From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database

Finnian Kelly[1], Anil Alexander[1], Oscar Forth[1], and David van der Vloed[2]
[1]*Oxford Wave Research Ltd., Oxford, UK*
[2]*Speech and Audio Research, Netherlands Forensic Institute, The Hague, Netherlands*
{finnian|anil|oscar}@oxfordwaveresearch.com,
d.van.der.vloed@nfi.minvenj.nl

The i-vector framework (Dehak et al., 2011) has served as the state-of-the-art in automatic speaker recognition for several years. Recently, a new approach to speaker recognition based on Deep Neural Networks (DNNs), namely the x-vector framework (Snyder et al., 2018), has been introduced. This new framework has been shown to provide significant speaker recognition performance improvements relative to the i-vector approach on a variety of databases. At the conceptual level, i-vector and x-vector approaches are similar – they both convert a speech recording into a fixed-length vector representing the speaker. While the i-vector approach achieves this through the factorisation of a Gaussian Mixture Model based Total Variability space (Dehak et al., 2011), the x-vector approach uses a feed-forward DNN (Snyder et al., 2018). Once extracted, x-vectors can be compared in the same way as i-vectors. In this paper, we introduce a VOCALISE (Alexander et al., 2016, Kelly et al., 2019) x-vector system and present some speaker recognition experiments on several challenging subsets of the forensically-relevant NFI-FRIDA database (van der Vloed et al., 2018). FRIDA (Forensically Realistic Inter-Device Audio) consists of recordings of 250 male, lower-educated native Dutch speakers that were acquired simultaneously by multiple recording devices. In this paper, a selection of challenging within- and cross-device FRIDA comparisons were evaluated with VOCALISE i-vector and x-vector systems.

Experiments and Results

A subset of 90 FRIDA speakers was used for testing. For each speaker, simultaneously acquired recordings from three devices were selected from two different recording sessions. The devices were: Shure WH20 HQ headset microphone (d1), Shure SM58 far microphone (d4), and intercepted telephone (d5). A separate subset of 45 speakers was used as a reference set for score normalisation. All possible combinations of within- and cross-device test speaker comparisons were evaluated using VOCALISE 2018A speaker recognition software in both x-vector and i-vector modes. For each device comparison combination, the number of same-speaker comparisons was 90, and the number of different-speaker comparisons was 8010. The x-vector session was trained with over 100,000 speech recordings, including 'augmentations', which are artificially noised versions of speech recordings (Snyder et al., 2018), and the i-vector session was trained with a similarly diverse set of approximately 47,000 speech recordings, without augmentation.

The Equal Error Rates (EERs) for each system are presented in Figure 1; for the least-challenging d1-d1 comparison, the EERs for all systems are low (<2%). For all other, more challenging comparisons, it is clear that the x-vector system consistently outperforms the i-vector system. For the d4-d4, d1-d5, and d4-d5 comparisons, there is an approximate halving of the EER between the i-vector and the x-vector-based systems.

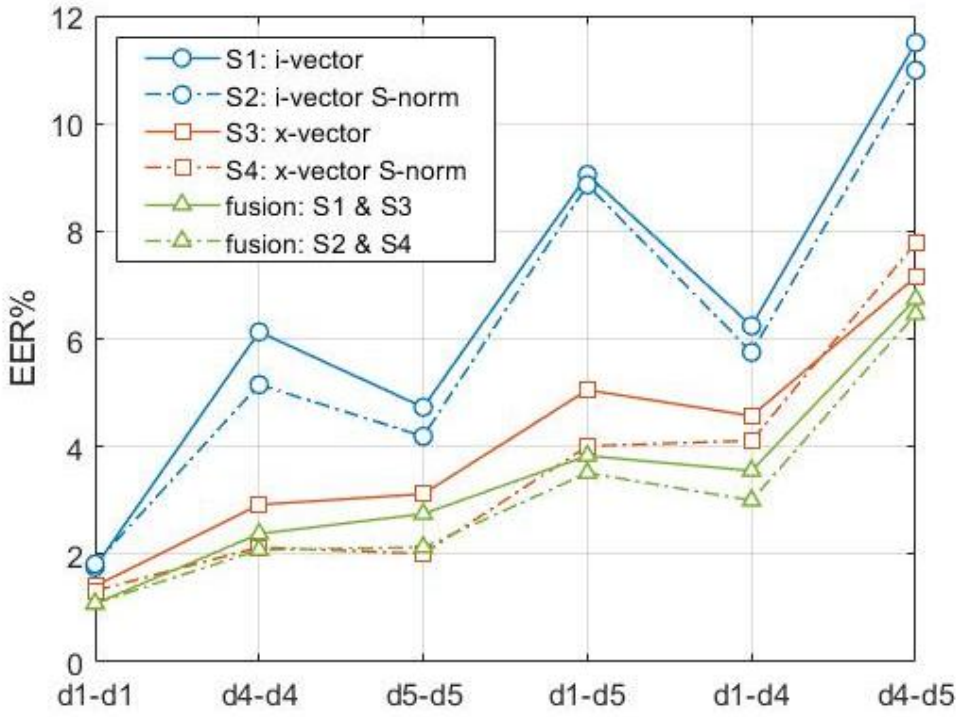

Figure 1: EERs for within- and cross-device comparisons using VOCALISE i-vector and x-vector systems, with and without S-norm, as well as cross-validation score fusion of i-vector and x-vector system. The devices are: **d1:** headset microphone, **d4:** far microphone, **d5:** intercepted telephone. Therefore 'd1-d5' denotes the comparison of headset microphone recordings to intercepted telephone recordings.

For both systems, symmetric normalisation (S-norm) generally reduces the EER. Finally, we note that further improvement is achieved via the fusion of i-vector and x-vector scores, particularly for mismatched (d1-d4, d1-d5, d4-d5) comparisons, highlighting the complementarity between the two systems. The consistent performance improvement achieved with x-vectors in forensically-relevant conditions further supports their position as a new state-of-the-art in speaker recognition.

References

Alexander, A., Forth, O., Atreya, A. A., and Kelly, F. (2016*). VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features*, Odyssey 2016.

Dehak, N., Kenny, P.J., Dehak, E., Dumouchel, P., Ouellet, P. (2011). *Front-end factor analysis for speaker verification*. IEEE Transactions on Acoustics, Speech, and Signal Processing. 19(14): 788–798.

Kelly, F., Forth, O., Kent, S., Gerlach, L., Alexander, A. (2019). *Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors*. Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. (2018), *X-Vectors: Robust DNN Embeddings for Speaker Recognition*, ICASSP 2018.

van der Vloed, D., Bouten, J., Kelly, F., and Alexander A. (2018). *NFI-FRIDA – Forensically Realistic Inter-Device Audio*, IAFPA 2018.