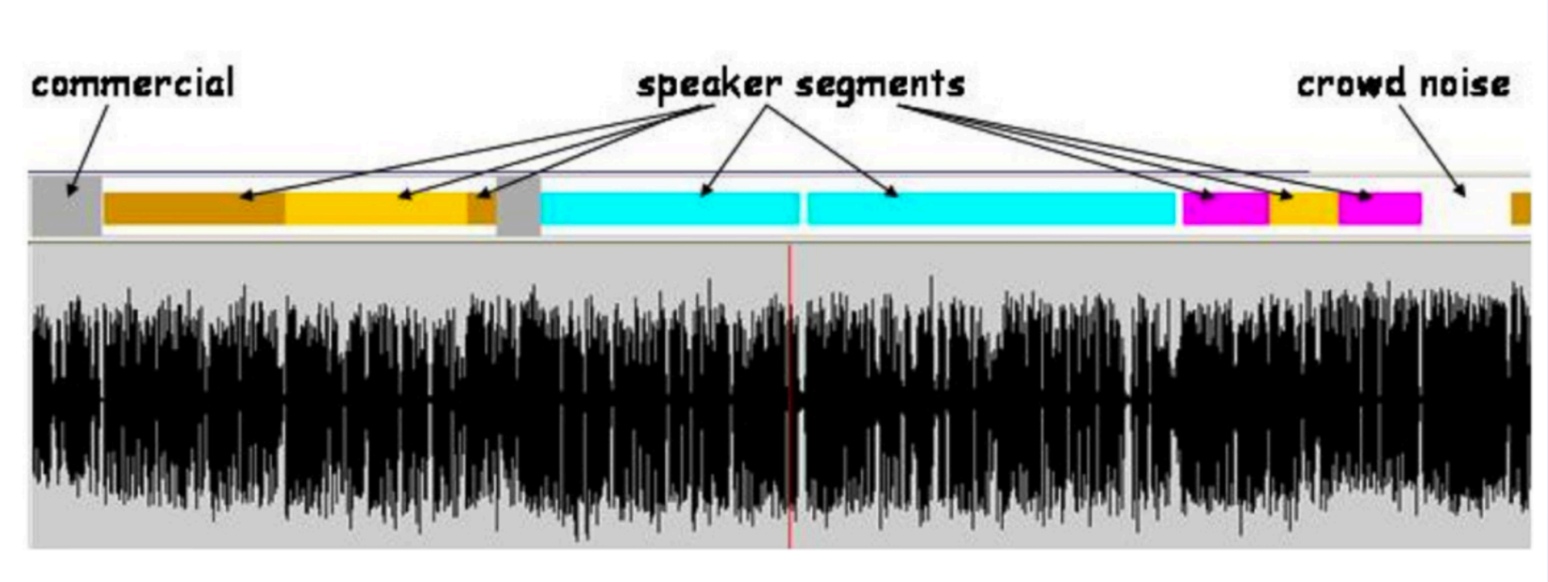
COM4511/COM6511 - Speech Technology

Lecture 16
Diarisation



Thomas Hain
t.hain@sheffield.ac.uk
Spring Semester

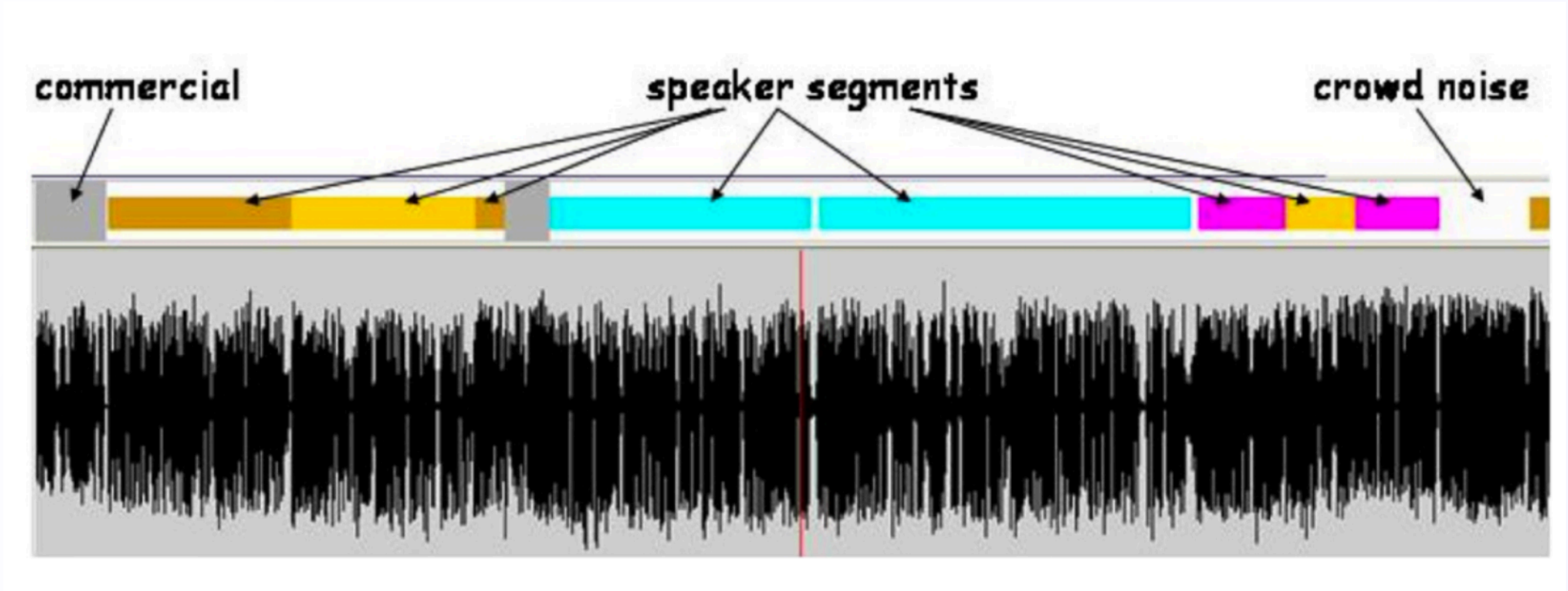
Terms

- Speaker identification – determine which of the set of enrolled speakers a test speaker matches
- Speaker verification – determine if a test speaker matches a specific speaker
- Speaker diarization – "who spoke when" segment and label a continuous recording by speaker

Dealing with multiple speakers

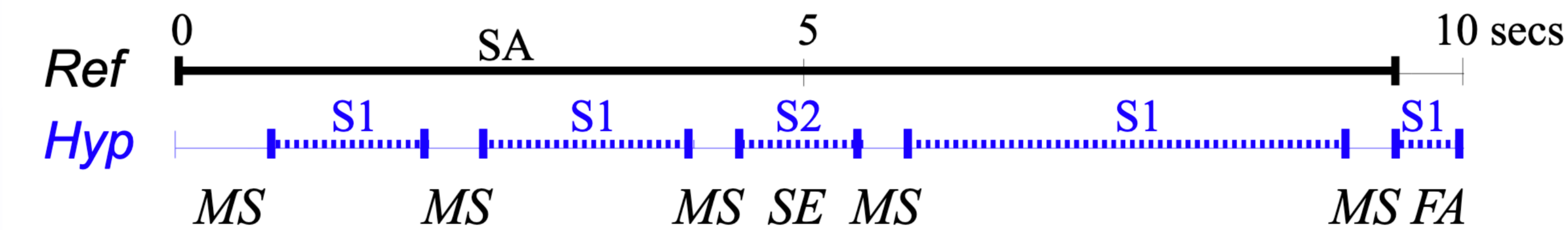
- Speaker diarization is the "who spoken when" task: given a recording, divide it into segments, where each segment corresponds to speech of a single speaker
- Each recording contains multiple speakers – unlike what we have assumed so far for speech recognition and speaker verification
- Multiple speakers in a recording is realistic – many possible domains, e.g.:
  - Broadcast media
  - Telephone conversations
  - Call centres
  - Meeting recordings

A basic system



- A basic approach to diarization:
  - Segment the recording into a sequence of short pieces, each assumed to be a single speaker.
  - Then treat as a speaker verification task between all pairs of segmented utterances
- Guaranteed to fail on segments with overlapping speakers!

Measuring speaker diarization – Diarization error rate



- There are three main type of error to consider in speaker diarization:
  - Missed speech (E_{miss}): system labels a segment as non-speech, but segment is attributed to a speaker in the reference
  - False-alarm speech (E_{fa}): system attributes segment to a speaker, but segment is labelled as non-speech in the reference
  - Speaker error (E_{spkr}): system attributes segment to a speaker different to the reference attribution
- These errors are computed in a time-based way: each is expressed as a fraction of the scored time in the reference
- The diarization error rate (DER) is computed as a sum of these errors

$$DER = E_{miss} + E_{fa} + E_{spkr}$$

Note that E_{miss} and E_{fa} arise from the speech activity detection

Segmental purity metrics

- Cluster purity, $p_{i.}$, of cluster i and the average cluster purity, acp, are:

$$p_{i.} = \sum_{j=1}^{N_s} \frac{n_{ij}^2}{n_{i.}^2}, \quad acp = \frac{1}{N} \sum_{i=1}^{N_c} p_{i.} n_{i.} \tag{1}$$

  $n_{i.}$ is the number of frames in cluster i, $n_{.j}$ is the number of frames uttered by speaker j, n_{ij} is the frame count in cluster i spoken by speaker j, N_c is the cluster count, N_s is the number of speakers and N is the number of frames.

- Speaker purity, $p_{.j}$, of speaker j and average speaker purity, asp, are:

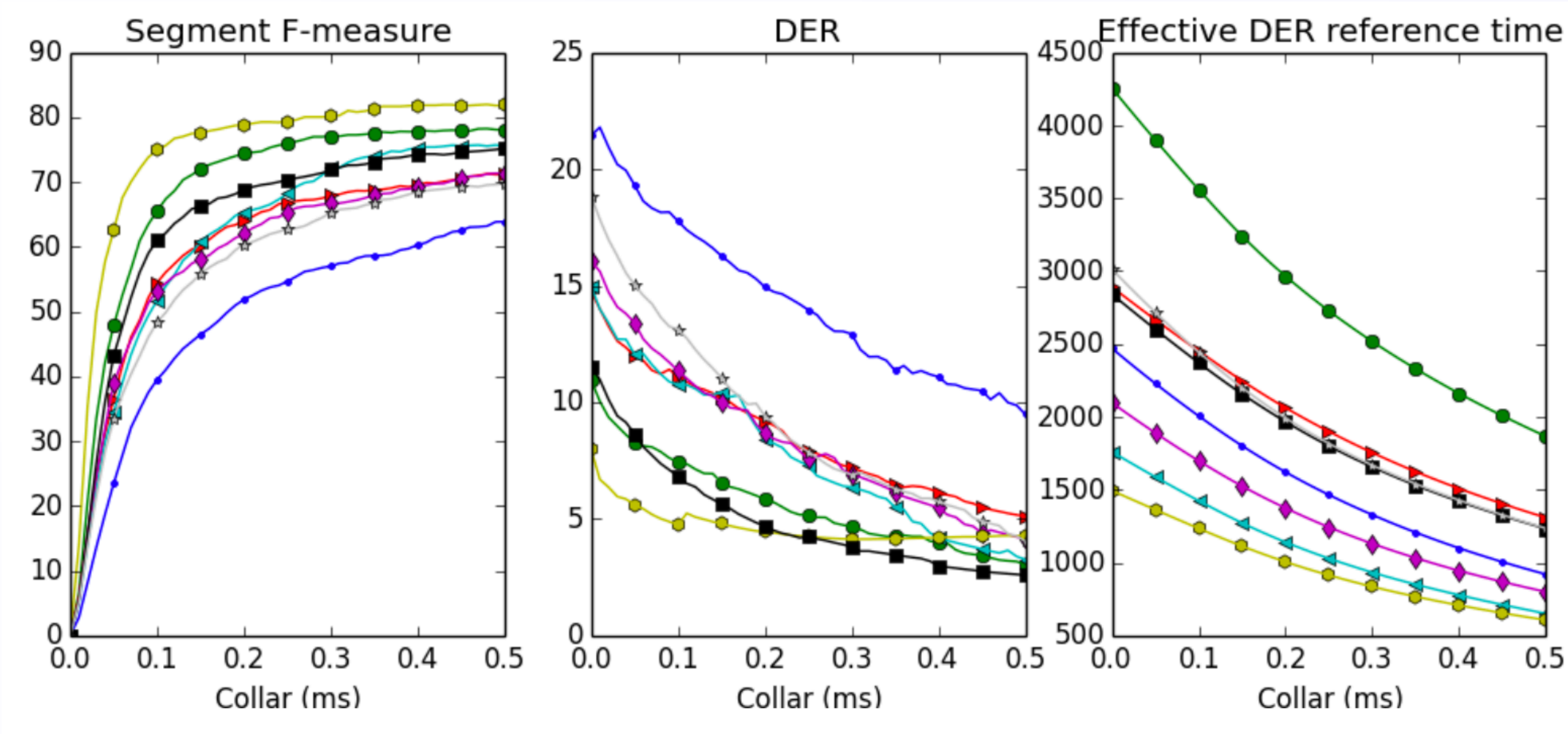
$$p_{.j} = \sum_{i=1}^{N_c} \frac{n_{ij}^2}{n_{.j}^2}, \quad asp = \frac{1}{N} \sum_{j=1}^{N_s} p_{.j} n_{.j} \tag{2}$$

- An overall purity calculation combines both cluster and speaker purity measures:

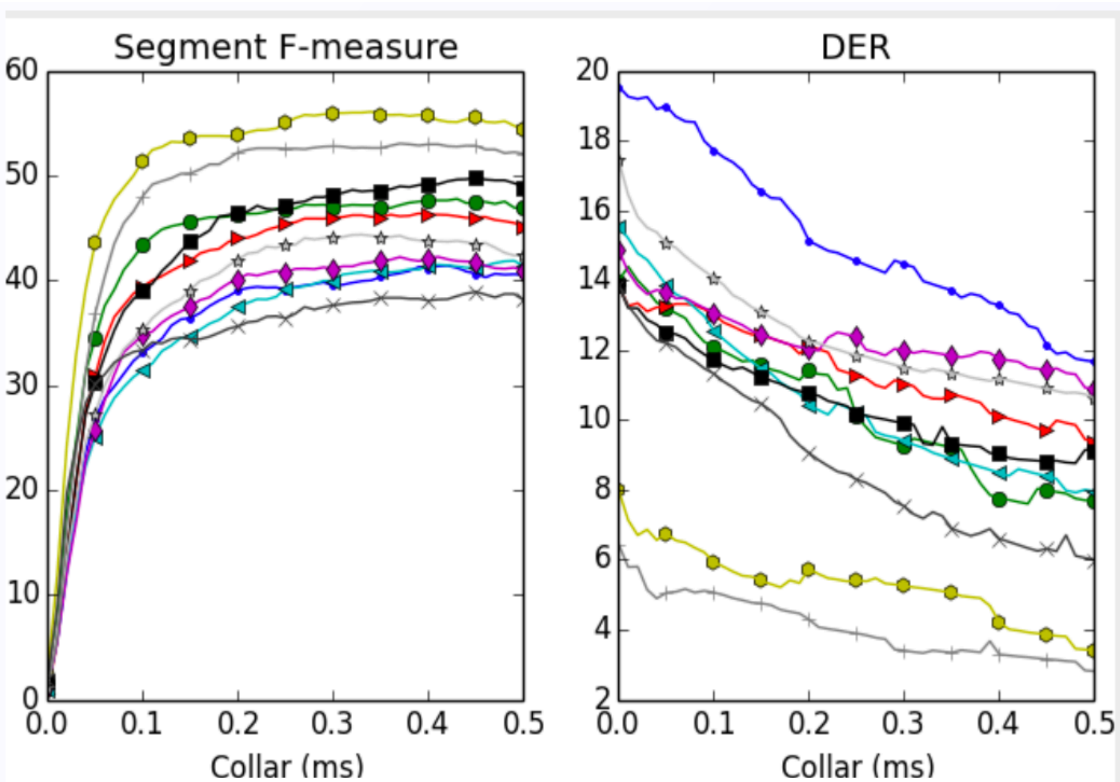
$$K = \sqrt{acp * asp} \tag{3}$$

Typical distribution of errors

Meetings

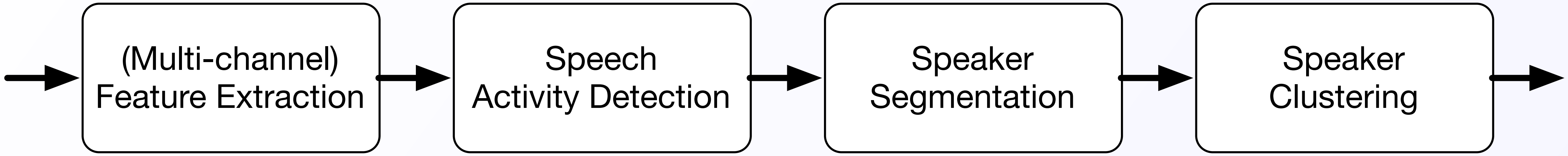


Media



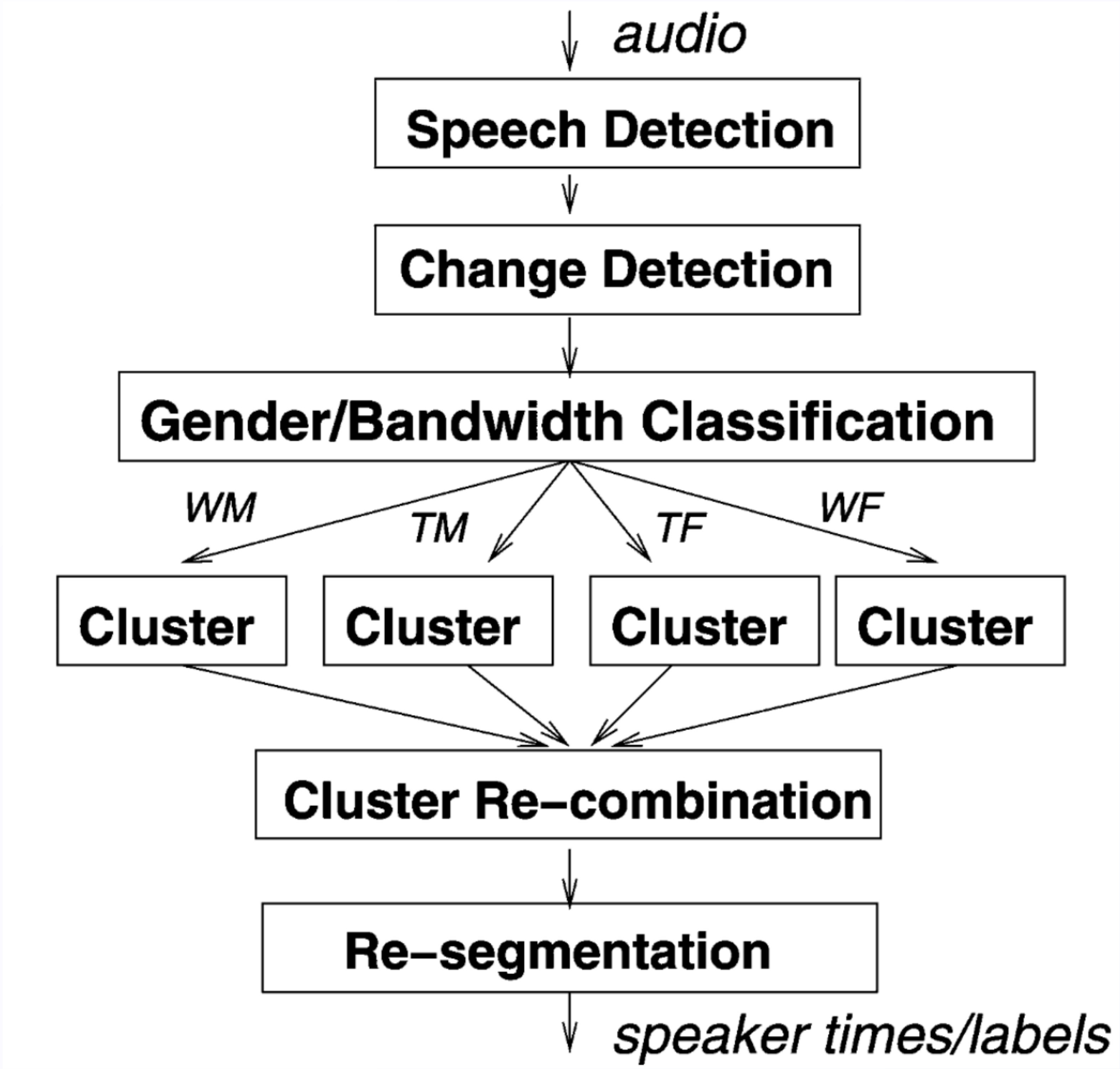
Collar is area around boundaries which is ignored in scoring

The tasks for diarisation

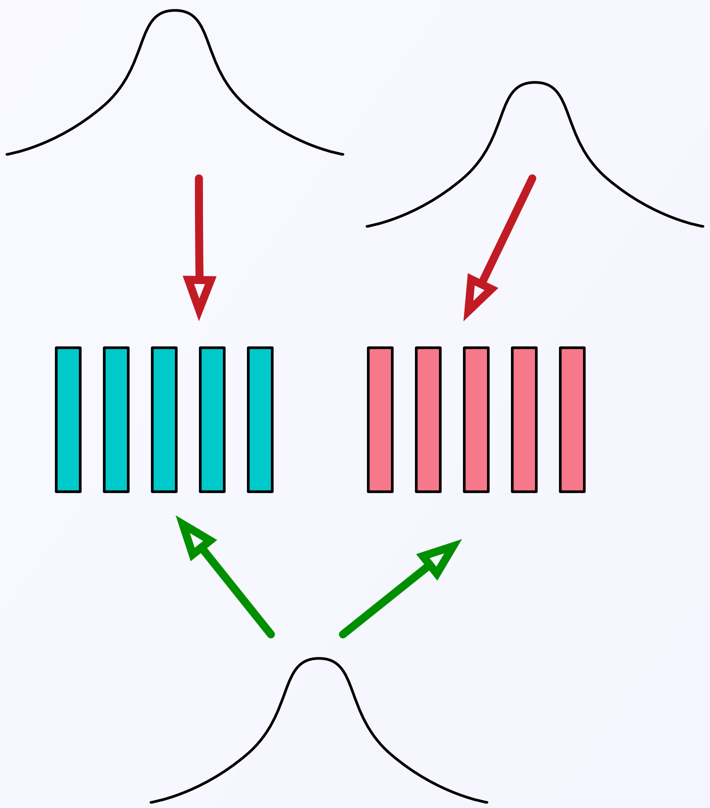


- Feature Extraction
  - Acoustic and location !
- Speech Activity Detection (SAD)
  - no speaker separation
- Speaker segmentation
  - timing information
- Speaker Clustering
  - Cope with small clusters and unknown cluster numbers

Tranter & Reynolds 2006



Classical - Bayesian Information Criterion

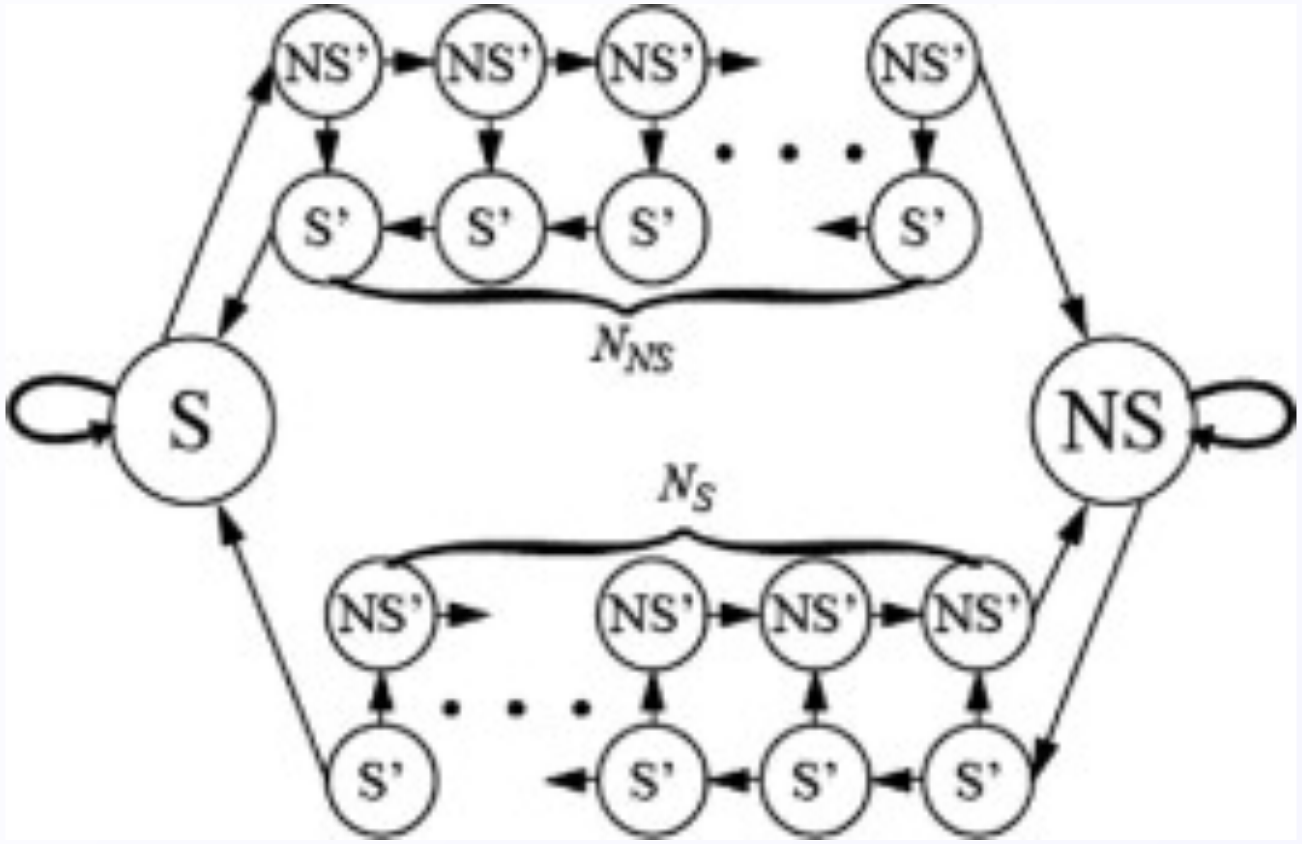


- Question: How to find out if data stems from two sources or one.
  - Compare using one model ore two to describe the data
- Issue two models == more parameters
  - compensate using number of parameters
  - k log(n) - k is number of parameters, n is number of data points

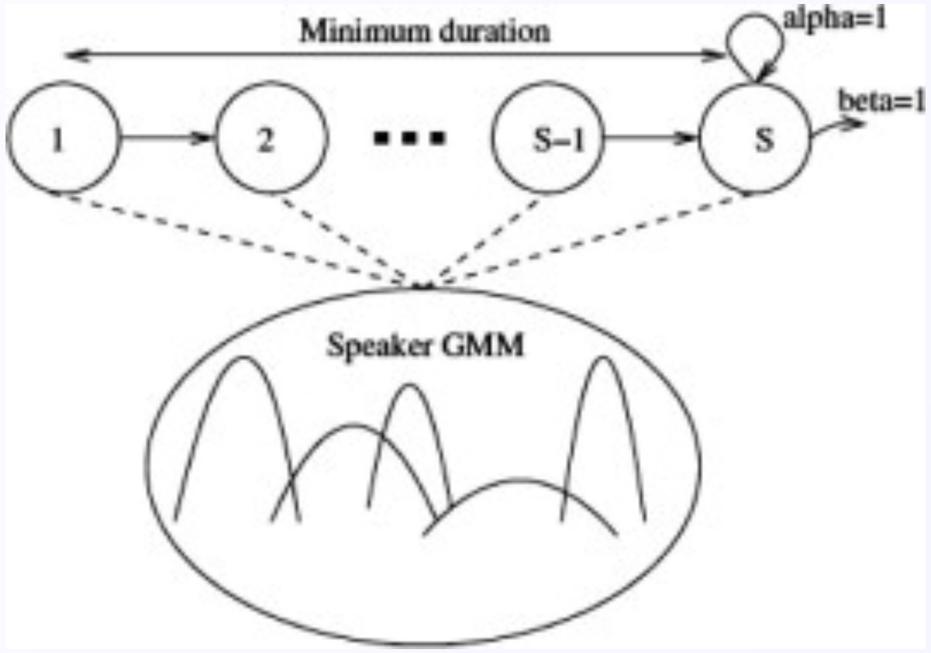
$$\Delta BIC = \frac{1}{2}[n_z \log(|\Sigma_z|) - n_x \log(|\Sigma_x|) - n_y \log(|\Sigma_y|)] - \lambda\left(\frac{d(d+3)}{4}\right) \log n_z$$

The ICSI-SRI approach

SAD

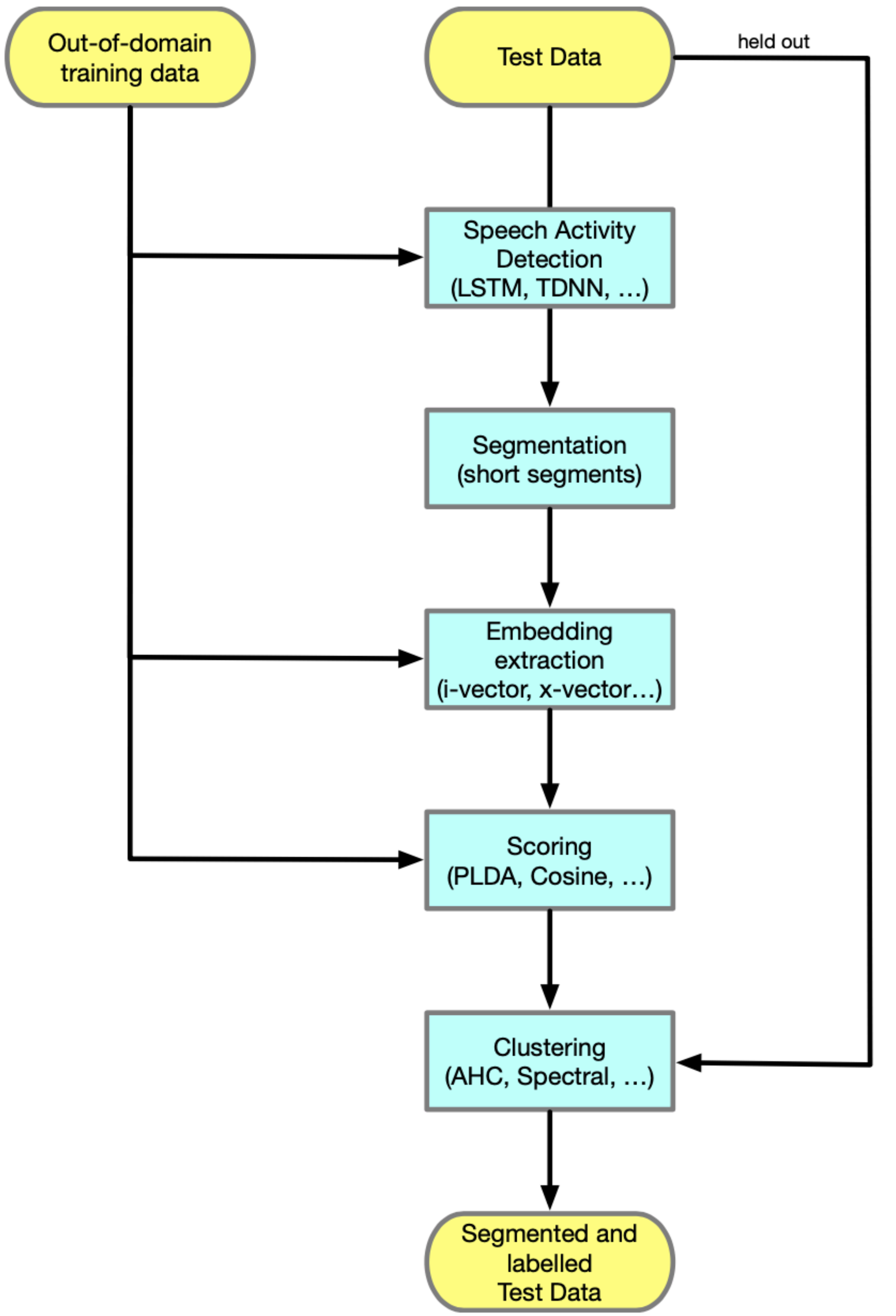


Clustering



Training and relabelling iteratively

Current Framework for Speaker Diarisation



Segment a recording, and attach a speaker label to each segment.

1. Split the recording into segments
2. Speech activity detection: identify whether each segment is speech or non-speech, discard non-speech
3. Represent the speech segments using some form of fixed length embedding: i-vector, x-vector, d-vector...
4. Compare all pairs of segments using a scoring metric such as PLDA
5. Cluster the segments using an algorithm such as agglomerative hierarchical clustering

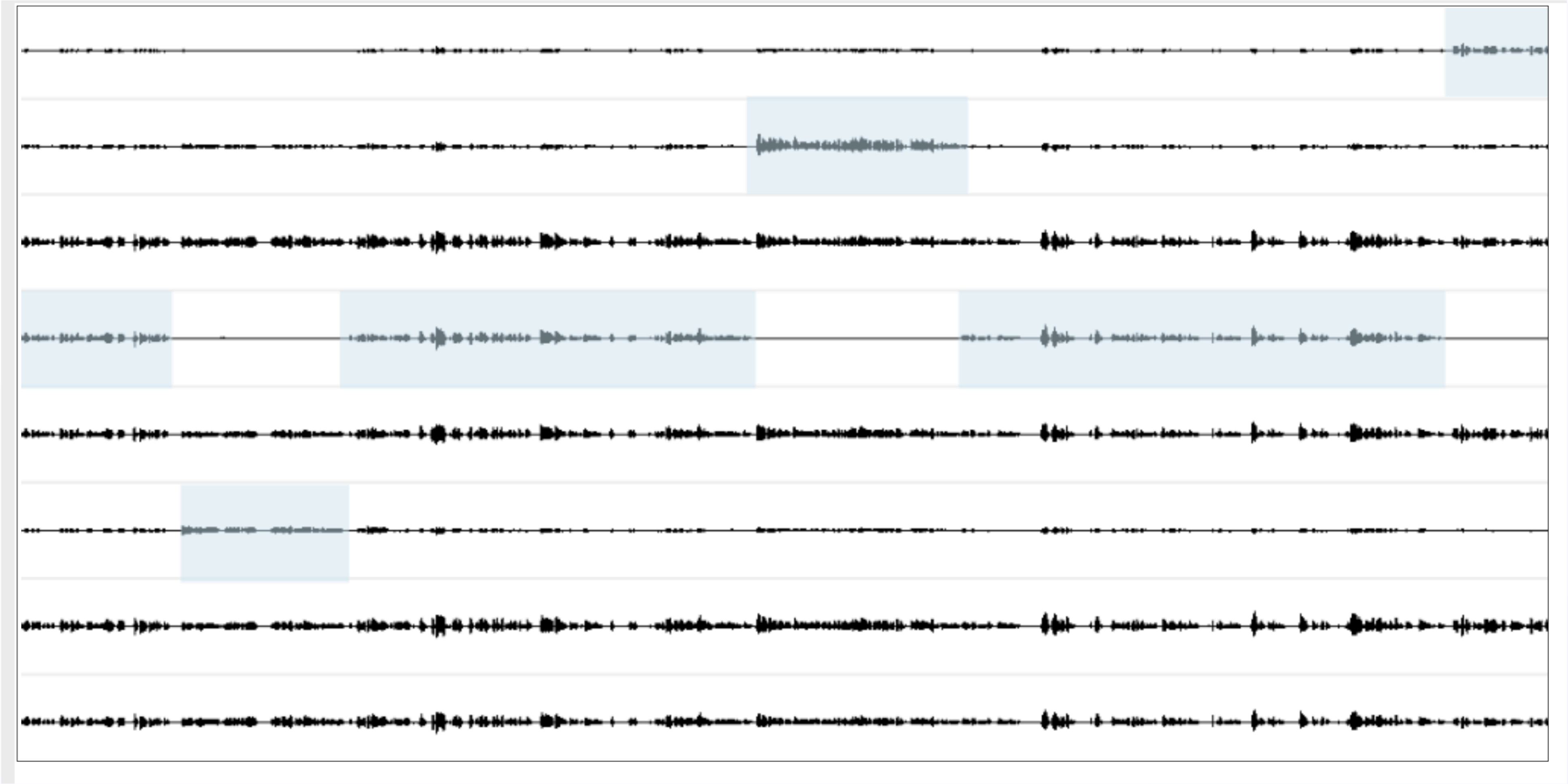
Segmentation and Speech Activity Detection

- Speech activity detection (SAD) typically carried out using an LSTM or TDNN neural network trained on a large amount of diverse data
  - Binary output: speech vs. non-speech
    Possibly with data augmentation – noise, reverb, etc.
- Following SAD, segment into short fixed-length segments (typically 2s) Assumes each segment contains speech from a single speaker
  - In practice can use overlapping segments (overlap by 0.5s at start and end) Relatively short segment duration for embedding computation

Speaker Embeddings and Clustering

- Compute a speaker representation for each segment
  i-vector - typically 64-128 dimension
  x-vector / d-vector - typically 128-256 dimension
  can reduce the dimension by performing PCA on the set of embeddings for a recording
- Score all segment pairs – typically use PLDA
  Cluster segments – many possible clustering algorithms: Agglomerative hierarchical clustering can work well
- Only need to compute pairwise segment scores once
  Score for a cluster pair is obtained by averaging the pairwise scores between the segments in each cluster
- Determine the number of clusters
  Clustering stopping criterion determines the number of clusters
  Define a prior distribution on the number of speakers, and apply to clustering Bayesian models with a prior on number of clusters – Variational Bayes (VB) HMM, Hierarchical Dirichlet Process (HDP) HMM, distance-dependent Chinese Restaurant Process (ddCRP), . . .

Multi-channel diarisation - crosstalk



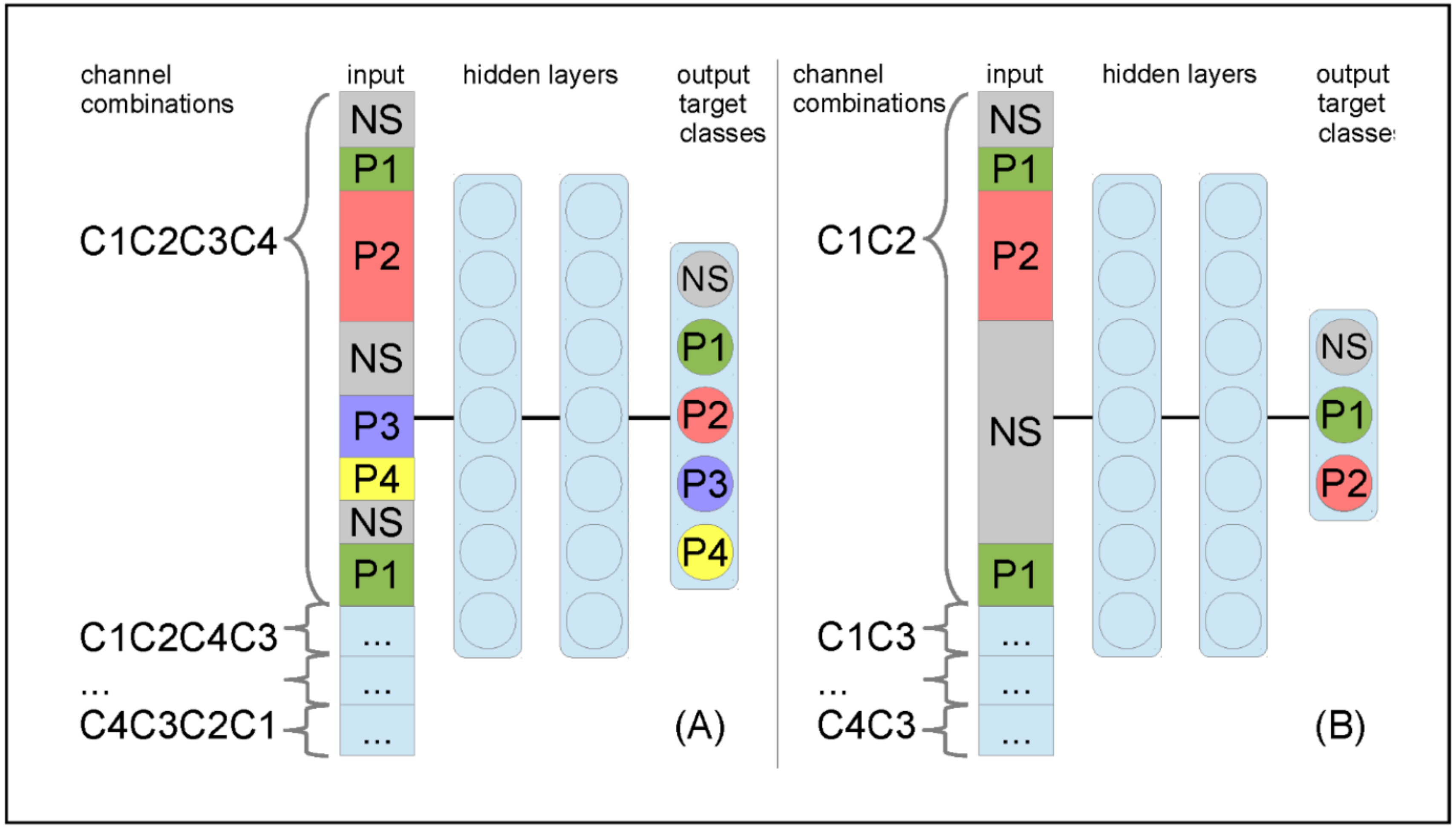
Scoring	Channel	#Segs	#Spkrs	DER%
Data: TBL				
NIST	SDM	2030	82	16.6
	IHM	8478	40	393.9
SHEF	SDM	2030	82	27.8
	IHM	8478	40	335.9

Solutions: Cross-talk features or combined modelling

Multi-channel diarisation - multi-channel models

Input are permutations of all channels (A)

Also pairwise option(B)



DNN			#Segs	MS%	FA%	SE%	DER%
TRN	OV	CT					
Data: TBL							
TBL	x		6732	4.3	2.4	1.2	8.0
TBL	x	x	7136	4.3	2.4	1.7	8.4
TBL			7269	4.3	2.5	1.5	8.3
TBL		x	2964	4.6	3.7	1.4	9.7

Rosanna Milner, Thomas Hain (2017). <u>DNN approach to speaker diarisation using speaker channels</u>, in Proc ICASSP 2017.

DIHARD

- R&D in speaker diarization has been very domain-dependent
  - 1990s – broadcast news (Hub4)
  - 2000s – multi-microphone meeting recordings (AMI, NIST RT)
  - 2010s – conversational telephone speech (Switchboard)
  - 2015 – general media
- Had the effect of fragmenting the field
- Since 2018 the DIHARD Challenge (https://coml.lscp.ens.fr/dihard/) has focused on “speaker diarization for challenging recordings where there is an expectation that the current state-of-the-art will fare poorly” – diverse set of data sets used

Some hot topics in diarization

- Overlapping speech – most systems do not explicitly deal with this
- Speech activity detection is still a significant cause of error
- Development of end-to-end systems
- Bayesian approaches (learning the number of speakers/ clusters from the data)
  - Dirichlet process GMMs
- Use of supervised learning

Some references

- D Garcia-Romero et al (2017), "Speaker diarization using deep neural network embeddings", ICASSP. https://ieeexplore.ieee.org/document/7953094
- G Sell et al (2018), "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge", Interspeech. https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1893.html
- K Church et al (2017), "Speaker diarization: A perspective on challenges and opportunities from theory to practice", ICASSP. https://ieeexplore.ieee.org/abstract/document/7953098