SYNVOX2: TOWARDS A PRIVACY-FRIENDLY VOXCELEB2 DATASET

Xiaoxiao Miao[1], Xin Wang[2], Erica Cooper[2], Junichi Yamagishi[2], Nicholas Evans[3] , Massimiliano Todisco[3] , Jean-Francois Bonastre[4,5] , Mickael Rouvier[5]

[1]Singapore Institute of Technology, [2]National institute of Informatics, [3]Eurecom, [4]Inria, [5]University of Avignon

Background

The widely-used large-scale multilingual VoxCeleb2 [1], with over 1 million utterances from nearly 7000 speakers, has become a standard ASV benchmark, cannot be downloaded from the official website [2] due to privacy issues.

Can we create SynVox2 with fewer privacy concerns while maintaining utility and fairness?

Requirements for a privacy-friendly synthetic speech database

Speaker privacy protection
Anonymized speech sounds dissimilar from original speech

Speaker diversity
Anonymized speech from the same speaker has a unique speaker identity

Speech intelligibility and naturalness
Anonymized speech satisfies the same distribution as original speech

Requirements are similar -> use language-robust OHNN-SAS [3] to create SynVox2

⟺

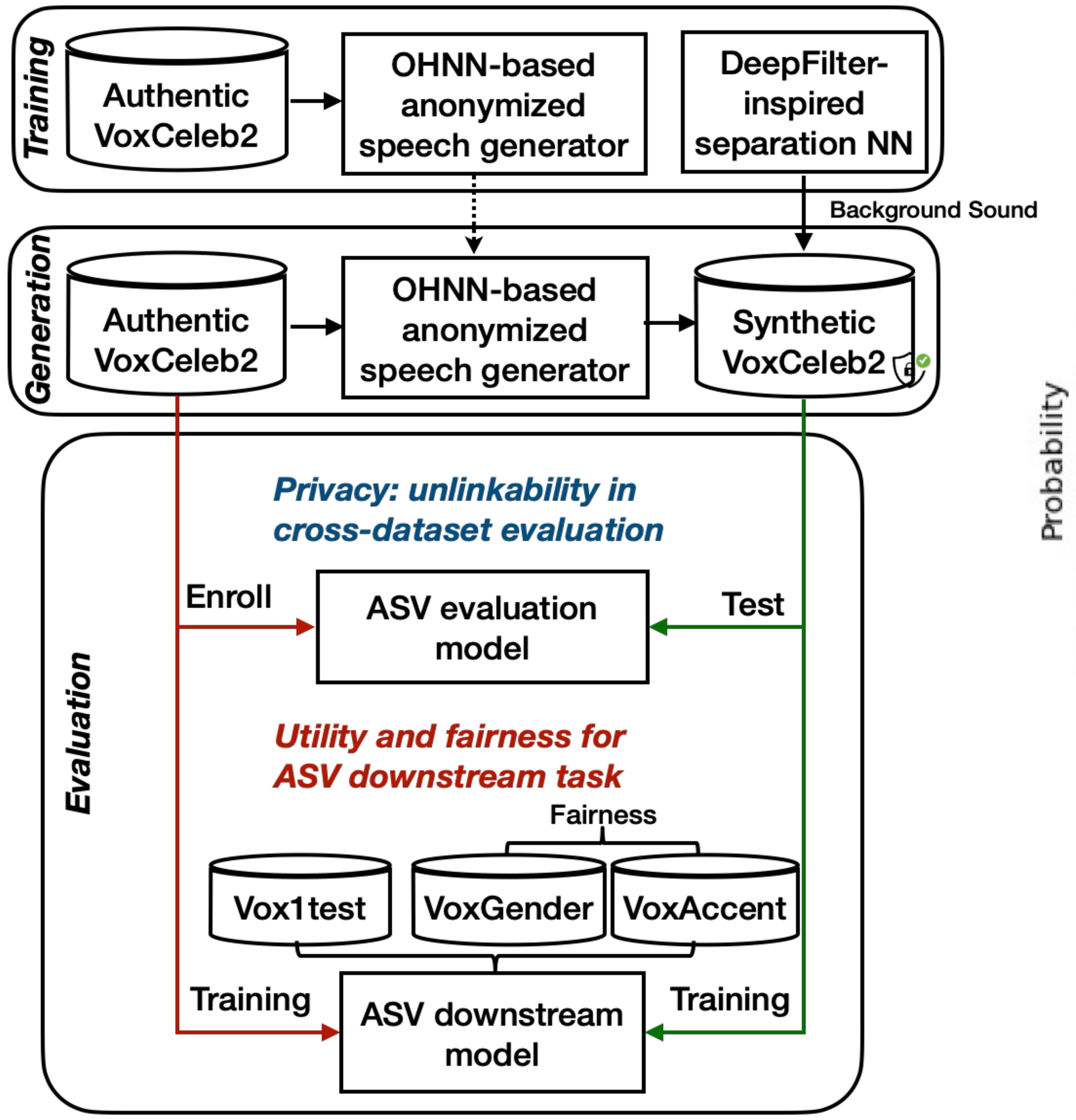
One Issue: OHNN-SAS, trained on clean speech, cannot generate wild VoxCeleb2
Solution: Extract background sounds and add them back to the speech.

Ensuring privacy
Synthetic speaker identity is unlinkable to its original identity

Maintaining utility
ASV model trained on synthetic data are expected to perform similarly to models trained using authentic data

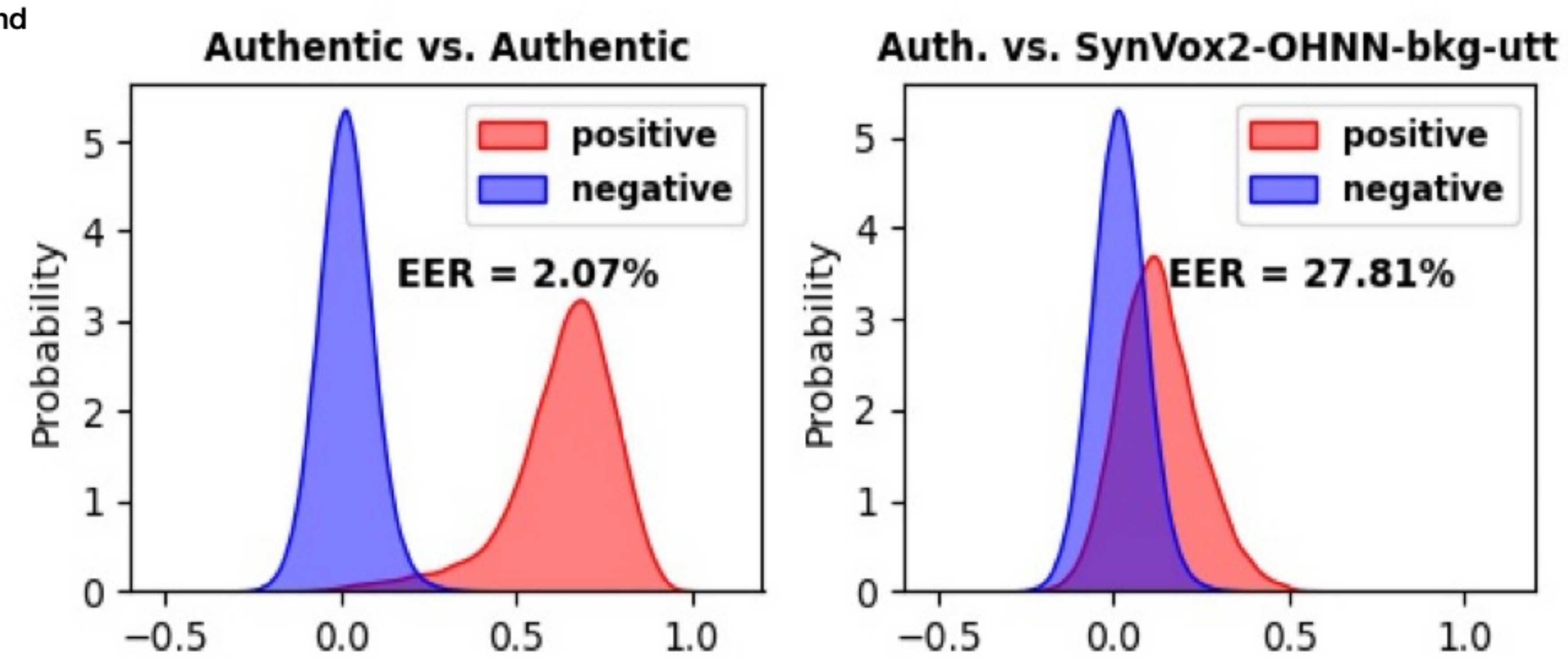
Increasing fairness
Not disfavor any particular group in the test set, e.g., genders, dialects

SynVox2 Generation Methods and Experiments



[1] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild", Computer Speech & Language, 2019
[2] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/
[3] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Speaker anonymization using orthogonal householder neural network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023

Q1: Do SynVox2 datasets protect speaker identity information? -> *Speaker privacy can be protected through anonymization*



Q2: Can SynVox2 datasets be used to train an ASV model? -> it's possible but..

Training dataset	EER(%) ↓
Authentic	1.33
SynVox2-OHNN-bkg-utt	7.58

Q3: Are the ASV models trained using SynVox2 datasets fair in terms of gender and accent? -> *Fairness degrades with the use of synthetic data*

Fairness Disrepancy Rate (FDR): Given decision threshold τ , the FDR considers the largest distance between false alarm rates (FAR) and false reject rates (FRR) over multiple groups $D = \{d_1 \ldots d_i \ldots\}$

$$FDR = 1 - [\alpha \times \max(|FAR^{d_i}(\tau) - FAR^{d_j}(\tau)|) + (1-\alpha)\times \max(|FRR^{d_i}(\tau) - FRR^{d_j}(\tau)|)]$$

FDR ↑	VoxGender	VoxAccent
Authentic	0.972	0.894
SynVox2-OHNN-bkg-utt	0.875	0.764

Q4: What is the bottleneck?
-> Obviously reduction in inter-speaker variation for the SynVox2