Bias & Fairness Evaluation Summary

Methods

I examined dataset and model biases in a LibriSpeech-based speaker verification pipeline using corpus metadata and evaluation scores. I conducted three complementary analyses.

1. I focused on LibriSpeech's `SPEAKERS.TXT` metadata, which provides each reader's gender, subset assignment, and total minutes spoken. To check bias, we performed aggregate, subset-level and distributional analysis.
2. I performed a prosodic coverage study that estimated per speaker median pitch F0 and speaking style proxies such as speech ratio and non-silent segments per second, then compared their distributions by sex to identify gaps that could degrade generalization.
3. I evaluated performance by subgroup using Detection Error Tradeoff curves and Equal Error Rate with thresholds, computed for the overall system, for gender groups, and for speaker duration categories.

Findings

1. **Gender representation:** Nearly balanced with a slight male tilt. Female speakers 1,201; male speakers **1,283**. Minutes: **female 28,553 (48.45%)**, male **30,385 (51.5%)**. Dev and test are ~50 to 50; some training partitions are ~52% male.
2. **Group performance:** Overall EER **0.48%** at threshold **0.4002**. Male EER **0.37%** at **0.4096**; female EER **0.53%** at **0.4053** (**~16%** relative gap favoring males). By duration: <10 minutes shows **~0%** EER at **0.5999** but is unreliable due to a tiny sample; 10 to 20 and 20 plus minutes are **~0.49%** EER at **0.4264** and **0.3981**.
3. **Prosodic coverage:** Median F0 by sex: **F ≈ 199.7 Hz**, **M ≈ 114.7 Hz**. Only one speaker (~0.4%) at ≥260 Hz, indicating sparse coverage of very high pitch voices.

Mitigation Strategies

1. **Data balancing and targeted collection:** Reweight or upsample female segments in skewed training, add high pitch voices at **≥260 Hz** including children, include spontaneous speech, and balance minutes per speaker.
2. **Group aware training:** Monitor subgroup losses, use reweighting, multi task or adversarial objectives, and apply distributionally robust optimization to protect worst case performance.
3. **Calibration and thresholds:** Calibrate scores per group when allowed, tune thresholds to reduce EER gaps, and track impacts on false positives and false negatives.

4. **Continuous audits and intersectional checks:** Run regular audits beyond gender to age, accent, and non native speech, combining metadata with audio features such as F0 bands and style proxies.

Fairness trade offs and stakeholder impact

Threshold tuning can reduce false rejects for underperforming groups but may raise false accepts and security risk. Group specific calibration narrows parity gaps but adds operational complexity and transparency duties. Targeted data collection closes prosodic gaps only with consent, privacy safeguards, and equitable representation. Key stakeholders are end users, system operators, and data subjects. The objective is to lift the worst case group without degrading overall safety, supported by continuous monitoring and clear communication.

Taken together, my analysis shows a nearly balanced corpus with a small male skew, a modest but measurable performance gap against female speakers, and a clear underrepresentation of very high pitch voices. The recommendations are supported by relevant, correctly cited references in the full report.