Self-Tuning Spectral Clustering for Speaker Diarization

Nikhil Raghav[1,3], Avisek Gupta[1], Md Sahidullah[1,3], Swagatam Das[1,2,4]

[1]Institute for Advancing Intelligence, TCG CREST, India

[2]Academy of Scientific and Innovative Research (AcSIR), India

[3]Department of Computer Science, RKMVERI, India

[4]Electronics and Communication Sciences Unit, Indian Statistical Institute, India

Unsupervised Speaker Diarization

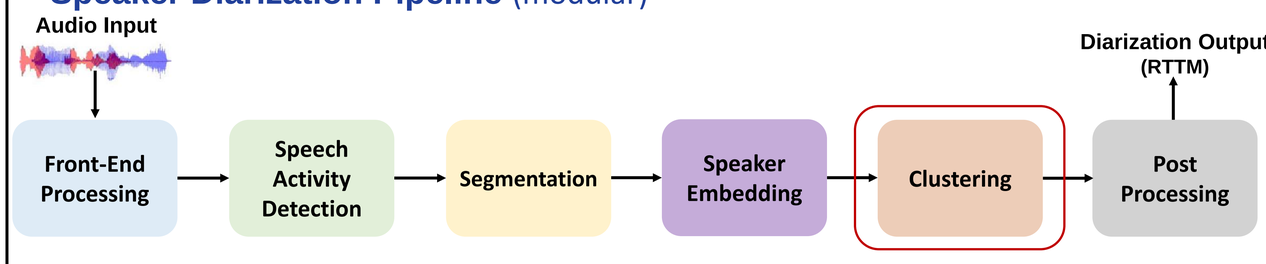
Speaker Diarization answers ***"who spoke when"*** in a multi-speaker audio conversation [1].

Applications Meeting transcription, cocktail party problem, forensics, etc.

Challenges

- ❑ Unknown number of speakers, Overlapping speech, Unbalanced speaker time
- ❑ Acoustic mismatch, Speaker variability, Scenario, etc.

Speaker Diarization Pipeline (modular)



Spectral Clustering Based Speaker diarization

Spectral clustering has proven effective in grouping speech representations for speaker diarization. Post-processing the affinity matrix remains difficult due to the need for careful tuning before the Laplacian.

Semi-supervised: Conventional spectral clustering (CSC)[2]

- ❑ It requires the Dev set to tune the pruning threshold
- ❑ Prunes a fixed number of elements in each row, equal to $\lfloor n(1-\alpha) \rfloor$

Unsupervised: Auto-tuning Spectral Clustering (ASC)[3]

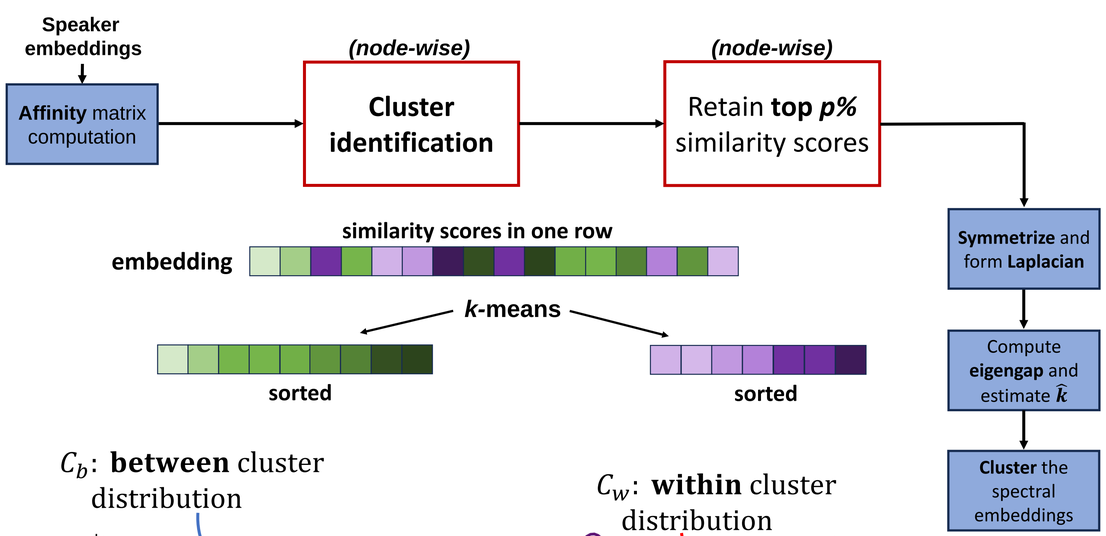
- ❑ It auto-tunes clustering parameters using normalized maximum eigengap (NME) values
- ❑ Eliminate the need for parameter tuning over the Dev set.
- ❑ Prunes a fixed number of elements from in each row.
- ❑ Its complexity is $O(Pn^3) = O(n^4)$,
  where $P = \lfloor n/4 \rfloor$ is the recommended number of binarized similarity matrices.

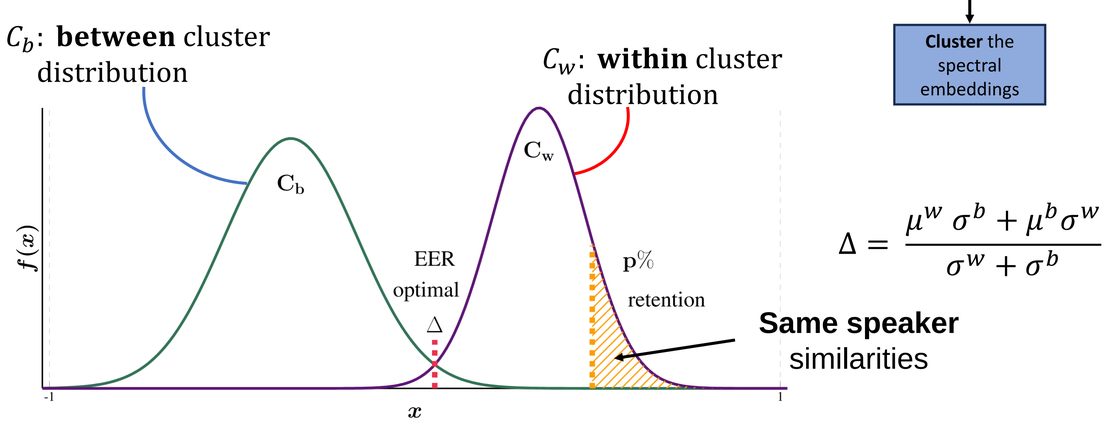
Proposed Methodology

Unsupervised: Spectral clustering with adaptive neighbours

- ❑ Speaker similarities in terms of two gaussian distributions is exploited using *k*-means clustering.
- ❑ A sparse affinity matrix called spectral clustering on p-neighbourhood retained affinity matrix (SC-pNA) is created.
- ❑ Initially, we chose EER-Δ as the pruning threshold.
- ❑ EER-Δ points where the false rejection rate and the false acceptance rate coincide, inspired from the equal error rate (EER).
- ❑ Later, we chose a more aggressive threshold over EER- Δ, by retaining top p% within speaker similarities.
- ❑ p = 20% is identified as the best selection for p.
- ❑ The complexity of SC-pNA is $O(n^3)$, more efficient than ASC.

SC-pNA Algorithm





$$\Delta = \frac{\mu^w \sigma^b + \mu^b \sigma^w}{\sigma^w + \sigma^b}$$

Experimental Setup

Dataset

DIHARD-III. It comprise data from 11 diverse domains.

Metric

Diarization error rate (DER) = $\frac{\textbf{Missed speech} + \textbf{False alarm} + \textbf{Speaker error}}{\textbf{Total speaking time}}$

SD system

SpeechBrain: an open source speech toolkit.

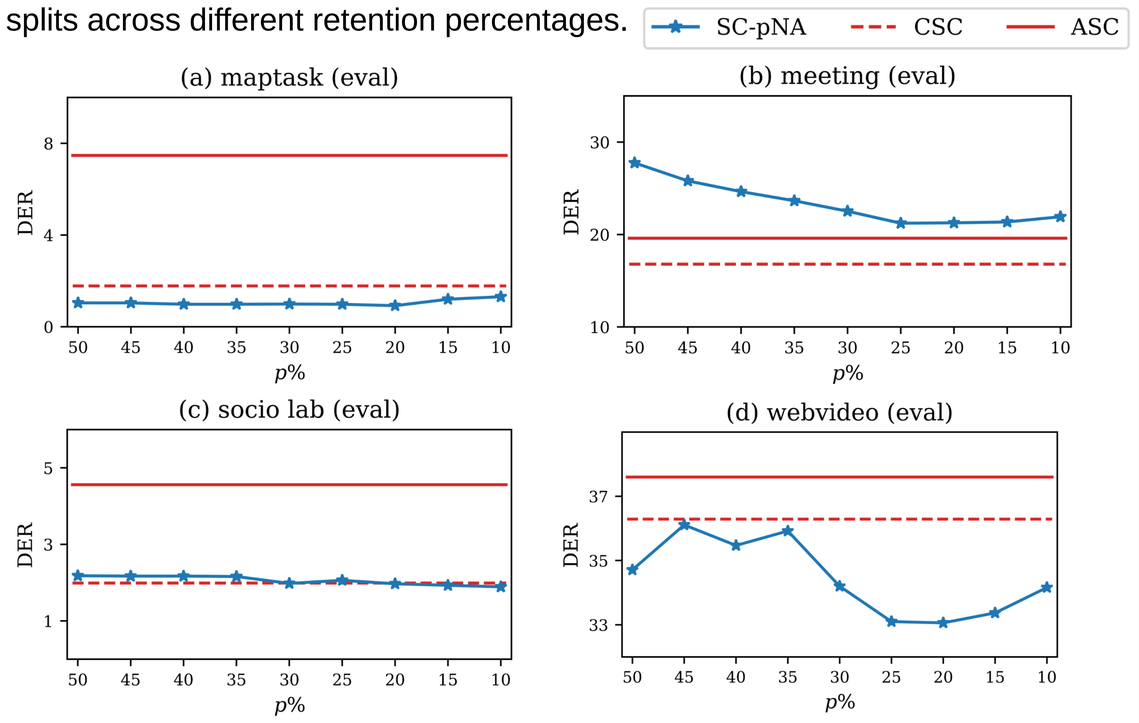
Results and Discussion

Speaker diarization system performance in terms of DER **(lower is better).** The semi-supervised CSC acts as a baseline along with unsupervised ASC, EER-Δ and the SC-pNA over DIHARD-III dataset Eval split.

	Semi-supervised (baseline)	Unsupervised		
Domain	**CSC**	**ASC**	**EER-Δ**	**SC-pNA**
broadcast interview	3.58	**3.67**	6.82	4.77
court	2.09	**2.73**	17.51	7.15
cts	6.58	7.3	12.66	**6.63**
maptask	1.78	7.47	5.25	**0.92**
meeting	16.79	**19.60**	40.84	21.26
socio lab	1.99	4.56	8.66	**1.97**
webvideo	36.29	37.60	36.52	**33.06**
restaurant	29.93	**33.35**	59.78	38.81
Audiobooks	0.50	27.53	0.18	**0.09**
clinical	4.34	10.62	31.23	**3.31**
socio field	5.18	10.30	16.57	**3.79**
Overall	9.97	13.11	20.08	**10.27**

Hyper-parameter selection

The variation in the speaker diarization performance in terms of DER across four data splits across different retention percentages.



Acknowledgements

The primary author expresses sincere gratitude to the Linguistic Data Consortium (LDC) for the LDC Data Scholarship, which enabled access to DIHARD-III dataset.

References

1. Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., Na, H. (2021) ECAPA-TDNN Embeddings for Speaker Diarization. Proc. Interspeech 2021.
2. Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, Shrikanth Narayanan, A review of speaker diarization: Recent advances with deep learning, Computer Speech & Language, 2022.
3. T. J. Park, K. J. Han, M. Kumar and S. Narayanan, "Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap," in IEEE Signal Processing Letters.