Speaker recognition system adaptation to unseen and mismatched recording devices in 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

Dealing with previously unseen or mismatched recording conditions between suspected speaker and questioned recordings poses a significant challenge in forensic automatic speaker recognition. FRIDA (van der Vloed et al., 2018) is a new database of forensically-realistic speech recordings that were acquired simultaneously by multiple recording devices. In this paper, we explore the effect of previously unseen and mismatched conditions, and propose a novel condition adaptation approach using only relatively small amounts of data from FRIDA in the unseen or mismatched conditions. We compare this approach with the more traditional reference (or score) normalisation for system adaptation.

Speaker recognition experiments: baseline

A subset of 40 speakers from FRIDA was used for testing. For each speaker, 3 (indoor, clean) close microphone recordings were compared against 1 (indoor, clean) recording from each of 5 different devices using the i-vector-based VOCALISE 2017B (Alexander et al., 2016). In VOCALISE, trained models and their associated parameters are stored as 'sessions'. Pre-trained and optimised models are provided with VOCALISE in the form of 'built-in' sessions. The user can also create custom sessions, either by adapting an existing session with supplementary data, or 'from-scratch', using exclusively their own data. Two built-in models were used for the experiments here – one trained on telephony only, and another 'general' session, trained on recordings from a wide variety of devices and channels (including telephony). Prior to comparison, transcriptions of the recordings, in the form of Praat TextGrids (Boersma & Weenink, 2018), were used to remove non-speech segments, resulting in 59 to 278 seconds of net speech. The resulting performance metrics for each experiment, obtained using Bio-Metrics software, are shown in Figure 1 ('baseline'). It can be observed that the general session outperforms the telephony-only session across all conditions. For both sessions, error rates increase in the presence of device mismatch.

Speaker recognition experiments: adaptation

A small subset of 15 speakers from FRIDA was used for system adaptation. For each speaker, 2 (indoor, clean) recordings from each of the 5 devices were extracted and the non-speech regions were removed in the same way as the test data. There was no overlap in the test and adaptation speaker groups. We consider two approaches for using this additional set for system adaptation: reference normalisation and condition adaptation.

A novel model-based approach to system adaptation supported in VOCALISE is condition adaptation. In this procedure, the recordings in the system adaptation set are used to update the internal LDA (Linear Discriminant Analysis) and PLDA (Probabilistic LDA) models in the VOCALISE session. This is achieved by extracting i-vectors for the new data, and using a weighted interpolation of the new i-vector and existing i-vector statistics to update the LDA transformation matrix. All (new and existing) i-vectors are transformed with the updated LDA model and are then used to re-estimate the PLDA model. The baseline experiments were repeated, using the adaptation recordings from the suspect recording device *and* questioned recording device to apply condition adaptation.

VOCALISE also supports ‘reference normalisation’ in the form of symmetric score normalisation (Shum et al. 2010). In this procedure, the suspect and questioned recordings in a test comparison are compared with all recordings in the system adaptation set, and the statistics of these scores are used to normalise the test comparison score The baseline experiments were repeated, using the adaptation recordings from the questioned recording device to apply reference normalisation.

Referring to Figure 1, for all mismatched comparisons, there is an improvement in performance after either (or both) reference normalisation and condition adaptation. With the telephony-only session (Fig. 1, left), the cross-device improvements with adaptation are more dramatic compared to the general session (Fig. 1, right). This is expected, given the much wider diversity of recording devices represented in the general session training data. Condition adaptation outperforms reference normalisation in cross-channel mic-tel comparisons (d1-d5), while performance is varied across the other mismatched within-channel mic-mic comparisons. Given the small number of speakers (15) used for condition adaptation, the improvements with the realistic FRIDA data are promising.

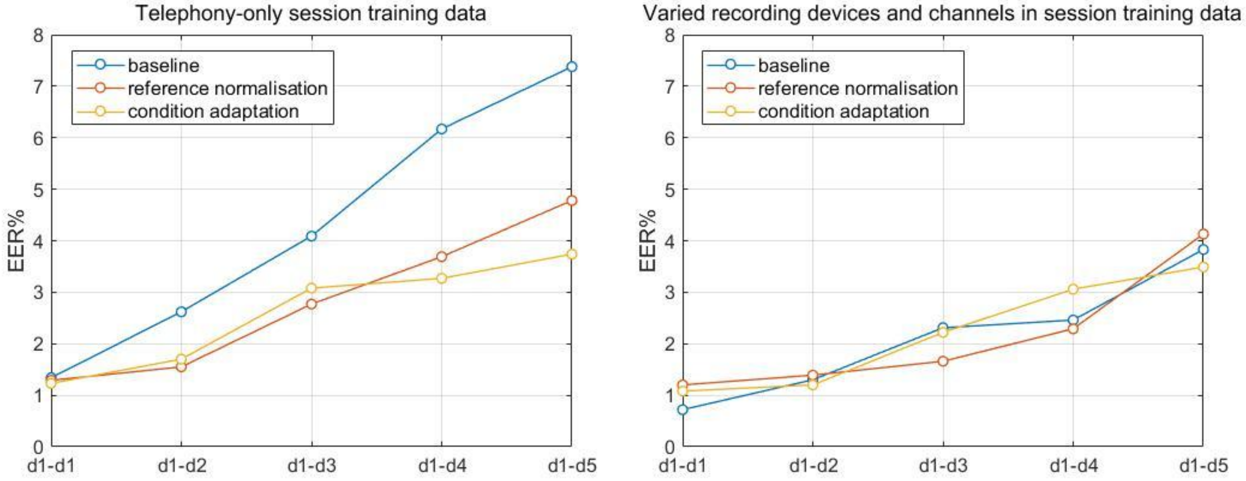

Figure 1: Convex Hull Equal Error Rates (EERs) for cross-device comparisons using VOCALISE sessions trained on recordings from telephone only **(left)**, and varied devices **(right)**. Each ‘d’ denotes a recording device, where, **d1:** Shure WH20 HQ headset, **d2:** Shure SM58 close mic, **d3:** AKG C400BL close mic, **d4:** Shure SM58 far mic, **d5:** Intercepted telephone. Thus d1-d1 is a matched comparison, and the remainder are mismatched. In each case, the number of same-speaker comparisons was 120, and the number of different-speaker comparisons was 4680. Reference normalisation used 30 recordings from 15 speakers on 1 device. Condition adaptation used 60 recordings from 15 speakers on 2 devices, with the exception of d1-d1, which used 30 recordings from 15 speakers on 1 device (d1).

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.*

Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems, *Digital Signal Processing*, vol. 10, no. 1–3, pp. 42–54, Jan. 2000.

Boersma, P., Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

Shum, S., Dehak, N., Dehak, R., and Glass, J. R. (2010). Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification, *Odyssey 2010*.

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