NFI-FRIDA – Forensically Realistic Inter-Device Audio database and initial experiments

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

an oral presentation (20 mins)

FRIDA is a new database of forensically-relevant speech recordings that were acquired simultaneously by multiple recording devices. Many forensically-relevant databases have been collected in recent years, including those by Ramos et al, (2008), Nolan et al (2009), Becker, (2012) and more. In Dutch, only NFI-FRITS (van der Vloed et al, 2014), which had a precursor in the NFI-TNO database (van Leeuwen, Bouten, 2004), has been collected up to now. For conditions other than telephone therefore, there was no forensically-relevant data available for Dutch. With FRIDA, we hope to fill that omission.

Outline of the database

250 speakers were recorded in 16 sessions of approximately 5 minutes each. The sessions were recorded over 2 days at least a week apart, with 8 sessions per day. All sessions consist of spontaneous telephone conversations with another participant, recorded across multiple recording devices simultaneously, as detailed in Table 1. Per-speaker, this amounts to an approximate gross speech duration of 80 minutes and an approximate gross recording duration of 6 hours. All speakers are lower educated, male L1 speakers of Dutch from Amsterdam. The speakers' backgrounds are native Dutch (50%), Turkish immigrant (25%) and Moroccan immigrant (25%). 80% of the speakers are between 18 and 35 years of age, 20% are older.

Table 1. Sessions per-day per-participant

Session	*Recording location*	*Background*	*Telephone used*	*Recording devices*
1	Indoor	Clean	Nokia 1280	1,2,3,4,5,6
2	Indoor	Clean	iPhone 4	1,2,3,4,5,6
3	Indoor	Noise	Nokia 1280	1,2,3,4,5,6
4	Indoor	Noise	iPhone 4	1,2,3,4,5,6
5	Outdoor	Calm	Nokia 1280	1,5,6
6	Outdoor	Calm	iPhone 4	1,5,6
7	Outdoor	Busy street	Nokia 1280	1,5,6
8	Outdoor	Busy street	iPhone 4	1,5,6

The recording devices are: (1) Shure WH20 HQ headset, (2) Shure SM58 Microphone close to the speaker, (3) AKG C400BL Microphone close to the speaker, (4) Shure SM58 Microphone at a 4 meter distance, (5) telephone intercept, (6) iPhone video at a 1 meter distance. Recording device (1) was chosen to provide a high quality recording of the speech; the other recording devices were chosen to reflect a wide variety of forensic conditions, including police interviews with reverberation or noise, intercepted telephone recordings, and, after appropriate processing, YouTube videos. The telephone recording (5) is a recording of telephone-transmitted speech, the others are direct microphone recordings.

The database is currently being transcribed, with the aim of completing this task by the end of 2018. The transcriptions are orthographic, using a protocol based on the one used in the Spoken Dutch Corpus (Goedertier & Goddijn, 2000) and are stored in Praat TextGrids (Boersma & Weenink, 2018). Transcriptions are based on the recordings from device 1, but are valid for all other (simultaneous) recordings in the same session.

Use of the database

The database provides a resource for validation research of mismatched recording conditions. It also provides reference population or training material for casework using (semi-)automatic methods. Furthermore, the transcriptions open up the database for validation research on specific speech features. They allow for edits of the speech recordings, and make them suitable for broader use, e.g. testing and training automatic speech-to-text systems. The transcriptions contain a lot of street language and language associated with young people and are a great resource for contemporary spoken Dutch.

Initial experiments

Initial speaker recognition experiments in matched conditions were completed using iVOCALISE 2017B (Alexander et al., 2016). In each experiment, 2 (indoor, clean) recordings from the same device were extracted for each of 40 speakers. The speech/non-speech temporal information from the transcriptions was used to remove non-speech segments, resulting in 59 to 253 seconds of net speech. The resulting error metrics, obtained using Bio-Metrics software, are shown in Table 2. Subsequent analysis in another contribution to this conference (XXX et al, 2018) will consider the effect of mismatched conditions, and explore the use of additional FRIDA data to compensate for this mismatch.

Table 2. Convex Hull Equal Error Rates (EER) and minimum log-likelihood ratio costs (Cllr-min) for a subset of matched comparisons on FRIDA. In each experiment, the number of same-speaker comparisons was 40, and the number of different-speaker comparisons was 1560.

Suspect recordings	*Offender recordings*	*EER%*	*Cllr-min*
Shure WH20 HQ headset	Shure WH20 HQ headset	0.86	0.017
Shure SM58 close	Shure SM58 close	0.12	0.002
AKG C400BL close	AKG C400BL close	1.29	0.030
Shure SM58 far	Shure SM58 far	6.47	0.171
Intercepted telephone	Intercepted telephone	3.12	0.077

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. *Speaker Odyssey 2016*, Bilbao, Spain.

Becker, T. (2012). Automatic forensic voice comparison (automatischer forensischer Stimmenvergleich). *The Journal of Speech, Language and the Law,* vol.19, pp 291-294.

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/

Goedertier, W., Goddijn, S. (2000) Orthographic Transcription of the Spoken Dutch Corpus. *In LREC-2000 (Second International Conference on Language Resources and Evaluation) Proceedings*. Vol II: 909-914.

van Leeuwen, D., Bouten, J. (2004) Results of the 2003 NFI-TNO forensic speaker recognition evaluation. *Proc. Odyssey 2004 Speaker and Language Recognition workshop*, June 2004, pp. 75-82 ISCA.

Nolan, F., McDougall, L., de Jong, G. and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 16(1), pp. 31-57.

Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Domingues, J., Lucena-Molina, J. 2008. Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. *Proc. Interspeech*, 2008, pp 1493-1496.

van der Vloed, D., Bouten, J., van Leeuwen, D. (2014). NFI-FRITS: A Forensic Speaker Recognition Database and Some First Experiments, *Proceedings of Odyssey 2014: The speaker and Language Recognition Workshop,* Joensuu Finland, 6-13.

Kelly F., Alexander A., Forth O. and Van der Vloed, D. (2018) Speaker recognition system adaptation to unseen and mismatched recording devices in the NFI-FRIDA database. *Submitted to The 27th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA).*