Identity Disambiguation in Common Voice: Enabling Fairness Evaluation Across Demographic Subgroups

Chenyi Lin[1], *Dāvis Šterns*[1], *Tom Bäckström*[1], *Nicholas Evans*[2]

[1]Department of Information and Communications Engineering, Aalto University, Finland
[2]Digital Security, EURECOM, France

chenyi.lin@aalto.fi, davis.sterns@aalto.fi, tom.backstrom@aalto.fi, nicholas.evans@eurecom.fr

Abstract

Speaker recognition (SR), which automatically identifies an individual based on their voice, is widely deployed in real-world applications, including banking authentication, forensic investigations, and access control systems. Despite its widespread adoption and high reported accuracy, studies have revealed biased performance across demographic subgroups such as age, gender, and ethnicity, raising significant concerns about fairness. For instance, models trained on the widely adopted VoxCeleb benchmark consistently demonstrate marked and systematic performance degradation for female speakers and non-US nationalities [1, 2]. Beyond biased performance, existing benchmark datasets are limited by insufficient demographic metadata. For example, VoxCeleb provides only binary gender and nationality labels, restricting the ability to analyze bias across a broader range of demographic dimensions. These limitations motivate the exploration of alternative datasets for more comprehensive fairness evaluation and the investigation of their potential to support more equitable SR systems.

The Mozilla Common Voice (CV) dataset presents a promising alternative, capturing a more diverse population and offering richer demographic variability than many traditional benchmark datasets [3, 4], including non-binary gender identities, a wider range of age groups, and detailed accent information. This makes it particularly suitable for investigating fairness in SR systems. However, the utility of CV for fairness research is currently undermined by identity heterogeneity, where one speaker is associated with multiple IDs or multiple speakers share the same ID. Such mislabeling can degrade both system performance and the validity of fairness evaluations, as SR systems fundamentally rely on accurate speaker identity annotations. To address this, we first improve the reliability of speaker labels to construct a cleaner resource, and then evaluate the fairness of SR systems on the CV corpus.

Prior work has addressed speaker ID inconsistencies using embedding-based methods combined with statistical analysis ([5, 6, 4]). However, these approaches rely on expert auditing, which limits scalability and introduces subjectivity. In addition, they often overlook cases where a single speaker is associated with multiple IDs. To mitigate these limitations, we propose a scalable and automated framework that eliminates the need for perceptual auditing. We first extract speaker embeddings using a pretrained ECAPA-TDNN model and compute pairwise similarity scores on the CV dataset. We then model the distribution of these scores and compute confidence scores for each verification pair, reflecting the likelihood of label inconsistency. These confidence scores are used as sample weights when retraining the ECAPA-TDNN model, down-weighting samples associated with potentially inconsistent labels.

Beyond inconsistent speaker labels, challenges also arise in leveraging demographic attributes for fairness evaluation. While the CV corpus provides rich metadata on age, gender, and accent, the accent attribute is not readily usable in its raw form. This is partly because accent information in CV is collected as self-reported free-text entries, resulting in substantial heterogeneity and noise, which prevents consistent grouping of speakers and reliable subgroup comparisons. Given evidence of accent-based bias in other speech domains [7] and its underexploration in SR, structuring accent information in CV is necessary for fairness evaluation. The same limitation applies to other potentially valuable but underexplored attributes, such as speech impediments, socio-economic status, and education level, which likewise lack standardized representations, hindering their systematic use despite their potential for more fine-grained fairness analysis. To overcome these limitations, we propose an automated pipeline that utilizes AI to normalize heterogeneous text entries, identify salient cues (e.g., accent region, L1/L2 status, speech characteristics), map them into a unified taxonomy, followed by a final human verification step to ensure accuracy, thereby enabling a reproducible framework for fairness evaluation across demographic subgroups.

Building on these data cleaning and demographic structuring steps, we conduct a multi-level evaluation of performance and fairness at the population, subgroup, and individual levels, following the SVEva speaker verification fairness framework [1]. Metrics include Equal Error Rate (EER), minDCF, and DET curve analysis, and t-SNE–based representation analysis.

Following the mitigation of speaker heterogeneity, the overall EER is hypothesized to decrease relative to the baseline established before correcting mislabeled client IDs. Drawing on prior fairness evaluations that identify training data imbalance as a primary driver of subgroup bias ([8]), we anticipate performance disparities across several demographic dimensions, particularly affecting (i) female and minority gender subgroups, including transgenders; (ii) older speakers, particularly those over the age of 50; and (iii) non-US English speakers, with an emphasis on non-native speakers. Furthermore, this analysis extends fairness evaluation to the individual level by examining specific speakers, such as polyglots. As their speech often reflects a mixture of multiple accents, they provide insight into how accent variability influences SR fairness.

In this work, our contributions are threefold: (1) a scalable method for mitigating speaker identity inconsistencies; (2) an automated pipeline for structuring heterogeneous demographic metadata; and (3) a comprehensive fairness evaluation of SR systems on CV. Together, these provide a robust foundation for fairness evaluation in SR and enable more reliable investigation of fairness–privacy trade-offs, as highlighted in prior work [9].

Index Terms: speaker recognition, common voice dataset, fairness, bias, demographic disparity, evaluation

1. Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme grant No 101168193.

2. References

[1] W. Toussaint and A. Y. Ding, "SVEva Fair: A Framework for Evaluating Fairness in Speaker Verification," Oct. 2022.

[2] W. T. Hutiri and A. Y. Ding, "Bias in Automated Speaker Recognition," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 230–247.

[3] G. Fenu, M. Marras, G. Medda, and G. Meloni, "Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 1892–1896.

[4] J. Hintz and I. Siegert, "CommonBench: A larger Scale Speaker Verification Benchmark," in *4th Symposium on Security and Privacy in Speech Communication*. ISCA, Sep. 2024, pp. 17–20.

[5] A. Farhadipour, J. Marquenie, S. Madikeri, and E. Chodroff, "TidyVoice: A Curated Multilingual Dataset for Speaker Verification Derived from Common Voice," Jan. 2026.

[6] M. Zhang, A. Farhadipour, A. Baker, J. Ma, B. Pricop, and E. Chodroff, "Quantifying and Reducing Speaker Heterogeneity within the Common Voice Corpus for Phonetic Analysis," May 2025.

[7] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," Apr. 2021.

[8] M. Baali, S. Bisht, F. Teixeira, K. Shapovalenko, R. Singh, and B. Raj, "Sveritas: Benchmark for robust speaker verification under diverse conditions," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 9714–9731.

[9] A. Leschanowsky and S. Das, "Examining the Interplay Between Privacy and Fairness for Speech Processing: A Review and Perspective," Sep. 2024.