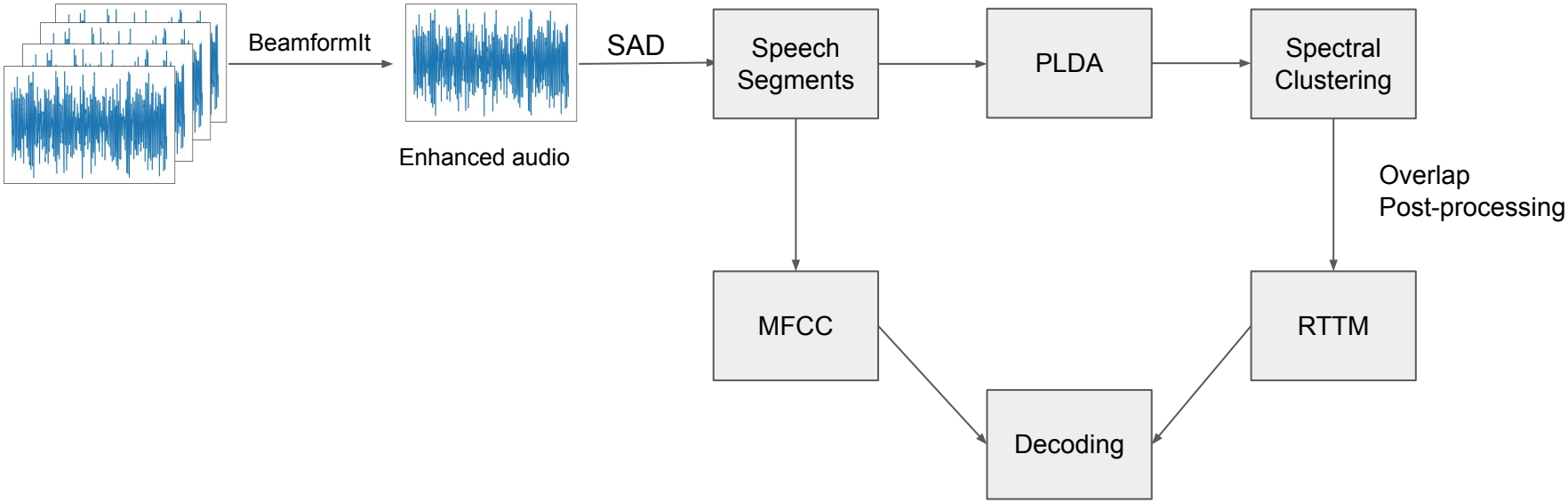
CUNY Speech Diarization System for the CHiME-6 Challenge

Zhaoheng Ni, Michael I Mandel

zni@gradcenter.cuny.edu mim@sci.brooklyn.cuny.edu

City University of New York

System Overview



Spectral Clustering

- Given the similarity matrix S from PLDA.

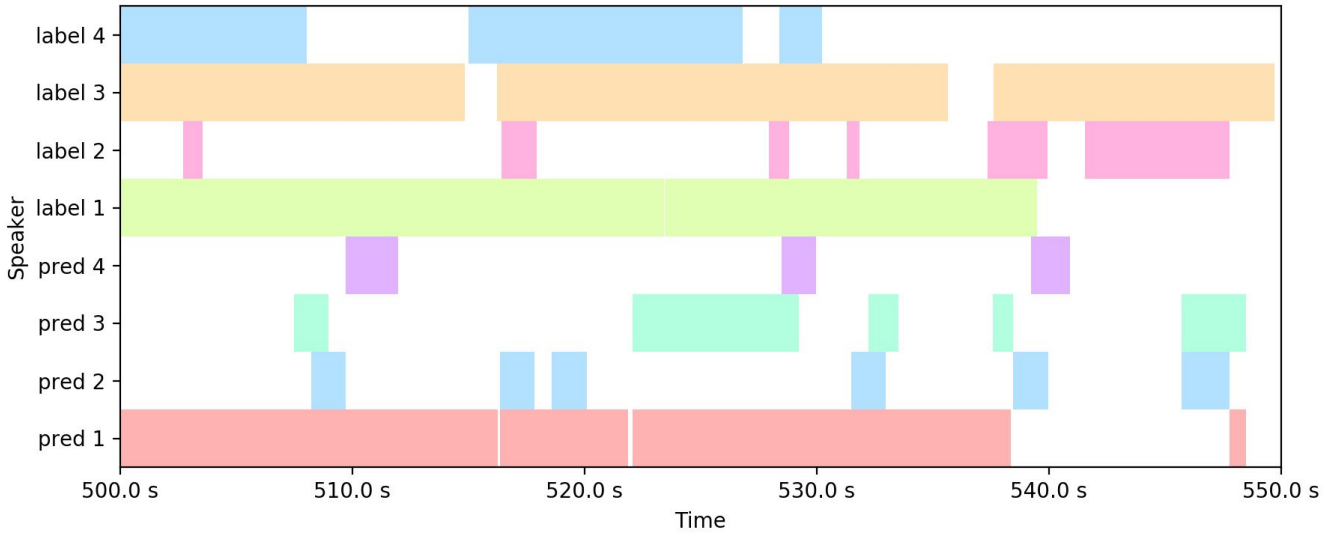
$$L_{\text{norm}} = D^{-1} \times (D - S), D_i = \Sigma_{j=1}^{n} S_{ij}$$

- Compute eigenvalues and eigenvectors of Lnorm.
- Use the first 4 smallest eigenvalues and corresponding eigenvectors as the embedding matrix.
- Apply KMeans clustering algorithm on the matrix.

Overlap Post-processing

- Clustering-based methods assign each frame to only one speaker
- Label overlaps for the x-vectors
  - If more than ratio of duration has overlap, label the x-vector 1
  - else label the x-vector 0
- Train a logistic regression classifier to classify overlaps
- Assign the x-vector to two closest speakers in the spectral clustering

Diarization Visualization



Visualization of the diarization prediction by spectral clustering with post processing (0.67 threshold). The upper 4 rows represent the 4 speakers in the original RTTM reference provided by the challenge, the lower 4 rows represent the 4 speakers in our diarization result.

DER Results: Dev

Reference	CHiME-5		Forced alignment	
Method	DER	JER	DER	JER
Baseline	61.56	63.42	69.75	70.83
SPC	57.15	61.77	57.55	61.18
SPC + PP (0.5)	54.60	52.53	78.83	57.79
SPC + PP(0.67)	51.67	54.45	63.81	57.20

SPC: spectral clustering

PP: post-processing

0.5 ratio to label overlap for x-vectors

DER Results: Eval

Reference	CHiME-5		Forced alignment	
Method	DER	JER	DER	JER
Baseline	61.96	71.40	68.20	72.54
SPC	60.64	65.59	66.29	65.48
SPC + PP (0.5)	70.18	59.72	96.71	63.60
SPC + PP (0.67)	61.51	60.51	77.75	62.75

SPC: spectral clustering

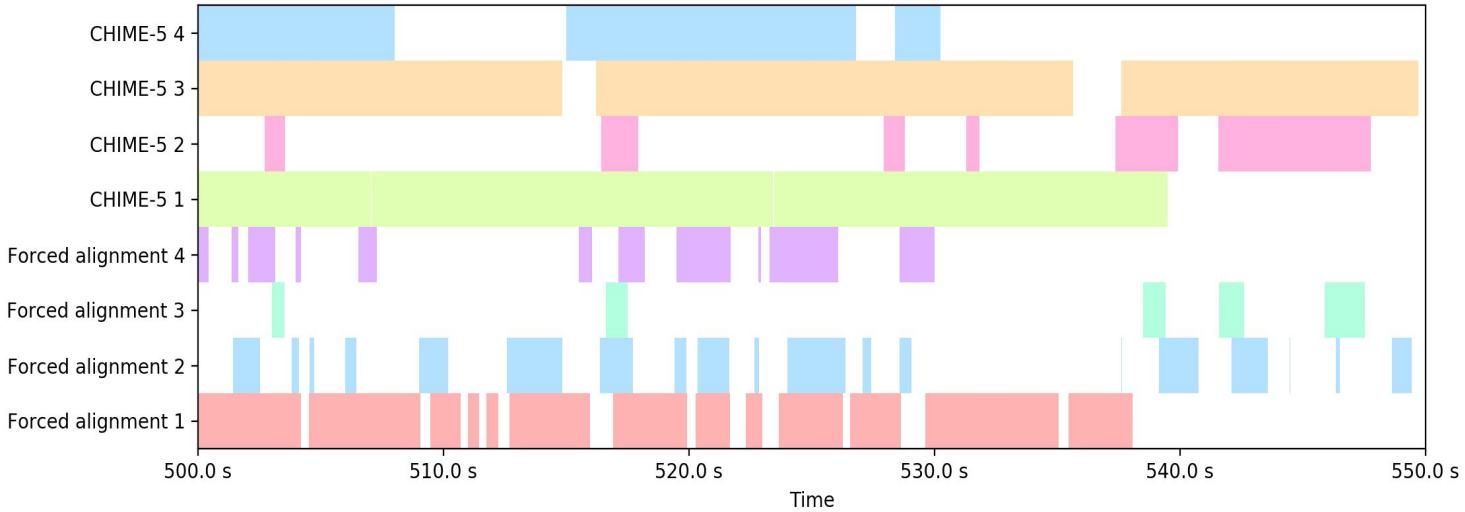
PP: post-processing

0.5 ratio to label overlap for x-vectors

WER Results

Method	Dev	Eval
Baseline	84.25	77.94
SPC	76.48	73.31
SPC + PP (0.5)	77.79	74.49
SPC + PP (0.67)	76.04	72.74
CHiME-5 reference	67.46	61.08
Forced align. reference	63.33	59.58

RTTM References Visualization



Visualization of the two RTTM references provided by the challenge. The upper 4 and lower 4 rows are from the original CHiME-5 reference and the binaural forced alignment reference, respectively.

References

Anguera, Xavier. "Beamformit, the fast and robust acoustic beamformer." (2006).

Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.

Ning, Huazhong, et al. "A spectral clustering approach to speaker diarization." *Ninth International Conference on Spoken Language Processing*. 2006.

Shum, Stephen, Najim Dehak, and James Glass. "On the use of spectral and iterative methods for speaker diarization." *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

Wang, Quan, et al. "Speaker diarization with lstm." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Lin, Qingjian, et al. "LSTM based similarity measurement with spectral clustering for speaker diarization." *arXiv preprint arXiv:1907.10393* (2019).

Thank you!