Improving Fairness in Speaker Verification Via Group-Adapted Fusion Network

PennState | amazon | science | amazon alexa @ ICASSP 2022

Hua Shen*, Yuguang Yang*, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, Andreas Stolcke

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)**.

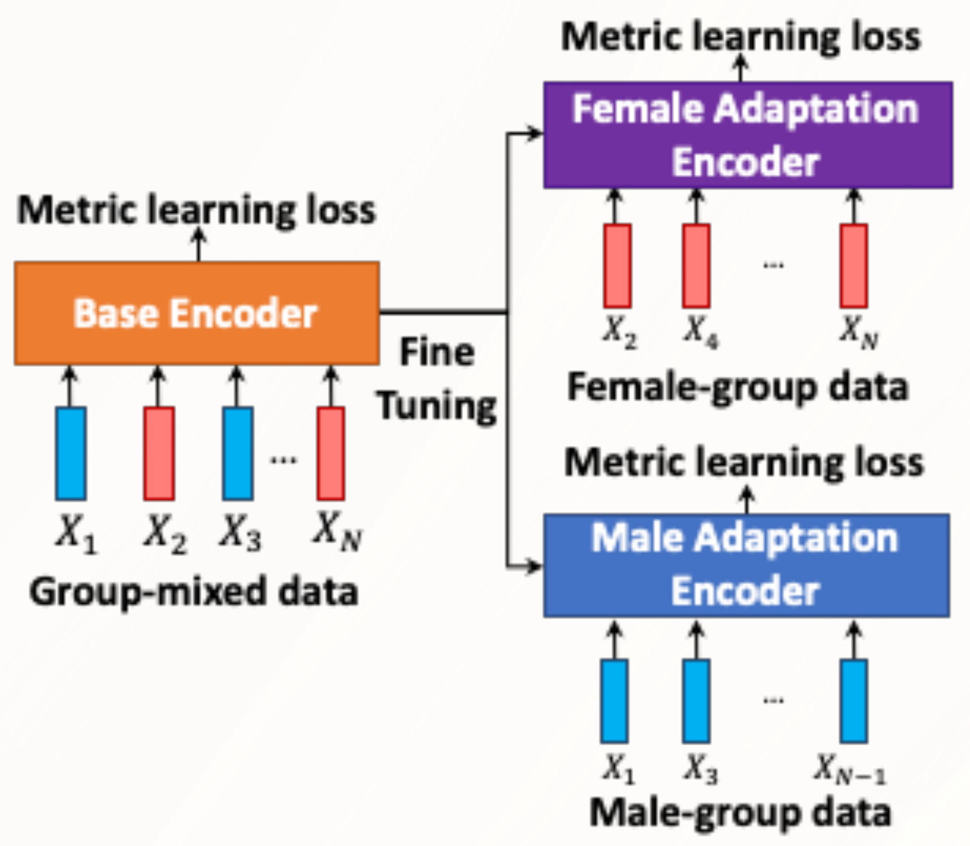
Model Unfair Performance

- Models are optimized to **differentiate arbitrary speakers**' voice characteristics in training.
- This learning process can lead to **model unfairness across groups**.

Contributions

- Create well-designed **training and evaluation data sets 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)



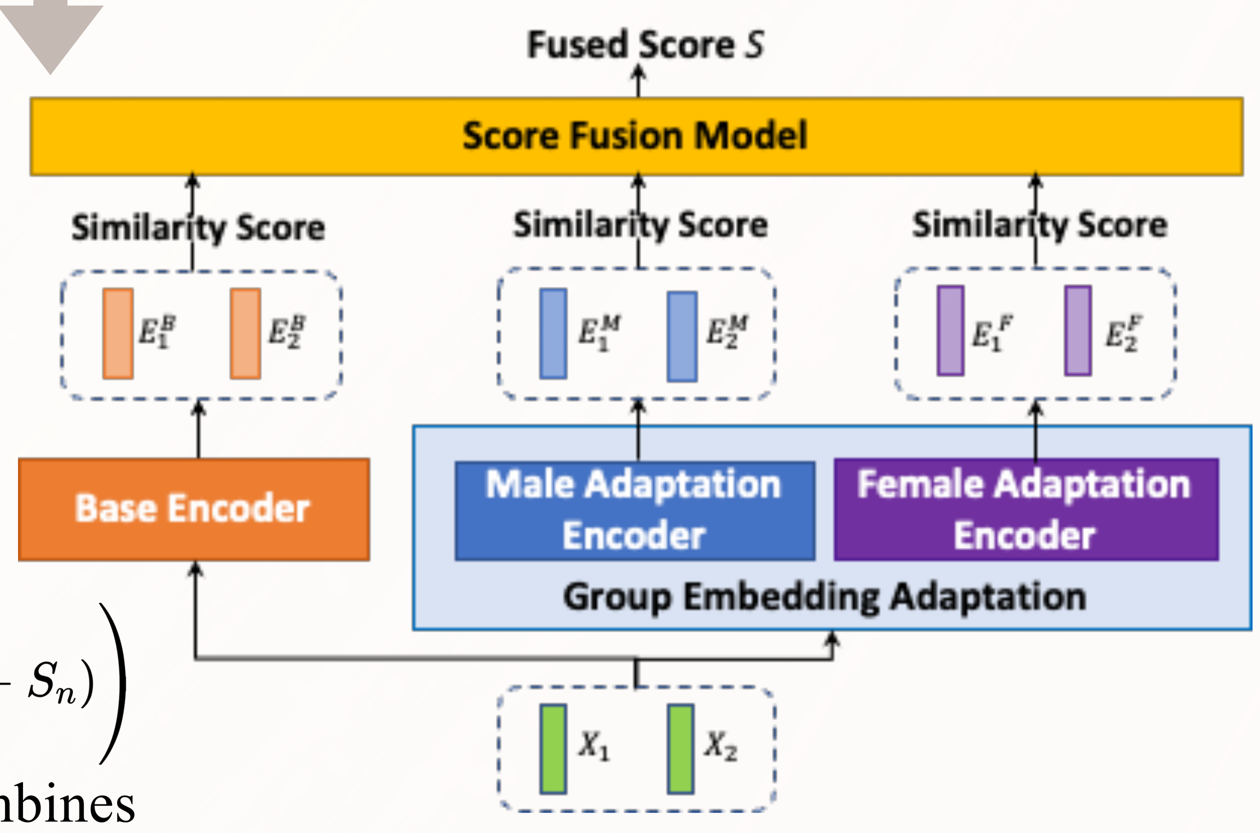
Front End — Group Embedding Adaptation

$$\mathbf{E}_i^B = \text{BaseEncoder}(\mathbf{X}_i), i = 1, 2$$
$$\mathbf{E}_i^F = \text{FemaleAdaptationEncoder}(\mathbf{X}_i), i = 1, 2$$
$$\mathbf{E}_i^M = \text{MaleAdaptationEncoder}(\mathbf{X}_i), i = 1, 2$$

- The front-end encoders extract base and group-adapted embeddings.

Back End — Score Fusion

$$S_b = \text{CosineSimilarity}(\mathbf{E}_1^B, \mathbf{E}_2^B)$$
$$S_f = \text{CosineSimilarity}(\mathbf{E}_1^F, \mathbf{E}_2^F)$$
$$S_m = \text{CosineSimilarity}(\mathbf{E}_1^M, \mathbf{E}_2^M)$$
$$S = \text{Sigmoid}(f([S_b, S_f, S_m]; \mathbf{w}))$$



Training Objectives

$$L = -\frac{1}{N}\left(\sum_{n\in\mathcal{P}_{pos}} y_n \log S_n + \sum_{n\in\mathcal{P}_{neg}} (1-y_n)\log(1-S_n)\right)$$

- The back-end score fusion model combines all scores for speaker verification.

3. Fairness Datasets and Evaluation Metrics

Front End — VoxCeleb2-GRC (Gender Ratio Controlled) — Training Sets

Gender Ratio F:M	Female speakers	Male speakers	Female utterances	Male utterances
9:1	2,250	250	387,322	45,181
4:1	2,000	500	341,500	95,157
1:1	1,250	1,250	214,919	228,823
1:4	500	2,000	86,616	372,133
1:9	250	2,250	43,482	419,853

Back End Sample **positive** (same speaker) and **negative** (different speakers) training pairs from VoxCeleb2-GRC for contrastive learning.

VoxCeleb1-F (Fairness) — Test Sets

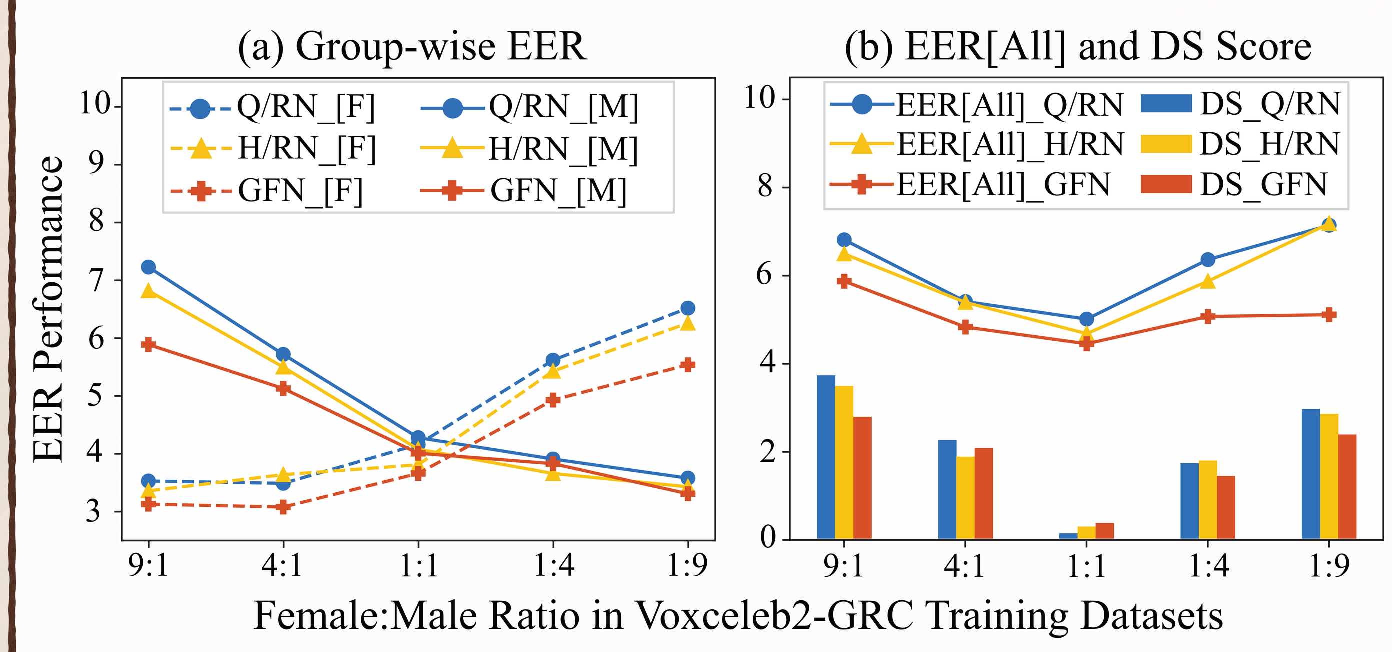
Gender Trials	Trial Count	VoxCeleb1-F [F]	[M]	[All]
Positive F-F	150,000	✓		✓
Negative F-F	150,000	✓		✓
Negative M-F	150,000	✓	✓	✓
Positive M-M	150,000		✓	✓
Negative M-M	150,000		✓	✓

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)**
  DS = |EER[F] - EER[M]|

4. Results and Findings

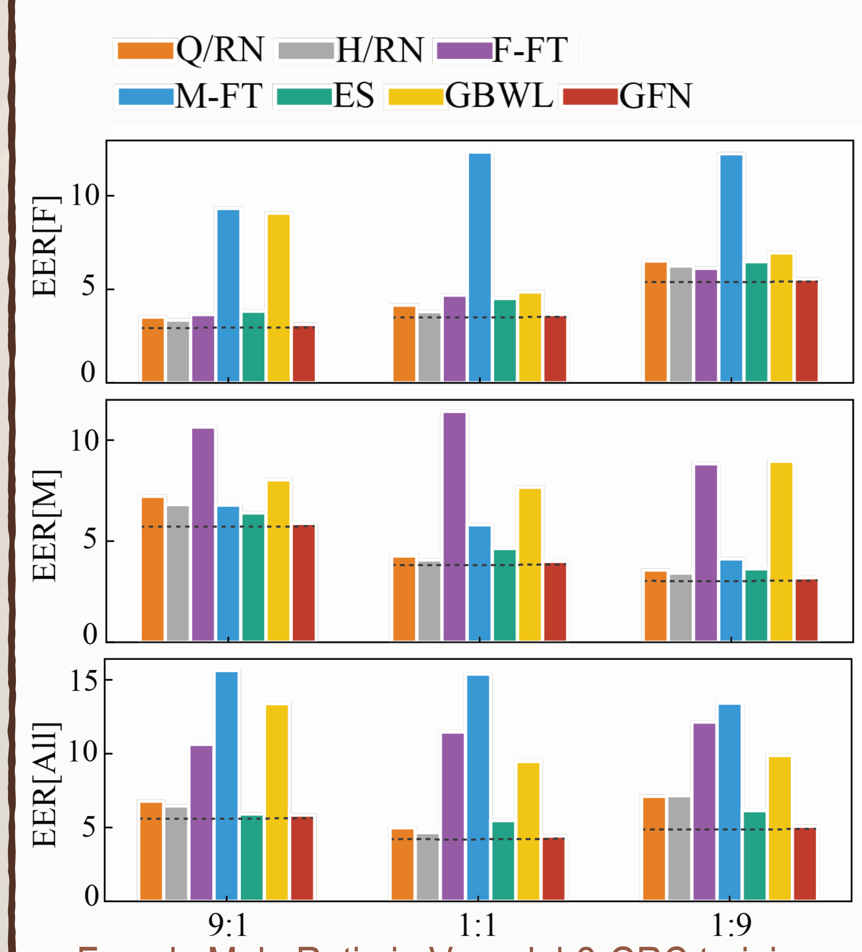


Cause of Model Unfairness

- 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.



Female:Male Ratio in Voxceleb2-GRC training sets

Ablation Study

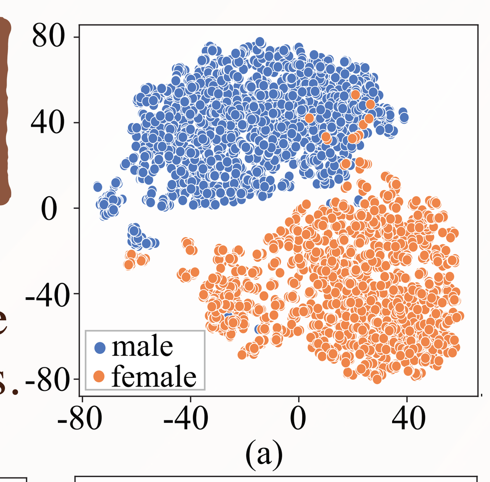
Among alternative embedding adaptation methods and baselines:

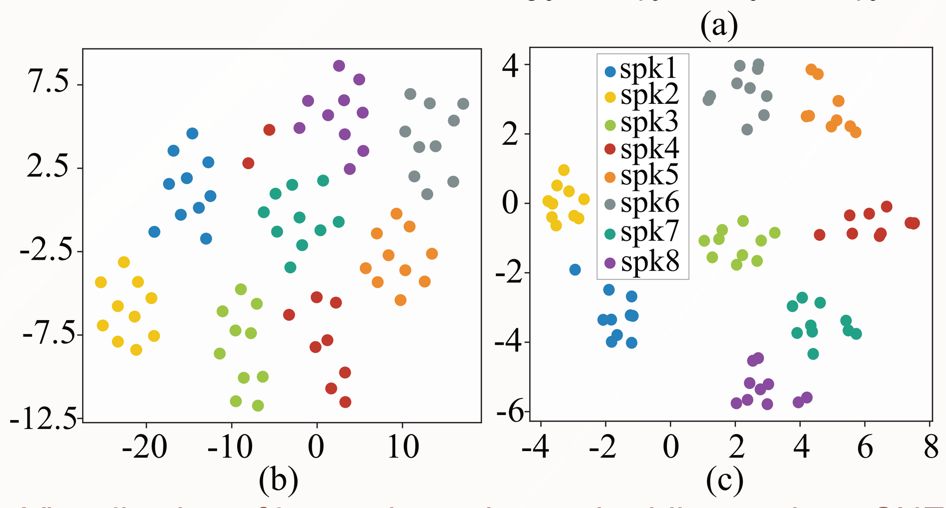
- F-FT,
- M-FT,
- ES,
- GBWL,
- Q/RN,
- H/RN,

Our GFN gets the best performance.

Embedding Analysis

- a) different genders in separate embedding regions.





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