Gender bias in voice recognition: An i-vector-based gender-specific automatic speaker recognition study

Thayabaran Kathiresan[1], Arjun Verma[1], and Volker Dellwo[1]

[1]Phonetics and Speech Sciences, Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

thayabaran.kathiresan@uzh.ch

Abstract

The performance of automatic speaker recognition (ASR) algorithms is continuously increasing. However, acoustic variabilities due to gender differences are challenges, and numerous methods have been proposed to handle them. Among many solutions, using gender-dependent features to train the ASR and testing with known gender trials [1][2] improves the overall recognition accuracy. However, the relative accuracy difference between gender-specific testing still exists. In this paper, we address the fundamental acoustic differences between gender concerning the ASR. We carried out a) an i-vector-based ASR experiment [3] on the TIMIT corpus (130 female and 290 male speakers) and b) the i-vector speaker embedding acoustic analysis. The i-vector extractor was trained on 7323 speakers (1211 speakers from the Voxceleb1 dataset and 6112 from the Voxceleb2 dataset [4]). The system was based on a UBM with 2048 Gaussian mixtures and a gender-independent total variability matrix with 400 total factors. We employed an i-vector length normalization (LN) to the 400-dimensional i-vector. Linear discriminant analysis (LDA) was used to alleviate intra-speaker variability further and reduce the dimension to 200. Finally, PLDA models with 200 latent identity factors were trained. On the trained i-vector system, we carried out the recognition experiment on the male and female speakers separately. The results show that the equal error rate (EER) for male speakers is 3.937% and for female speakers, 5.128%.

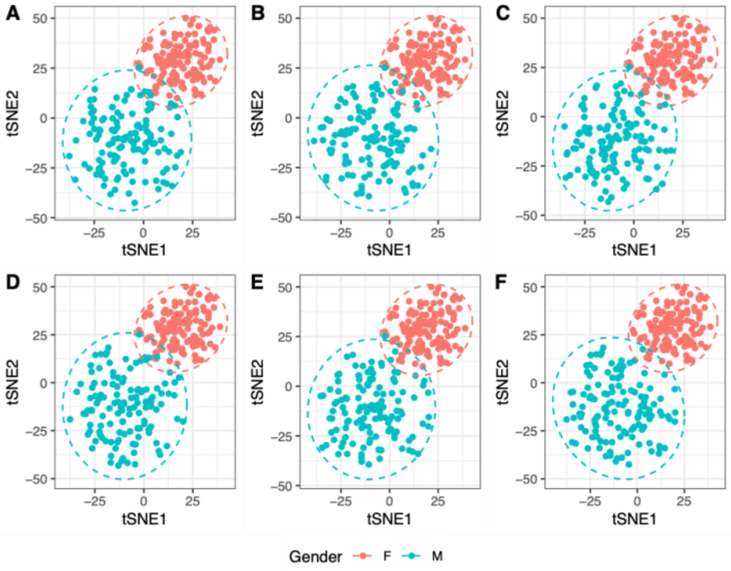

Fig. 1 Distribution of i-vector speaker embeddings (400 dimension) in the t-SNE indexical space (2 dimension). Subfigures A-F indicates the random sampling (6-fold) of male speakers (130 out of 290), and the female speaker count kept constant (130).

We used t-distributed stochastic neighbor embedding (t-SNE), a dimension reduction technique, to reduce the 400-dimensional i-vector to 2-dimension features. To keep the balanced speaker count in both genders, we randomly sampled male speakers to match the female speakers count, and the females' speaker count was kept constant (Fig. 1). The distributed area of both male and female speaker embedding in the 2-dimensional indexical space was measured, as shown in Fig. 2. The male speakers are relatively sparsely distributed (the mean area is 136.2) than female speakers (the mean area is 49.48).

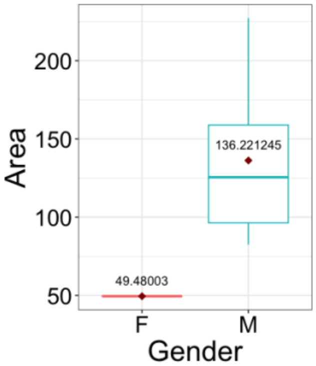

Fig. 2 Area distribution of i-vector speaker embeddings of both male and female speakers in t-SNE indexical space (2 dimension).

We will discuss the results in the context of two plausible explanations: (a) It is possible that voice recognition technology has been developed predominantly based on male voices. Thus, recognition works better for male compared to female voices. Such a gender bias in research knowledge is well known in many scientific disciplines [5], and traditionally research on speech has been predominantly carried out on male voices. (b) It is possible that the acoustics of male voices offer a wider variety of indexical cues to identity and can thus be easier recognized. This would be interesting, particularly from an evolutionary perspective in which voices of different genders could have been attributed different roles in terms of their recognizability [6][7].

References

[1] S. Cumani, O. Glembek, N. Brummer, E. De Villiers, and P. Laface, “Gender independent discriminative speaker recognition in i-vector space,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 1, pp. 4361–4364, 2012, doi: 10.1109/ICASSP.2012.6288885.

[2] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, “Mixture of PLDA models in I-vector space for gender-independent speaker recognition,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 25–28, 2011.

[3] T. Johns, “TIME DELAY DEEP NEURAL NETWORK-BASED UNIVERSAL BACKGROUND MODELS FOR SPEAKER RECOGNITION David Snyder , Daniel Garcia-Romero , Daniel Povey Center for Language and Speech Processing & Human Language Technology Center of Excellence,” no. 1232825, pp. 92–97, 2015.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxceleB2: Deep speaker recognition,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018, doi: 10.21437/Interspeech.2018-1929.

[5] S. Adani and M. Cepanec, “Sex differences in early communication development: Behavioral and neurobiological indicators of more vulnerable communication system development in boys,” *Croat. Med. J.*, vol. 60, no. 2, pp. 141–149, 2019, doi: 10.3325/cmj.2019.60.141.

[6] R. Joseph, “The evolution of sex differences in language, sexuality, and visual- spatial skills,” *Arch. Sex. Behav.*, vol. 29, no. 1, pp. 35–66, 2000, doi: 10.1023/A:1001834404611.

[7] S. E. Yoho, S. A. Borrie, T. S. Barrett, and D. B. Whittaker, “Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology,” *Attention, Perception, Psychophys.*, vol. 81, no. 2, pp. 558–570, 2019, doi: 10.3758/s13414-018-1635-3.