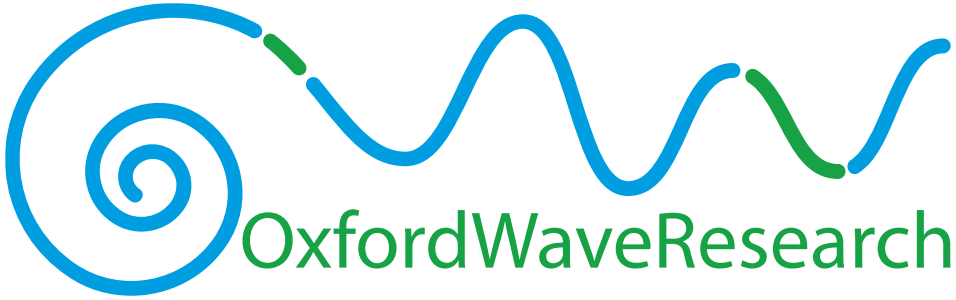


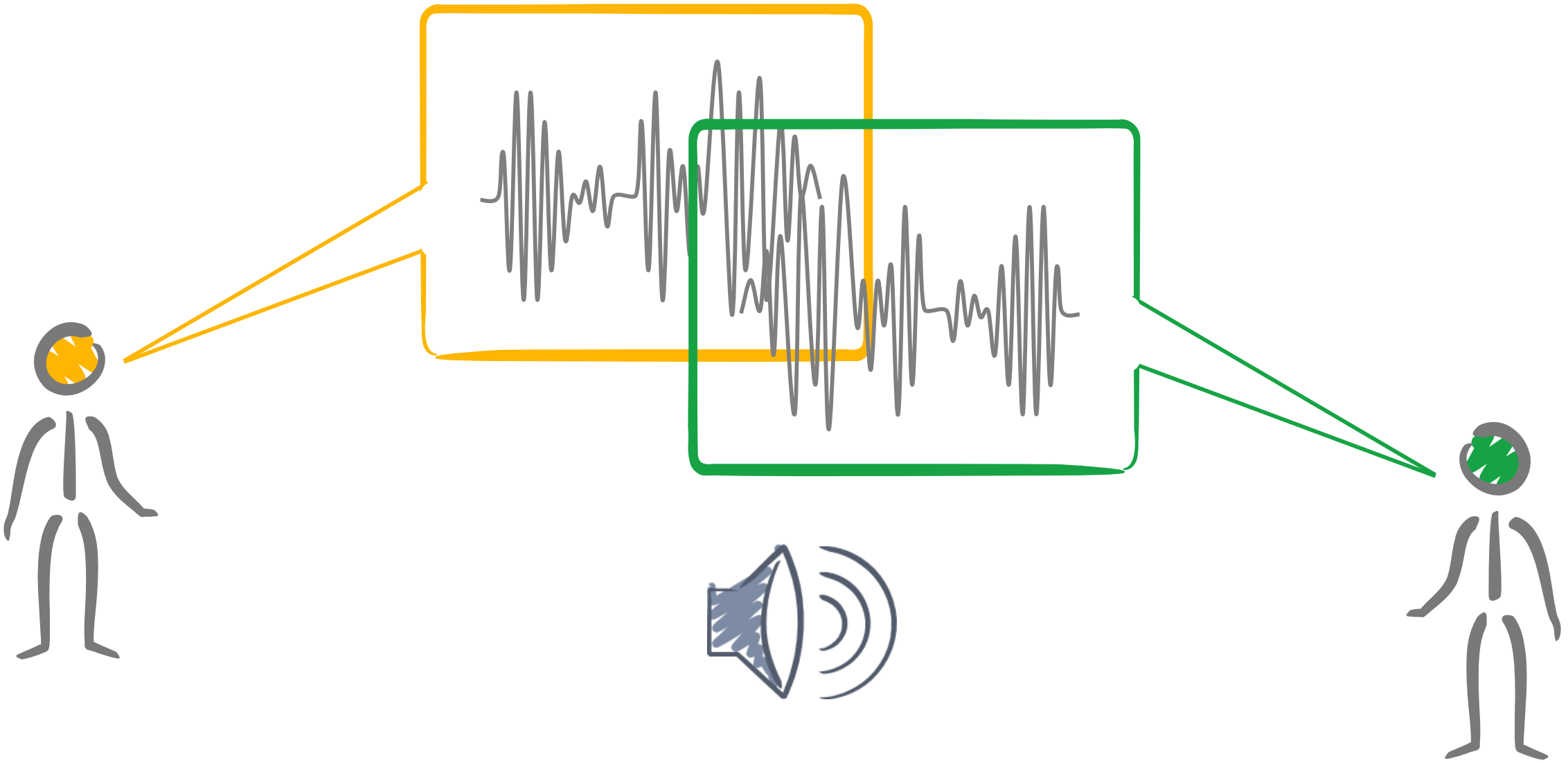
One out of many: A sliding window approach to automatic speaker recognition with multi-speaker files

Linda Gerlach[1], Finnian Kelly[2], Anil Alexander[2]

[1]Philipps-Universität Marburg, [2]Oxford Wave Research Ltd.

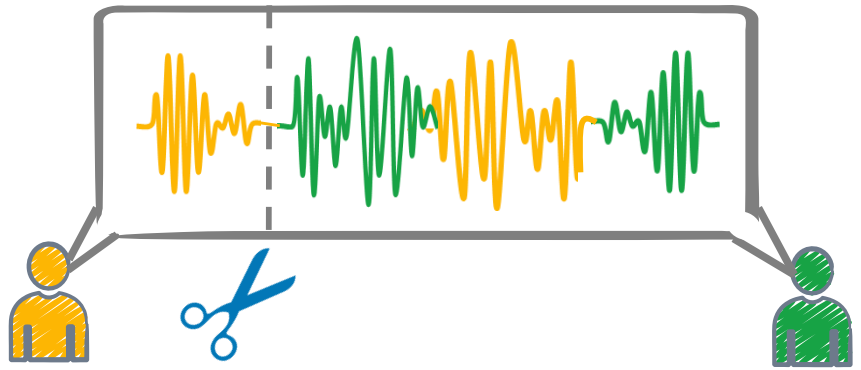
{gerlach8@students.uni-marburg.de, anil|finnian@oxfordwaveresearch.com}

IAFPA conference Istanbul, 14.-17.07.2019



Background

- Situation: a multi-speaker recording has to be analysed
- A typical first step: speaker diarisation
- Problem: it is time consuming
- Especially when dealing with large numbers of multi-speaker recordings, manual or semi-automatic diarisation is not feasible.



Real-case motivation

- Alexander et al. (IAFPA 2017):
  - Dutch police:

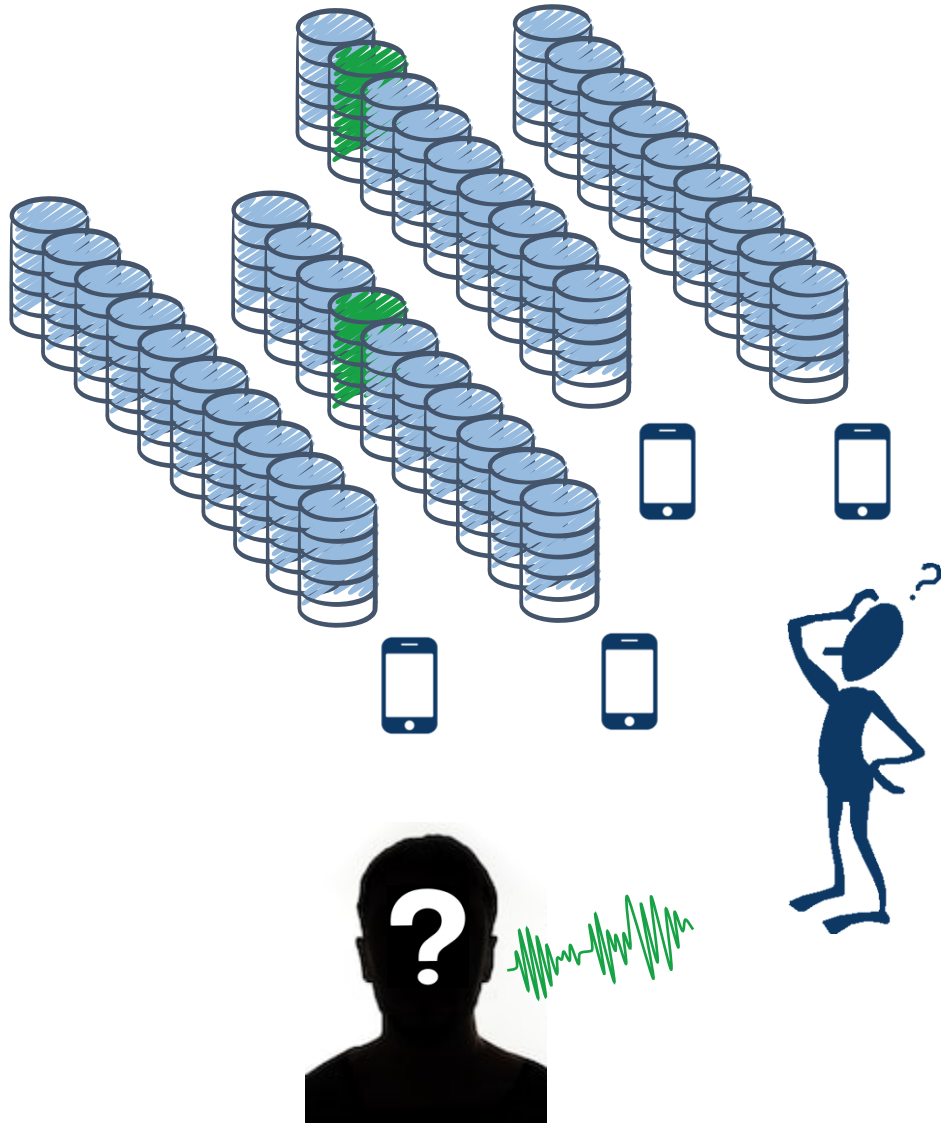
    4 years of telephone intercept recordings comprising about 1000 files
  - Question:

    In which calls may a known suspect be present?
  - Options:

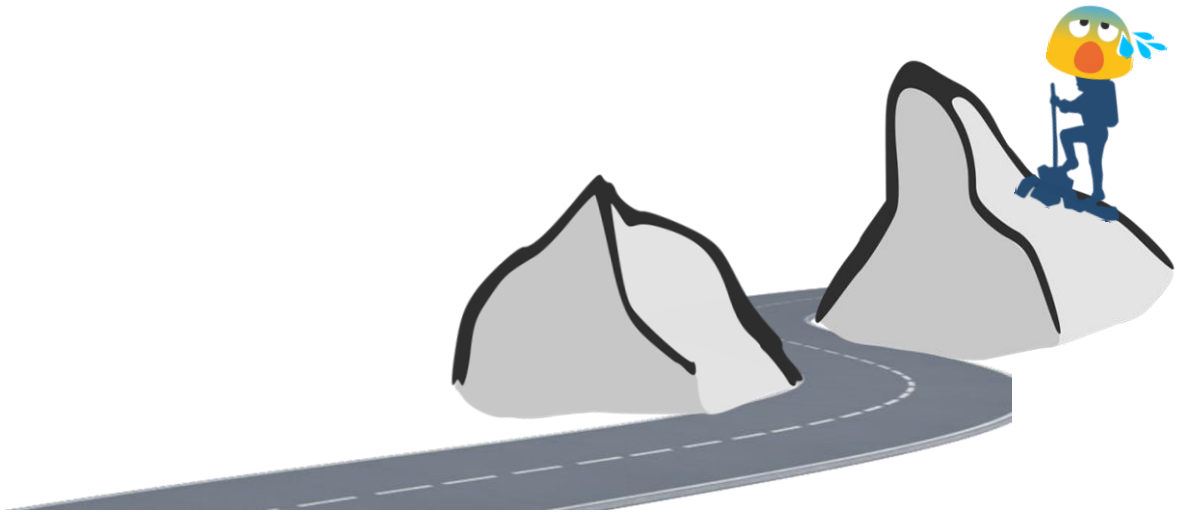
    Manual → takes too long (x4)

    or

    automatic diarisation → varying accuracy; human assistance required



Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2017). *Not a lone voice: Automatically identifying speakers in multi-speaker recordings*. Presentation at IAFPA 2017.



Is it possible to bypass speaker diarisation?

(by adopting a simple sliding window approach to speaker recognition)

Previous approaches

- Approach in Alexander et al. (IAFPA 2017):
  - They used a segmental approach within the i-vector framework, where the overall score was made up by the average of the three highest scores across the whole recording.
  - Disadvantage: Tricky for analysis of live recordings.

- The present study explores a further simplified method and compares the performance of the i-vector and x-vector frameworks.

Approach

1. Short segments from the given multi-speaker recording are extracted.
2. Each segment is modelled within the VOCALISE i-vector or x-vector framework.
3. The speaker models of each segment are compared with the model of the target speaker.
4. The comparison scores obtained across all segments are compared to each other.
5. The final match score is provided by the maximum score of the segment comparisons.

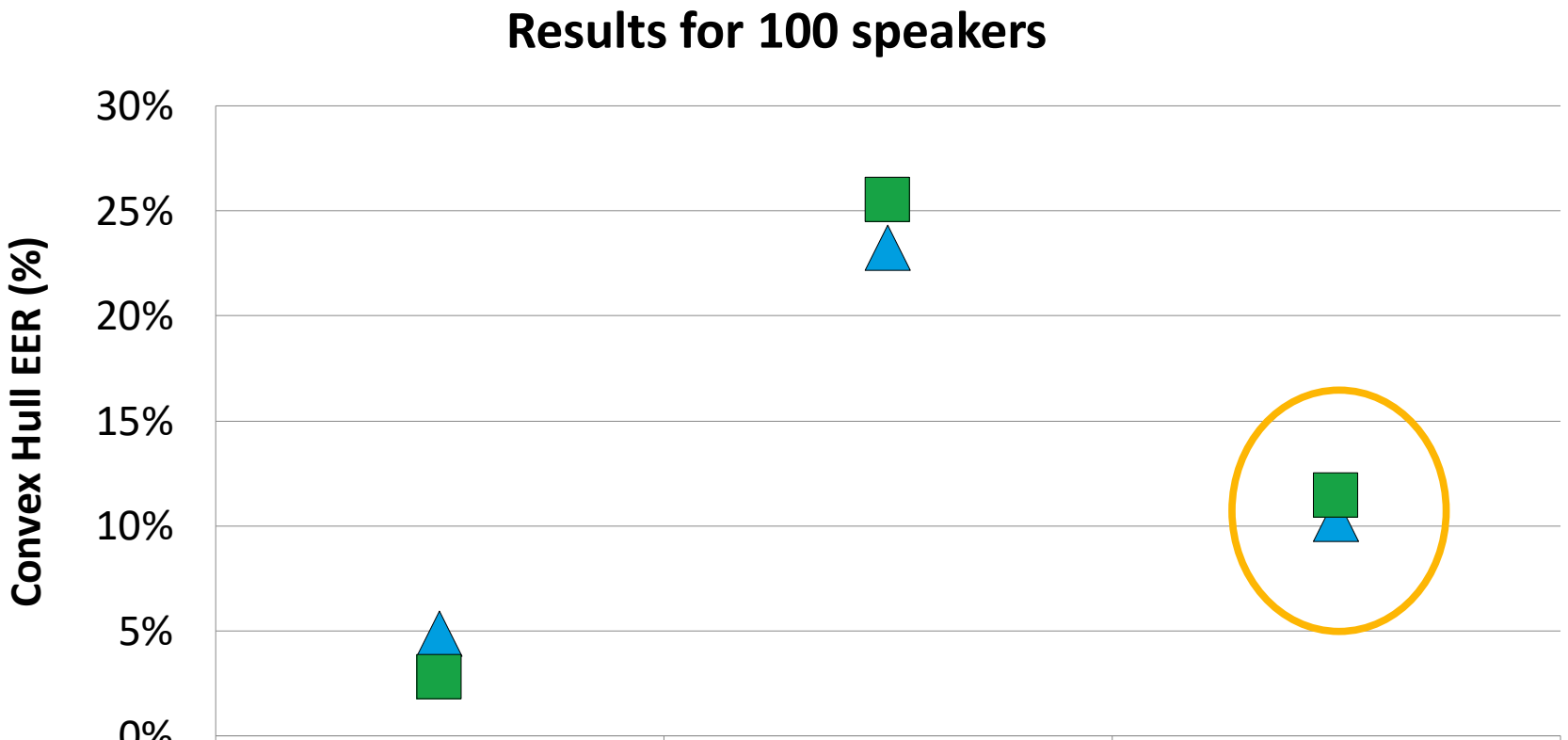
Two conditions (controlled versus real-world) were tested.

Sliding window approach

Experimental data – controlled conditions

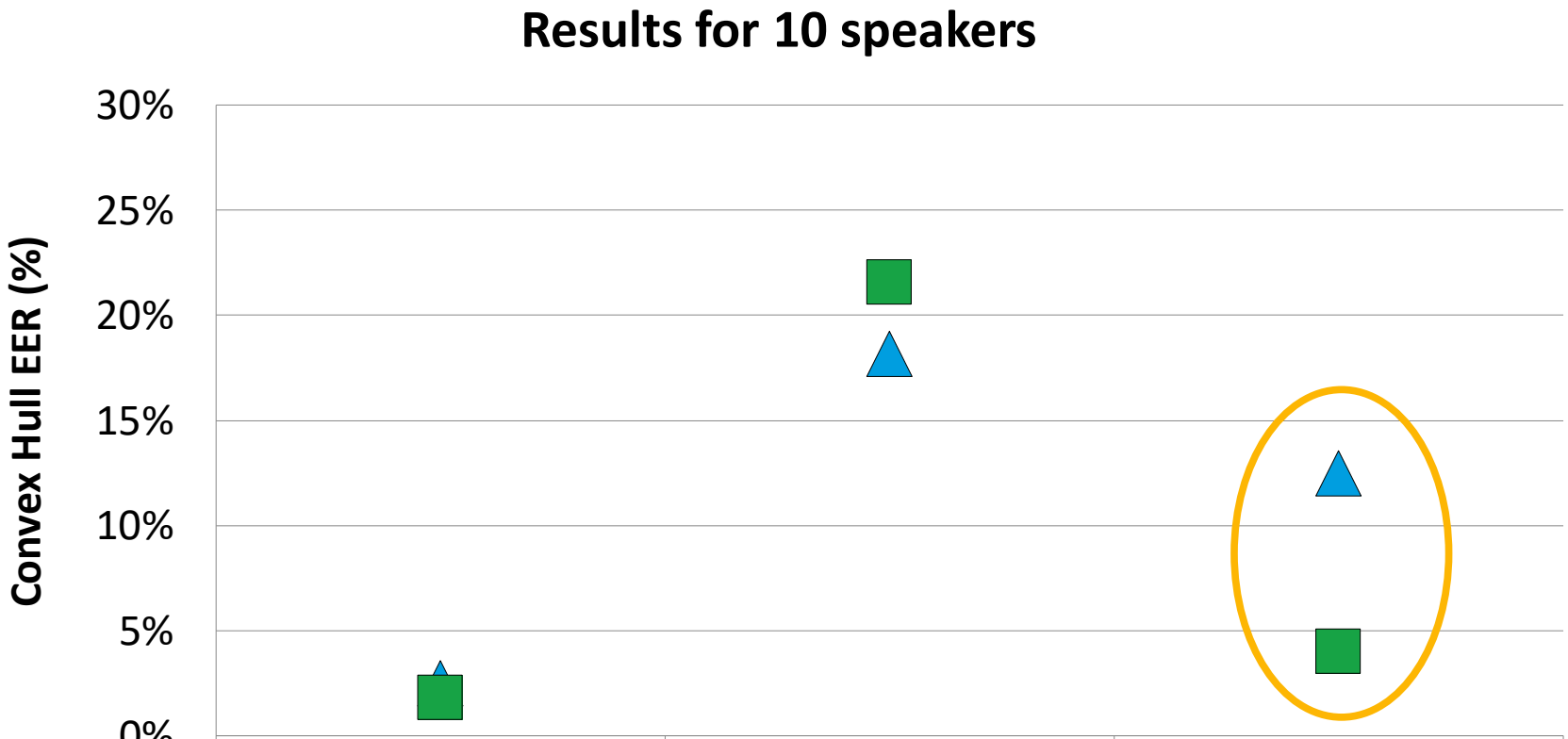
- Corpus: DyViS (Nolan, 2011)
  - Task 1 (Interview) as a source of two-speaker data
    - **Channel 1**: predominantly target speaker (+ bleeding from channel 2)
    - **Channel 2**: predominantly interviewer (+ bleeding from channel 1)
    - **Merged channels**: speech from both speakers
  - Task 3 (Report), containing only the target speaker
  - 100 speakers in total; 10 speakers used for sliding window approach

EERs for DyViS Task 1 (channel 1, channel 2, merged channels) vs Task 3:



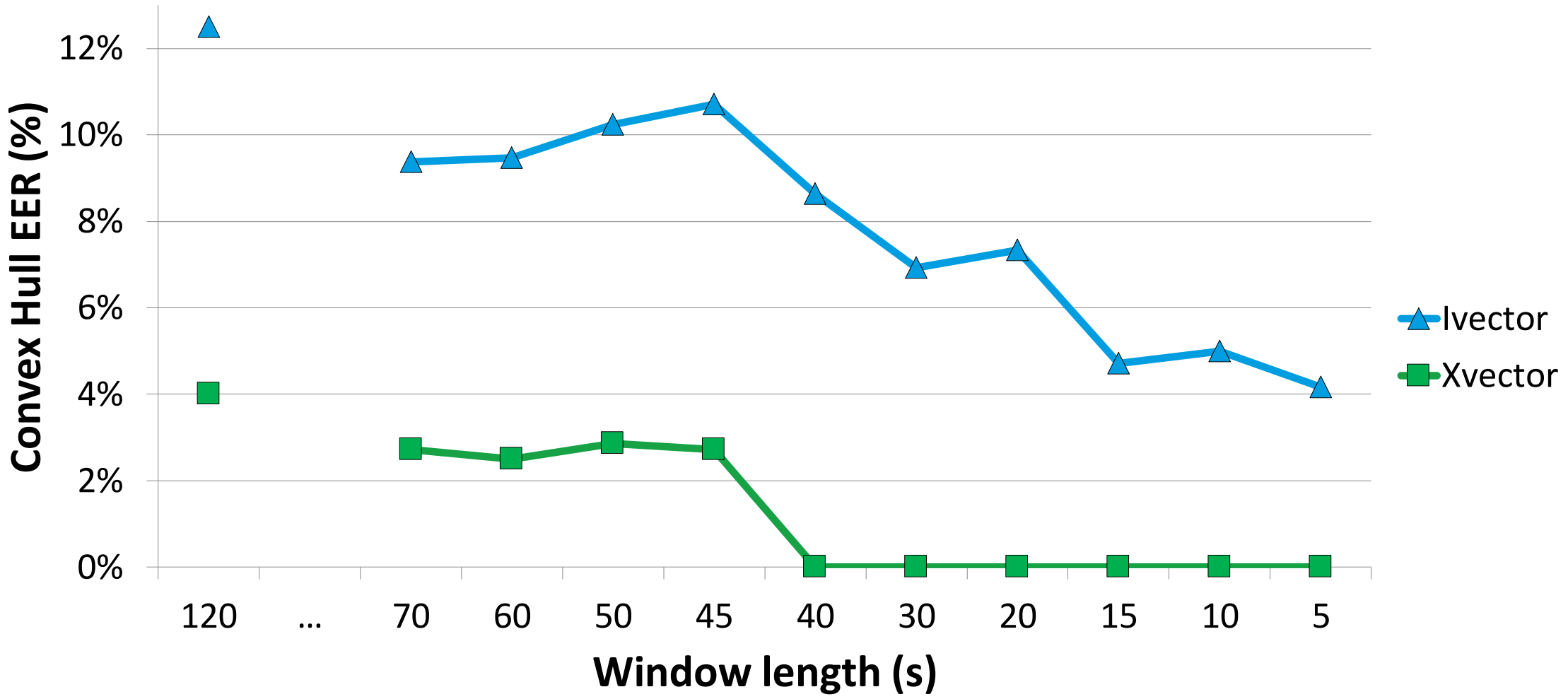
	Channel 1	Channel 2	Merged
▲ i-vector	4,85%	23,25%	10,32%
■ x-vector	2,78%	25,54%	11,45%

EERs for DyViS Task 1 (channel 1, channel 2, merged channels) vs Task 3:



	Channel 1	Channel 2	Merged
▲ i-vector	2,50%	18,16%	12,50%
■ x-vector	1,82%	21,58%	4,00%

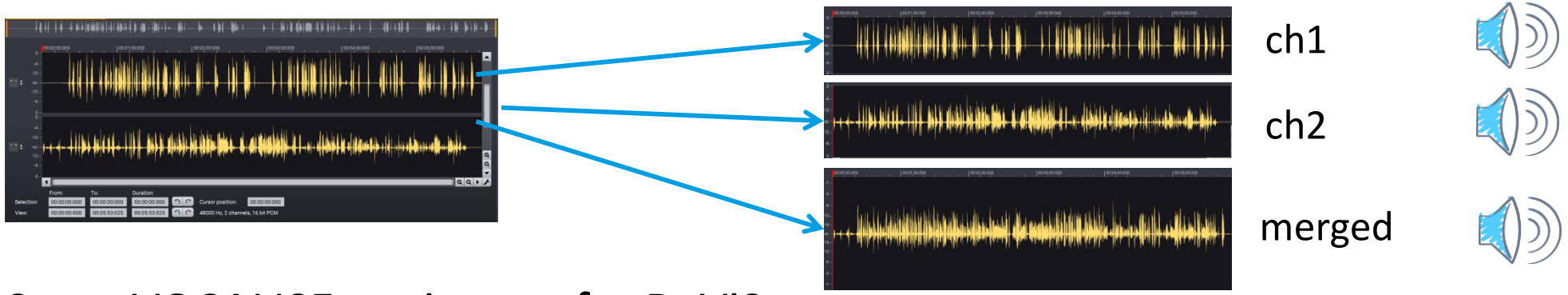
EERs for DyViS Task 1 (merged channels with sliding window approach) vs Task 3: 10 speakers



Experimenting with FRIDA – uncontrolled condition

- Corpus: NFI-FRIDA
- 10 stereo recordings containing 5 speaker pairs
- Background noise: wind, birds
- Both channels extracted represent a speaker of interest
- Merged channels were subjected to the sliding window procedure
- Extracted channels contain bleeding from the other channel to some extent

***Thank you, David van der Vloed (NFI)**

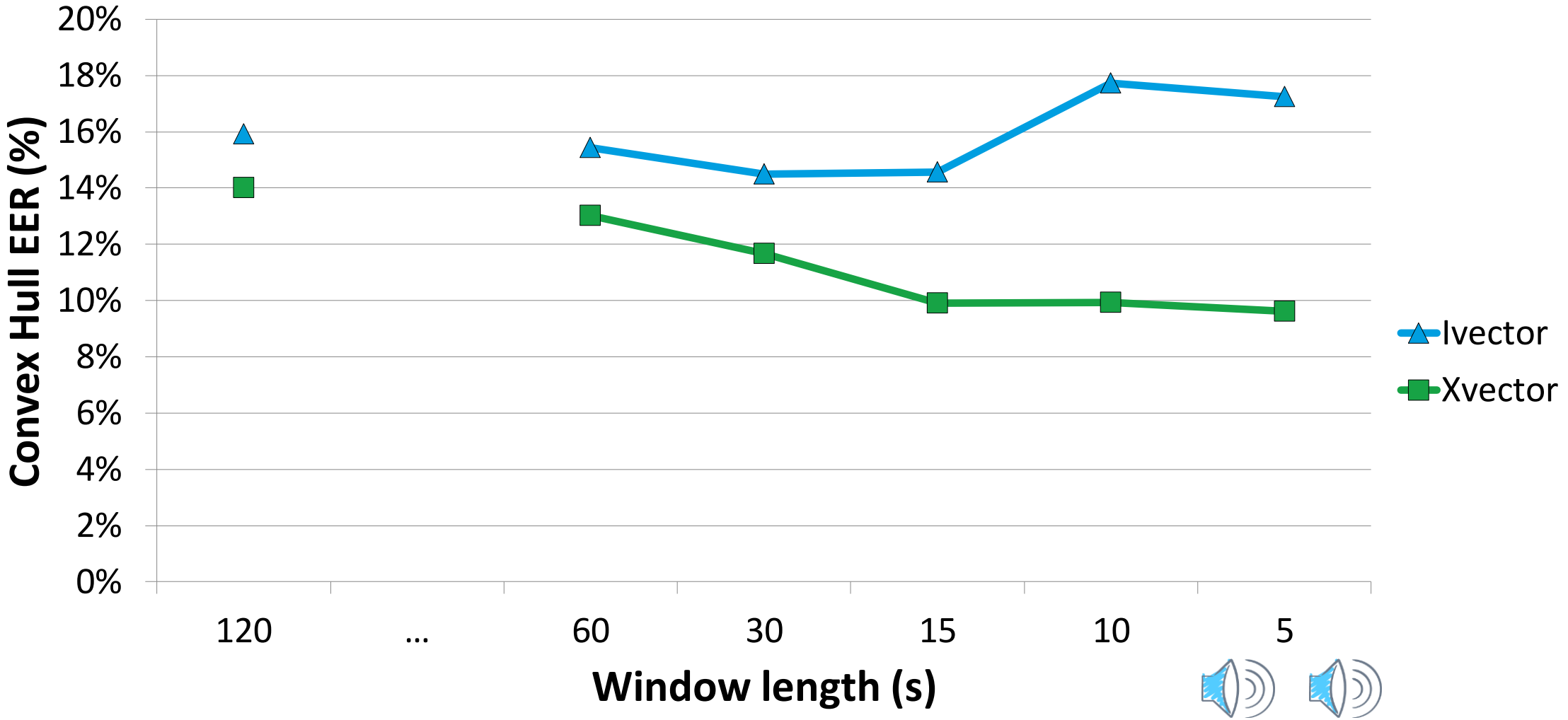


- Same VOCALISE settings as for DyViS

8 conversations per day, background noisiness

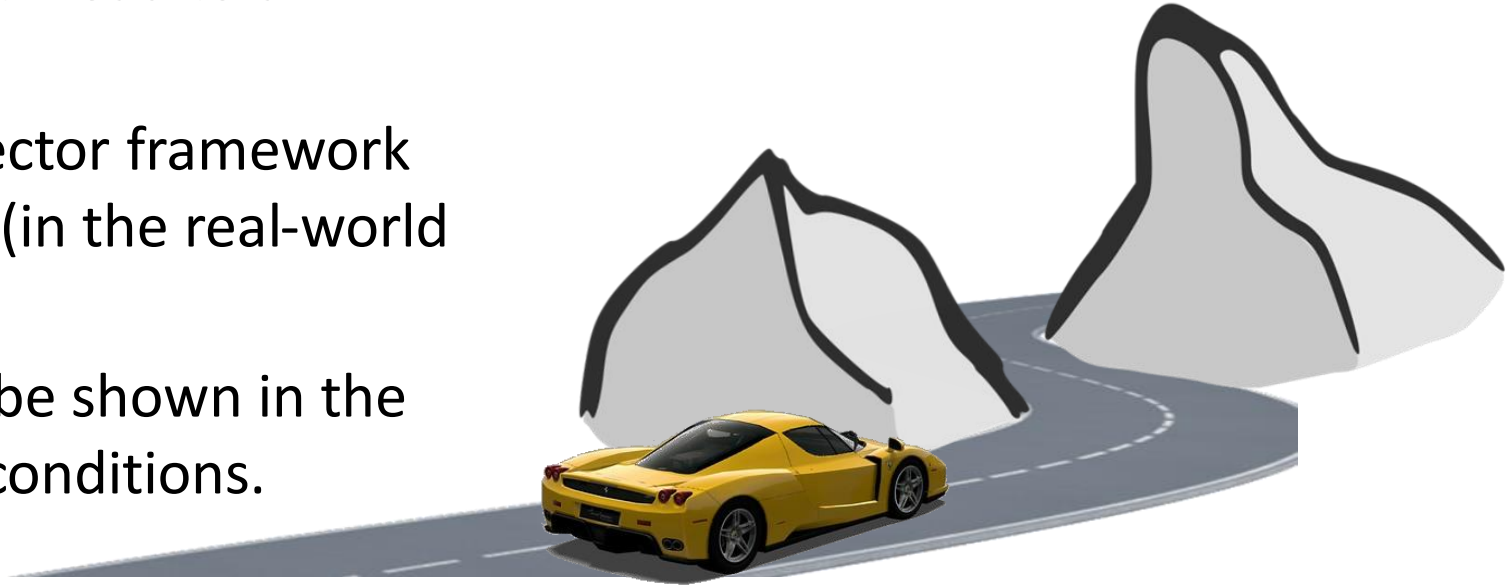
Van der Vloed, D., Bouten, J., Kelly, F. & Alexander, A. (2019). Forensically Realistic Inter-Device Audio (NFI-FRIDA) and initial experiments. Nederlands Forensisch Institut.

Results for merged file comparisons using FRIDA



Can we hope to relieve the analyst?

- Decreasing the window length lead to an increase in discrimination.
- An initial decrease in the i-vector framework was followed by an increase (in the real-world condition).
- A continued decrease could be shown in the x-vector framework in both conditions.



- The sliding window approach looks promising and would be an effective, light-weight approach that could be used for analysing multi-speaker files and could be extended to live comparisons.

Special Thanks for Data

Netherlands Forensic Institute

University of Cambridge

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.
- Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2017). *Not a lone voice: Automatically identifying speakers in multi-speaker recordings*. Presentation at IAFPA 2017.
- Nolan, F. (2011). *Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006-2007*. [data collection]. UK Data Service. SN: 6790, http://doi.org/10.5255/UKDA-SN-6790-1
- 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.
- van der Vloed, D., Bouten, J., Kelly, F., and Alexander A. (2018). NFI-FRIDA – Forensically Realistic Inter-Device Audio. IAFPA 2018.

Questions?