variants:

- **DARe-FT**. The debiasing module is inserted, and the original feature extractor is *duplicated* to serve as both content and group encoders; only these duplicated extractors, along with the debiasing submodules $(f_g, G2C, C2G, W)$, are trainable. All downstream ASR layers remain frozen, preserving linguistic priors while enforcing fairness constraints.
- **DARe-Full**. Identical to **DARe-FT** except that *all* backbone parameters are unfrozen. This configuration explores whether simultaneous optimization of recognition and debiasing objectives yields further gains.

Optimization Details: Backbones employ learning rates tuned to their parameter scales: 10^{-5} (WAV2VEC2), 10^{-7} (SEW), and 10^{-5} (WHISPER). Training proceeds for 60, 50, and 40 epochs, respectively, each selected via validation WER. The mini–max objectives are realized through *alternating* updates: two gradient steps maximize $\widetilde{\mathcal{L}}_{ASR}$ with respect to ψ (weight predictor) followed by one step minimizing all losses with respect to θ_{ASR}, ϕ_c, ϕ_g and auxiliary parameters. Hugging Face processors specific to each architecture handle tokenization, feature extraction, and batching of raw waveforms and transcripts.

After training, we performed inference on the test set and got individual and overall WERs. Then we fit Poisson models on the individual WERs, performed the drop-in-deviance test on the Poisson models, and got the p-values.

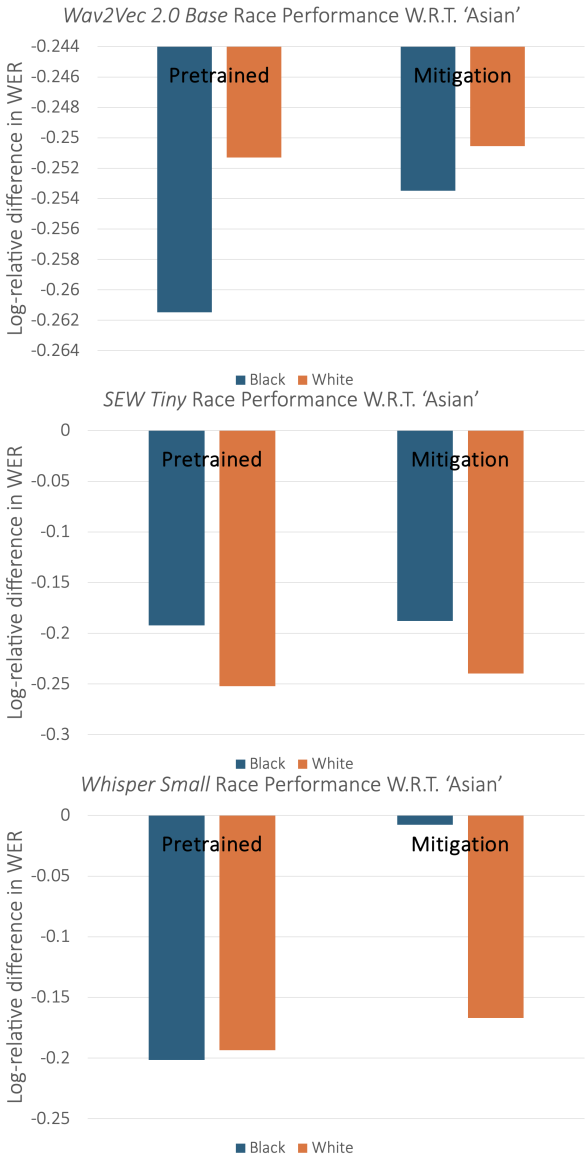

Figure 2: The log-relative difference in expected WER between other races and Asian race.

6. Results

In Table 1, a silhouette score closer to 1 means that the data point fits well within its own cluster and is clearly separated from other clusters. So, for our case, the higher the silhouette score, the better. For the embedding clusters to align well with the race labels, the ratio ρ between the between-group and the within-group distances should be high. The higher this ratio is, the better the system is for our cause. And finally, the higher the PLDA classification accuracy, the more demographic information the embeddings will have. So, considering all these, we picked NVIDIA TitaNet-Large $(\#5)$ for our experiments.

Table 2 presents word-error rates (WER) and Poisson drop-in-deviance p-values for each of the five system configurations discussed in Section 5.3.

	w2v2				
	BL	FT	Full	DARe-FT	DARe-Full
WER	0.31	0.29	**0.21**	0.30	0.28
P-val	0.0073	0.0086	0.0075	0.0103	**0.057**
	sew				
	BL	FT	Full	DARe-FT	DARe-Full
WER	0.48	0.49	**0.35**	0.48	0.48
P-val	0.0007	0.0005	0.0011	0.0021	**0.0024**
	whisper				
	BL	FT	Full	DARe-FT	DARe-Full
WER	0.27	0.21	**0.10**	0.21	0.14
P-val	0.0457	0.111	0.4111	0.071	**0.8301**

Table 2: Fairness evaluation of the ASR systems.

Several patterns emerge:

- **Fairness Gains.** Both debiasing variants (DARE-FT and DARE-FULL) increase p-values across all backbone recognizers, indicating a substantial attenuation of race-dependent error disparities. The DARE-FULL configuration achieves the highest p-values, thereby offering the strongest statistical evidence of fairness improvement.
- **Accuracy Preservation.** Relative to the frozen baseline (BL), debiasing incurs negligible, or occasionally favorable, changes in overall WER, demonstrating that fairness can be advanced without sacrificing transcription fidelity. When compared with the feature-tuned baseline (FT), DARE-FT and DARE-FULL often match or barely fall short in WER, underscoring the benefit of demographic disentanglement even under limited adaptation budgets.
- **Trade-off with Full Fine-Tuning.** The fully tuned baseline (FULL) records the lowest WERs but simultaneously yields the smallest p-values, signifying amplified racial disparity. This outcome reflects a common optimization bias: unconstrained fine-tuning gravitates toward over-represented speech varieties, thereby improving aggregate accuracy at the expense of equitable performance.

From a business perspective, the value of this work is less about squeezing out the last decimal of average WER and more about reducing tail risk and building durable trust. By improving accuracy for the worst-performing groups, DARe reduces compliance risks, lowers support costs and user loss in underserved groups, and provides more consistent performance when demographics change across regions or over time. These gains translate into stronger brand equity and access to broader markets, often outweighing the modest aggregate-WER advantage of full fine-tuning that chiefly benefits over-represented speech.

We also observe the coefficients, as shown in Figure 2, of the races after fitting the Poisson model for BL (Pretrained) and DARe-FT (Mitigation). Asian was the reference category, so the coefficient for a race is the log-relative difference in expected WER between that race and Asian speakers. So, the lower the numbers in Figure 2, the closer the race performances are to the Asian category. And, the lower the gap between the coefficients of the races, the closer their performances are to each other. So, ideally, the bars after bias mitigation should be shorter and closer to each other than the baseline. That is the case for w2v2 and sew. For whisper, the performance gap with Asian decreases, but the gap between Black and White races increases. This could be the result of Black and White categories becoming closer in performance to the Unknown and Multiracial categories.

7. Conclusion

We proposed a novel model-agnostic, demographic label-agnostic bias mitigation approach for ASRs called DARe. The approach disentangles content and group characteristics, uses pseudo labels through unsupervised clustering, classification to enforce group information in group embedding, as well as adversarial reweighting and min-max optimization to train the debiasing module and the ASR. Using a Poisson regression-based fairness evaluation approach to evaluate the performance of the approach, our experiments show that the fairness significantly improves, indicated by higher p-values, while maintaining similar or lower WERs. We also observe the log-relative difference in expected WERs, which shows that in most cases, the performance gap was improved.

8. Acknowledgements

This project was supported by NSF Award 2147350.

9. Bibliographical References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Proc. Interspeech 2005*, pages 2205–2208.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Guy Bailey and Erik Thomas. 2021. Some aspects of african-american vernacular english phonology. *African-American English*, pages 85–109.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Bjoern Schuller. 2023. Speech-based age and gender prediction with transformers. In *Speech Communication; 15th ITG Conference*, pages 46–50.

Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. 2021. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5954–5958.

Ralph O. Coleman. 1976. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research*, 19(1):168–180.

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.

Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran, and Sidney D'Mello. 2025. "it feels like we're not meeting the criteria": Examining and mitigating the cascading effects of bias in automatic speech recognition in spoken language interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Comput. Speech Lang.*, 84(C).

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition.

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.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender representation in french broadcast corpora and its impact on asr performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, page 3–9, New York, NY, USA. Association for Computing Machinery.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the impact of gender representation in ASR training data: a case study on librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92, Online. Association for Computational Linguistics.

James M. Hillenbrand and Michael J. Clark. 2009. The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, 71:1150–1166.

Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. Unveiling performance bias in asr systems: A study on gender, age, accent, and more. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Maliha Jahan, Helin Wang, Thomas Thebaud, Yinglun Sun, Giang Ha Le, Zsuzsanna Fagyal, Odette Scharenborg, Mark Hasegawa-Johnson, Laureano Moro Velazquez, and Najim Dehak. 2024. Finding spoken identifications: Using GPT-4 annotation for an efficient and fast dataset creation pipeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7296–7306, Torino, Italia. ELRA and ICCL.

Minho Jin, Chelsea Ju, Zeya Chen, Yi Chieh Liu, Jasha Droppo, and Andreas Stolcke. 2022. Adversarial reweighting for speaker verification fairness. In *Interspeech 2022*, interspeech_2022. ISCA.

Tyler Kendall and Charlie Farrington. 2023. The corpus of regional african american language. The Online Resources for African American Language Project.

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

Chunxi Liu, Michael Picheny, Leda Sari, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022a. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. 2022b. Model-based approach for measuring the fairness in asr. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6532–6536.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. FAccT '22, page 521–534, New York, NY, USA. Association for Computing Machinery.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, pages 626–630.

Brandon Merritt and Tessa Bent. 2022. Revisiting the acoustics of speaker gender perception: A gender expansive perspectivea). *The Journal of the Acoustical Society of America*, 151(1):484–499.

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.

Mozilla. Common voice. https://commonvoice.mozilla.org/.

Hyeong-Ju Na and Jeong-Sik Park. 2021. Accented speech recognition based on end-to-end domain adversarial training of neural networks. *Applied Sciences*, 11(18).

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Leda Sarı, Mark Hasegawa-Johnson, and Chang D. Yoo. 2021. Counterfactually fair automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3515–3525.

Majdi Sawalha and M Abu Shariah. 2013. The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus.

Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. 2021. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6783–6787.

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.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*, pages 934–938.

Silero Team. 2021. Silero voice activity detector (vad). https://github.com/snakers4/silero-vad. Accessed: 2025-05-20.

Erik R. Thomas. 2007. Phonological and phonetic characteristics of african american vernacular english. *Language and Linguistics Compass*, 1(5):450–475.

Irina-Elena Veliche and Pascale Fung. 2023. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Felix Wu, Kwangyoun Kim, Jing Pan, Kyu J. Han, Kilian Q. Weinberger, and Yoav Artzi. 2022. Performance-efficiency trade-offs in unsupervised pre-training for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7667–7671.

Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2020. See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA. Association for Computing Machinery.

Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proc. Interspeech 2022*, pages 3168–3172.