Exploring Demographic Effects on Speaker Verification

IEEE CNS 21 Poster

Sophie Si
Computer Science and Engineering
University of California, Riverside
Riverside, CA
ssi003@ucr.edu

Zhengxiong Li
Computer Science and Engineering
University of Colorado Denver
Denver, CO
zhengxiong.li@ucdenver.edu

Wenyao Xu
Computer Science and Engineering
University at Buffalo, SUNY
Buffalo, NY
wenyaoxu@buffalo.edu

2021 IEEE Conference on Communications and Network Security (CNS) | 978-1-6654-4496-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CNS53000.2021.9729038

***Abstract*—Voice biometrics (e.g., Speaker Verification) is a critical type of biometrics based on human voice characteristics and is known for security and user-friendliness. It has been widely applied in worldwide applications, such as voice assistants and online banking. However, a concern is raised rapidly about the demographic fairness that different subgroups may have different speaker verification performance due to the inherent voice characteristics. And little work done investigates this concern. A diverse group of 300 speakers by race and gender is recruited for exploration. After running some speaker verification evaluations, three conclusions were reached. Firstly, the Latinx are performed the worst among the four major races in the US (White, Black, Latinx, and Asian) in speaker verification. Secondly, that gender shows little difference in performance between men and women. Thirdly, that high entropy voices performed better than low entropy voices in speaker verification performance.**

***Index Terms*—voice biometrics, speaker verification, fairness**

I. Introduction

Voice biometrics utilizes human voice characteristics and voice patterns to recognize the speakers. With the advantages of security, user-friendliness, and low cost, it (e.g., Speaker Verification, SV) has been widely deployed in a lot of applications for hundreds of thousands of users from different backgrounds around the world, such as voice assistants and online banking [2]. Currently, there isn't a lot of work done to investigate the connection between the inherent voice characteristics (i.e., entropy) and speaker verification performance which is a concern because different groups of people with different vocal structures may have different vocal properties, such as voice entropy that may lead them to have different barriers to enjoy the bio-metrics. Voice entropy, in particular, is noted because it represents the information capacity in the voice, and low entropy voices may have fewer notable traits to identify. By investigating this area, we discover a new unnoticeable fairness issue in voice-based biometrics.

METHODOLOGY

Speaker Verification Models: Three state-of-the-art representative speaker verification products are utilized in this work, the Xvector-TDNN model, the ECAPA-TDNN model and the DTW model.

Voice Entropy Metrics: Two representative biometric entropy metrics are utilized: PDF entropy and perm entropy.

Voice Datasets: The data used in this work is a subset of mPower—a smartphone-based clinical observational study [1]. The voice recording methodology is close to the real practice condition in voice biometrics. To explore the racial and gender disparities in voice biometric, we set two matched datasets on race and gender, respectively. In the racial dataset, there are four sub-groups, including White, Black, Latinx, Asian. 75 speakers with 512 snippets are collected for each sub-group. Besides, in the gender dataset, there are two sub-groups, female and male. 150 female speakers and 150 male speakers are recruited.

Data Analysis and Metrics: To illustrate the disparity or significant difference among different subgroups, ANOVA statistics and POST-HOC statistics with Turkey Honest Significant Difference (HSD) are employed, comparing scores between demographics. All voices are tested against each enrolled voice of each demographic. To analyze speaker verification performance in race and gender, the raw scores of all comparisons between different speakers are grouped by demographic and compared. To analyze SV performance in relation to entropy, the entropy of all voices tested are grouped as True-Positive (TP), True-Negative (TN), False-Positive (FP) and False-Negative (FN) and compared. The p-value derived from the resulting test in combination with the mean value calculated is used. The p-val<0.5 for ANOVA represents existence of significant differences in general while for HSD represents existence of significant difference between two specific groups. HSD mean difference represents the difference between two groups with negative meaning that the first group is larger. For ECAPA and Xvector, and as opposed to DTW, lower scores represent speaker mismatch and therefore accuracy.

TABLE I
ANOVA/HSD RESULTS FOR DEMOGRAPHICS

statistic	group1	group2	meandiff	p-val
dtw	female	male	-2478.9929	0.001
ecapa	all gender	all gender	N/A	0.920781665
xvector	female	male	-0.2602	0.001
dtw	asian	black	-1559.6426	0.0066
dtw	asian	latinx	-4414.3964	0.001
dtw	asian	white	-2122.635	0.001
dtw	black	latinx	-2854.7538	0.001
dtw	black	white	-562.9924	0.6291
dtw	latinx	white	2291.7614	0.001
ecapa	asian	black	0.0055	0.8053
ecapa	asian	latinx	0.0225	0.0024
ecapa	asian	white	-0.0237	0.0012
ecapa	black	latinx	0.0171	0.0385
ecapa	black	white	-0.0292	0.001
ecapa	latinx	white	-0.0462	0.001
xvector	asian	black	-0.0064	0.9
xvector	asian	latinx	0.1194	0.001
xvector	asian	white	-0.0108	0.9
xvector	black	latinx	0.1258	0.001
xvector	black	white	-0.0044	0.9
xvector	latinx	white	-0.1302	0.001

TABLE II
ANOVA/HSD RESULTS FOR ENTROPY

statistic	group1	group2	meandiff	p-val
pdf	FN (dtw)	FP (dtw)	0	0.9
pdf	FN (dtw)	TN (dtw)	0.0104	0.001
pdf	FN (dtw)	TP (dtw)	0.0619	0.001
pdf	FP (dtw)	TN (dtw)	0.0104	0.001
pdf	FP (dtw)	TP (dtw)	0.0619	0.001
pdf	TN (dtw)	TP (dtw)	0.0515	0.001
perm	FN (dtw)	FP (dtw)	0	0.9
perm	FN (dtw)	TN (dtw)	0.0229	0.001
perm	FN (dtw)	TP (dtw)	-0.0003	0.9
perm	FP (dtw)	TN (dtw)	0.0229	0.001
perm	FP (dtw)	TP (dtw)	-0.0003	0.9
perm	TN (dtw)	TP (dtw)	-0.0232	0.5966
pdf	FN (ecapa)	FP (ecapa)	0	0.9
pdf	FN (ecapa)	TN (ecapa)	0.0027	0.5792
pdf	FN (ecapa)	TP (ecapa)	0.0677	0.001
pdf	FP (ecapa)	TN (ecapa)	0.0027	0.5792
pdf	FP (ecapa)	TP (ecapa)	0.0677	0.001
pdf	TN (ecapa)	TP (ecapa)	0.065	0.001
perm	all ecapa	all ecapa	N/A	0.814116585
pdf	FN (xvec)	FP (xvec)	0	0.9
pdf	FN (xvec)	TN (xvec)	-0.0172	0.001
pdf	FN (xvec)	TP (xvec)	0.0415	0.0471
pdf	FP (xvec)	TN (xvec)	-0.0172	0.001
pdf	FP (xvec)	TP (xvec)	0.0415	0.0471
pdf	TN (xvec)	TP (xvec)	0.0587	0.0014
perm	FN (xvec)	FP (xvec)	0	0.9
perm	FN (xvec)	TN (xvec)	-0.0541	0.001
perm	FN (xvec)	TP (xvec)	-0.0573	0.0882
perm	FP (xvec)	TN (xvec)	-0.0541	0.001
perm	FP (xvec)	TP (xvec)	-0.0573	0.0882
perm	TN (xvec)	TP (xvec)	-0.0032	0.9

II. EVALUATION

A. Gender Effect

As shown in Table. I, the difference in scores between genders is found to be significant when using Xvector and DTW models but not when using the ECAPA model. Moreover, in both Xvector and DTW, men are more likely to have a lower score. This means that under the Xvector model women are more likely to be falsely accepted as a match, that under the DTW model the opposite was true, and under the ECAPA model, neither group performed better. *Considering the contradictory results between models, it is concluded that there is little difference on gender in general in terms of performance in SV systems. This may also mean that gender does not majorly impact adversarial attacks as well.*

B. Race Effect

As shown in Table. I, Latinx scores are shown to be significantly different from other racial groups in all three models. In the DTW model, Latinx would have a significantly lower score than all other categories, while the opposite is generally true in the Xvector and ECAPA model. This would mean that within all three models a Latinx person would generally be the most likely category to be falsely accepted into the system all system. *In conclusion, when comparing two mismatching voices, a false positive result is more likely to be reached if a Latinx voice is involved. This may imply that Latinx are the most susceptible to attacks among tested races.*

C. Entropy Analysis

As shown in Table. II, generally, accurate results (TP, TN) will have a higher entropy than inaccurate results (FP, FN). *This leads me to the conclusion that voices with high entropy will produce more accurate results. This may imply that high entropy voices are more protected against adversarial attacks.*

III. CONCLUSION

In this work, we discovered a new unnoticeable fairness issue in voice biometrics. Different subgroups have different speaker verification performances due to their inherent voice characteristics, such as the Latinx subgroup has the worst performance compared to other subgroups. Besides, it is time to rethink this technology and prevent it from causing potential hazards or bias toward particular subgroups.

ACKNOWLEDGMENT

This work was supported by the U.S. National Science Foundation Research Experiences for Undergraduates (NSF REU) Site Program under Grant CNS-2050910.

REFERENCES

[1] Bot, Brian M., et al. ”The mPower study, Parkinson disease mobile data collected using ResearchKit.” Scientific data 3.1 (2016): 1-9.

[2] Chen, Guangke, et al. ”Who is real bob? adversarial attacks on speaker recognition systems.” IEEE Oakland (2021).