Exploring the impact of face coverings on x-vector speaker recognition using VOCALISE

Tom Iszatt, Ekrem Malkoc, Finnian Kelly, and Anil Alexander
Oxford Wave Research, Oxford, UK
{tom|ekrem|finnian|anil}@oxfordwaveresearch.com

Introduction

With face coverings becoming commonplace as a result of the COVID-19 pandemic, biometric recognition systems based on face and voice have encountered new challenges. In some cases, face recognition performance is severely affected when faces are occluded with coverings (Damer et al., 2020; Ngan, Grother, & Hanaoka, 2020), yet a previous study has shown i-vector automatic speaker recognition systems to be fairly robust to speech under face coverings (Saeidi et al., 2015). This study explores the effect of face coverings on speaker recognition using VOCALISE, an x-vector speaker recognition system, and two independent datasets of face covering recordings.

Methods

A new face covering dataset, the OWR Audio Face Covering Corpus (OWR-AFCC), was collected for use in this study. Short speech samples were recorded by eight participants, six male and two female, on their phones. Recordings were made without a face covering, with a fabric covering and with a surgical mask, twice indoors and twice outdoors (Figure 1). The indoor recordings without a face covering were compared to all outdoor recordings using VOCALISE in x-vector PLDA mode. As a point of comparison, audio speech data from the "Audio-Visual Face Cover Corpus" (NF-AVFCC), a gender-balanced 10-speaker corpus containing short utterances under a range of face covering conditions recorded indoors in a studio (Fecher, 2012; Figure 1), were concatenated, creating a single speech recording for each speaker under each face covering condition. The recordings without a face covering were compared to the recordings with different types of face coverings, again using VOCALISE in x-vector PLDA mode. In both cases, Equal Error Rates (EERs) were calculated using Bio-Metrics performance metrics software.

Figure 1. Photos of face covering conditions from OWR-AFCC (left) and NF-AVFCC (right; reproduced with permission from Fecher, 2012).

Results

An EER of 0.00% was observed for all the comparisons made on OWR-AFCC, indicating a perfect discrimination of speakers. Same-speaker scores were lower under the fabric covering condition, but not to a sufficient extent to introduce errors, while the same-speaker scores with a surgical mask and without a face covering were similar (Figure 2). The EERs for NF-AVFCC differed across face covering conditions. The surgical mask condition had an EER of 0.00%, as did the hoodie/scarf combination condition, the closest analogue to the fabric coverings in OWR-AFCC, though again, the surgical mask same-speaker scores were the highest. The balaclava (one hole), niqāb, and rubber mask conditions also resulted in no errors

(0.00% EER). On the extreme end of face coverings, the motorcycle helmet condition resulted in a 4.00% EER, and the taped-up mouth condition resulted in a 14.67% EER. This degraded performance was due to lowered same-speaker scores – in this case, different-speaker scores were also lowered, but to a lesser degree (Figure 2). Thus, these results suggest that everyday face coverings do not negatively affect x-vector speaker recognition performance.

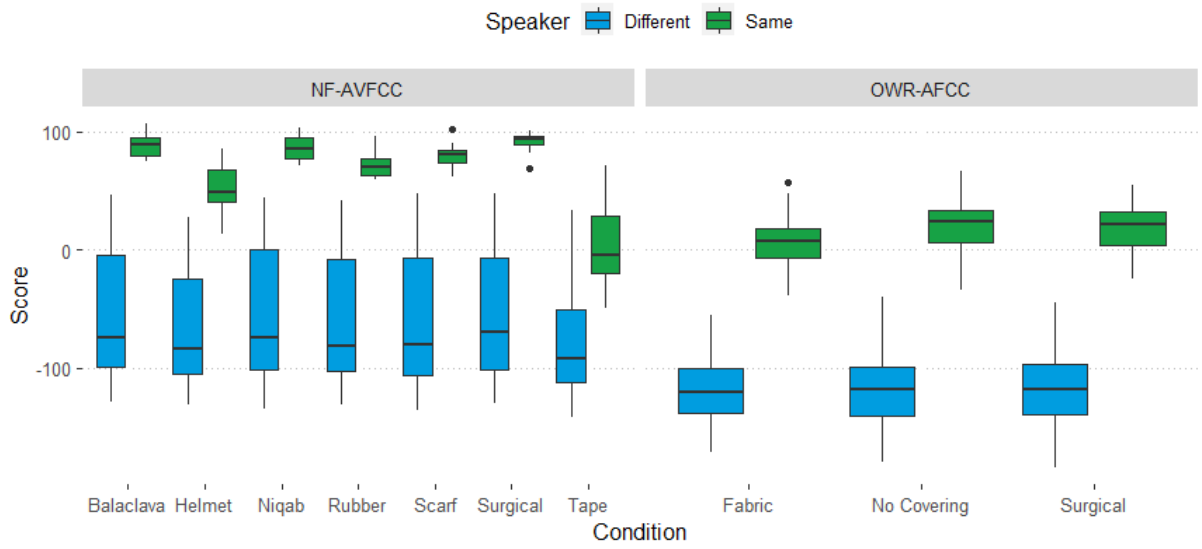

Figure 2. Box plot comparing same-speaker and different-speaker scores in VOCALISE under a range of face covering conditions from NF-AVFCC and OWR-AFCC.

References

Damer, N., Grebe, J. H., Chen, C., Boutros, F., Kirchbuchner, F., and Kuijper, A. (2020). The effect of wearing a mask on face recognition performance: An exploratory study. *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1-6.

Fecher, N. (2012). The "audio-visual face cover corpus": Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear. *INTERSPEECH*, 2250–2253.

Ngan, M., Grother, P., and Hanaoka, K. (2020). Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms. *NISTIR 8311.*

Saeidi, R., Niemi, T., Karppelin, H., Pohjalainen, J., Kinnunen, T., and Alku, P. (2015). Speaker recognition for speech under face cover. *INTERSPEECH*, 1800–1804.