NOT A LONE VOICE: AUTOMATICALLY IDENTIFYING SPEAKERS IN MULTI-SPEAKER RECORDINGS

Anil Alexander, Oscar Forth, Alankar A. Atreya & Finnian Kelly

Research and Development

Oxford Wave Research Ltd

IAFPA 2017, Split, Croatia

MOTIVATION

- Law enforcement audio recordings such as interviews, telephone intercepts and surveillance recordings often contain speech from more than one speaker.
- Identifying speakers of interest within these multi-speaker recordings first involves editing to extract the speech of a single speaker.
- This editing process, in which extraneous noises and other speakers are removed, can either be performed manually or assisted using speaker diarisation software.
- However, if a large number of such files need to be analyzed in a short period of time, it may not be practical to involve a human in the loop.

TELEPHONE INTERCEPTS

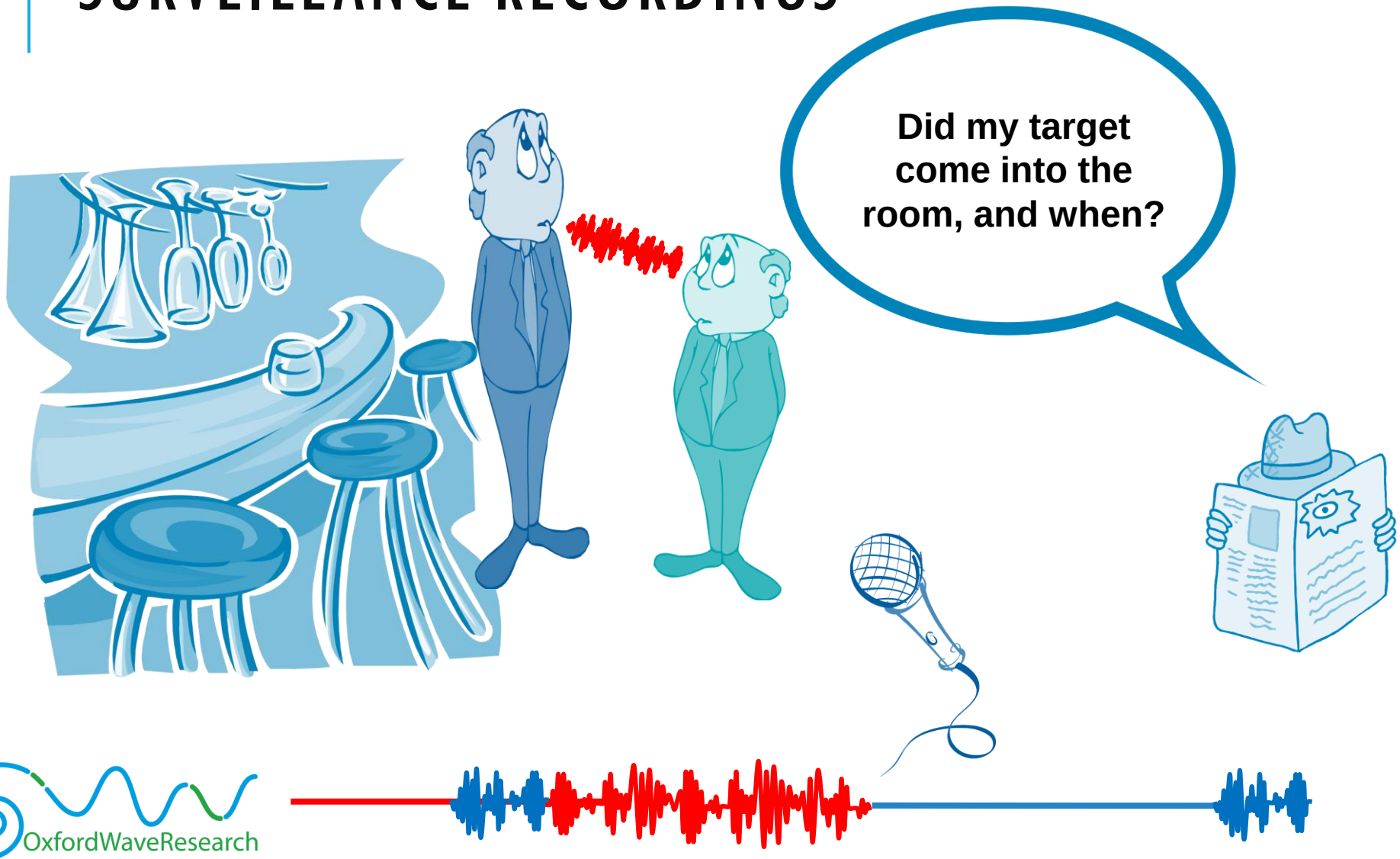
Telephone conversations may be recorded as

- Mono – both speakers on one channel
- Stereo – each speaker on one channel
- Dual Mono - stereo, but two speakers each channel

In order to perform speaker recognition you need to separate out speakers either by manually cutting up the files or by using automatic speaker diarization tools.

OxfordWaveResearch

SURVEILLANCE RECORDINGS



REAL-CASE MOTIVATION

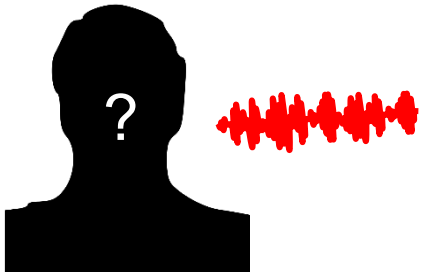
Netherlands (Dutch Police) *

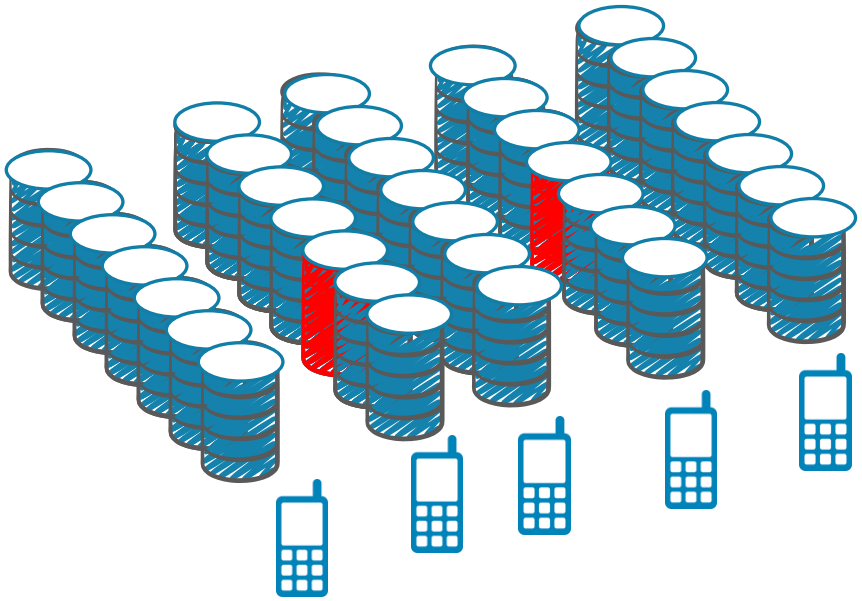
The police have four years of intercept recordings.

These are all two-wire (mono) recordings with two or more speakers per file.

There are approx. 1000 files.

The police have a known suspect, and they want to find which calls he may be present in.





*Thanks David van der Vloed, NFI

OxfordWaveResearch

THE ALTERNATIVES – MANUAL SPLITTING OR SPEAKER DIARISATION

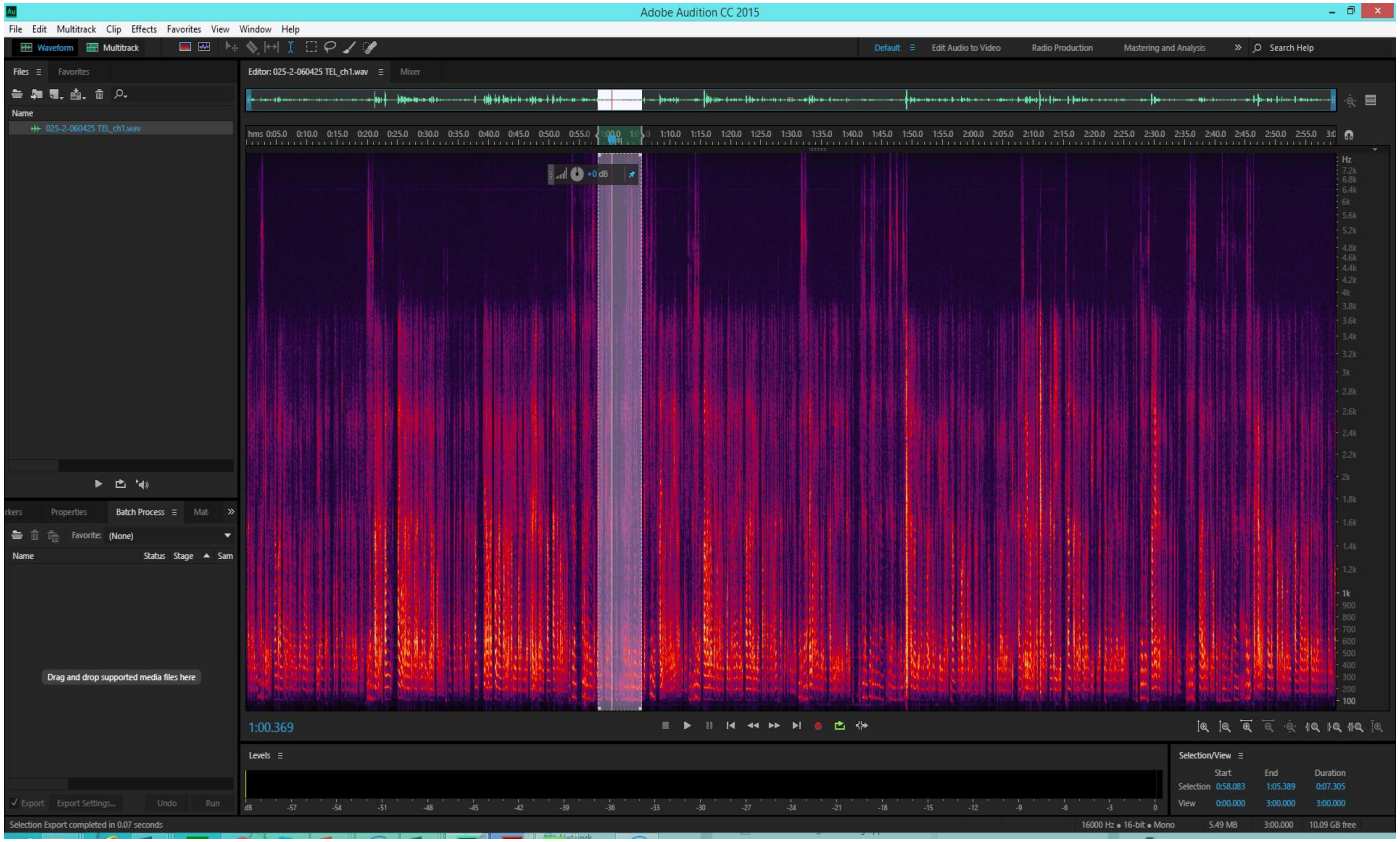
We estimate that manually diarizing an audio file takes at least 4-5 times the duration of the file. (e.g. a one hour file could take four hours or more to diarize even by an experienced practitioner).

The other option is to use automatic diarization software:

- In certain cases they can perform well blindly
- In other cases they are helped by human intervention.
- With diarization you don't know which of the files might be your target speaker

OxfordWaveResearch

MANUAL DIARIZATION

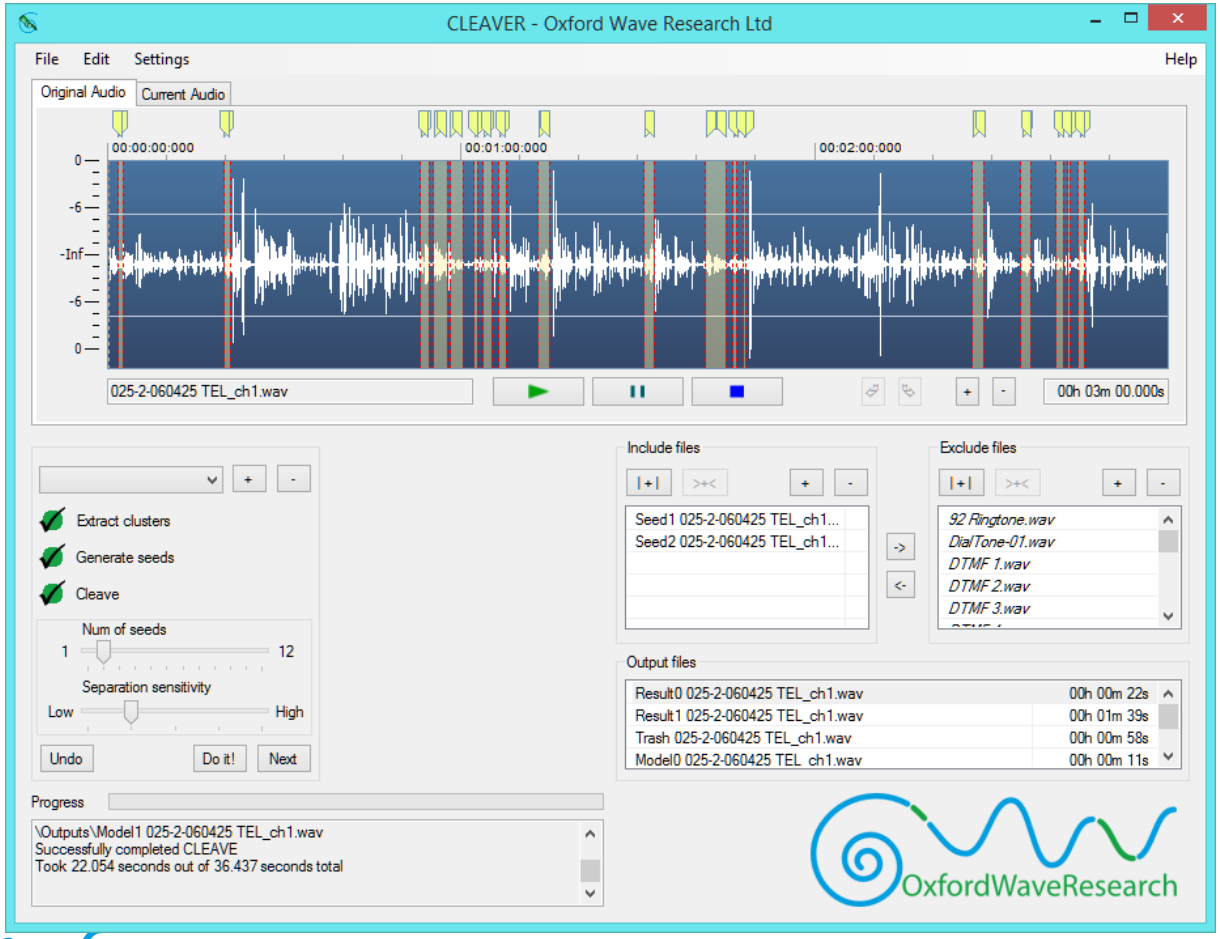


Both Speakers

Interfering Speaker

OxfordWaveResearch

SPEAKER DIARIZATION (BLIND OR ASSISTED)



Speaker 1

Speaker 2

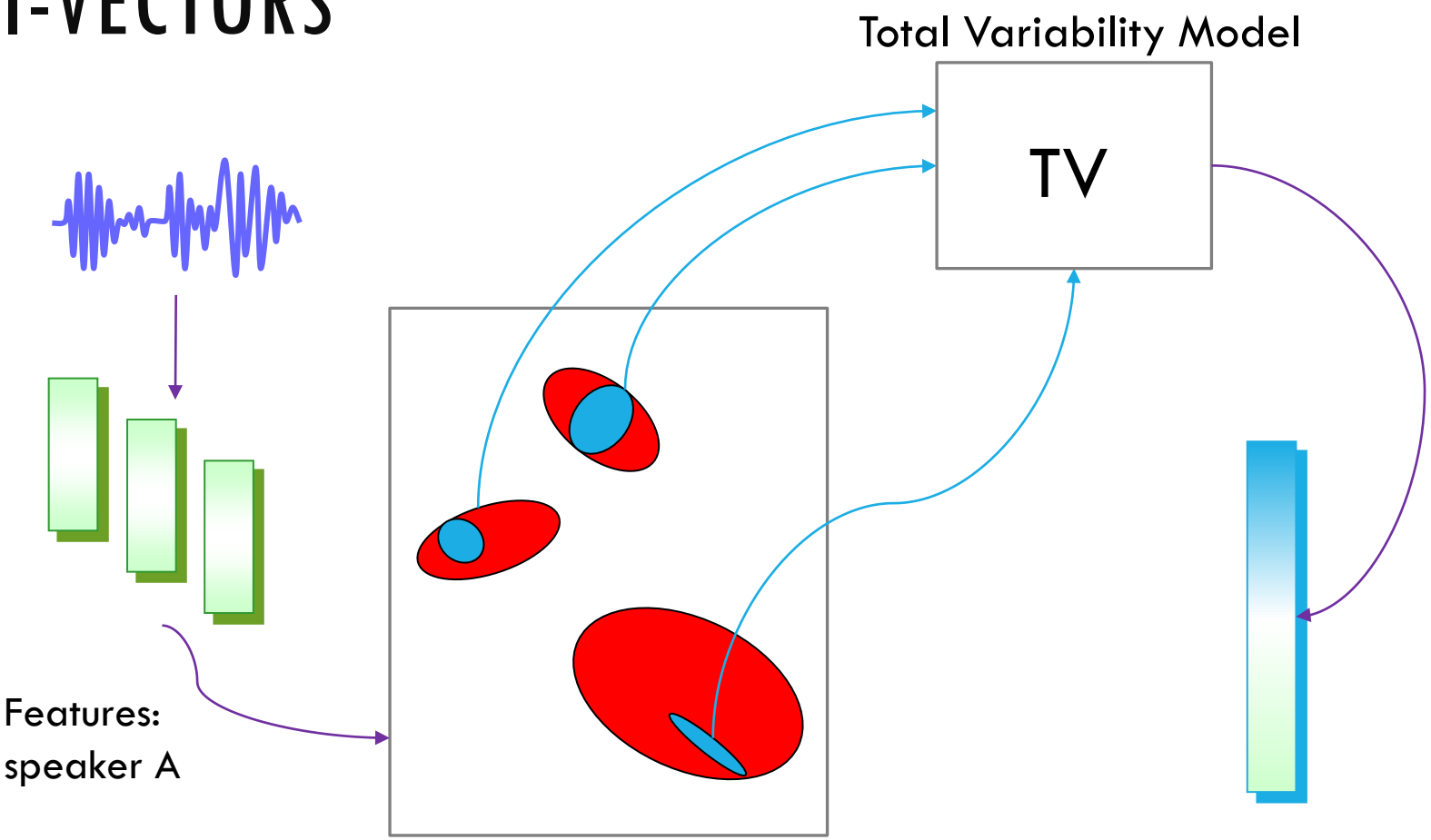
OxfordWaveResearch

APPROACH

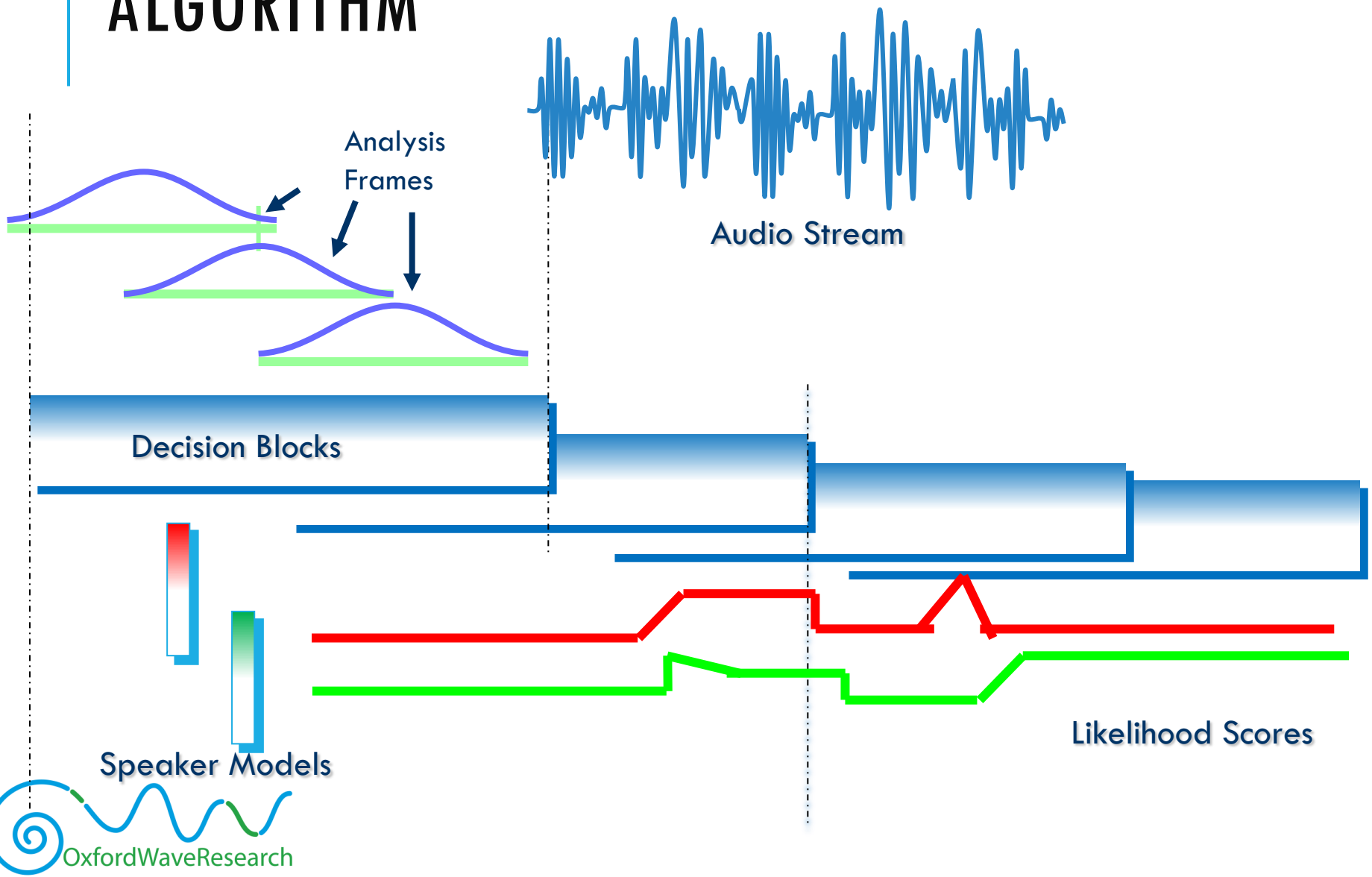
- A simple but effective approach in which short overlapping segments of the multi-speaker recording are extracted and modeled within an i-vector framework.
- The i-vector approach converts a recording into a fixed length, low-dimensional representation of the speaker's voice.
- The i-vectors for each overlapping segment (e.g. 10s segments, with 5s overlap) are compared with the i-vector for the target speaker file.
- The match scores obtained across all overlapping segments are first smoothed to reduce the effect of outliers, and then an average of the three maximum scoring segments provides a match score for the file.

OxfordWaveResearch

I-VECTORS


Total Variability Model
TV
Features:
speaker A
Adapted GMM-UBM:
Speaker A
Large, universal
speaker space
i-vector: Speaker A
Small, highly speaker-
dependent space

ALGORITHM

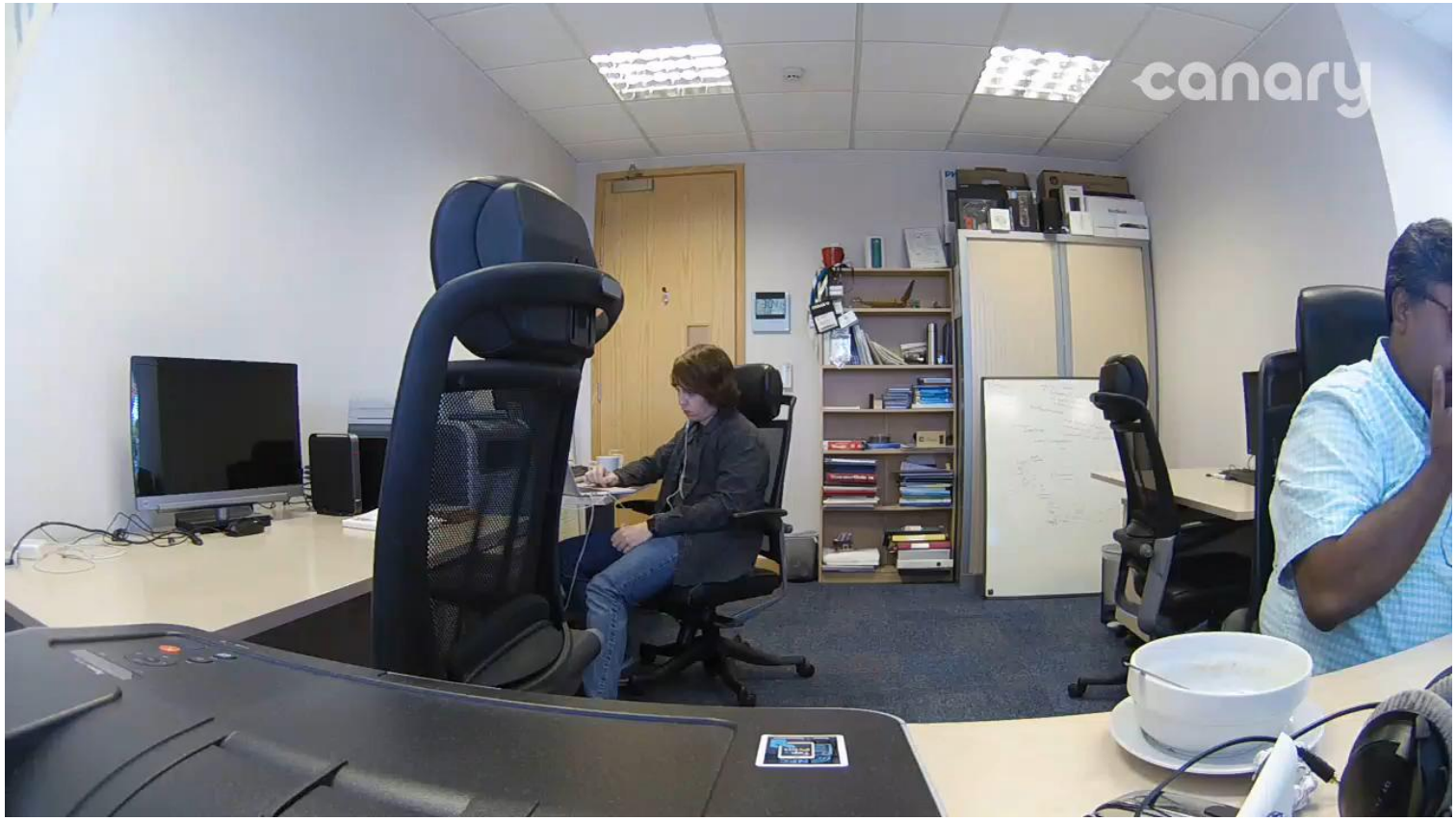

Analysis
Frames
Audio Stream
Decision Blocks
Speaker Models
Likelihood Scores
OxfordWaveResearch

THE BLOCK-BASED MULTI-SPEAKER RECOGNITION APPROACH

- MFCC Features are extracted from the audio
- Voice Activity Detection (VAD) is applied to the audio and frames are marked as speech and non-speech.
- Features normalised using cepstral mean subtraction (CMS) and cepstral mean and variance normalisation (CMVN) and Delta-Deltas.
- The features are split into 10 seconds chunks with a 5 second overlap.
- In each block, any non-speech frames are then removed.
- Each block is modelled using i-vectors and compared to each target speaker's i-vector to obtain scores.
- The score trajectories for each target speaker are smoothed using a three frame running average.
- The average of the top 3 scores for each target speaker is chosen as the match score.

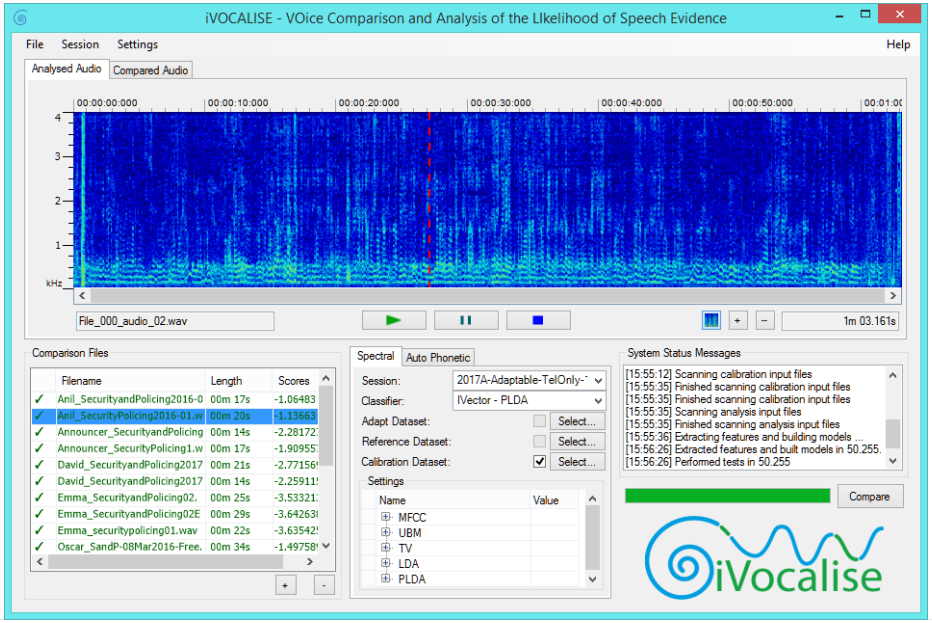
OxfordWaveResearch

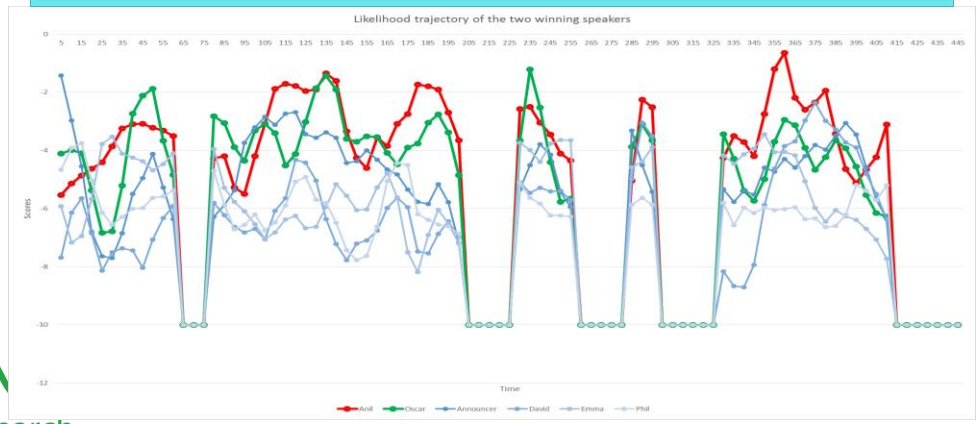
SIMULATED SURVEILLANCE RECORDINGS



OxfordWaveResearch

BLOCKED COMPARISON RESULTS



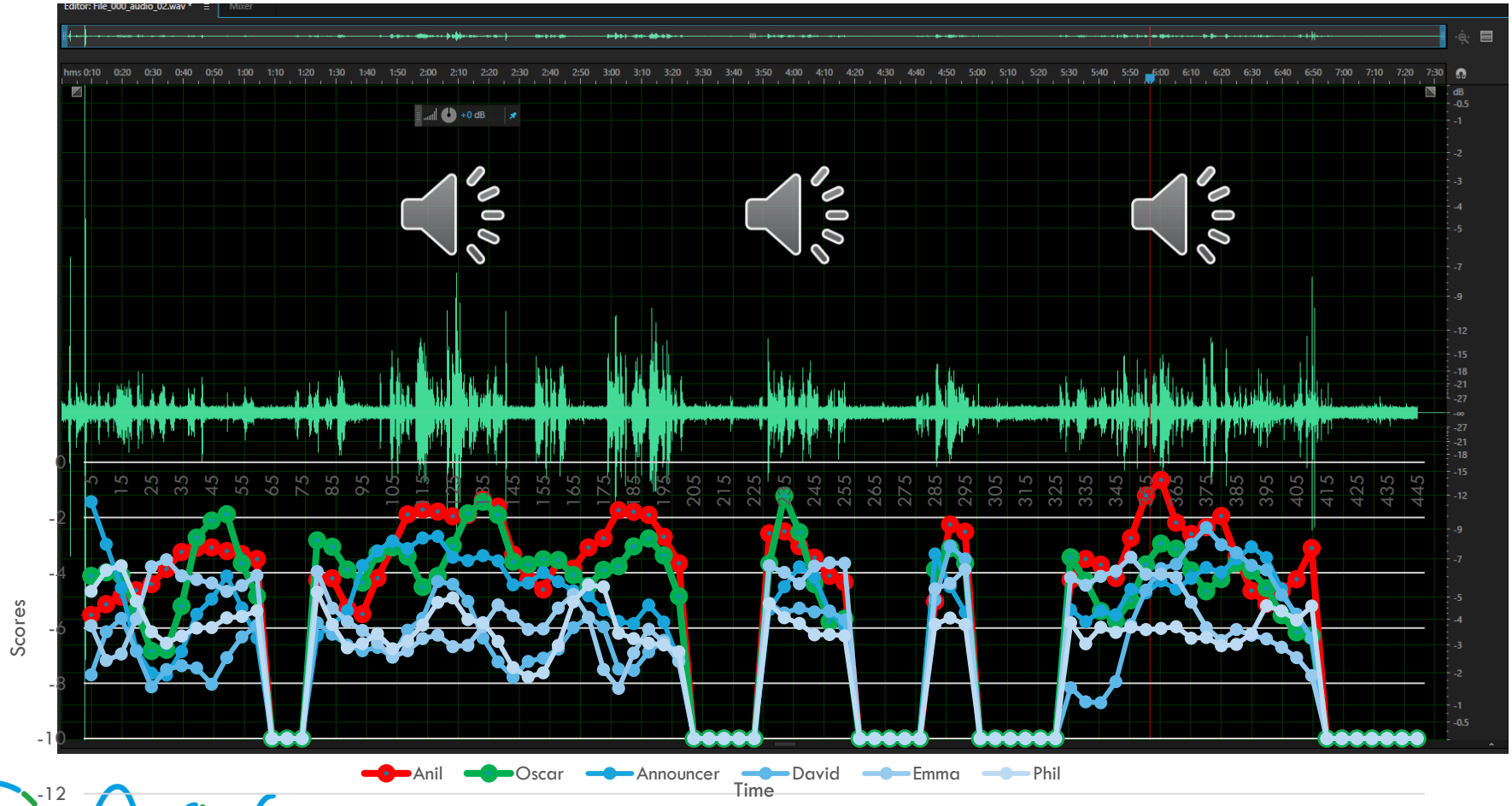


Running blocked Comparisons with score calibration on VOCALISE (2017A)

Scores obtained comparing blocks

OxfordWaveResearch

SIMULATED SURVEILLANCE RECORDINGS



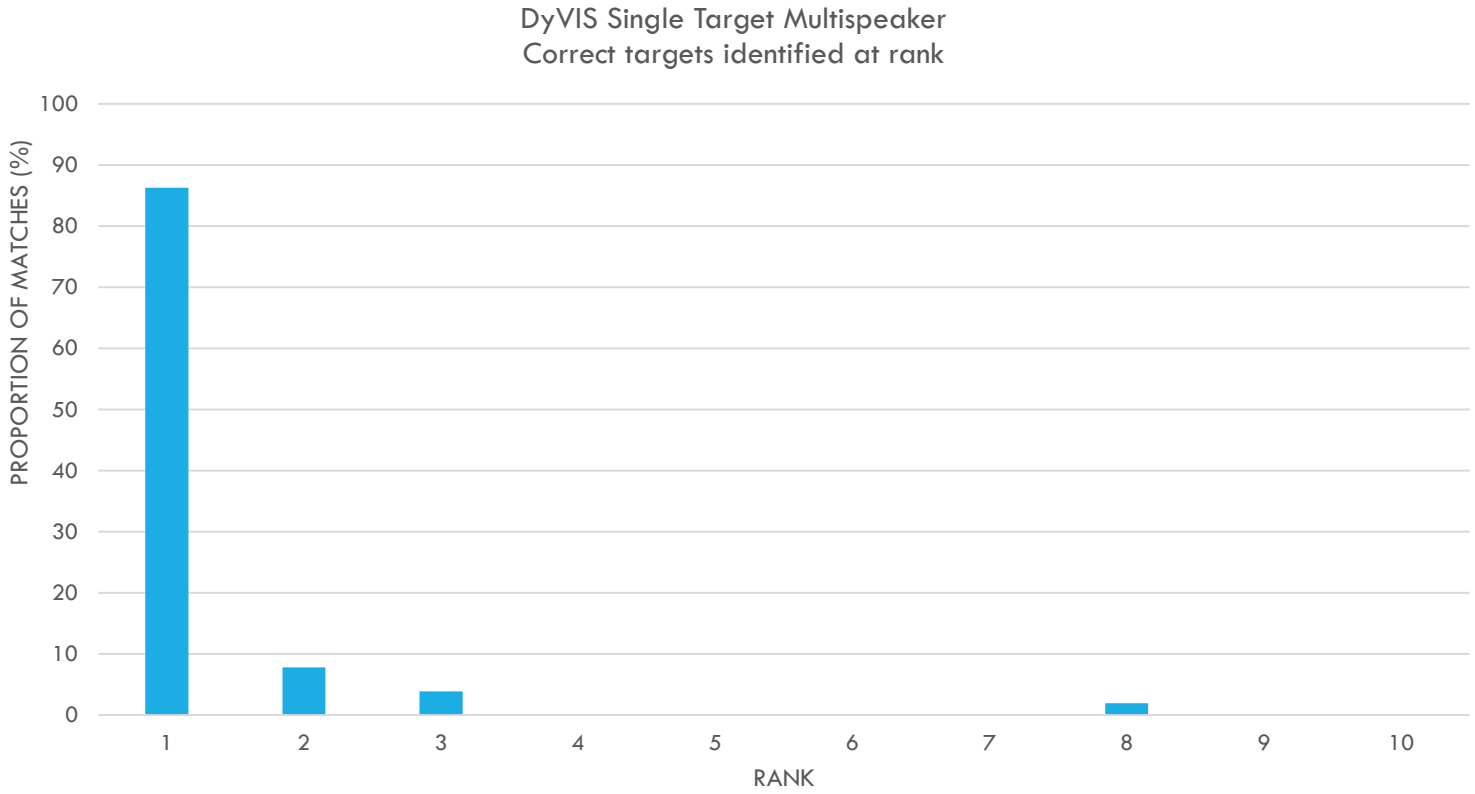
OxfordWaveResearch

EXPERIMENTS CONTROLLED CONDITIONS

- We tested our approach with controlled laboratory data as well as real telephone intercept data. We used a multi-speaker modified version of the VOCALISE speaker recognition software (Alexander et al, 2014).
- For our experiments with laboratory data, we used interview and intercept recordings in same- and cross-channel conditions from the DyVIS database (Nolan et al, 2009).
- For 'single target' cross-channel comparisons, we used 51 files containing two speakers from the intercept task and compared them with 59 single speaker files from DyVIS Task 3 (report and report recall).

OxfordWaveResearch

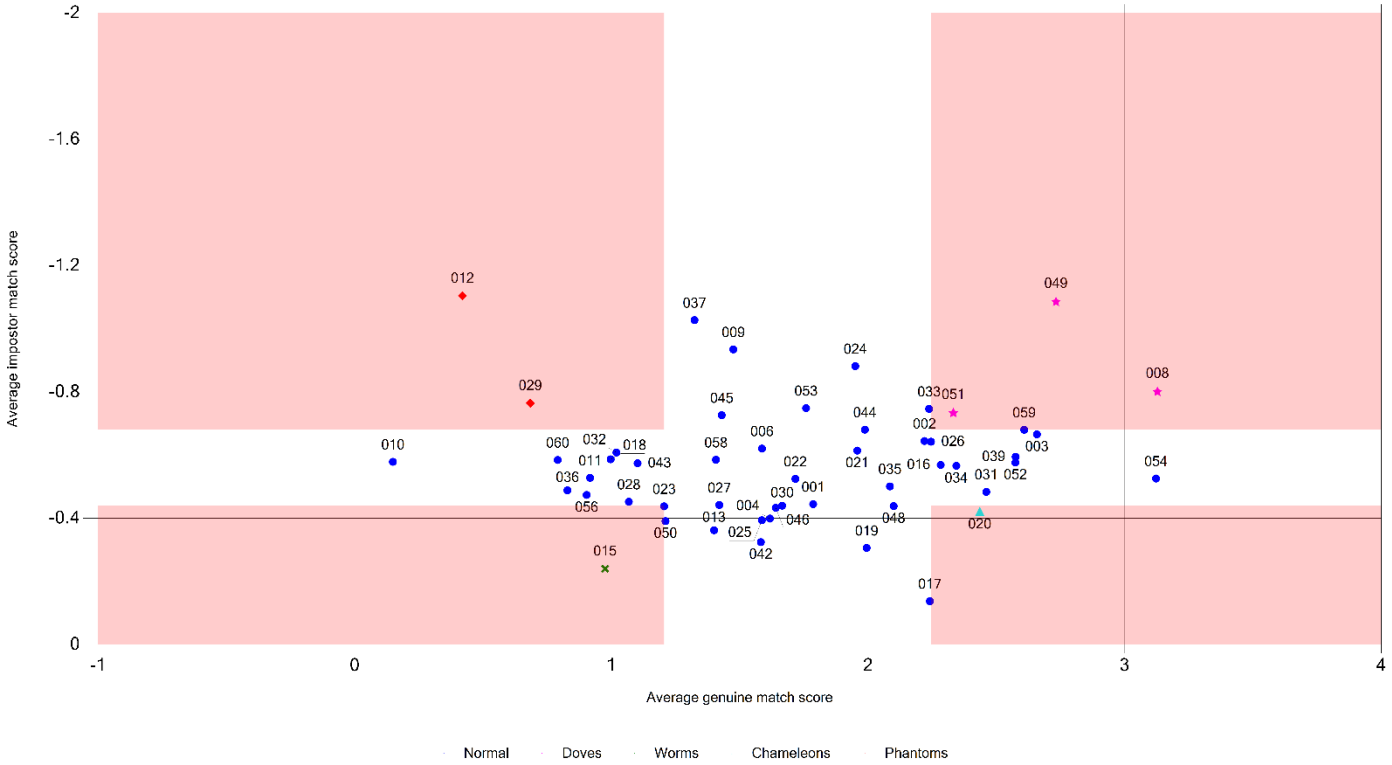
RESULTS DYVIS



For each multi-speaker recording, the majority (94.1%) of corresponding target speakers were identified at rank one or two of the match score list

OxfordWaveResearch

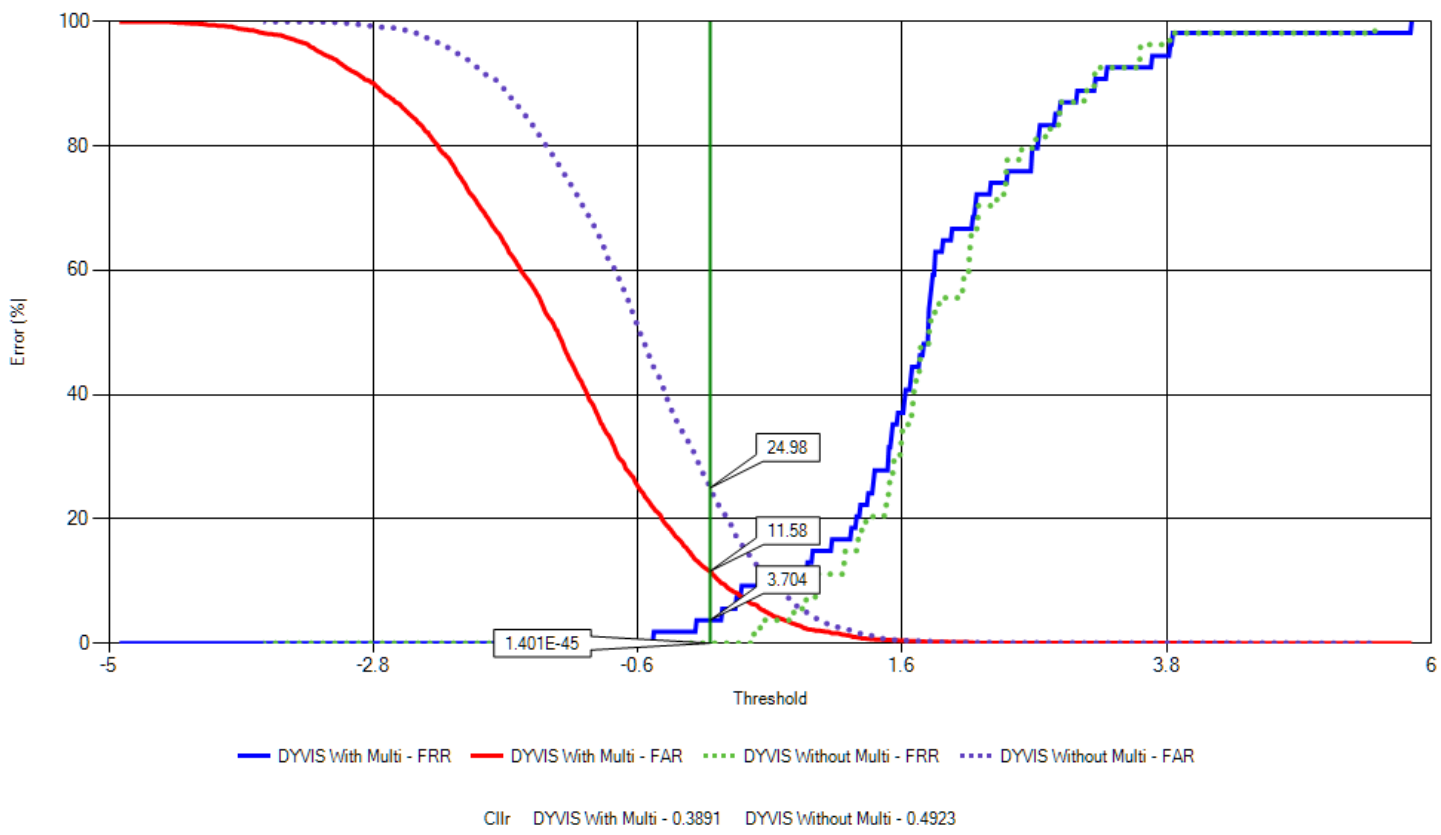
MULTI-SPEAKER DYVIS RESULTS ZOOPLOTS



Equal error rates with (3.9% -ref norm)

OxfordWaveResearch

DYVIS – MULTI VS DIRECT COMPARISON



Without Multi
EER: 7.48%
CLLR:0.492

Multi-speaker
EER: 5.56%,
CLLR: 0.389

OxfordWaveResearch

Analysis Files (single-speaker): DyVis Task 3 Report and Report Recall, Speakers 001 - 060
Reference Normalisation Files (single-speaker): DyVis Task 3 DyVis Task 3 Report and Report Recall, Speakers 061 - 121 (61 files)
Comparison Files (multi-speaker): DyVis Task 1 Interview, Speakers 001 - 060 (54 files)

AUTOMATICALLY DIARISED AUDIO

OxfordWaveResearch

Obtains better accuracy of 0.78% !

EXPERIMENTS REAL INTERCEPT DATA

- For uncontrolled real telephone intercept data, we have worked with a subset of the FRITS database (van der Vloed et al, 2014).
- All tests were conducted by and at the Netherlands Forensic Institute (NFI). This subset consisted of 11 multi-speaker conversations (mostly two, and in some cases, more speakers) and a set of 32 target speakers.
- Both channels were combined together by NFI for these experiments.

OxfordWaveResearch

NFI-FRITS DATABASE

- Data comes from real police intercepts.
- Data is anonymized by editing.
- Availability is very limited due to the sensitivity of the data.
- Speaker ID is 'by proxy'.
- People tasked to judge which file belonged to which speaker were given telephone number, and the possibility to listen through every recording from that number.

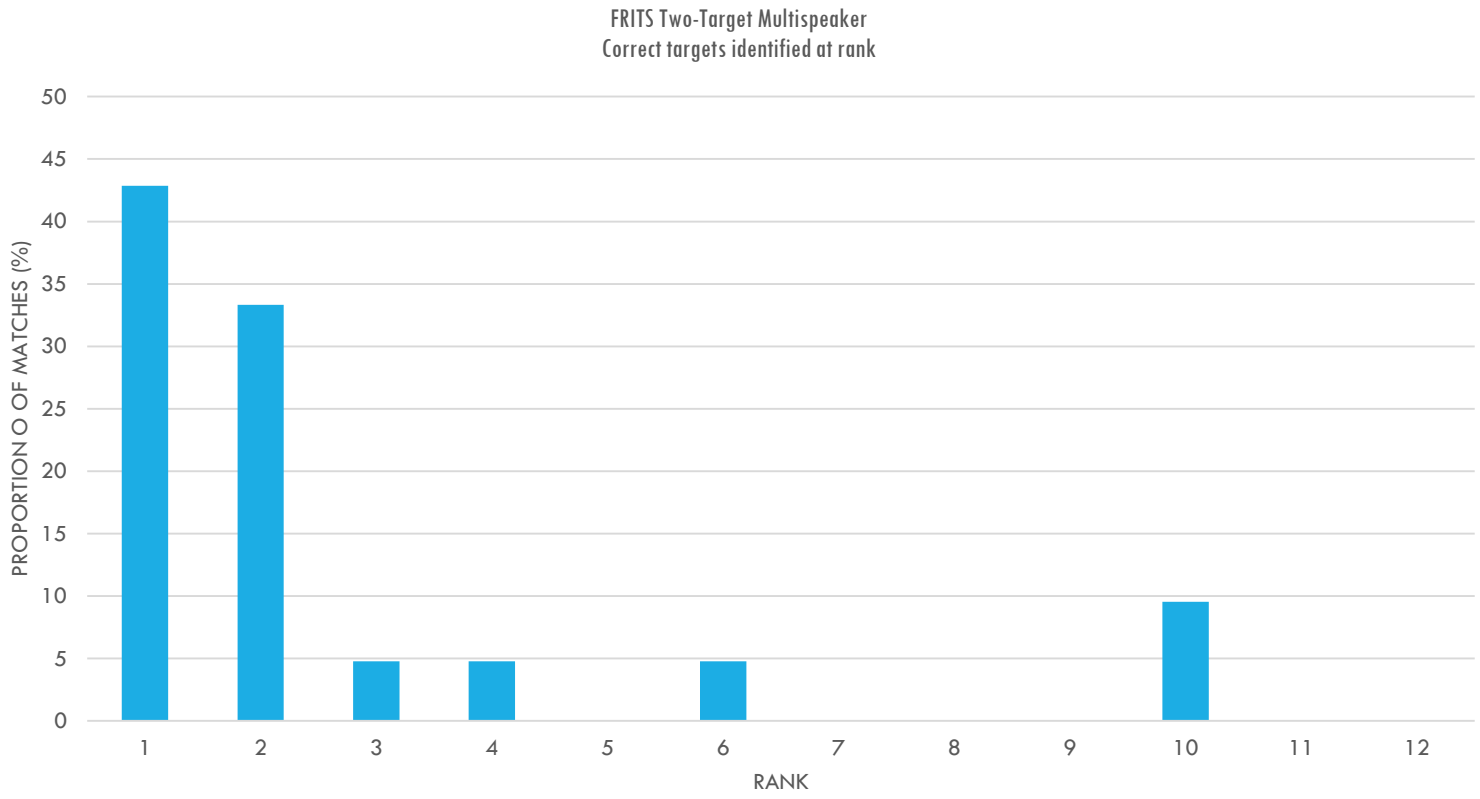
General numbers

604 speakers in 4188 recordings, 165 hours of speech

177 female / 427 male speakers
1068 female / 3120 male recordings

Phone conversations from January 2008 to March 2013

REAL CASE DATA RESULTS (FRITS)



OxfordWaveResearch

REAL CASE DATA RESULTS (FRITS)

- For each multi-speaker recording the majority of corresponding target speakers were identified at rank one or two of the match score list (76.1%).
- Conversely, for each target, a matching multi-speaker file containing that speaker was identified at rank one or two, 80% percent of the time

OxfordWaveResearch

CONCLUSIONS

- We observe that the total duration of speech and the relative speaker mix for each target in a multi-speaker file are important for accurate recognition.
- Despite these challenges, this approach shows promise for automatically processing large volumes of real-world multi-speaker files.
- Automatic diarization or manual segmentation will provide higher accuracy results.
- The chunked approach provides an effective means for detection of speakers of interest from multi-speaker recordings without requiring manual segmentation or automatic diarization

OxfordWaveResearch

SPECIAL THANKS

University of Cambridge

Netherlands Forensic Institute

OxfordWaveResearch

REFERENCES

- A. Alexander, O. Forth, A. A. Atreya, and F. Kelly (2016). "VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features", *Odyssey 2016 Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016.
- D. L. Van der Vloed, J. S.Bouten, and D.A. Van Leeuwen (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments, *Proceedings of Odyssey Speaker and Language Recognition Workshop 2014*, Joensuu, Finland, pp. 6-13, 2014.
- N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P.Ouellet (2011). Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- F. Nolan, K. McDougall, G. de Jong & T. Hudson (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16: 31–57, 2009.

OxfordWaveResearch