Blind Speaker Clustering Using Phonetic and Spectral Features in Simulated and Realistic Police Interviews

ANIL ALEXANDER AND OSCAR FORTH

Oxford Wave Research Ltd, United Kingdom

{anil|oscar}@oxfordwaveresearch.com

IAFPA

International Association for Forensic Phonetics and Acoustics (IAFPA)

2012 Annual Conference

Santander, Spain

August 5th-8th, 2012

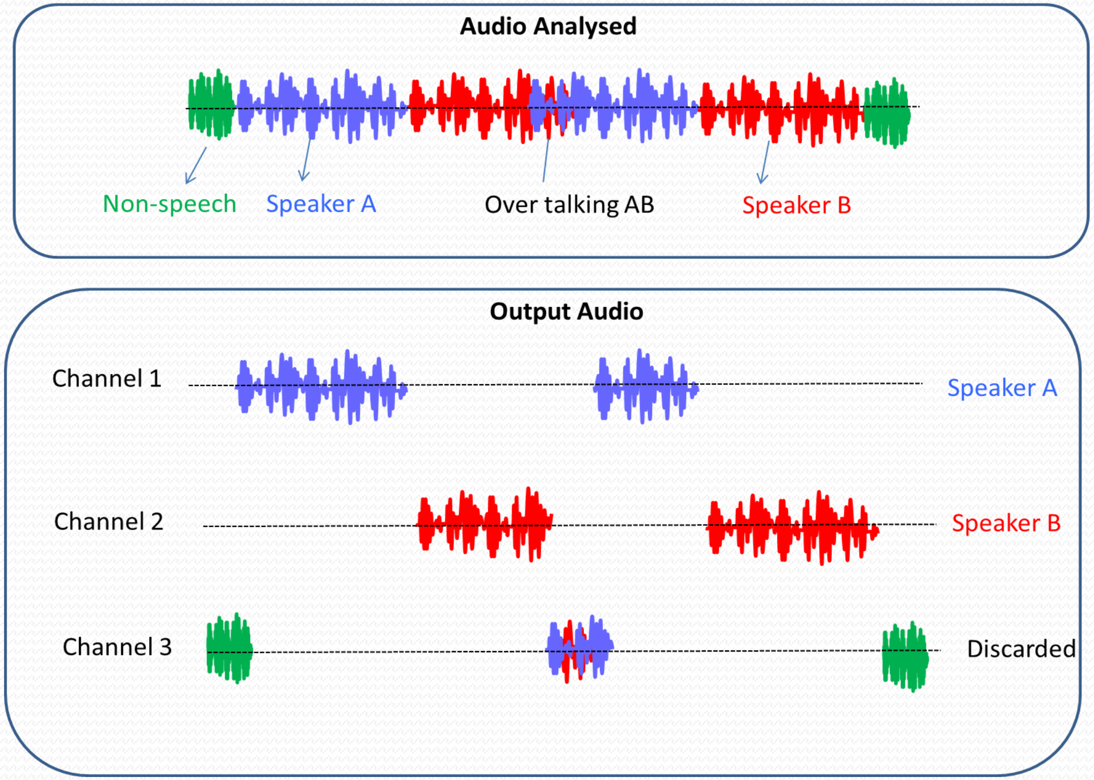
OxfordWaveResearch

Motivation

Benefits of automatically separating out the speech of individual speakers from a multi-speaker conversation

- Harvesting of the speech of a single speaker for use in phonetic speaker recognition
- Gleaning quick intelligence in surveillance recordings
- Phonetic research in areas such as long-term formant analysis and vocal profiling

Problem Overview



“Is it possible, with little or no user interaction, to separate out the good quality speech of individual speakers from a recording?”

OxfordWaveResearch

Prior Work

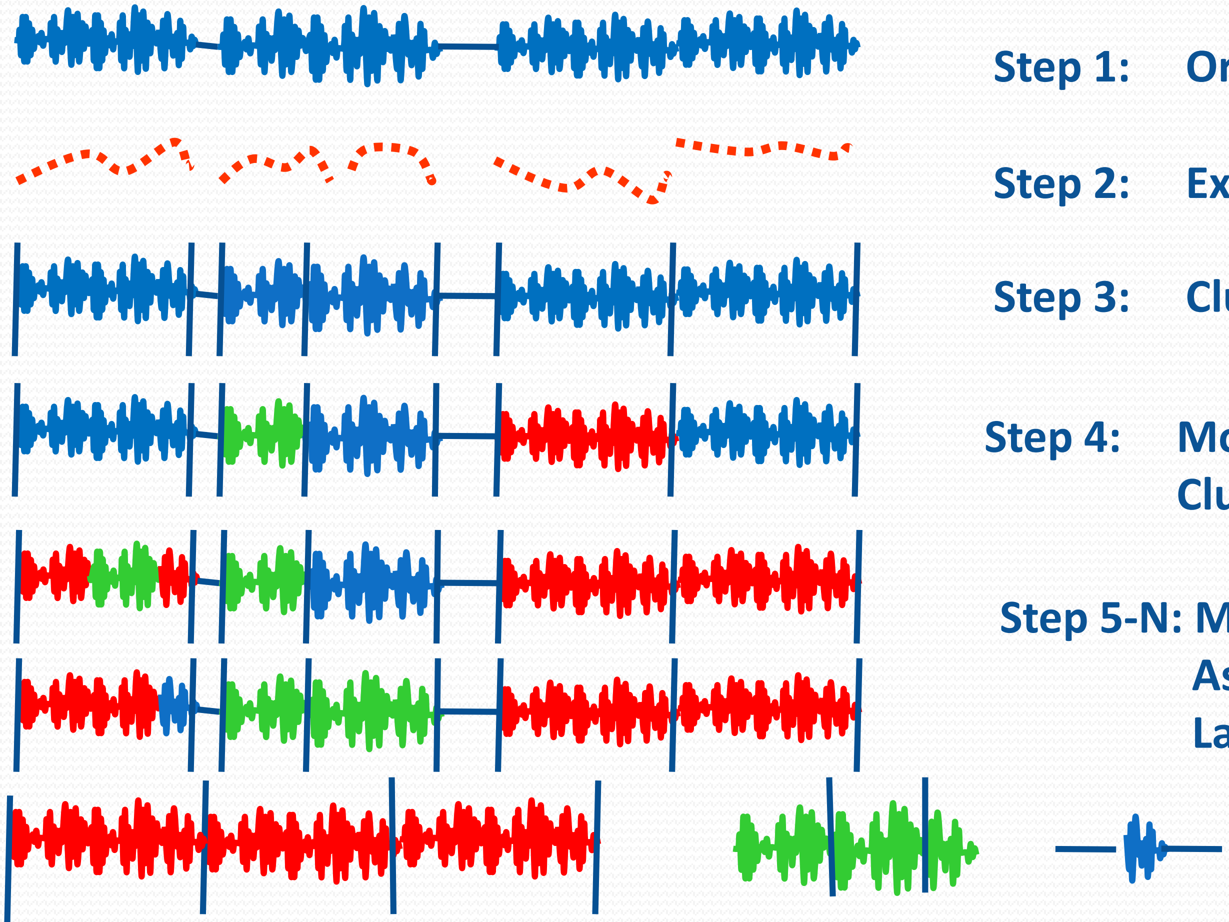
- Human-assisted automatic speaker diarisation applied to the disguises of the voices of vulnerable witnesses in police interview *[Alexander, Forth and How, IAFPA 2009, AES 2010]*
  - Provides a means of separating speakers from a police interview recording
  - Requires the user to provide training data (recommended duration 15-30s)
  - End goal is to preserve all speech data for each speaker (including over-talking, non-speech, etc.)
- Traditional speaker diarisation methods mainly use spectral features and do not consider phonetic measures

OxfordWaveResearch

Proposed Approach Overview (1/2)

- Two-tier approach using little or no user-interaction
  1. Clustering based on higher-level phonetic information
     - Continuous pitch track found to be a good indicator within an utterance of speaker identity
     - Adaptations for speakers of similar pitches
  2. Short-term spectral features like Mel Frequency Cepstral Coefficients (MFCCs)
     - Incorporating iterative agglomerative training and Gaussian mixture modelling to harvest features
     - Considering temporal information in MFCCs such as delta, delta-delta features

Proposed Approach Overview (2/2)



Step 1: Original Speech

Step 2: Extracted Pitch Track

Step 3: Clustering Performed

Step 4: Most Divergent Clusters Selected

Step 5-N: Most Similar Clusters Assigned Speaker Labels Iteratively

Final Assignments

Speaker A Speaker B Discarded

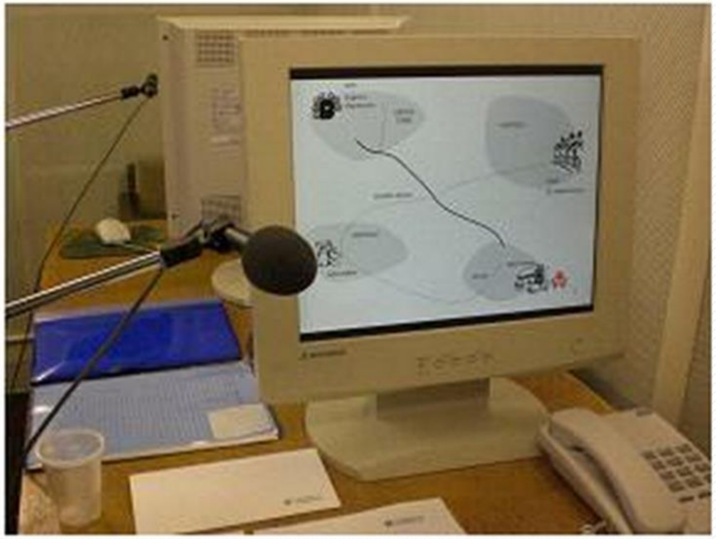
OxfordWaveResearch

Blind Speaker Clustering - Databases

- Two interview databases
  - Simulated police interview from the DyVIS database (Nolan et al 2009)
  - Realistic police interview data was recorded in a vulnerable witness interview room at a police station in London (RETAPE – Reduced Effort Transcription of Audio Product as Evidence)
- Recording quality: 16bit, 44,100 Hz uncompressed mono files in Microsoft WAV file format.

Simulated Police Interview (DyVIS)

- Subset of 20 speakers from the DyVIS database
- Simulated police interviews in Task 1.
- Task was developed ‘to elicit spontaneous speech in a situation of cognitive conflict’ (Nolan et al 2009).
- The DyVIS database is recorded in relatively ideal circumstances with high quality microphones, in a noise-controlled environment
- A provides a good approximation of a police interview, albeit one of unusually high quality.

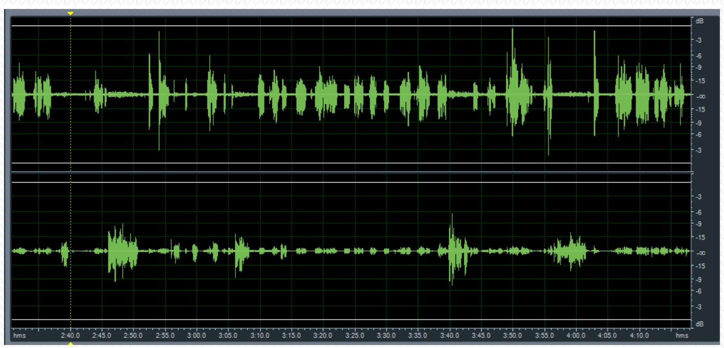


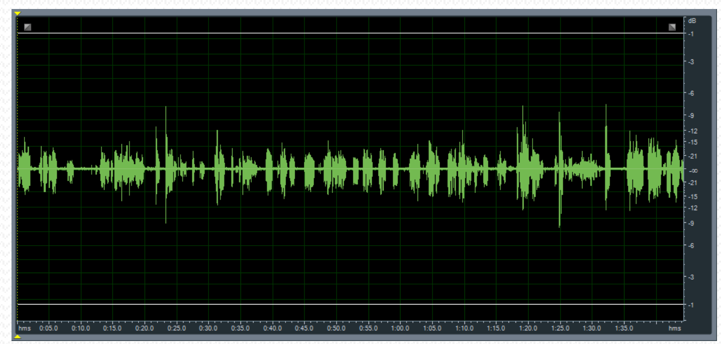
http://www.ling.cam.ac.uk/dyvis/database/experiment.png

OxfordWaveResearch

Simulated Police Interview - DyVIS

- **As the recordings were stereo recordings with each channel containing the speech of both speakers at different levels, we mixed the two channels (50% from each) into a single channel waveform.**





Realistic Police Database (RETAPE)

- Recorded in a vulnerable witness interview room at a police station in London.
- Small amount of electrical interference, overdriven audio and background noise present – considered representative of real-world conditions.
- The room was reasonably well sound-proofed with soft-furnishings, carpets and sofas.
- Test data used (Free speech -1 hour 33mins of simulated police interviews)
- Subjects
  - Police officers and staff (2 female + 1 male)
  - Member of public (1 male)
  - Children (1 female + 1 male)

Fig: Two views of the witness interview room

OxfordWaveResearch

Phonetic and Spectral Features

PHONETIC FEATURES

- Pitch F0
  - Relatively stable for each speaker within an utterance
  - Autocorrelation-based
- Assumptions
  - Voice onset time (80ms)
  - Maximum unvoiced region(500ms)
  - Decision criteria for clustering depends on the square of frequency difference and time difference

SPECTRAL FEATURES

- Mel Frequency Cepstral Coefficients – MFCCs (12)
- Delta features
- Delta-Delta features tried
- Energy Coefficients optionally used
- Frequency range considered (50-16,000 Hz)

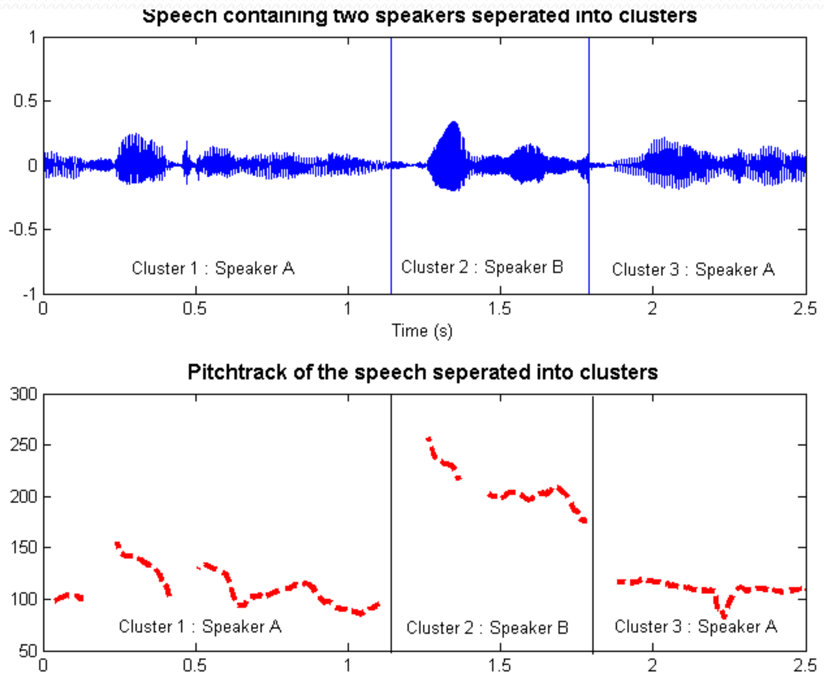
OxfordWaveResearch

Clustering

Pitch tracks for voiced segments are extracted using the autocorrelation-based pitch tracker in Praat (Boersma, 1993).

A continuous 'run' of similar values in the pitch track is used as 'zones of reliability' for the identity of a speaker.

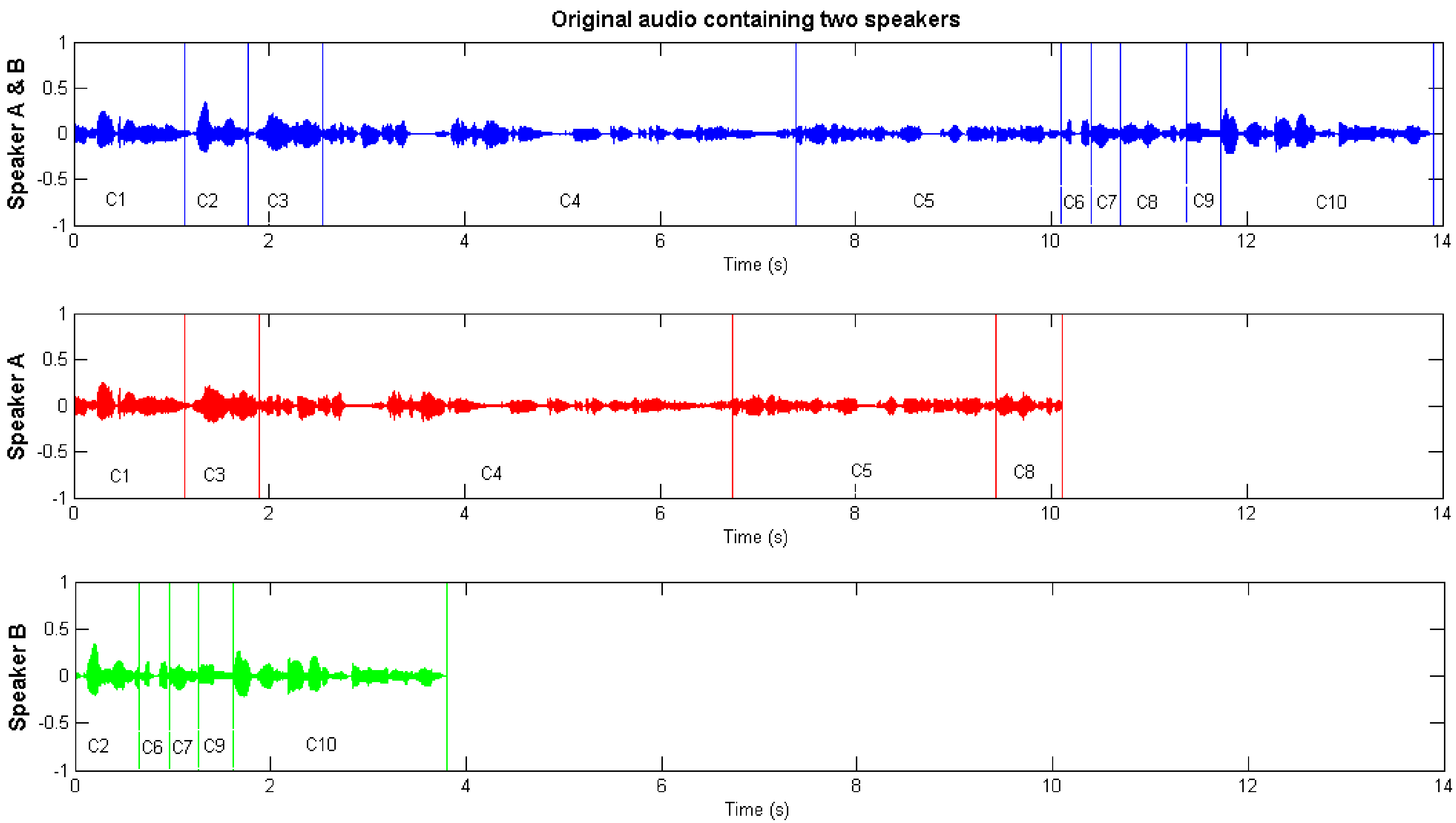
Any significant discontinuities either in time or frequency, is used to define a candidate transition point between speakers and a cluster .



DyVIS

OxfordWaveResearch

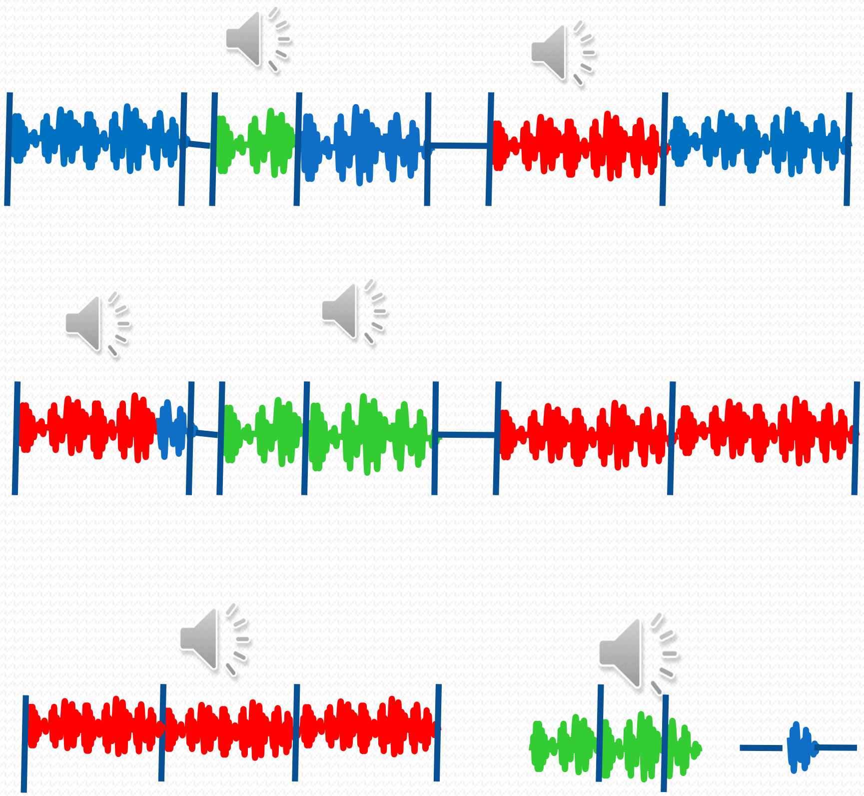
Clustering Results –DyVIS



OxfordWaveResearch

Algorithm

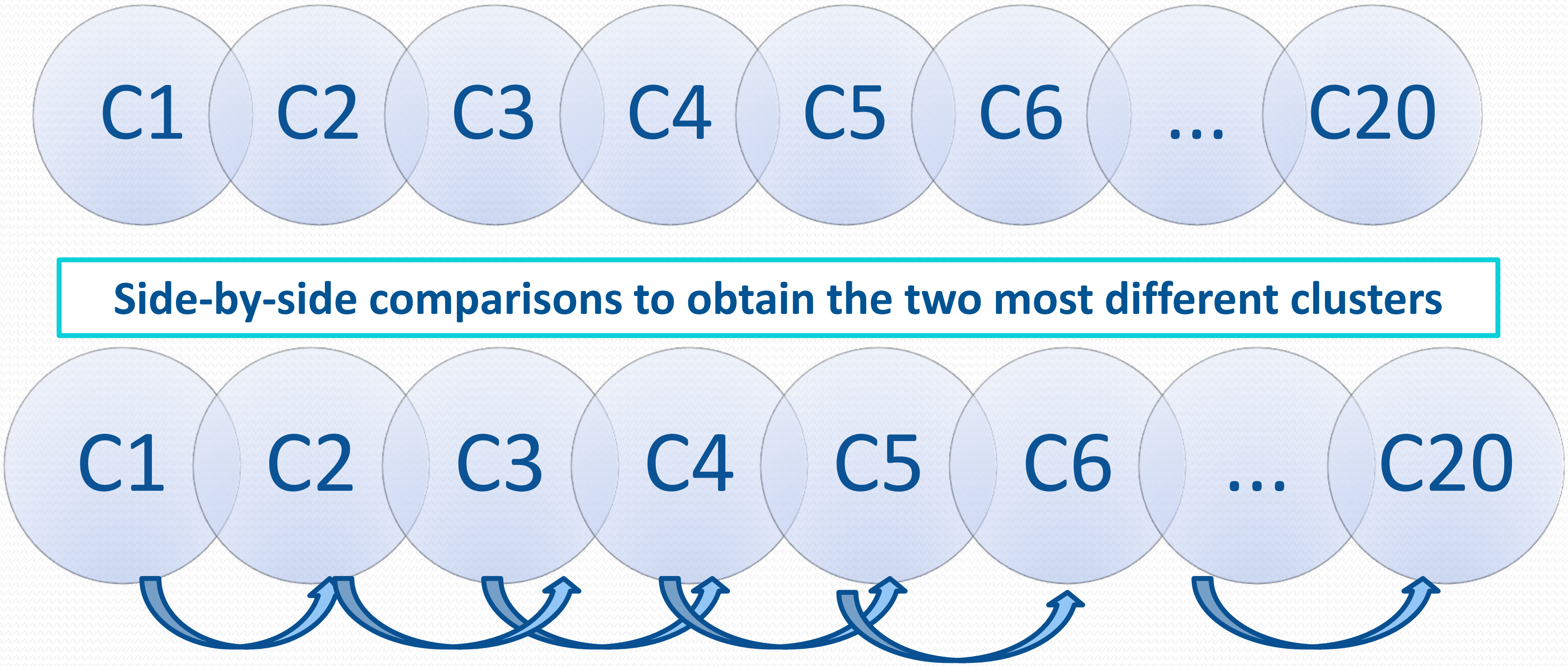
- Cluster creation based on pitch track discontinuities
- Selection of the two most divergent seed 'clusters' (can be ½s or 1s long)
- Agglomerative clustering
- Iterative improvement of the seed to obtain models for each speaker
- Harvesting of high-probability clusters for each speaker



Results

OxfordWaveResearch

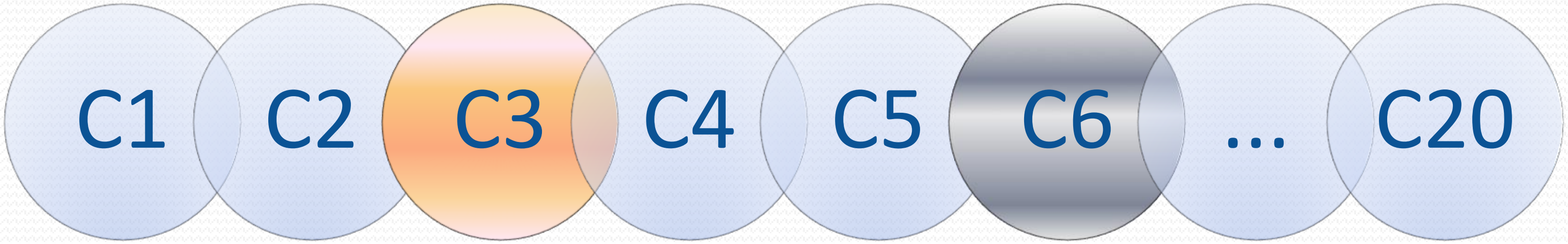
Clustering Algorithm



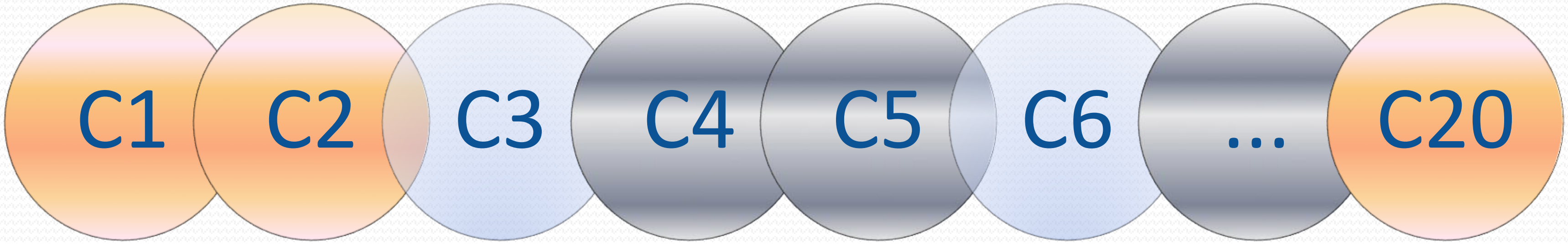
RETAPE

Clustering Algorithm

Once the pair of most different speech models are identified, they are used as training to collect the most similar clusters

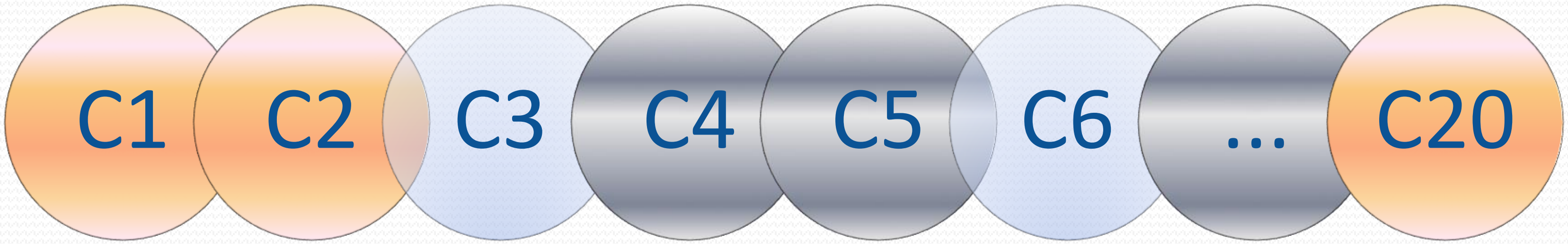


The longest clusters that are most similar to the seed clusters are then assigned to each speaker

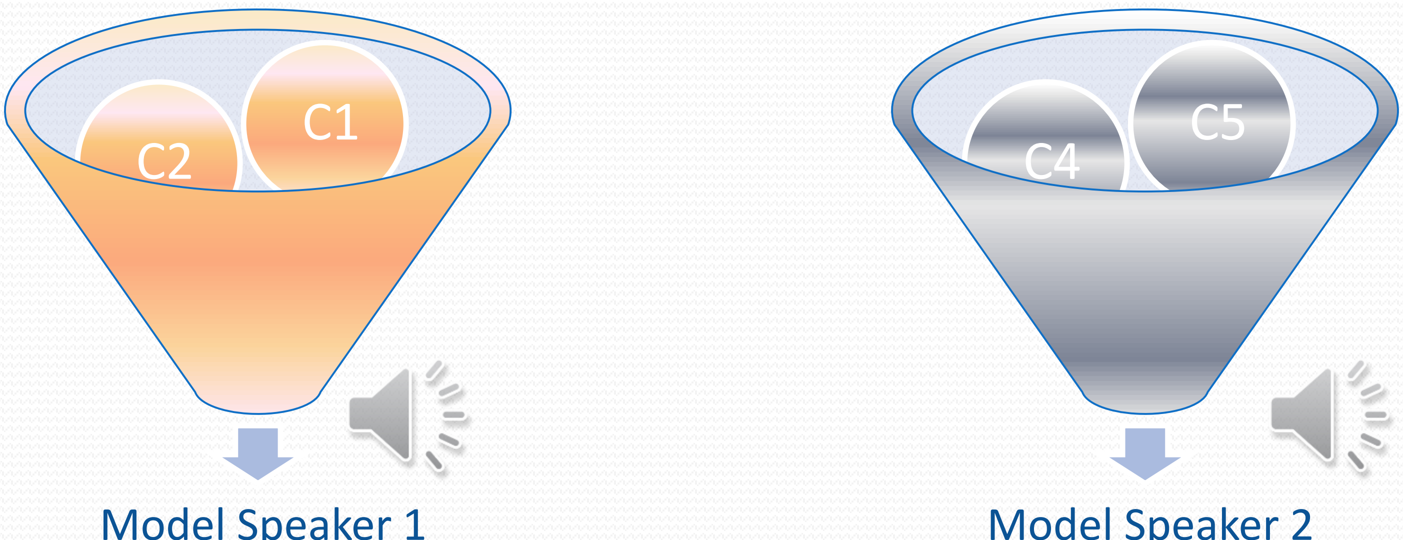


OxfordWaveResearch

Clustering Algorithm

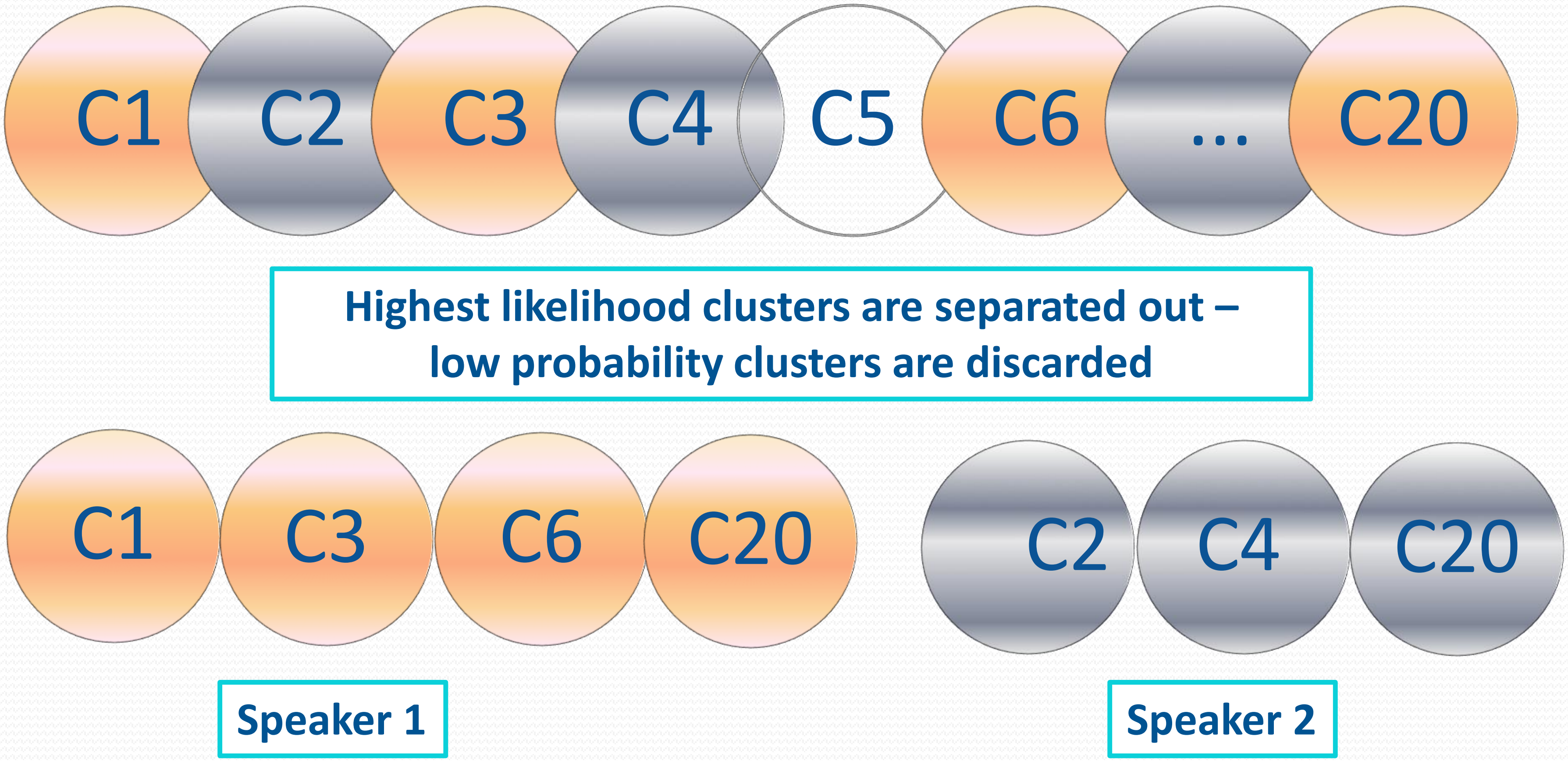


Clustering is done iteratively refining the models with each iteration



OxfordWaveResearch

Clustering Algorithm



Clustering Results – MET RETAPE

- Noisier recordings
- Different speaking styles within the recording (e.g. exclamations, excited speech)
- Same sex-speaker recordings were more difficult than different
- For some recordings, user interaction required to merge seeds of different kinds of speech

Similar-Sounding Speakers and Over-talking

- Similar pitch and accent
  - Galleon FBI wiretap
  - Two male speakers of Indian and Sri Lankan origin (Galleon group founder Raj Rajaratnam and Rajat Gupta – Goldman Sach's director)
  - Similar sounding speech and pitch ranges that are close to each other
- Over-talking

Practical Problems and Solutions (1/2)

- Similar voices more challenging – male/male or female/female
  - Only slight differences in pitch between speakers
  - Leads to each cluster of audio being 'impure'
  - Cluster purity is central to our method
  - More stringent criteria for splitting audio into clusters was required
    - More sensitive to Pitch track discontinuities and unvoiced gaps
- Spectral features now include temporal information
  - MFCCs with delta, delta-delta (derivatives) now used

Practical Problems and Solutions (2/2)

- Speed
  - Divergent cluster search limited to the largest clusters -> limits the number of comparisons in an order (N^2) calculation
  - Significant improvement in accuracy and seed generation
- Accuracy
  - Frame-based voting and winning-based cluster assignment
  - Refinement of models for each speaker run only using large and reliable clusters

Results With Both Databases

DyVIS: Simulated Police Interview Database

- Mainly only two speech ‘events’ in the recording
- Generally male-female speakers so clear distinction
- Majority of files could be processed using with no intermediate user interaction

MET: Realistic Police Interviews

- Noisier recordings
- Multiple speaking styles (exclamation, excited speech, etc.)
- No user-interaction required for some files
- Minimal user interaction required to merge seeds of different kinds of speech

Applications

- Easy extraction of only vowel data for a speaker – long-term formant analysis?
- Used as pre-processing for an automatic speaker recognition system
- Forensic phonetic research
- Quick intelligence gleaning from a recording

Conclusions

- Blind speaker separation approach is able to accurately extract the speech of individual speakers with minimal mislabelled speaker assignments.
- Approach shows reasonable robustness to noise and works well even with voices of speakers with close pitch ranges.
- Most challenging problem encountered was of over-talking between speakers.
- Capability of being able to collect quantities of the speech of individual speakers from a multi-speaker conversation would not only be of use to automatic speaker identification systems and phonetic analysis, but also in phonetic research in areas such as long-term formant analysis and vocal profiling.

References

- P. Boersma (1993), Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in Proc. of the Institute of Phonetic Sciences, University of Amsterdam, vol. 17, pp. 97 – 110.
- B. Narayanaswamy, R. Gangadharaiah and R. Stern (2006) Voting for Two-Speaker Segmentation, Proceedings of Interspeech, (ICSLP), pp. 2086–2089.
- Nolan, F., McDougall, K., de Jong, G., 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.
- A. Alexander, O. Forth and R. How (2010) 'Voice carving in Police Dialogue: Forensic application of Automatic Speaker Segmentation' in Proceedings of the AES 39th International Conference: Audio Forensics, Denmark
- J. M. Walsh, Y. E. Kim, and T. M. Doll (2007), Joint Iterative Multi-Speaker Identification and Source Separation using Expectation Propagation, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 2007, 283 – 286
- P. Delacourt and C. J. Wellekens (2000), DISTBIC: A speaker-based segmentation for audio data indexing, Speech Communication, vol. 32, 111-126.

Special Thanks

- London Metropolitan Police, for allowing us to use this data
  - Johanna Morley
- Department of Linguistics, University of Cambridge, for kindly providing data from the DyVIS database
  - Francis Nolan
  - Kirsty McDougall