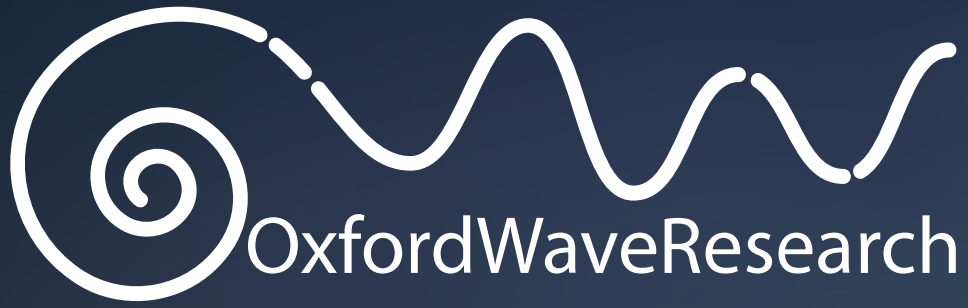


From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database

Finnian Kelly[1], Anil Alexander[1], Oscar Forth[1], and David van der Vloed[2]

[1]*Oxford Wave Research Ltd., Oxford, United Kingdom*

[2]*Speech and Audio Research, Netherlands Forensic Institute, The Hague, Netherlands*

15th July 2019, IAFPA conference, Istanbul

Introduction

- Deep Neural Networks (DNNs) mark a new phase in the evolution of automatic speaker recognition technology, providing a powerful way to extract highly-discriminative speaker-specific features from a recording of speech
- The latest version of VOCALISE supports the DNN-based ‘x-vector’ framework, a state-of-the-art approach that uses a DNN to extract compact speaker representations
- The x-vector version of VOCALISE aims to preserve the ‘open-box’ philosophy of its predecessors, offering the forensic practitioner flexibility in the configuration, training, and adaptation of all parts of the speaker recognition pipeline
- This presentation will introduce the x-vector framework in VOCALISE, and demonstrate its performance capabilities on challenging comparisons within the forensically-relevant NFI-FRIDA database

Timeline of automatic speaker recognition

1990

- Gaussian Mixture Models: **GMM**

Reynolds, D. A., Rose, R. C., *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE trans. speech and audio processing, 3(1), 72-83, 1995

2000

- Adapted Gaussian Mixture Models: **GMM-UBM**

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., *Speaker verification using adapted Gaussian mixture models*, Digital signal processing, 10(1-3), 19-41, 2000

2010

- Factor Analysis: **i-vectors**

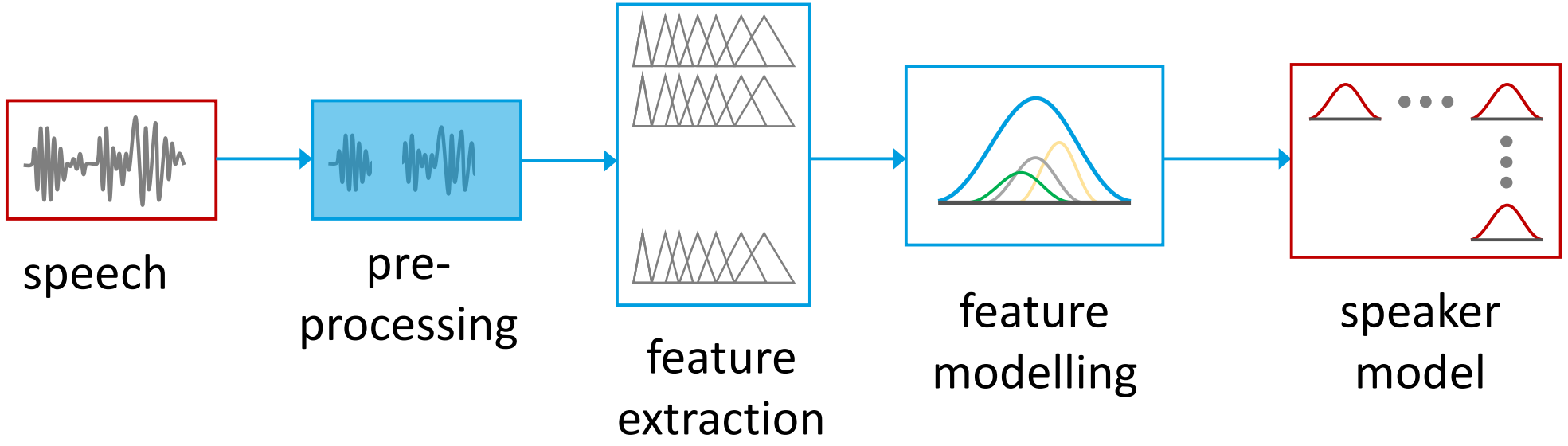
Dehak, N., Kenny, P. J. Kenny, Dehak, R., Dumouchel, P., Ouellet, P., *Front-End Factor Analysis for Speaker Verification*, IEEE trans. audio, speech, and language processing, 19(4), 788-798, 2011

2018

- Deep Neural Networks: **x-vectors**

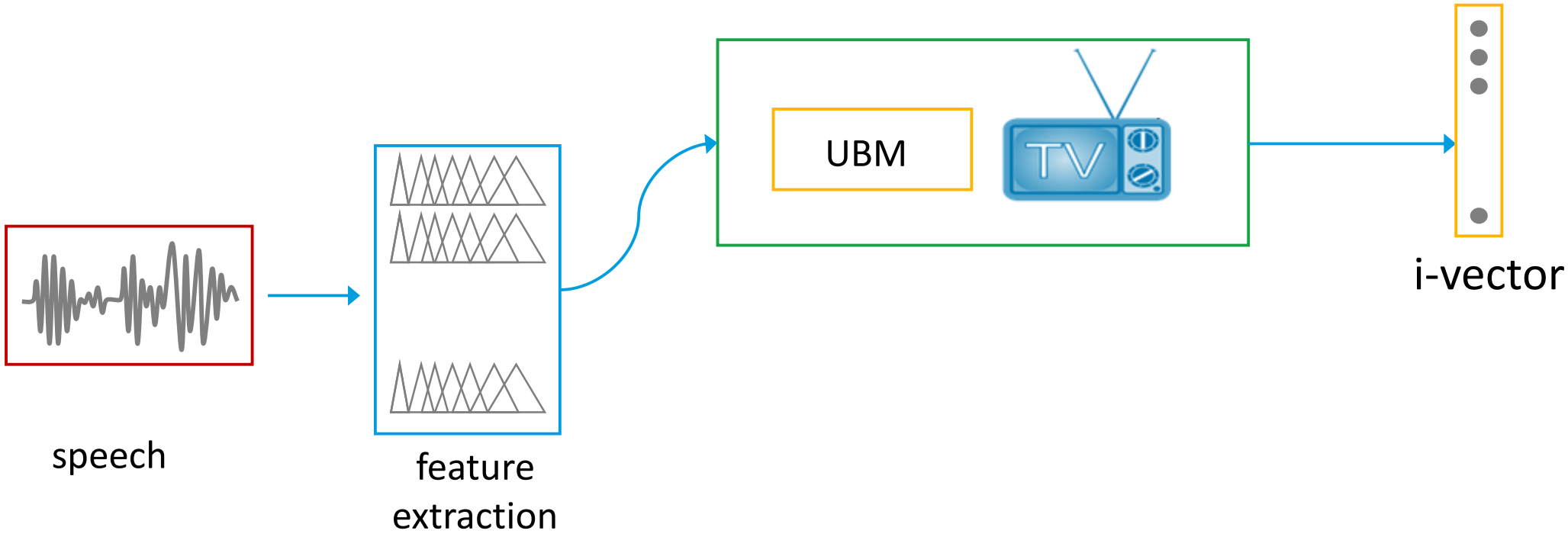
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., *X-vectors: Robust DNN Embeddings for Speaker Recognition*, ICASSP **2018**

An automatic speaker recognition pipeline

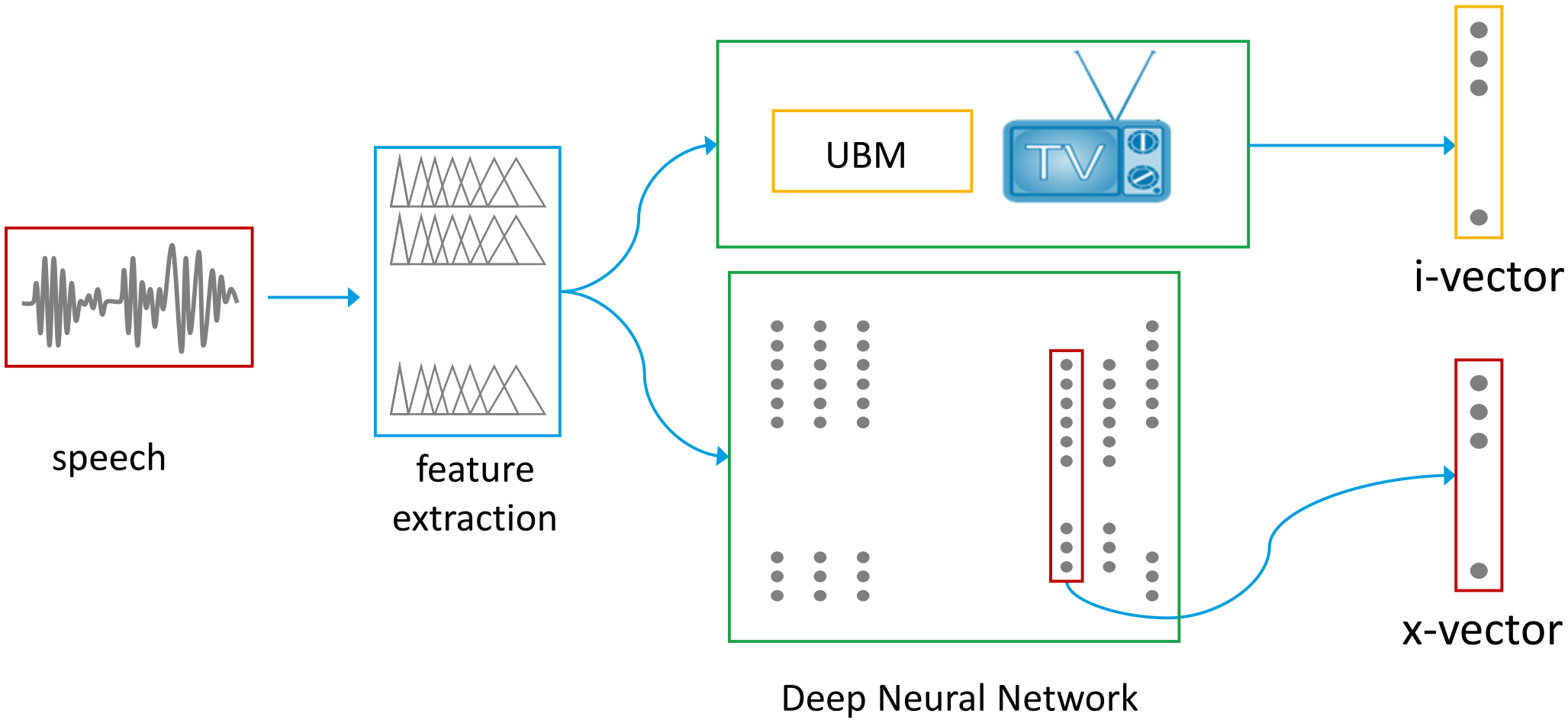


The technology has evolved, but the general pipeline has remained consistent

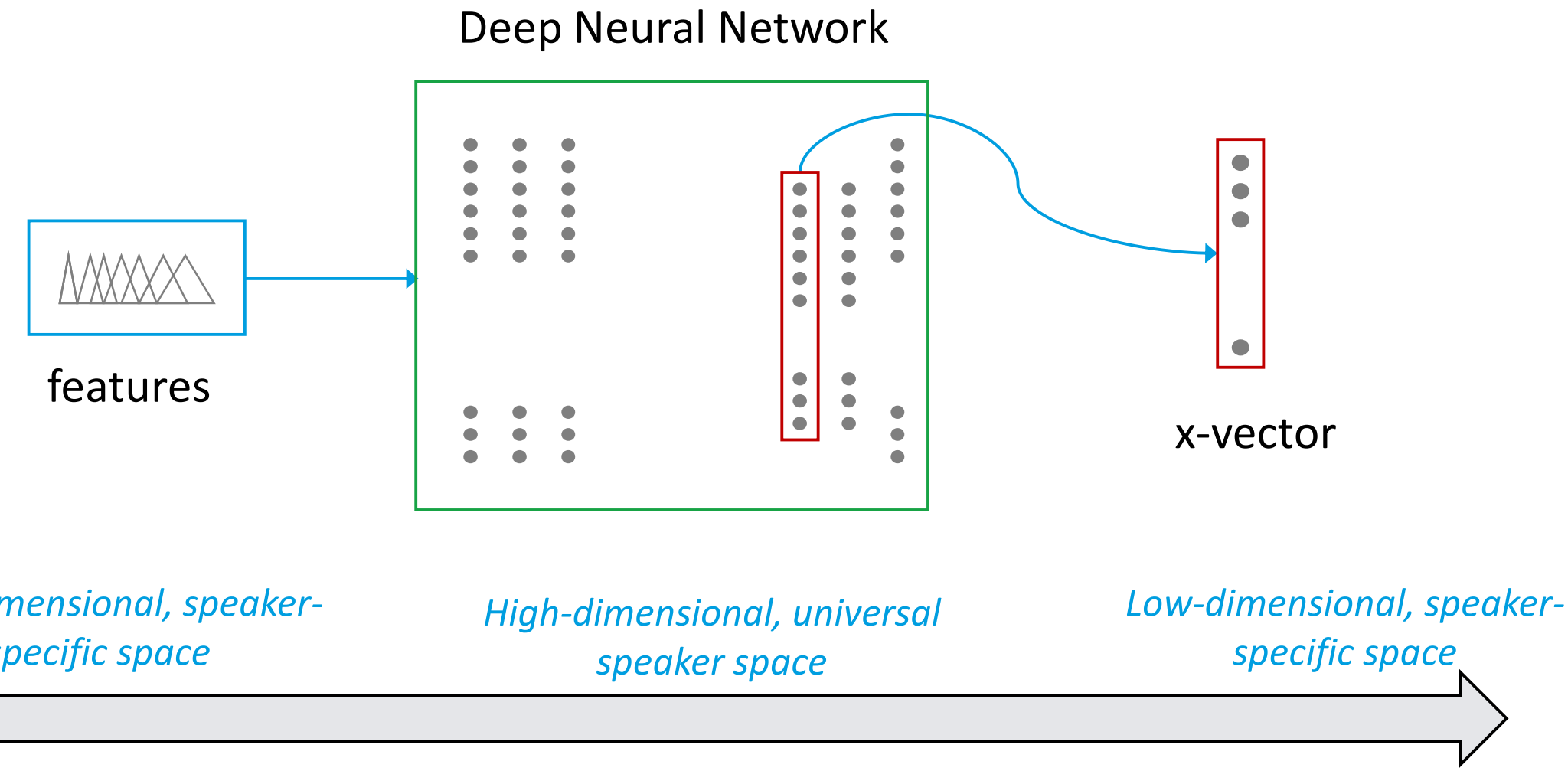
The i-vector and x-vector pipelines



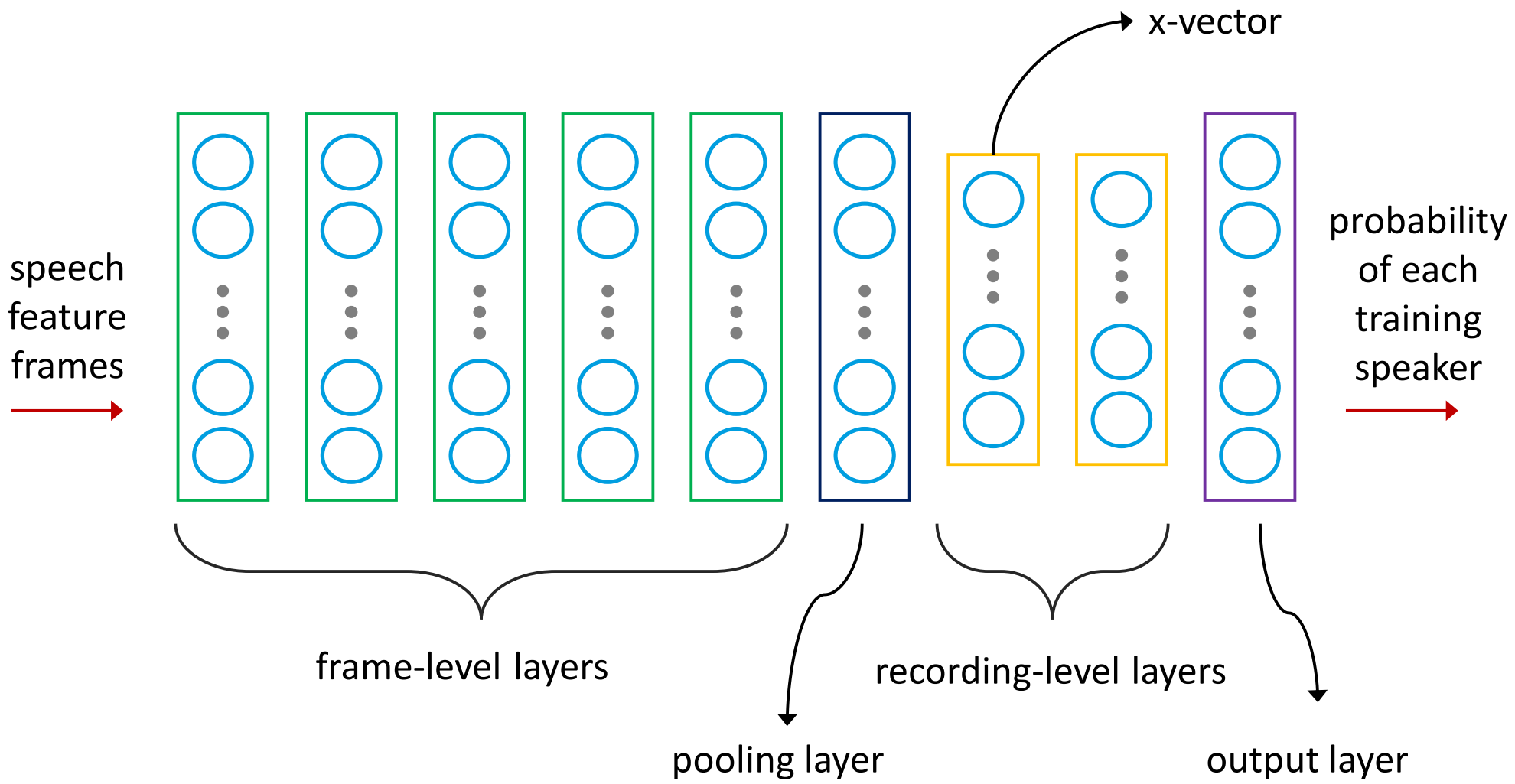
The i-vector and x-vector pipelines



The x-vector pipeline

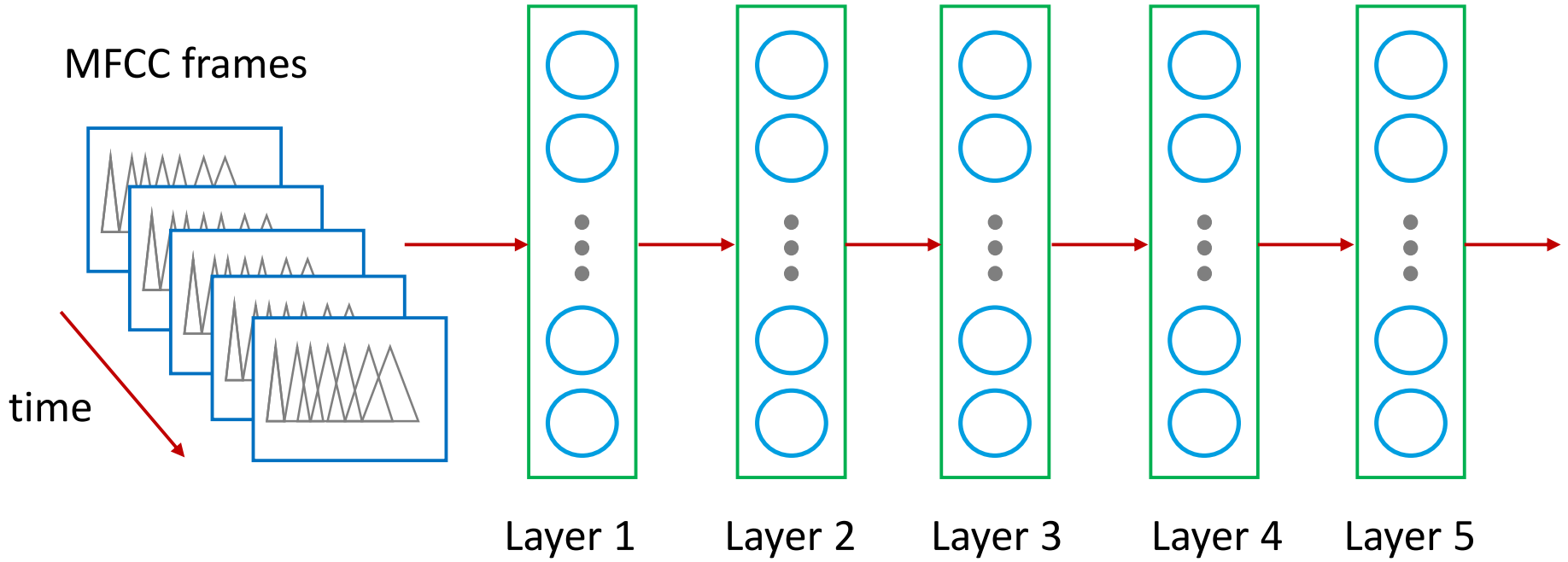


The x-vector DNN



Frame-level layers: capture temporal information

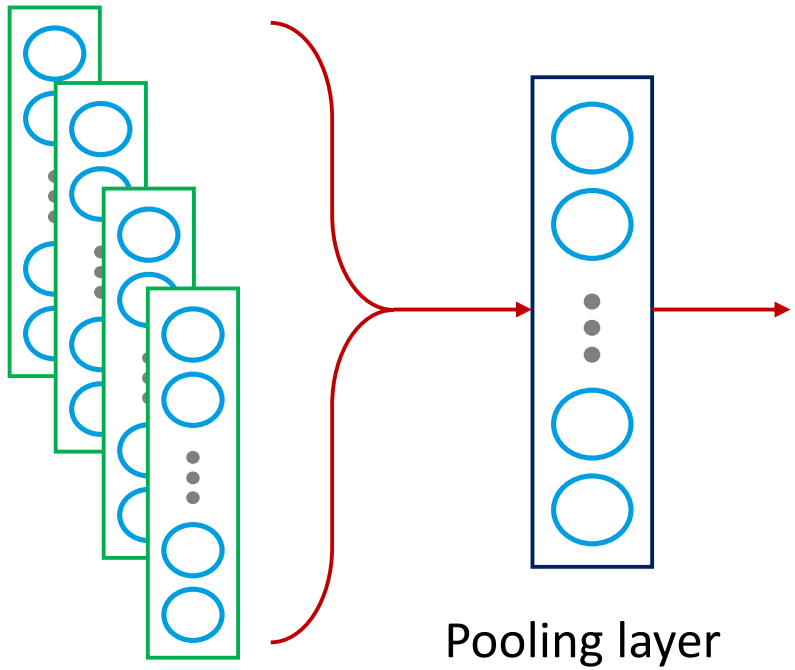
- The input features cascade through the layers, temporal information is captured by increasing the time context of the frames being modelled
- Both static and dynamic characteristics of the MFCCs are captured; therefore no Δ or ΔΔ coefficients are required



Pooling Layer: aggregate information across frames

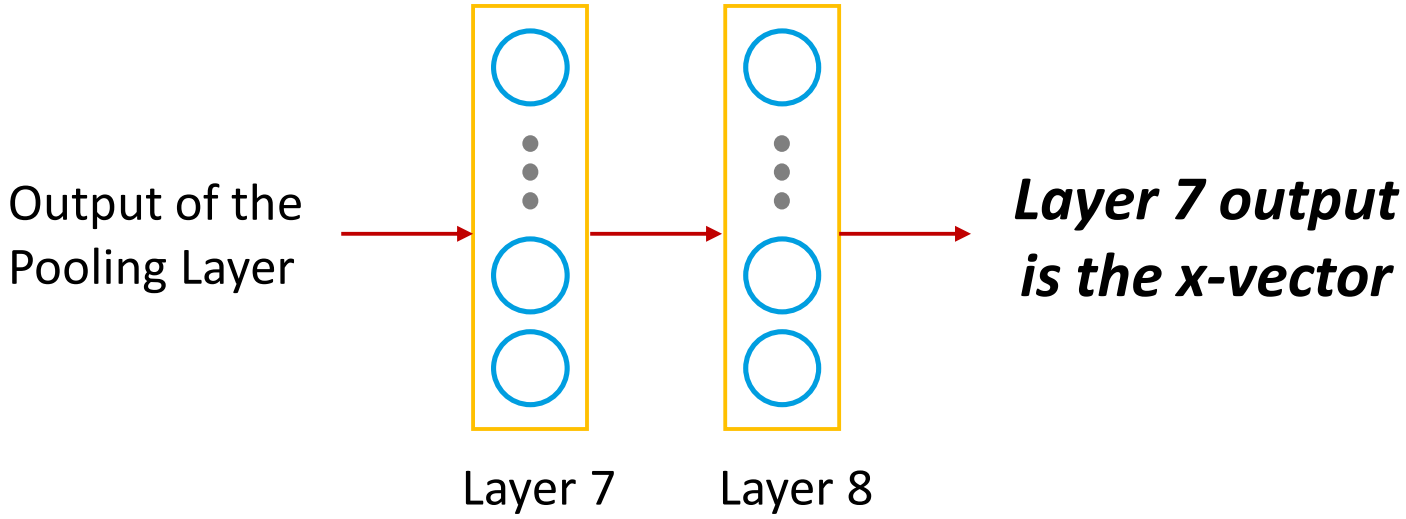
- The Pooling Layer calculates the mean and standard deviation of the Layer 5 outputs across all frames in the recording
- The Pooling Layer therefore converts frame-level information into recording-level information

Layer 5 outputs for all frames in the recording



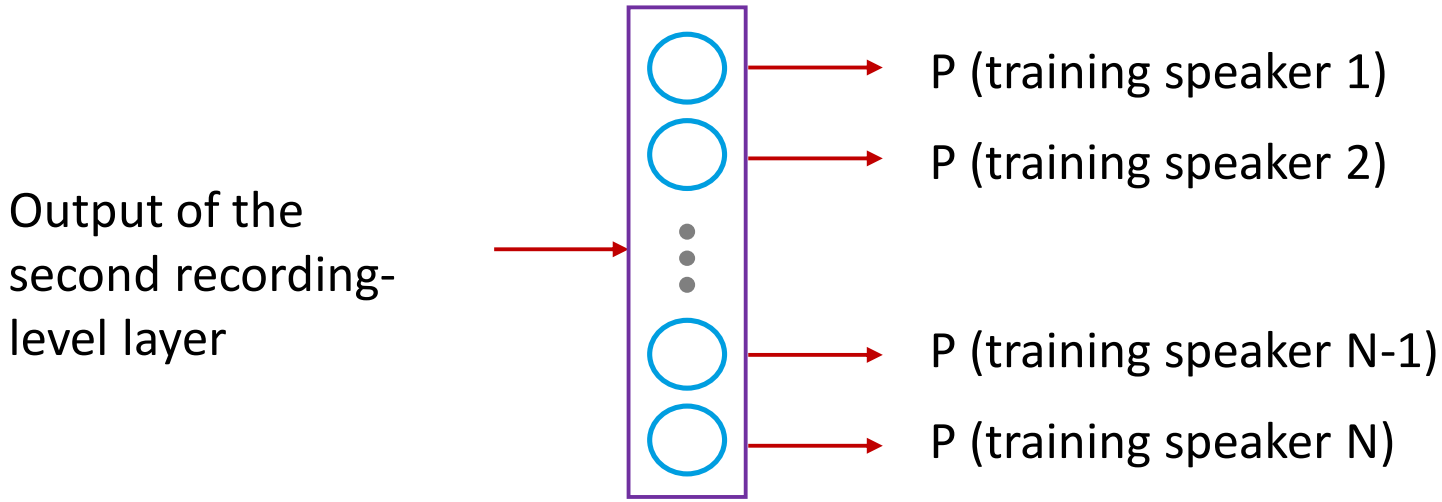
Recording-level layers: the speaker embeddings

- The information in the recording-level layers represent the whole utterance.
- They are smaller in size (fewer nodes) than the previous layers, and therefore provide dimension reduction.
- Both Layers 7 and 8 can be regarded as speaker representations or **speaker embeddings**; layer 7 is typically taken as the x-vector
- The size of the x-vector is defined by the size of these layers, and is typically **512** values.

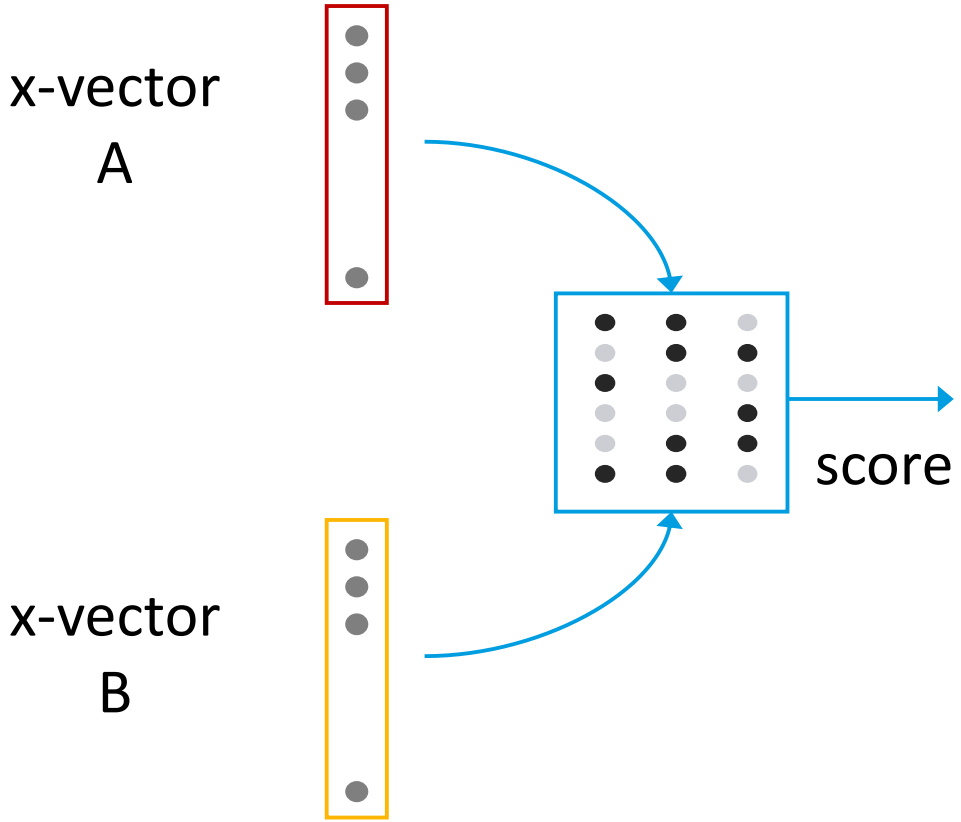


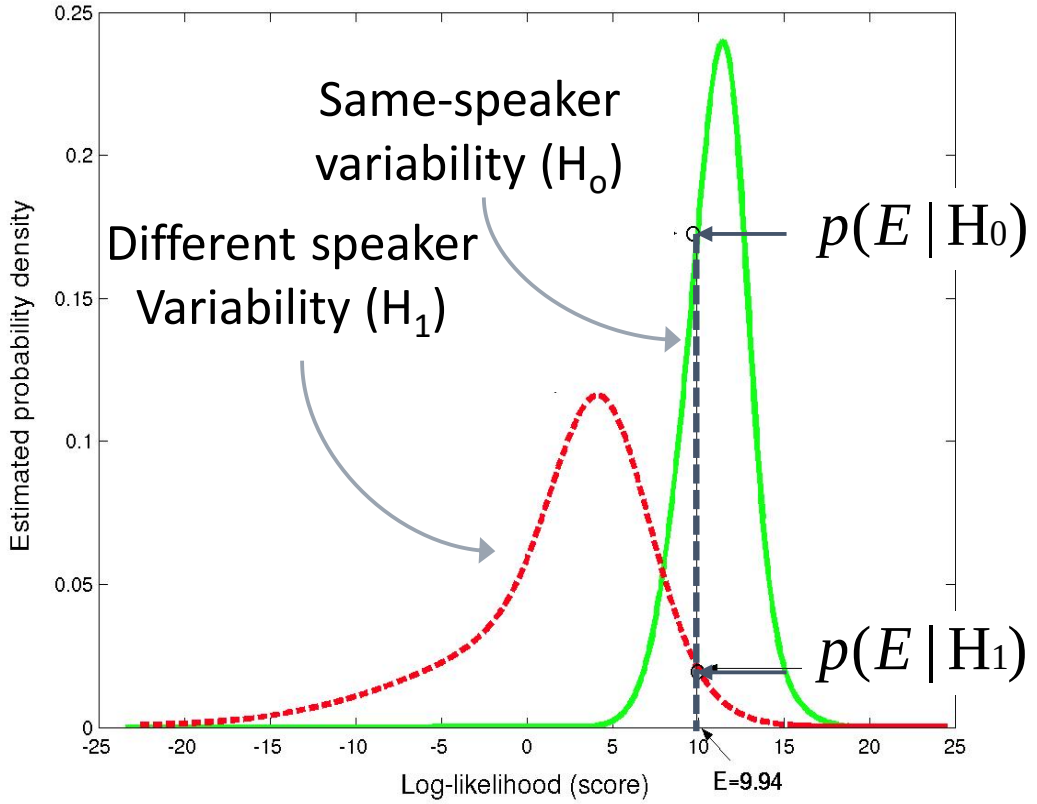
Output Layer: relevant for training only

- The Output Layer takes as input the second recording-level layer and outputs the probability of each of the training speakers given the input recording.
- The Output Layer probabilities are used during training to optimise the weights in each of the layers; the Output Layer probabilities are not relevant during testing, as they concern the training speakers only.



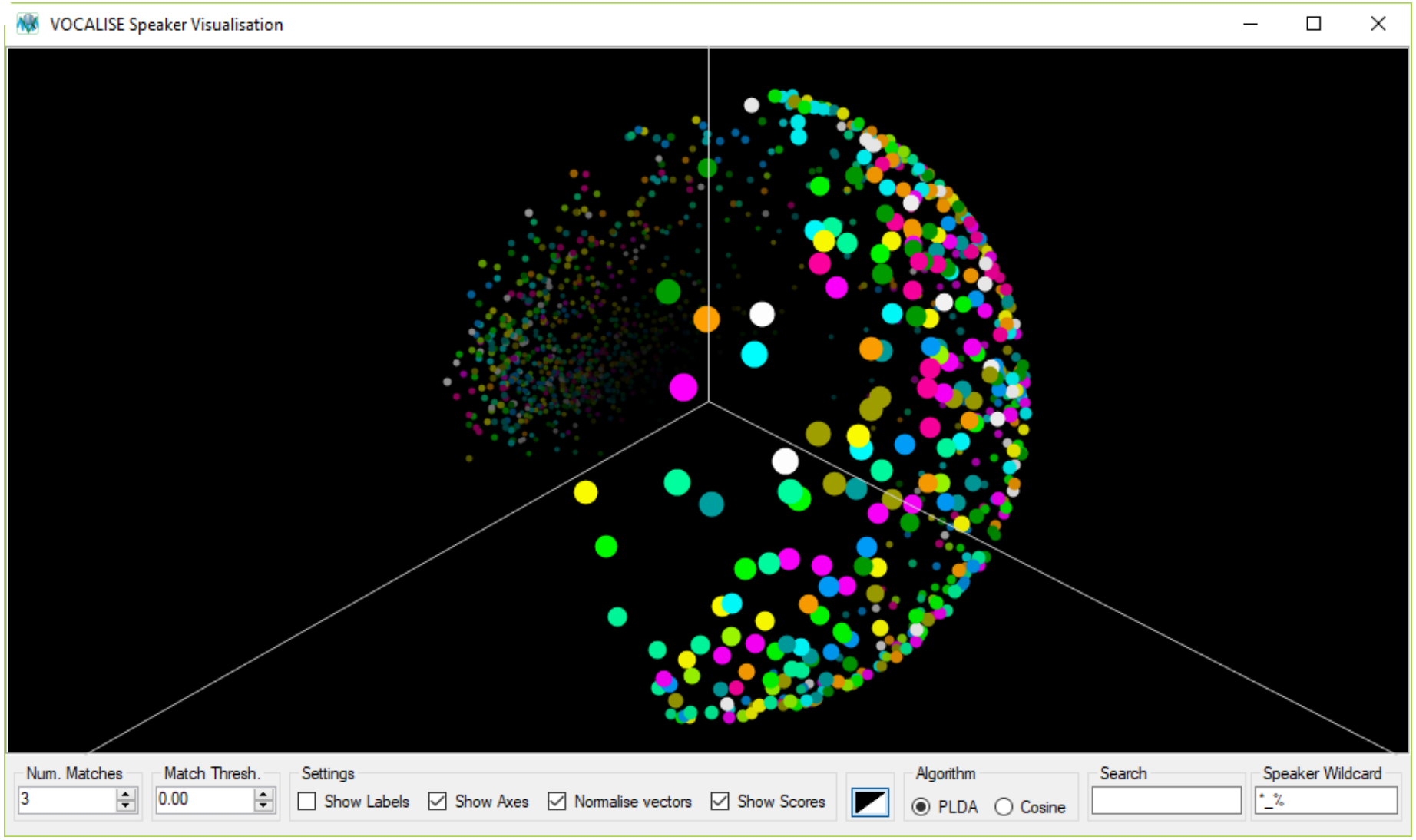
Comparing x-vectors





Using same-speaker and different-speaker score distributions to estimate a likelihood ratio given a comparison score

Visualising x-vectors



The success of x-vectors

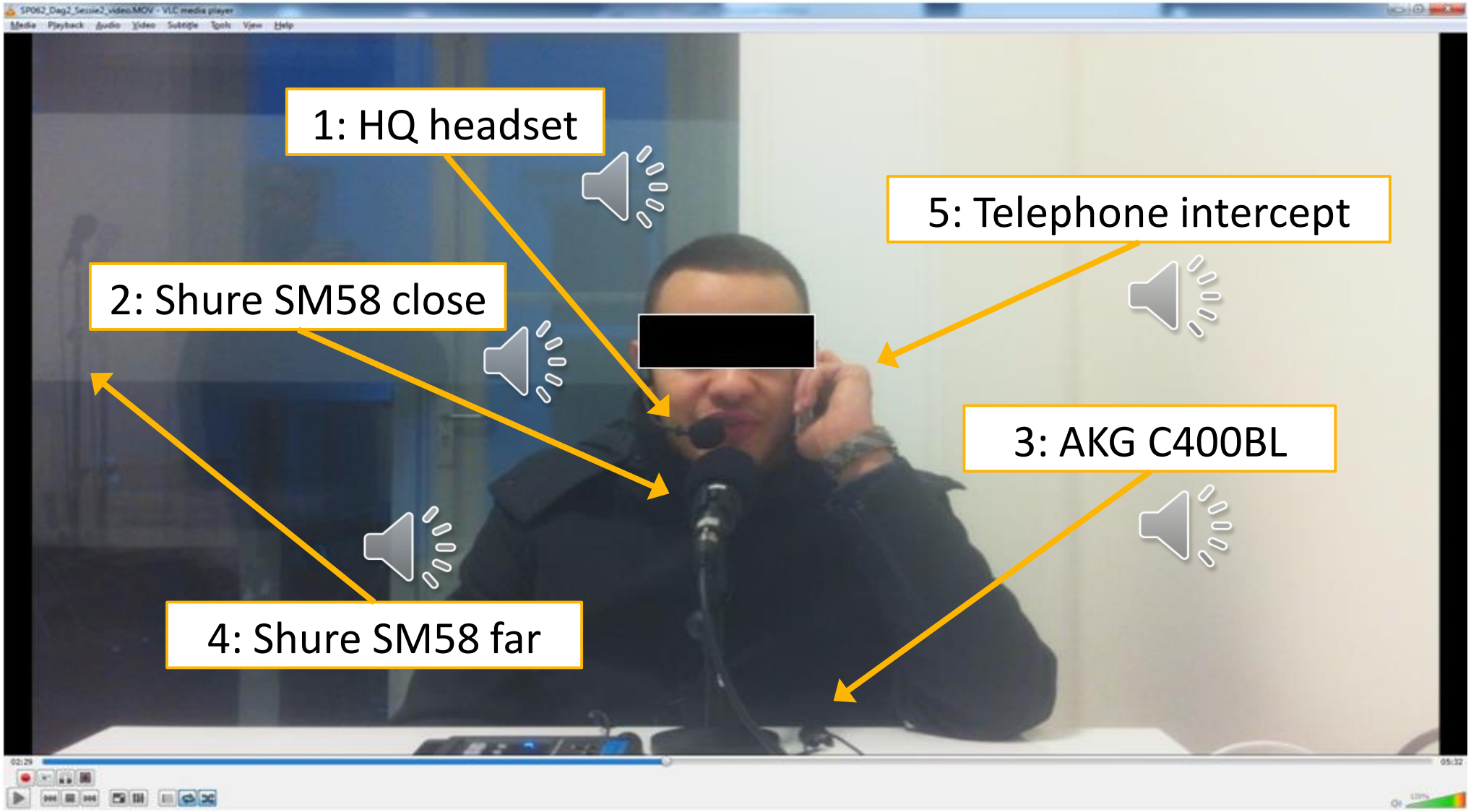
- The performance of x-vectors has been demonstrated to significantly outperform that of i-vectors, particularly at short durations.
- The x-vector DNN is discriminatively trained using speaker labels.
- The x-vector DNN is capable of exploiting **larger amounts of training data** than the i-vector framework, which saturates after a certain quantity of training data.
- This also facilitates a method of boosting the quantity and diversity of training data referred to as **data augmentation.** This process adds noise and reverb to the training samples and includes them in training alongside the original samples.
- The ability to use the **same front-end** (feature extraction) **and back-end** (vector comparison) for both i-vector and x-vector systems facilitates system integration and allows for more direct comparison between the two modelling approaches.

NFI-FRIDA database (van der Vloed et al., 2018)

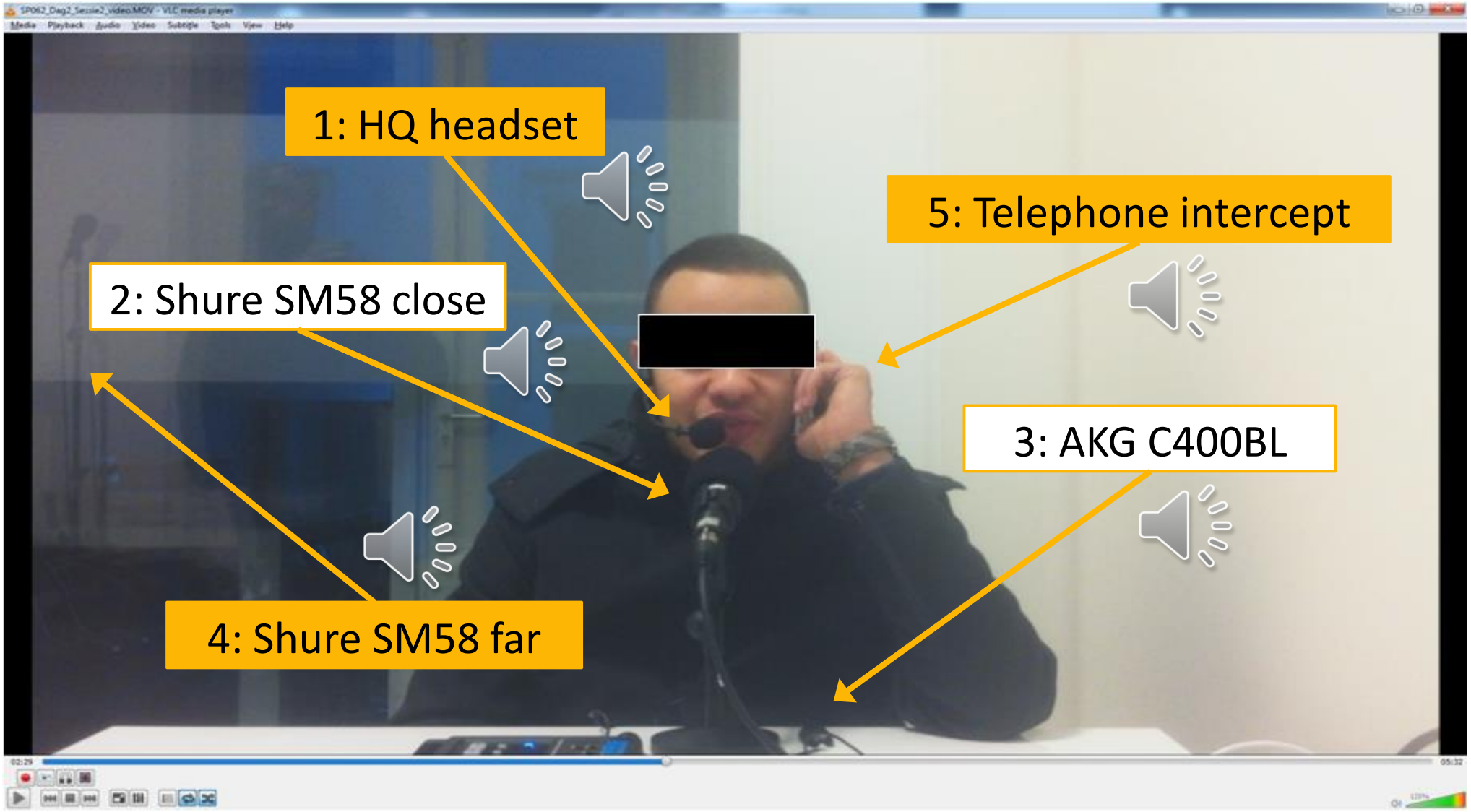
- FRIDA (Forensically Realistic Inter-Device Audio) database:
  - 250 male Dutch speakers
  - Recordings **simultaneously acquired** by multiple recording devices
  - Multiple spontaneous conversations recorded on two separate days
  - Recordings have been manually transcribed
- We evaluate a selection of challenging within- and cross-device comparisons using VOCALISE x-vector and i-vector systems.

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

NFI-FRIDA collection



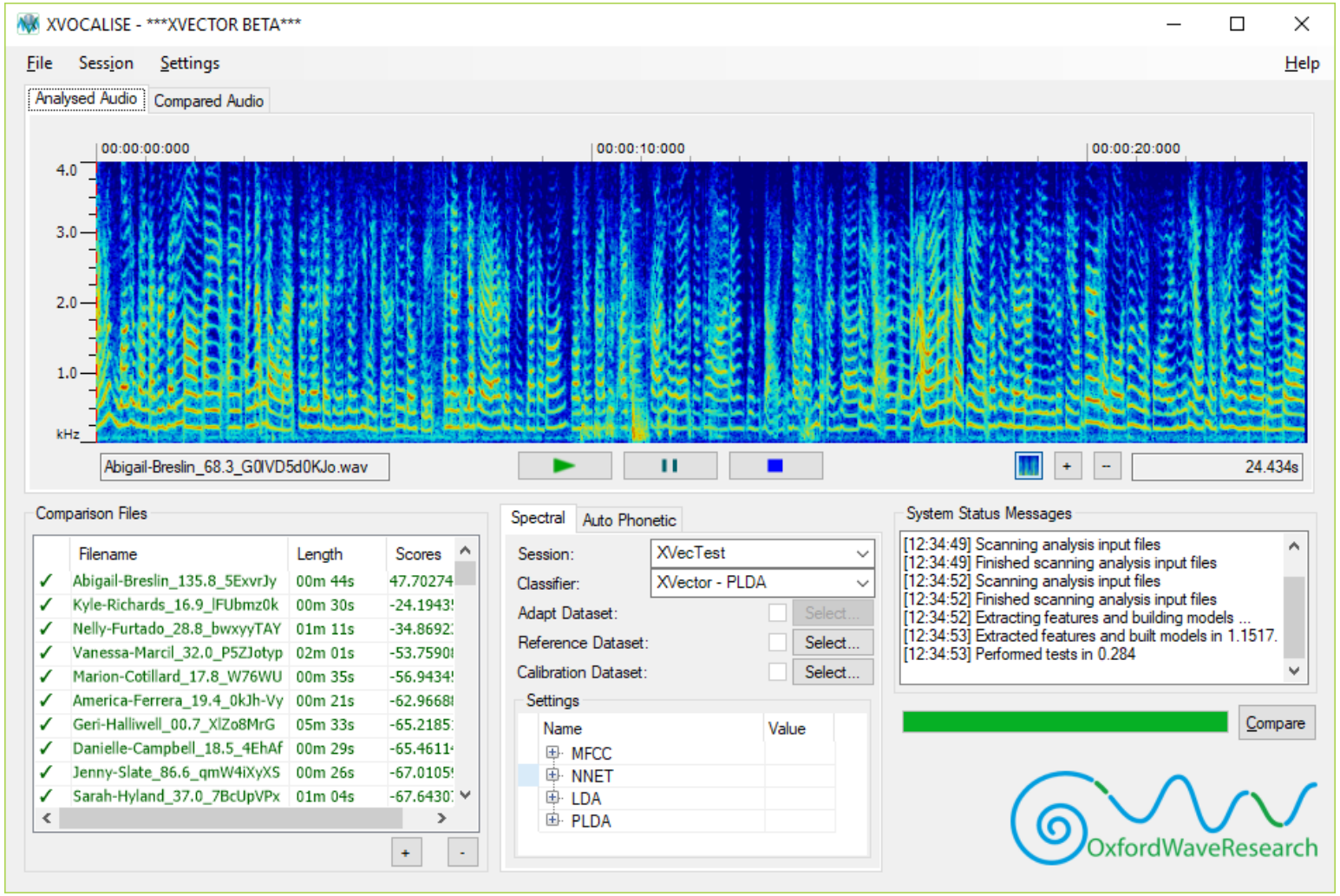
NFI-FRIDA collection



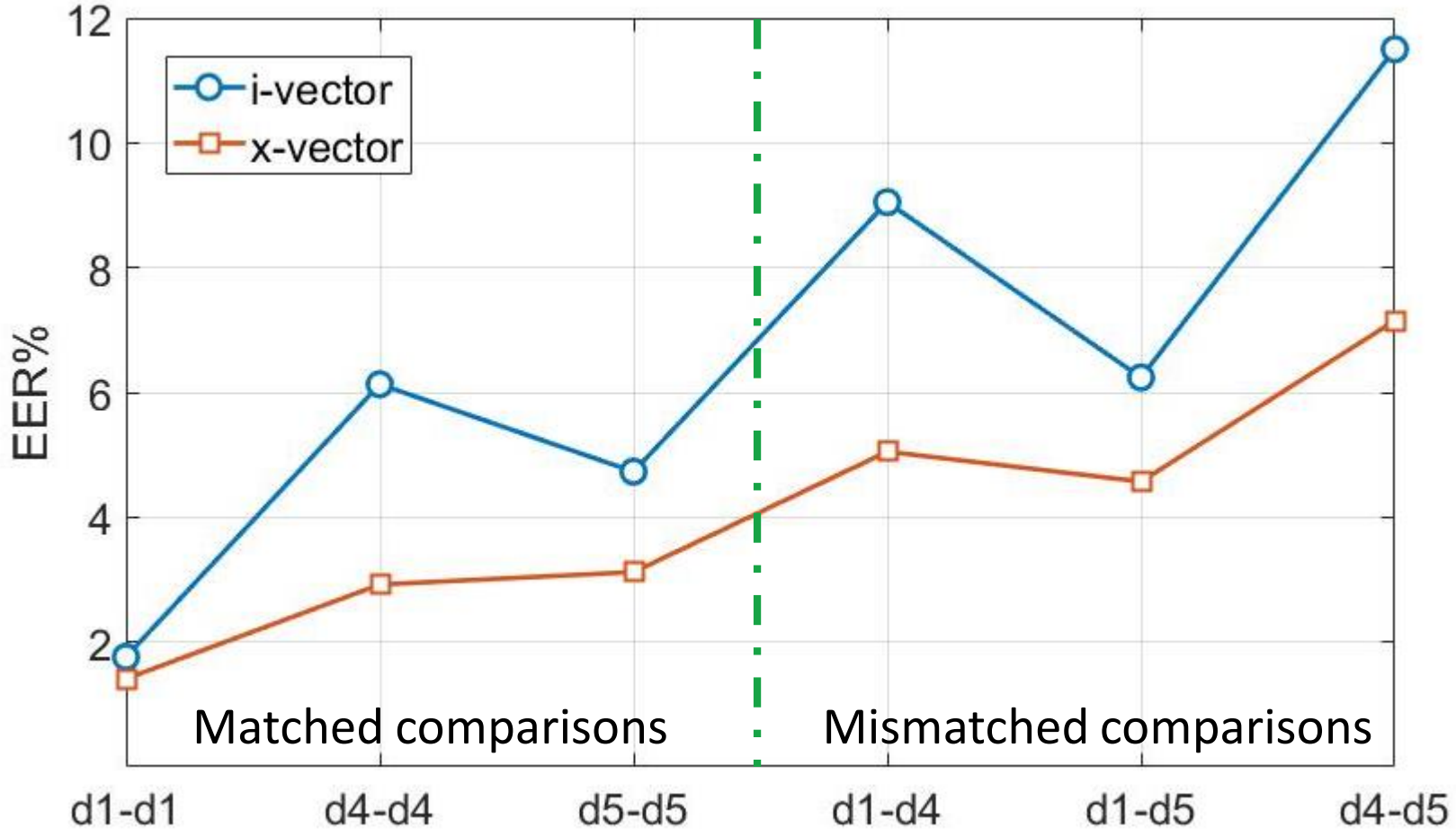
NFI-FRIDA experiments

- A subset of 90 speakers was selected
- For each speaker, two recordings (made on separate days) from each recording device were selected
- The following recording devices were considered:
  - **d1**: Close (headset) microphone
  - **d4**: Far microphone
  - **d5**: Telephone intercept
- The duration of all recordings was 30 seconds (**net speech**)
- All within- and cross-device comparisons between day 1 and day 2 recordings were evaluated using VOCALISE i-vector and x-vector systems
- A separate set of 45 speakers was reserved for score normalisation

VOCALISE – Voice Comparison and Analysis of the Likelihood of Speech Evidence



NFI-FRIDA results: i-vector vs x-vector

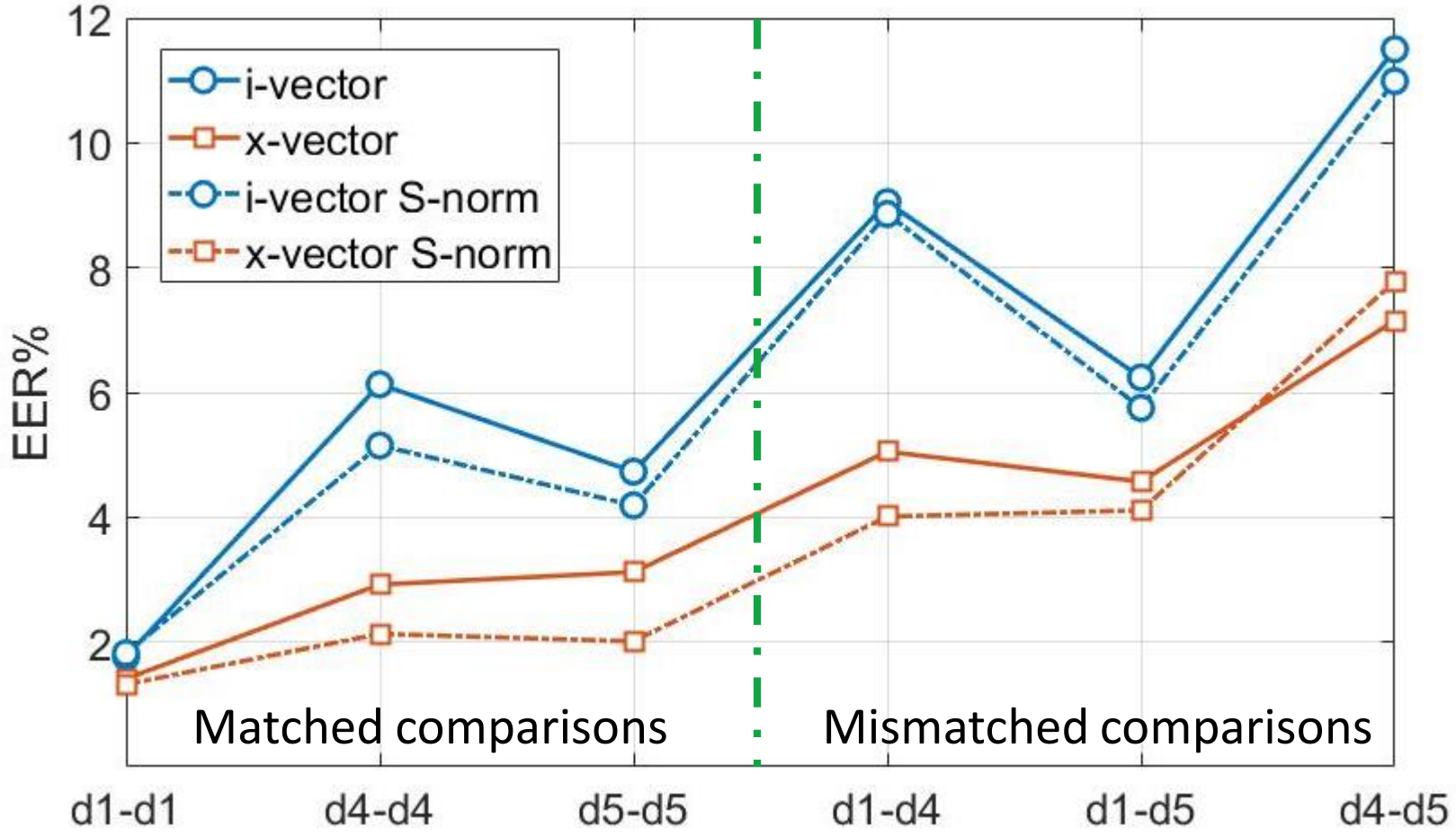


d1: Close mic

d4: Far mic

d5: Tel intercept

NFI-FRIDA results: normalisation

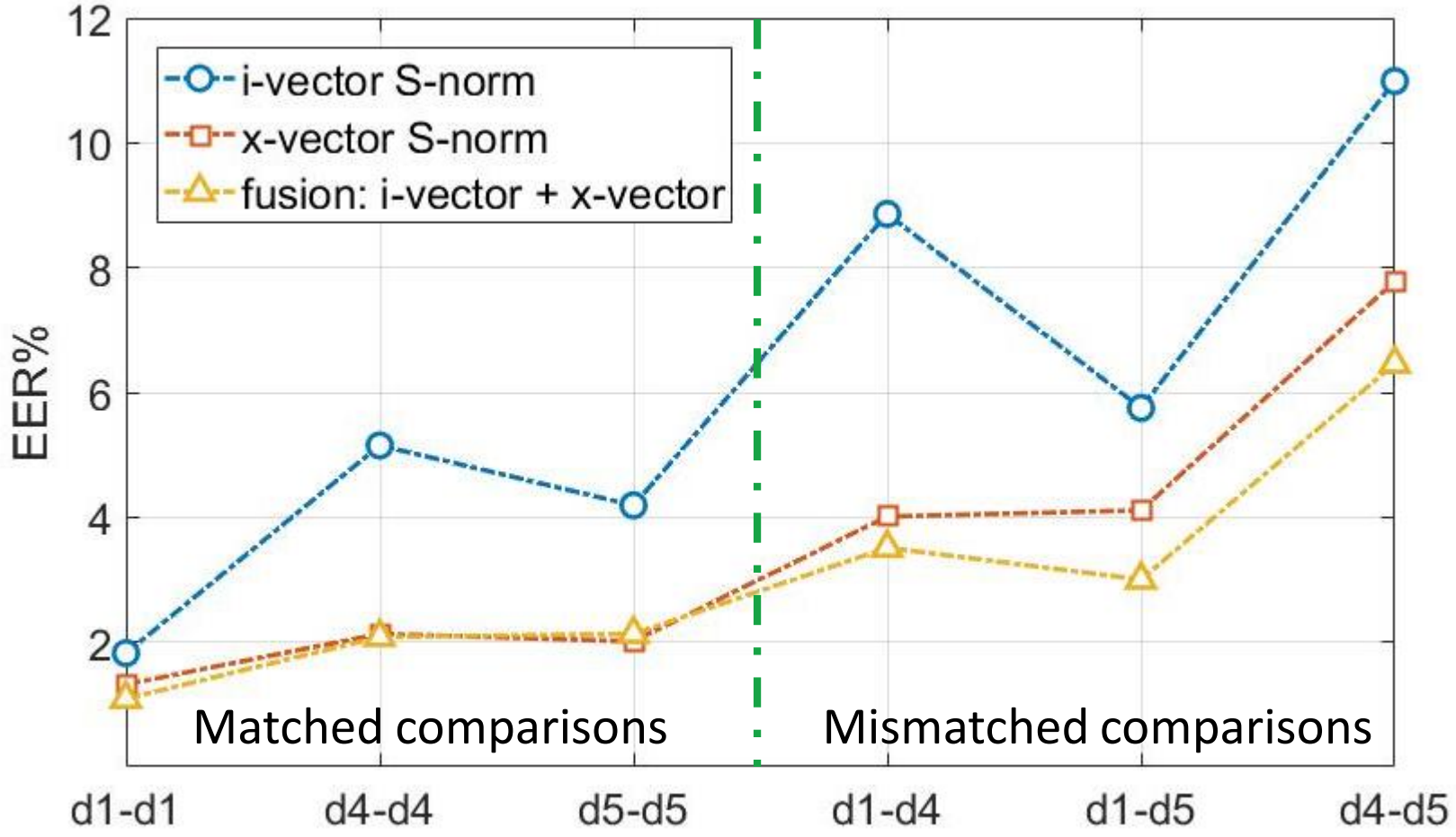


d1: Close mic

d4: Far mic

d5: Tel intercept

NFI-FRIDA results: fusion of i-vector and x-vector



d1: Close mic

d4: Far mic

d5: Tel intercept

Sample Experiments: GBR-ENG*

- 6000 telephone recordings from 600 speakers.
- One side of a landline or mobile telephone conversation of 3-6 minutes duration.
- English speech, recorded across three different accent regions in England
- Within- and cross-condition comparisons with 2134 landline recordings (from 387 speakers), and 3349 mobile recordings (from 534 speakers)
- A separate set was reserved for condition adaptation, consisting of 281 landline recordings and 236 mobile recordings from 50 speakers

* GBR-ENG: *A telephonic speech database collected for the UK Government for evaluating speech technologies*. Further details on application.

Sample Experiments: GBR-ENG

	Condition	Baseline EER%
x-vector	Landline-Landline	0.94
	Mobile-Mobile	1.68
	Mobile-Landline	3.30
i-vector	Landline-Landline	2.38
	Mobile-Mobile	15.33
	Mobile-Landline	15.67

Gerlach, L., Kelly, F., and Alexander A. (2018). *More than just identity: speaker recognition and speaker profiling using the GBR-ENG database*, IAFPA 2019

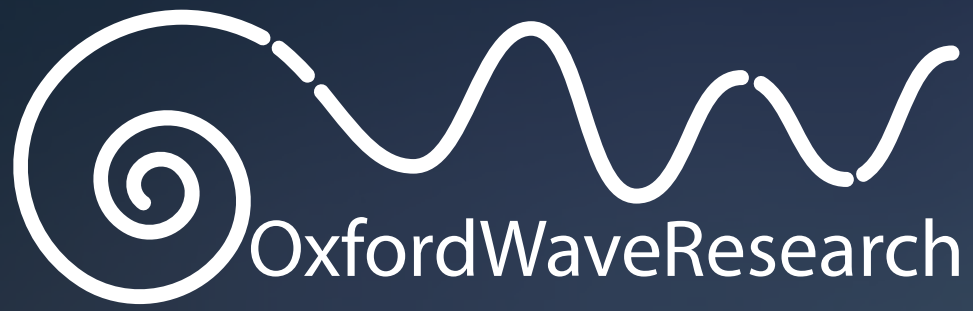
Sample Experiments: GBR-ENG

	Condition	Baseline EER%	Adapted EER%
x-vector	Landline-Landline	0.94	0.71
	Mobile-Mobile	1.68	1.40
	Mobile-Landline	3.30	3.02
i-vector	Landline-Landline	2.38	2.05
	Mobile-Mobile	15.33	5.80
	Mobile-Landline	15.67	7.23

Gerlach, L., Kelly, F., and Alexander A. (2018). *More than just identity: speaker recognition and speaker profiling using the GBR-ENG database*, IAFPA 2019

Conclusions

- The new DNN-based version of VOCALISE using x-vectors provides a powerful, flexible tool for automatic speaker recognition
- It maintains an open-box philosophy and allows the forensic practitioner to interpret their speaker recognition results in a likelihood-ratio framework.
- Significant performance improvements are observed using the new VOCALISE x-vector framework ‘out of the box’ with challenging FRIDA data
- Further improvements observed using VOCALISE reference normalisation, condition adaptation, and fusion



Questions?