1. Motivation

Speaker Verification (SV) Models
- The performance of speaker verification (SV) models has dramatically improved due to deep learning algorithms and large-scale datasets.
- SV models typically have two stages: encoding speech embeddings (front-end) and scoring function (back-end).

Model Unfair Performance
- Models are optimized to speakers' voice characteristics in training.
- This learning process can lead to SV model unfairness.

Contributions
- Create well-designed training and evaluation datasets and metrics for analyzing SV model fairness (using gender as a test case) (Section 3)
- Evidence that imbalanced dataset composition leads to SV model unfairness to under-represented groups. (Section 4)
- Propose a flexible, modular model to alleviates model unfairness. (Section 2)

2. Method: Group-adapted Fusion Network (GFN)

Front End

Group Embedding Adaptation
- Efr = BaseEncoder (Xf), i = 1, 2
- Efm = FemaleAdaptationEncoder (Xf), i = 1, 2
- EAm = MaleAdaptationEncoder (Xm), i = 1, 2
- The front-end encoders extract base and group-adapted embeddings.

Back End

Score Fusion
- Sf = CostSimilarity(Efr, Efm)
- Sm = CostSimilarity(Em, EAm)
- S = $\text{Sigmoid}(f(S_f, S_m; w))$
- The back-end score fusion model combines all scores for speaker verification.

3. Fairness Datasets and Evaluation Metrics

VoxCeleb2-GRC
- Gender Ratio Controlled
  - Gender Ratio F:M
  - Female speakers
  - Male speakers
  - Female utterances
  - Male utterances
- Sample positive (same speaker) and negative (different speakers) training pairs from VoxCeleb2-GRC for contrastive learning.

VoxCeleb1-F (Fairness)
- Gender Trials
  - Trial Count
  - VoxCeleb1-F
    - Group-wise EER
      - Female-group: EER[F], Male-group: EER[M]
    - Overall EER: EER[All]
    - Disparity Score (DS): $\text{DS} = |\text{EER[F]} - \text{EER[M]}|

Evaluation Metrics
- We define three model fairness metrics based on Equal Error Rate (EER).
  - Group-wise EER
  - Female-group: EER[F], Male-group: EER[M]
  - Overall EER: EER[All]
  - Disparity Score (DS)
  - $\text{DS} = |\text{EER[F]} - \text{EER[M]}|$

4. Results and Findings

Cause of Model Unfairness
- Increasing dominance of one gender group in training set (e.g., 4:1 and 9:1) leads to increasing performance gap (DS scores) and model unfairness.

Improving Fairness with GFN
- Proposed GFN model achieves better group-wise and overall EER than baselines.

Ablation Study
- Among alternative embedding adaptation methods and baselines:
  - F-FT, M-FT, ES, GBWL, Q/RN, H/RN
  - Our GFN gets the best performance.

Embedding Analysis
- Visualization of learned speaker embeddings using t-SNE.
- GFN (c) generates more compact embedding clusters than the baseline (b).

Acknowledgments
- We thank our colleague Victor Rozgic and Amazon Alexa Speaker Understanding team members and managers for their input and support.

Open-sourced Datasets: https://github.com/huashen218/Voxceleb-Fairness.git