Improving Fairness in Speaker Verification via Group-Adapted Fusion Network

Hua Shen*,[1,2] Yuguang Yang*,[2] Guoli Sun,[2] Ryan Langman,[2] Eunjung Han,[2]
Jasha Droppo,[2] Andreas Stolcke[2]

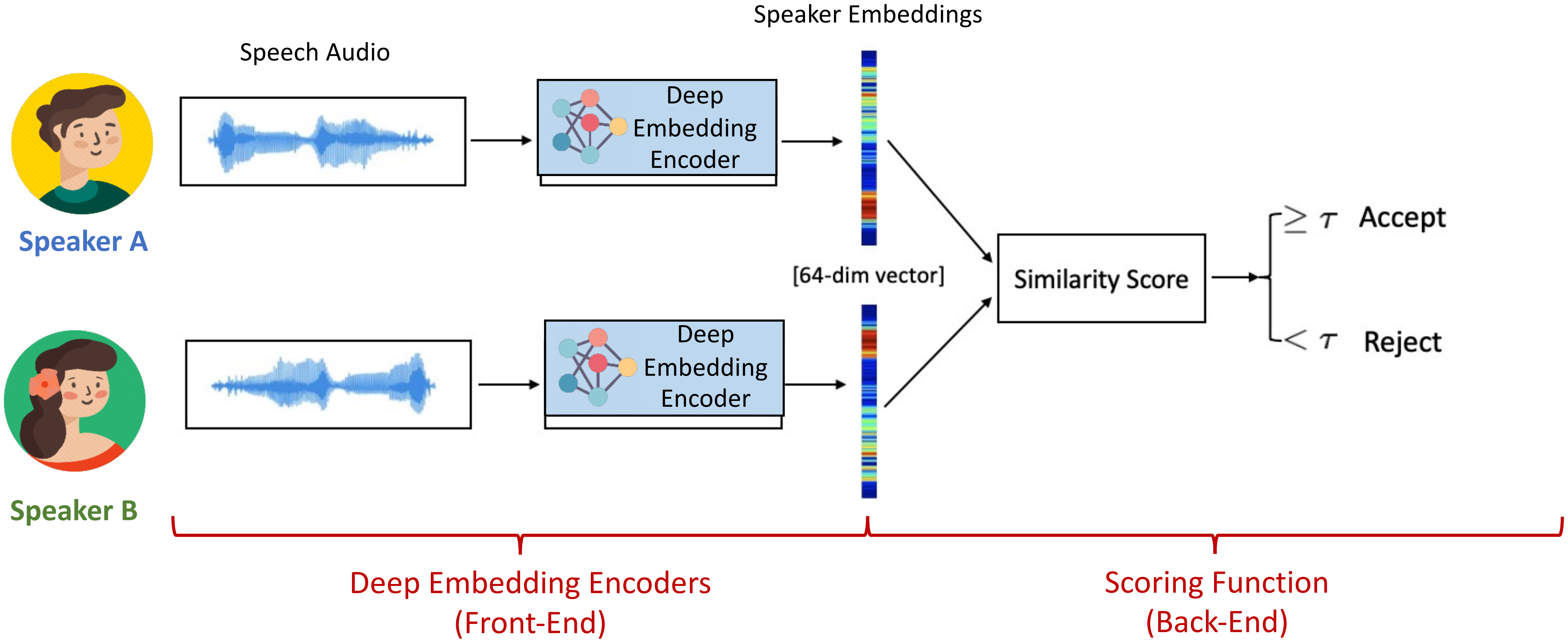
[1] The Pennsylvania State University, [2] Amazon Alexa AI

Background

Model Architecture

The performance of **speaker verification systems** has dramatically improved due to both **deep learning algorithms** and **large-scale datasets**. The state-of-the-art **speaker verification models** typically have two stages:

1. **Deep embedding encoders (Front-end):** compute speaker embeddings from speech audio;
2. **Scoring function (Back-end):** compute similarity score between two embeddings.

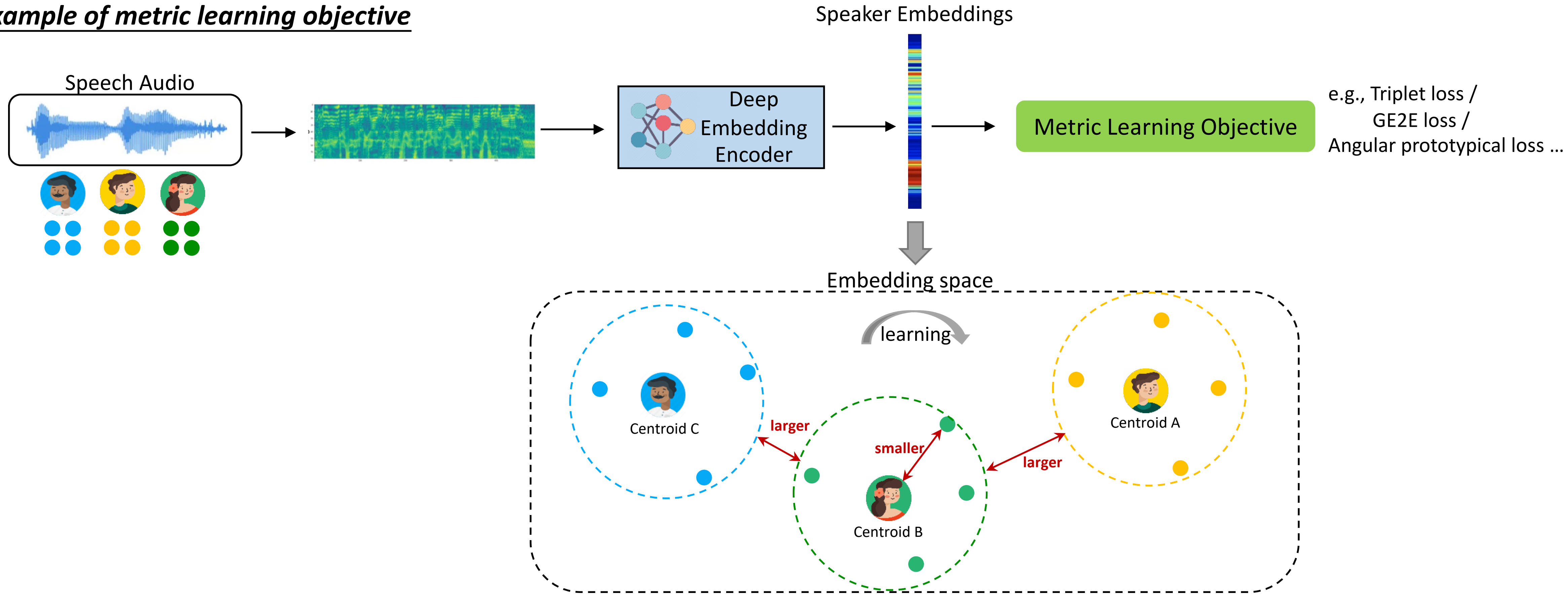


Background

Training Process

We commonly **train** the Front-end **deep embedding encoders** with **classification** or **metric learning** objectives.

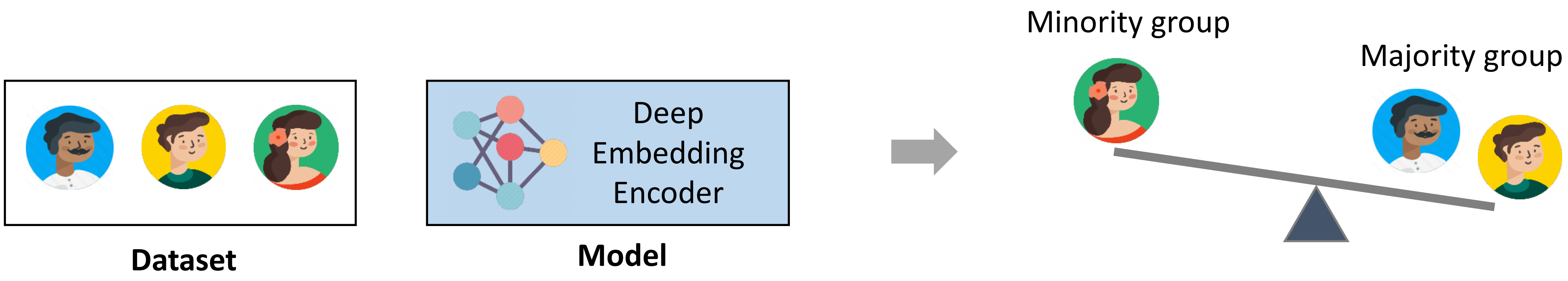
Example of metric learning objective



Learn to optimize the embedding to get:

- **smaller** distance between **same** speakers
- **larger** distance with **different** speakers.

Motivation



However, this learning process can potentially lead to **model unfairness across groups, because:**

- **Training:** Models **minimize average loss** over the full datasets, which might ignore the voice characteristics of **underrepresented groups;**
- **Evaluation:** The **performance metrics** (e.g., EER) typically measure **overall performance**, which does **not reflect performance over different subgroups**.

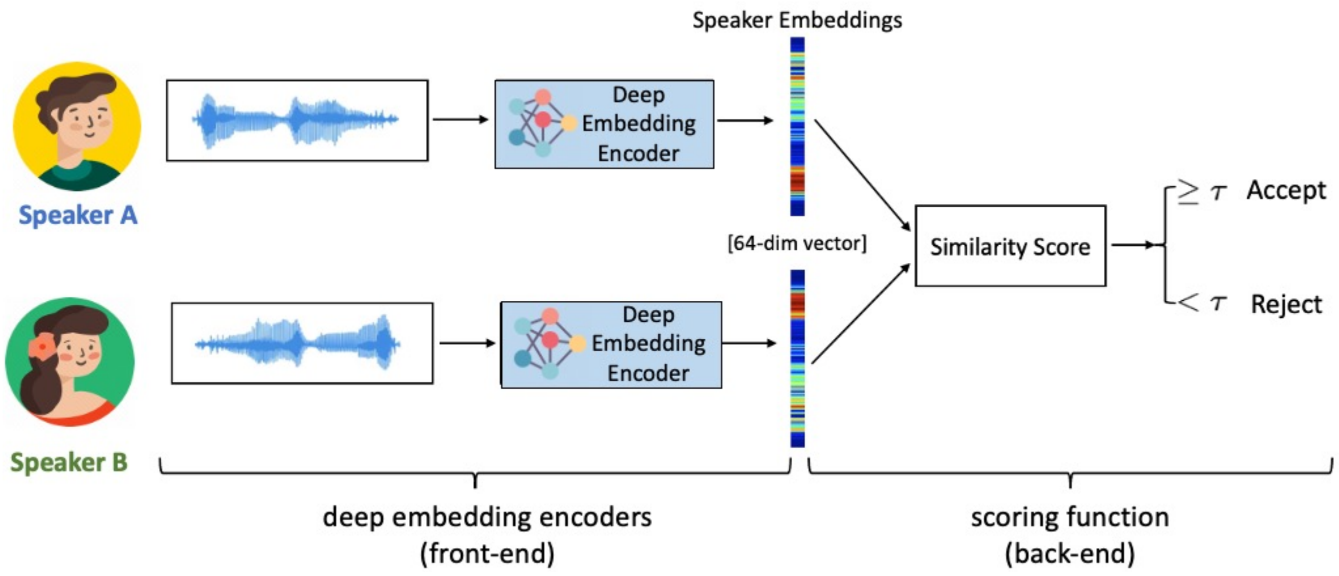
Research Objective

Rigorously **analyze model unfairness** in speaker verification systems and offer a generalizable **solution to alleviate model unfairness**.

Contributions

1. We originally **crafted training and evaluation datasets,** and **evaluation metrics,** to rigorously evaluate and analyze model fairness performance.
2. We provide direct evidence showing that **group-imbalanced training dataset can lead to model unfairness** to underrepresented groups.
3. We **propose a flexible, modular model** based on group embedding adaptation and score fusion to **alleviate model unfairness**.

Core Idea of the Proposed Method



Group Embedding Adaptation

Use **group-wise adaptation encoders** to capture fine voice characteristics specific to each group

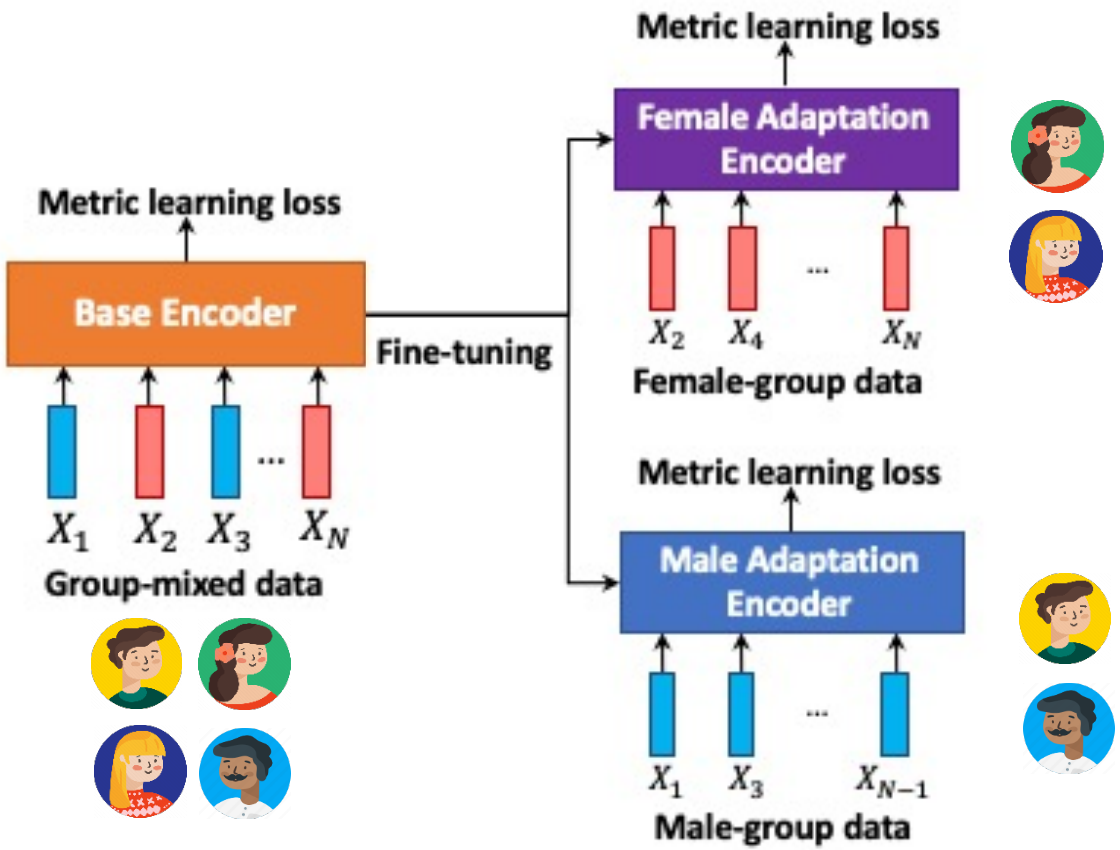
Score Fusion Strategy

Use **score fusion strategy** to aggregate scores from all group-specific embeddings

Group-adapted Fusion Network (GFN)

Group-adapted Fusion Network (GFN)

Front-end



Group Embedding Adaptation

$$\mathbf{E}_i^B = \mathrm{BaseEncoder}(\mathbf{X}_i),\ i = 1, 2$$

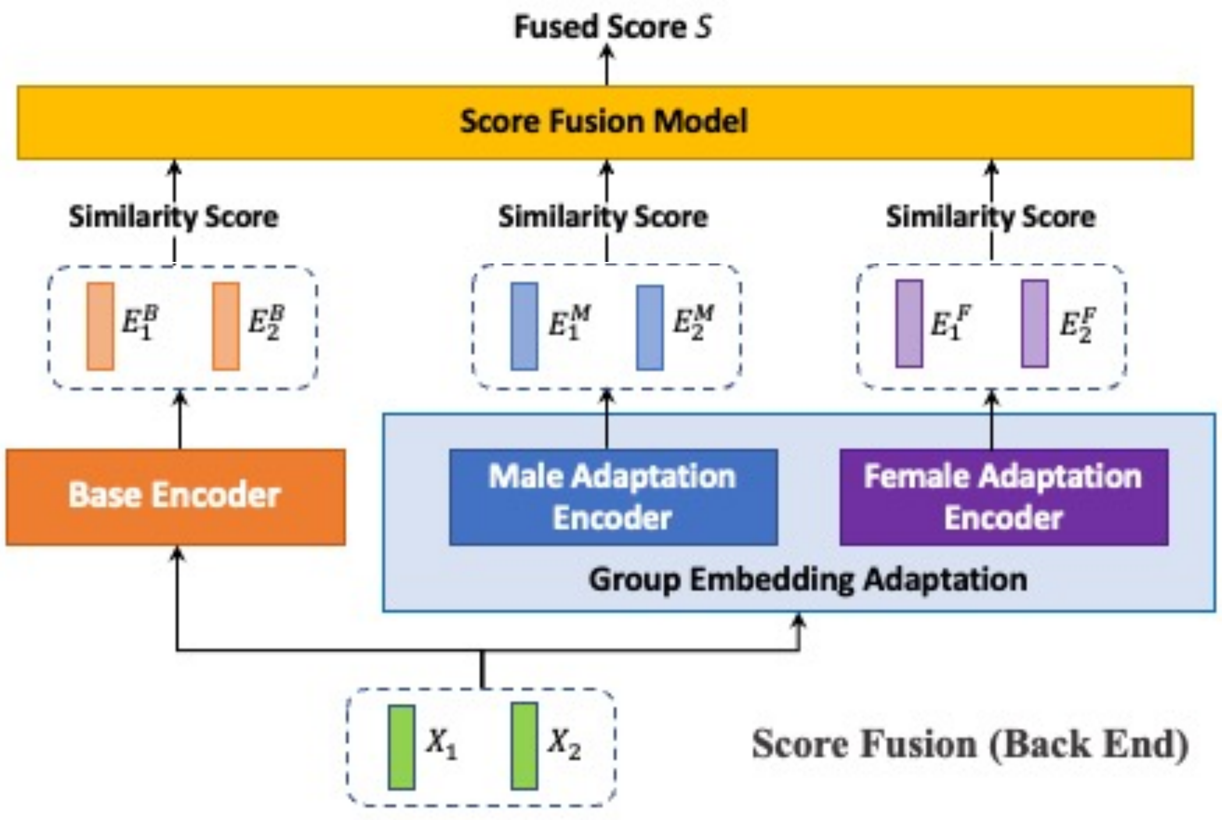
$$\mathbf{E}_i^F = \mathrm{FemaleAdaptationEncoder}(\mathbf{X}_i),\ i = 1, 2$$

$$\mathbf{E}_i^M = \mathrm{MaleAdaptationEncoder}(\mathbf{X}_i),\ i = 1, 2$$

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

Group-adapted Fusion Network (GFN)

Back-End



Score fusion model

$$
\begin{aligned}
S^B &= \text{CosineSimilarity}(E_1^B, E_2^B), \\
S^F &= \text{CosineSimilarity}(E_1^F, E_2^F), \\
S^M &= \text{CosineSimilarity}(E_1^M, E_2^M) \\
S &= \text{Sigmoid}(f([S^B, S^F, S^M]; W)).
\end{aligned}
$$

Neural Network

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

Training objective

Binary cross-entropy loss
with positive and negative training pairs

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

Crafted Datasets and Metrics for Fairness

Training sets

- Voxceleb2-GRC (Gender Ratio Controlled) Dataset

Front-End

Gender Ratio (Female:Male)	Female Speakers	Male Speakers	Female Utterances	Male Utterances
9:1	2250	250	387,322	45,181
4:1	2000	500	341,500	95,157
1:1	1250	1250	214,919	228,823
1:4	500	2000	86,616	372,133
1:9	250	2250	43,482	419,853
-	**Total Speakers: 2500**		-	

unbalanced

balanced

unbalanced

Back-End

Sample positive (same speaker) and negative (different speakers) training pairs from VoxCeleb2-GRC for metric learning.

Test sets

- Voxceleb1-F (Fairness) Dataset

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		✓	✓

Crafted Datasets and Metrics for Fairness

Evaluation metrics

Equal error rate (EER) is one of the most common metrics to evaluate speaker verification models, denoting the rate where *False accept rate (FAR) = False rejection rate (FRR).*

Model fairness evaluation via three metrics:

(1) **Group-wise EERs:** monitor group-specific performance

- **Female**-group: $EER[F]$
- **Male**-group: $EER[M]$

(2) **Overall EERs:** monitor performance across all groups

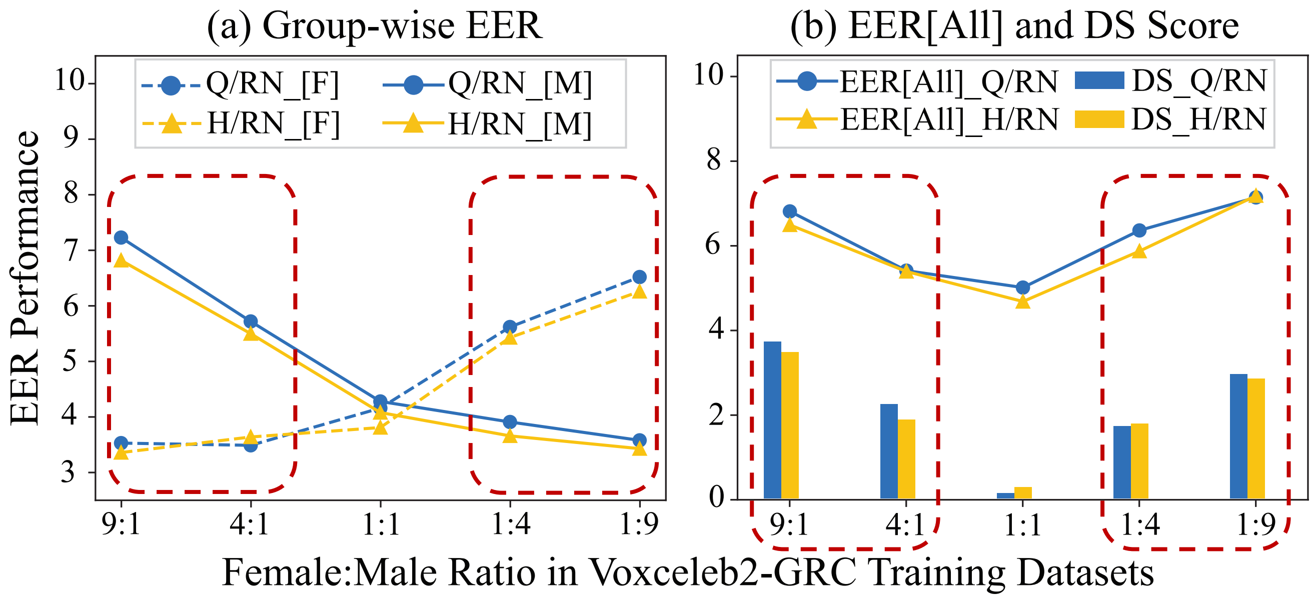
- Overall EER: $EER[\mathrm{All}]$

(3) **Disparity Score (DS):** model performance gap between groups

- $\mathrm{DiparityScore}\ (\mathrm{DS}) = |EER[F] - EER[M]|$

Evaluation Results

RQ1: Does imbalanced group size in training dataset cause model unfairness?



Baselines:

- **Q/RN: Q**uarter-channel **R**es**N**et-34
- **H/RN: H**alf-channel **R**es**N**et-34;

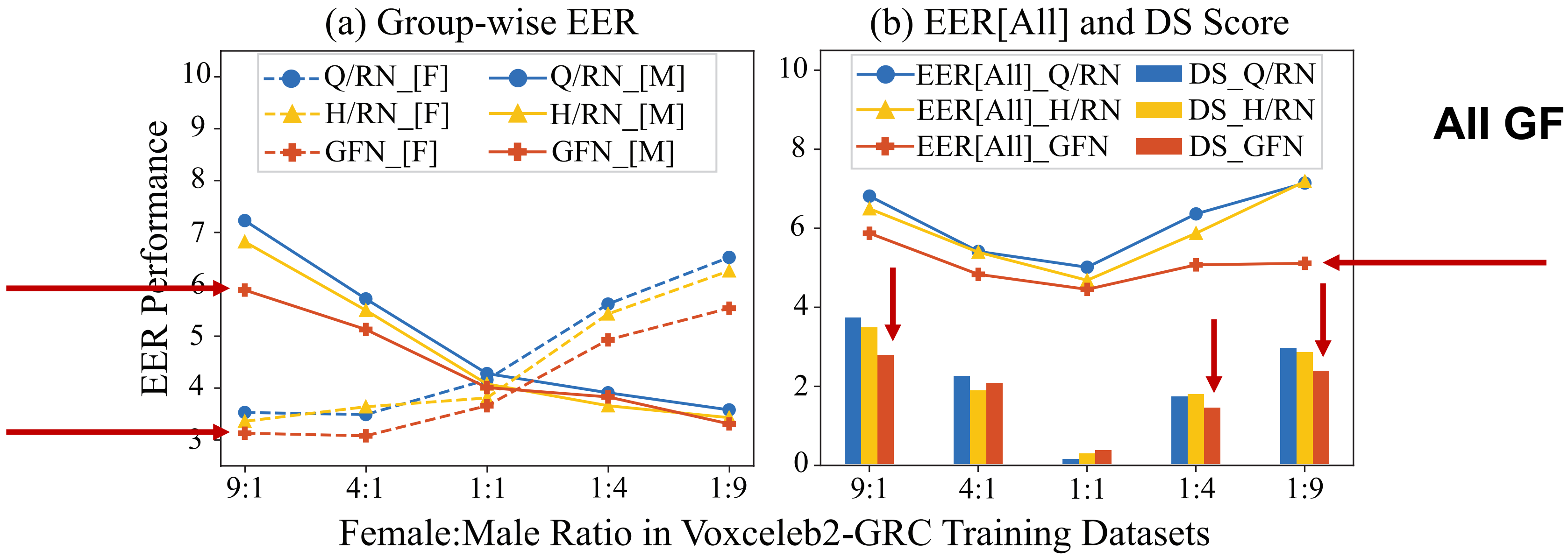
Findings:

- Training with same total speaker numbers (i.e., 2500), the dominant group achieves better group-wise EER than the underrepresented group.
- Increasing dominance of one gender group (e.g., 4:1 → 9:1) leads to increasing performance gap (DS score) and overall EER, indicating increasing model unfairness and worse overall performance, respectively.

Imbalanced group ratios in training sets can lead to model unfairness towards underrepresented groups.

Evaluation Results

RQ2: Can Group-adapted Fusion Network (GFN) improve model fairness?



All GFN's encoders: Q/RN

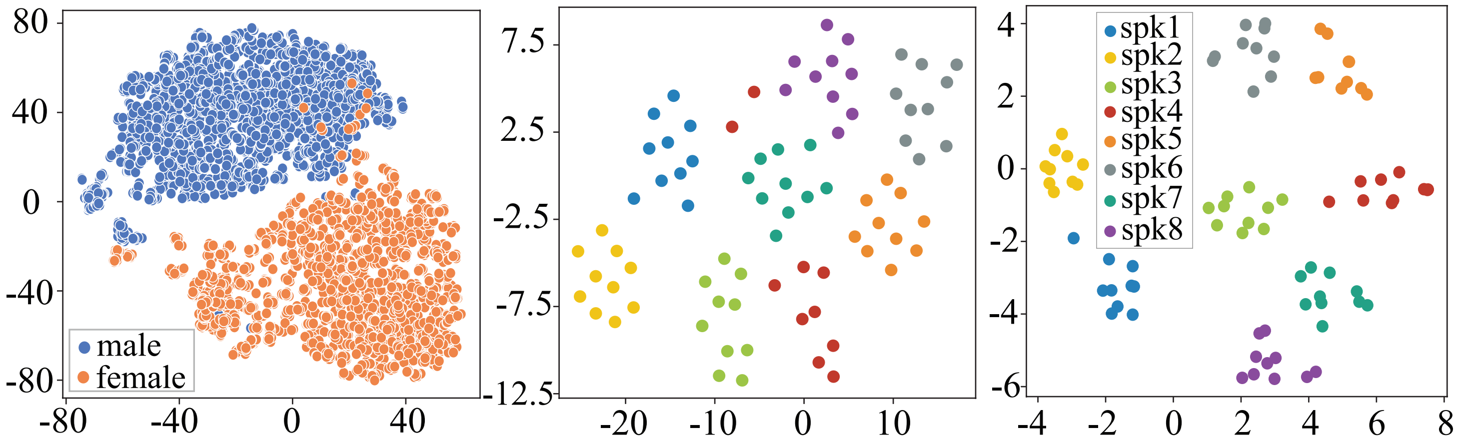
Findings:

- GFN model achieves better group-wise and overall EERs than baselines, regardless of gender group imbalances.
- The GFN also reduces the performance gap (DS Score) in 9:1, 1:4 and 1:9 gender ratio settings.

GFN model can improve gender-specific EER over baselines, and further reduces the performance gap in most imbalanced group ratio settings.

Evaluation Results

RQ3: Embedding visualization and analysis



(a) (b) Q/RN (c) GFN

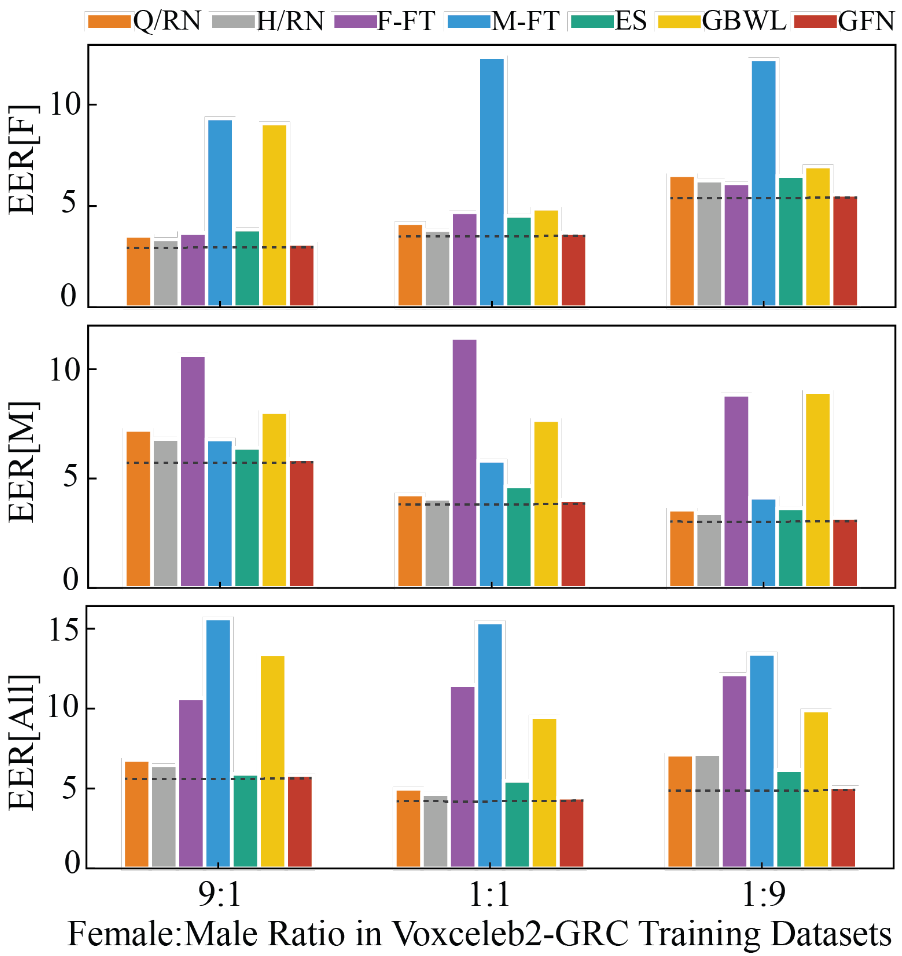
t-SNE projection

Genders tend to aggregate in different regions of the embedding space.

GFN encoder tends to generate higher quality embeddings compared with Q/RN baseline (more compact for the same speakers and separate for different speakers)

Evaluation Results

RQ4: Ablation Study



Listing Methods:

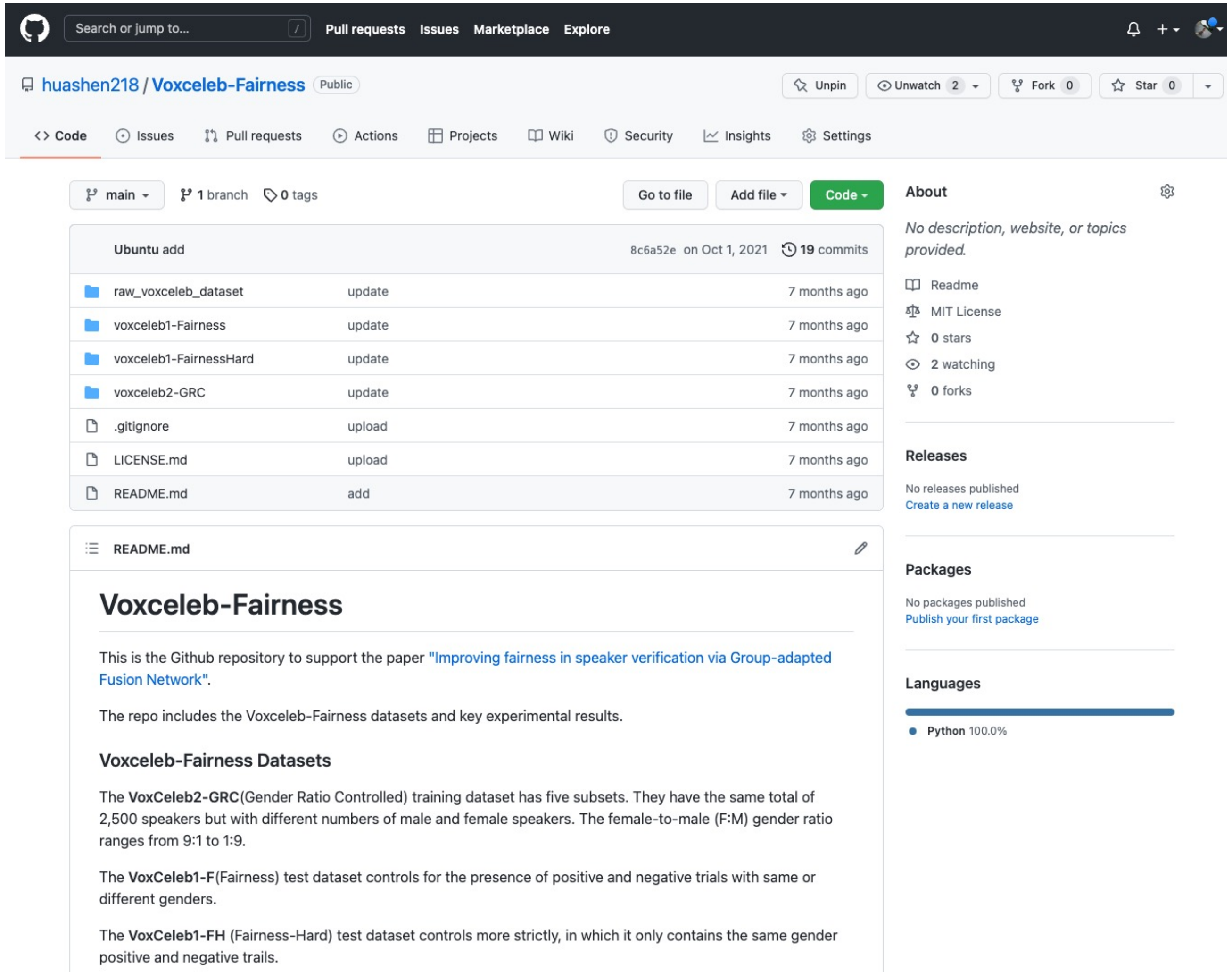
- Gender Batching with Weighted Loss (GBWL);
- Equal Score (ES);
- Female-FineTuned (F-FT);
- Male-FineTuned (M-FT);
- Q/RN Baseline;
- H/RN Baseline.

GFN achieves the best performance among all methods.

Key Takeaways

- We use **evaluation metrics** and **datasets with defined group (male/female) ratios** to analyze model fairness performance.
- We provide the direct evidence that **imbalanced group presence can lead to model unfairness** to different subgroups, specialized in gender-group settings.
- We **propose Group-adapted Fusion Network (GFN),** based on group embedding adaptation and score fusion, to counteract model unfairness.
- We demonstrate that **GFN reduces group-disparity** for imbalanced training scenarios, while **reducing overall speaker verification EER.**

Github: https://github.com/huashen218/Voxceleb-Fairness



Check out our open-source **VoxCeleb2-GRC** and **VoxCeleb1-Fairness** datasets at Github!

Acknowledgement

Alexa Speaker Understanding Members

Long Chen, Zeya Chen, Roger Cheng, Minho Jin, Chelsea Ju,

Terry Tan, Christine Han, Ruirui Li, Hongda Mao, Matt Meng

Managers

Oguz Elibol, Mohamed El-Geish, Itay Teller

Principal Applied Scientist

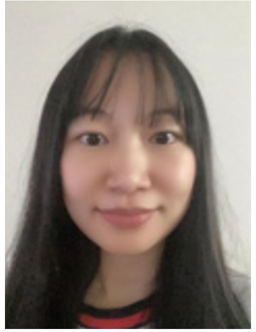
Victor Rozgic

amazon | science

Authorship

Hua Shen

Yuguang Yang



Guoli Sun

Ryan Langman

Eunjung (Christine) Han

Jasha Droppo

Andreas Stolcke

PennState

amazon | science

amazon alexa