Voice Doppelgänger Susceptibility among Racial and Gender Groups

IEEE CNS 21 Poster

Emily Wu
Computer Science and Engineering
University of California, Riverside
Riverside, CA
ewu020@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/CNS53000.2021.9729035

***Abstract*—Voice doppelganger describes a person or thing that sounds like another, which is considered a potential security risk for voice biometrics. Considering that racial or gender groups have different biological vocal structures, it is possible that these subgroups have different vulnerabilities to the voice doppelganger and voice biometrics. This study investigates if racial and gender disparities exist in the security risk of the voice doppelganger towards the voice biometric. We used three different metrics to measure voice similarity: fundamental frequency, MFCCs, and pitch. The result was that persons within gender and racial subgroups do indeed sound more similar to each other, with racial subgroups displaying more similarity than gender.**

***Index Terms*—voice biometrics, voice doppelganger, fairness**

I. Introduction

There are many applications of voice biometrics that are used broadly, including: automated call center routing, IoT devices, automatic translation, closed captioning, virtual assistants, speech dictation, and of course security access. Systems which employ speech for either useability or security also increase their accessibility. However, this raises the concern if persons within specific subgroups are naturally more prone to sound similar to each other. Little work explores the relationship between voice doppelganger and these demographic factors.

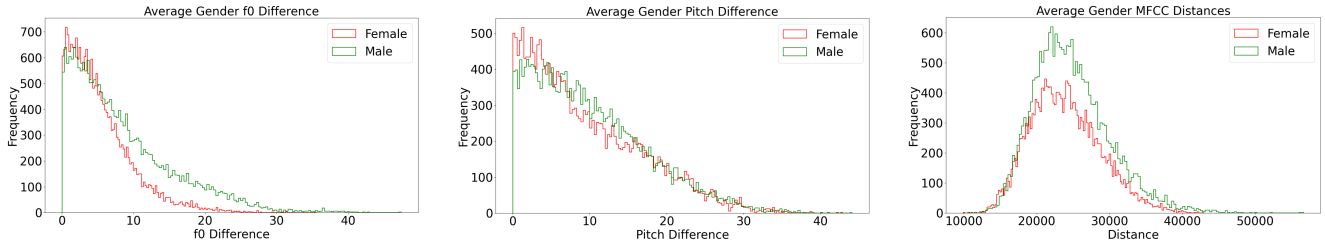
II. METHODOLOGY

In this study, we chose to measure speaker similarity by comparing the differences of distances between speakers in terms of vocal features. As aforementioned, the three voice similarity features used were fundamental frequency (f0), Mel-frequency Cepstrum Coefficients (MFCCs), and pitch. Based on the source-filter theory [1], human speech can be characterized as a glottal source signal filtered through the vocal tract. The features were chosen with this concept in mind: f0 represents the frequency at which the vocal cords vibrate during speech, and originates from the glottis. Likewise, the pitch is the perception of the f0, and therefore also glottal. MFCCs describe the vocal timbre, which are the characteristic qualities of a sound apart from its pitch/intensity and result from the shape and spacings of the vocal tract filters. Through analyzing features of both glottal and vocal tract origin, more specific biological disparities between racial and gender subgroups may be identified.

MFCCs are one of the most widely used features in speech signal processing and vocal recognition. They are cited in numerous studies [2] and give good results describing speech signals, which are useful for speaker similarity comparison. f0 and pitch were also used in the above cited studies as well as many other articles as features for speech signal comparison/processing.

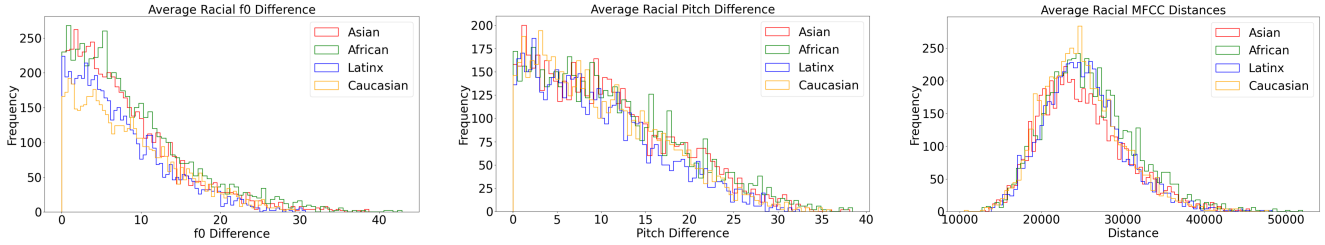
We used a gender and race matched dataset consisting of 75 participants in each race: Caucasian, African, Asian, and Latinx. In particular, these are the predominant racial groups in the US. There are also two gender groupings: male and female, each consisting of 150 participants. The average age of White, Black, Latino, Asian are 32.02±11.46 y, 30.93±10.36 y, 27.31±8.45 y, 28.57±9.23 y, respectively. The average age of female and male are 32.18±13.19 y, and 32.93±12.39 y, respectively. The recordings in the dataset are /ɑ/ (‘a’) vocalization snippets from mPower dataset, 10 seconds long each. All recordings have been manually checked for excessive noise or other unsuitable properties.

Average speaker distances were measured by calculating the MFCCs of each .wav file using librosa [3], then employing fastDTW to compare each speaker against every other speaker in the dataset, from which the mean distances and standard deviations are derived. Average speaker pitch was detected using the YIN pitch detection algorithm, and speaker pitches were similarly compared to each other. Average speaker f0 was estimated using a fast harmonic comb filter, mean differences between speaker f0’s calculated as before. Using the mean values and standard deviations, multiple coefficients of variation, CV = $\frac{\sigma}{\mu}$,were calculated and used to generate results for each subgroup. Finally, frequency histograms of the speaker distances were generated. The results were analyzed using a one-way analysis of variance (ANOVA).



(a) f0 (b) Pitch (c) MFCC

Fig. 1. Three three voice similarity features: (a) f0; (b) Pitch; (c) MFCC.



(a) f0 (b) Pitch (c) MFCC

Fig. 2. Voice Doppelganger evaluation on gender and race subgroups with three primary voice features in voice biometrics: (a) f0; (b) Pitch; (c) MFCC.

III. Results

A. Gender Effect

As shown in Fig. 1 and 3, the p-values for all gender subgroups are $p < 0.01$ and 95% CI does not contain 0, which implies that the differences in vocal features between the various inter-subgroups is statistically significant. Additionally, gender-matching in intra-subgroup evaluation did not prove to result in as much similarity as race-matching.

B. Race Effect

As shown in Fig. 2 and 3, the p-values for all race subgroups are $p < 0.01$ and 95% CI does not contain 0, which implies that the differences in vocal features between the various inter-subgroups is statistically significant. However, for intra-group, overall, the Caucasian speakers were the most similar to each other and thus may be more vulnerable to security risks in voice biometrics. Based on each individual feature, Caucasians, Asians, and Latinx were most similar in f0, pitch, and MFCC distance, respectively.

ANOVA								
f0			Pitch			MFCC Distance		
Gender	df_b	1	Gender	df_b	1	Gender	df_b	1
	df_w	44698		df_w	44698		df_w	44698
	F	3070.217		F	69.30691		F	303.1443
Race	df_b	3	Race	df_b	3	Race	df_b	3
	df_w	22196		df_w	22196		df_w	22196
	F	46.57979		F	47.25538		F	44.63407

Fig. 3. ANOVA of the results

Conclusions

Based on the results above, there seems to be a greater probability of voice doppelgangers within these same race or gender subgroups. The persons within these subgroups proved to be significantly similar in terms of vocal frequency, pitch, and timbre. Therefore, these demographic subgroups are more likely to suffer voice biometric vulnerabilities and security risks. Besides, race seems to be a bigger factor in the probability of voice similarity than gender. In other words, persons of the same race are more likely to sound similar than persons of the same gender. Especially, the Caucasian speakers were the most similar to each other.

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] San Segundo, E., Tsanas, A., Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. Forensic Science International, 270, 25–38. https://doi.org/10.1016/j.forsciint.2016.11.020

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, August 1980, doi: 10.1109/TASSP.1980.1163420.

[3] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015.