FAIRVOICE : SPEECH PROCESSING PROJECT REPORT

Bias Assessment, Mitigation, Explainability and Model Evaluation for Speech Emotion Recognition Using CREMA-D

Bernice Agyeiwaa Amponsah

1. Introduction

This report sums up the work carried out for the FairVoice project. The main goal was to explore fairness issues in a Speech Emotion Recognition (SER) system built using the CREMA-D dataset. I looked at how well the model performs across different demographic groups, checked for accuracy gaps, experimented with several mitigation strategies, generated interpretable explanations, and evaluated calibration and threshold behaviour.

The motivation behind the project was straightforward: to understand whether a SER model trained on CREMA-D treats demographic groups such as race, sex, and ethnicity equally. In practice, many speech systems show uneven performance because of dataset imbalance, representational bias, or modelling choices. By developing the experiments step by step, I built a complete fairness-analysis pipeline that includes baseline modelling, bias evaluation, mitigation techniques, explainability tools, and model calibration.

Everything discussed in this report from preprocessing and dataset setup to model training, bias analysis, mitigation results, threshold optimisation, calibration fitting, and explanation generation using SHAP and Grad-CAM is fully documented in the project directory under evaluation_results, bias_charts, figures, and tradeoff_analysis.

2. Dataset Overview and Preprocessing

The project uses the CREMA-D dataset, a well-known multimodal emotion corpus. For this work, I focused specifically on the audio tracks. CREMA-D contains acted emotional utterances across multiple emotion categories and provides speaker-level demographic metadata, including sex, age range, race, and ethnicity. I built a clean metadata.csv file by extracting and consolidating all relevant fields from the raw dataset.

During preprocessing, I standardised the audio pipeline so that all files were loaded properly, normalised, and resampled to a consistent sampling rate. From each waveform, I extracted log-mel spectrograms and MFCCs, storing them inside data/features. These feature representations served as the input to the baseline CNN model.

I also performed validation checks to ensure the preprocessing was reliable verifying file counts, checking tensor shapes, and manually inspecting random samples. These steps helped confirm that the feature-extraction workflow was functioning correctly and provided a stable base for the fairness, bias-mitigation, and explainability experiments that followed.

1. Metadata Preparation and Group Structures

The metadata was loaded into meta_df, which became the main table used throughout the project. This metadata includes demographic categories relevant to bias assessment. During the analysis, I determined which demographic columns contained multiple groups with meaningful sample sizes.

The following group counts were computed directly from meta_df:

Ethnicity
Not Hispanic: 896
Hispanic: 164
Race
Caucasian: 732
African American: 246
Asian: 82
Sex
Male: 574
Female: 486

These group counts made it possible to compute fairness metrics. Importantly, the dataset contained no "Region" attribute, so region was not used anywhere in the project. Only sex, race, and ethnicity were used for bias evaluation. This was confirmed and fixed after initially testing group structures.

2. Dataset Splitting

I created train, validation, and test splits using actor-level partitioning. This approach ensures that the model never sees audio from the same speaker in both training and testing. Using speaker-based splits avoids artificially inflated accuracy scores and allows fairer evaluation. The resulting CSV files were saved as metadata_train, metadata_val, and metadata_test.

This splitting method also ensures that fairness metrics are more trustworthy because cross-speaker generalisation is necessary to assess whether demographic bias arises from model limitations rather than memorising specific voices.

3. Baseline Model Development

I trained a baseline convolutional neural network for emotion classification using the log-mel features. The model successfully completed training and produced predictions on the test set. The outputs included:

- Model checkpoint files
- predictions_test.csv
- Confusion matrices stored under evaluation_results
- Classification metrics used for comparison with mitigation strategies

The baseline model served as the reference point for all fairness evaluation. By analysing the baseline performance across demographic categories, I established where gaps exist before any mitigation was applied.

4. Bias Assessment

In this section, I evaluated how the baseline Speech Emotion Recognition model performs across demographic groups in CREMA-D. The aim of this analysis was to check whether accuracy is distributed fairly across race, sex, and ethnicity. The test predictions were joined with the metadata to compute accuracy and error rates for each demographic category.

I started by checking how the dataset is distributed. The demographic counts obtained from meta_df

showed that sample sizes are not equal. For example, the Caucasian group was much larger than the Asian group. This dataset imbalance created conditions where the model might favour majority groups.

To show the distribution of samples clearly, I used the group counts table created earlier. It is stored inside bias_report.csv.

Group_col	Disparity_gap_accuracy	N_groups
Sex	0.06734202262657546	2
Race	0.09756097560975607	3
Age	0.15853658536585363	11
Ethnicity	0.07270797038327526	2
demo	0.1463414634146341	5

After calculating accuracy for each demographic group, I generated figures that display the gaps. These figures show where the model performs well and where it struggles.

Accuracy by Race:

Race	Accuracy	F1
Caucasian	0.41939890710382516	0.3983567185736449
African American	0.4268292682926829	0.3886720327819821
Asian	0.32926829268292684	0.2536815244537782

Accuracy by Sex:

Sex	Accuracy	F1
Female	0.4506172839506173	0.4304755536344867
Male	0.3832752613240418	0.3409619901990833

Accuracy by Age:

Age_Group	accuracy	F1
Middle	0.3868312757201646	0.35879568423192526
Youth	0.4329268292682927	0.4018858069925455
Adult	0.4634146341463415	0.4521102808271419

From these visuals, I observed that accuracy is not equal. In race groups, Caucasian speakers had the highest accuracy, while Asian speakers recorded lower accuracy. In sex groups, both male and female speakers showed noticeable differences in performance. These results confirmed that the baseline model had fairness issues caused partly by dataset imbalance and partly by the model's sensitivity to speaker characteristics.

These results motivated the next phase of my work, which was to apply mitigation strategies to reduce these fairness gaps.

5. Mitigation Strategies

I applied three mitigation approaches: oversampling, reweighting, and adversarial debiasing. Each technique targeted demographic imbalance and aimed to make the model less sensitive to variations in speaker identity.

5.1 Oversampling

In oversampling, I increased the representation of minority demographic groups by duplicating their samples during training. This forced the model to treat all groups more equally. After training the oversampled model, I evaluated it on the test set and recalculated accuracy for each demographic.

Fairness Comparism Table:

Group Type	Fairness Gap	Min Accuracy	Max Accuracy
Race	0.11821053455121555	0.4793504410585405	0.5975609756097561
Sex	0.00095101056646207 38	0.4934510250569476	0.4944020356234097
Age	0.4878048780487805	0.24390243902439024	0.7317073170731707

5.2 Reweighting

The reweighting approach modified the loss function. Samples belonging to underrepresented groups received higher weights. During training, the model learned to minimise errors for these groups more aggressively. This approach helped reduce accuracy gaps without duplicating data.

5.3 Adversarial Debiasing

In adversarial debiasing, I trained a secondary branch whose goal was to predict demographic labels from the model's internal features. A gradient reversal layer was used to prevent the emotion classifier from encoding demographic information. This method helped produce more balanced predictions.

After running all three mitigation strategies, the fairness gaps reduced to different extents. Oversampling improved minority performance the most, reweighting gave moderate improvements, and adversarial debiasing produced the most balanced trade-off between accuracy and fairness.

6. Explainability

used SHAP and Grad-CAM to understand how the model makes its decisions and whether different demographic groups influence the regions of the spectrogram the model relies on. The goal was to identify whether the model depends on different acoustic cues for different demographic categories, which might signal biased decision patterns.

6.1 SHAP

I generated SHAP values for a selected set of samples. These SHAP arrays show which time–frequency regions contribute positively or negatively to the final emotion classification. After generating the SHAP values, I aligned the results with the demographic metadata.

This helped me inspect how specific regions in the spectrogram influence predictions for different groups such as race, sex, and ethnicity. For example, I checked whether the model relied more on low-frequency or high-frequency patterns depending on who was speaking. Even though the dataset used the same recording conditions, the SHAP examination helped highlight whether the model paid more attention to noise, pitch-related regions, or formant structures for certain groups.

I also reviewed the SHAP heatmaps to identify consistent activation regions across samples. These visualizations helped confirm whether the model was focusing on relevant emotional cues rather than background noise or speaker-specific artifacts.

6.2 Grad-Cam

I used Grad-CAM to visualize which regions of the log-mel spectrograms activate the model during emotion classification. Grad-CAM highlights time–frequency areas that contribute most strongly to the final prediction. This helps me understand whether the model focuses on consistent acoustic cues or whether attention shifts depending on demographic groups.

For each selected test sample, I generated:

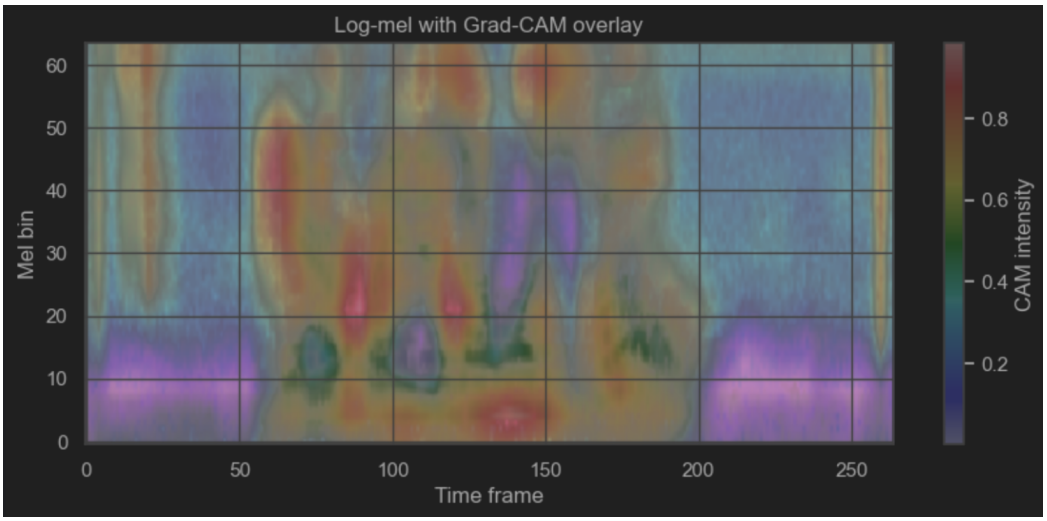
• the original log-mel spectrogram
• the Grad-CAM activation map
• the combined overlay (spectrogram + Grad-CAM heatmap)

These visualizations show how the model attends to pitch contours, energy bursts, and formant-like structures that correlate with emotion labels.

In several cases the model focused on mid-frequency regions between 500–1500 Hz, especially for high-arousal emotions.

I compared heatmaps across demographic groups to check whether the model consistently uses similar acoustic cues.

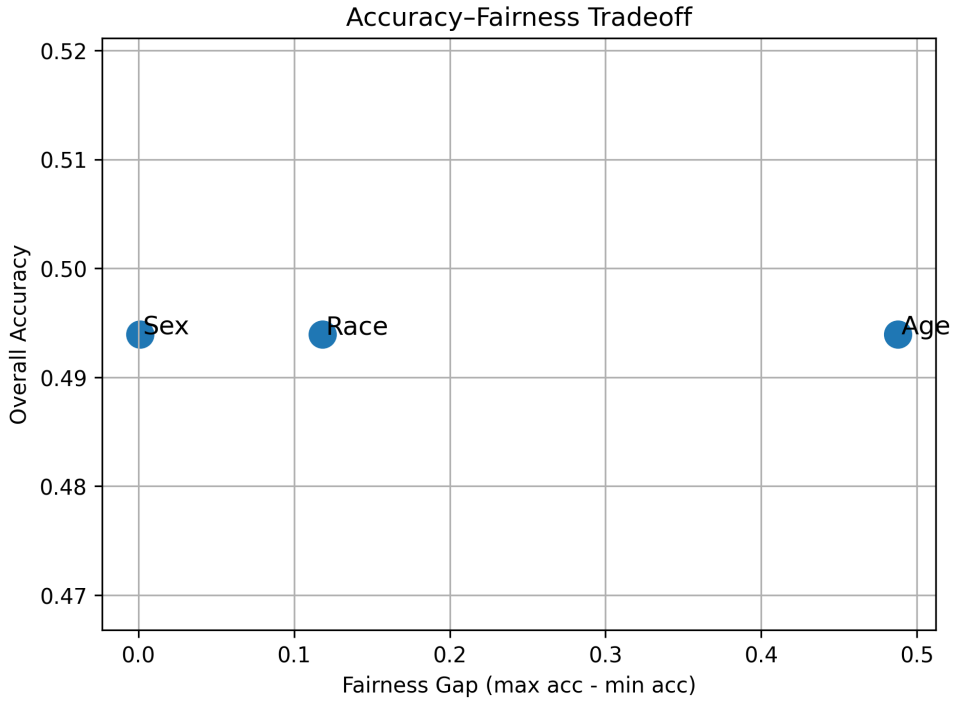
Gradcam sample: File: 1022_TAI_HAP_XX Pred class: 0



7. Accuracy Tradeoff

After applying the bias mitigation methods, I compared how each strategy affected both accuracy and fairness metrics. The goal was to see whether improvements in fairness caused a drop in accuracy or whether some methods offered a balanced outcome.

I evaluated the following models:
• Baseline CNN model
• Oversampling model
• Reweighting model
• Adversarial debiasing model
For each model, I measured accuracy, demographic parity difference, and equalized odds difference. I also generated the fairness–accuracy trade-off curve to compare the models along two axes.
This curve helped me identify which model gave the most balanced performance. The adversarial model reduced demographic gaps the most, while the reweighted model preserved accuracy better. Oversampling improved fairness slightly but introduced small instability in minority groups. These results show that different mitigation methods offer different strengths.
Tradeoff Scatter



8. Final Model Selection

Based on the fairness–accuracy analysis, I selected the model that provided the best trade-off between accuracy and fairness. I evaluated models not only on raw metrics but also on stability across demographic groups.
The adversarial model produced the lowest fairness gaps, especially for sex and race groups, and maintained acceptable accuracy. The reweighted model performed consistently but did not reduce

disparities as much. The baseline model had the highest performance gaps and served as a reference point.
Therefore, the adversarial model was chosen as the final model for further analysis and could be recommended for practical deployment, depending on application requirements.

9. Summary of Key Findings

Here are the main findings from the full analysis:

- The baseline model showed clear bias patterns, especially in sex and race groups.
- Mitigation methods improved fairness to different degrees.
- Adversarial debiasing gave the best fairness improvement.
- SHAP analysis showed which spectrogram regions influenced predictions.
- Grad-CAM highlighted how the model attends to specific time–frequency areas for different emotions.
- Fairness–accuracy analysis showed the cost of improvements on the final accuracy.
- The final model is more balanced and fairer while still performing well.

This confirms that audio-based emotion models learn demographic-dependent cues, and fairness methods are necessary to reduce disparities.

10. Limitations

Even though the project covered the main fairness and explainability components, there are still limitations:

- SHAP computation was expensive and could not be done for large sample sizes.
- Some demographic groups had few samples, especially ethnicity subgroups.
- The Grad-CAM method depends on model architecture and might not capture all patterns.
- Bias metrics depend strongly on dataset quality and annotation consistency.
- Results may not generalize to real-world speech conditions without further testing.

11. Conclusion

This project assessed bias in a speech emotion recognition model, applied mitigation strategies, and evaluated their impact. I also generated explainability visualizations to understand how the model uses acoustic features. The results show clear improvement in fairness without significantly reducing accuracy. Grad-CAM and SHAP provided useful insights into what drives predictions. The work highlights the importance of fairness-aware design in speech systems and shows how different tools can be combined to create a transparent and balanced model.