Multi-class Spectral Clustering with Overlaps

for Speaker Diarization

Desh Raj, Zili Huang, Sanjeev Khudanpur

Center for Language and Speech Processing, Johns Hopkins University

Motivation

What is speaker diarization?

Task of “who spoke when”

Input: ***recording containing multiple speakers***

Output: ***homogeneous speaker segments***

Xavier Anguera Miro et al., “Speaker diarization: A review of recent research,” IEEE Transactions on Audio, Speech, and Language Processing, 2012.

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Motivation

What is speaker diarization?

Task of “who spoke when”

Input: ***recording containing multiple speakers***

Number of speakers may be unknown

Overlapping speech may be present

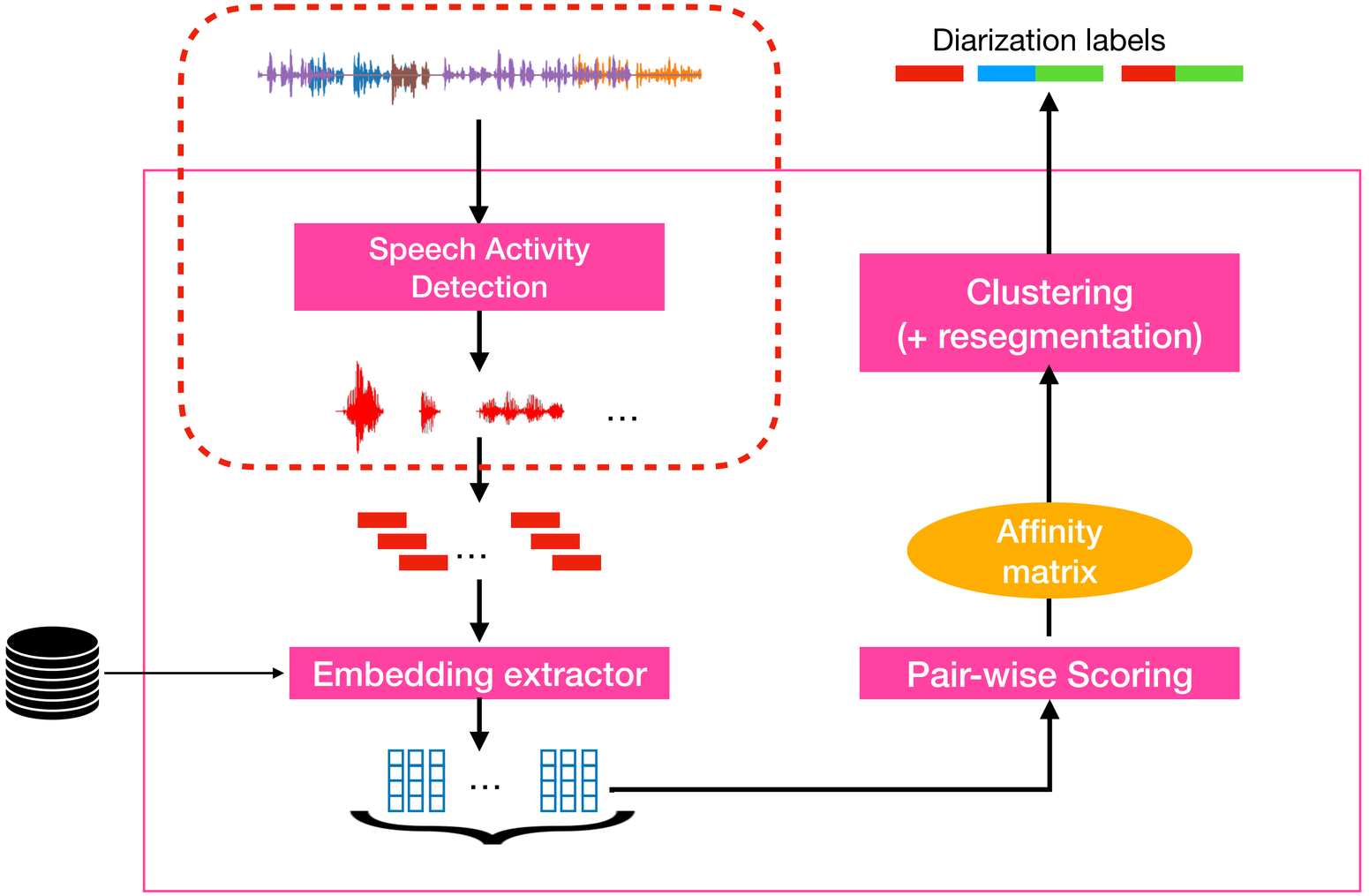
Output: ***homogeneous speaker segments***

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Traditional clustering-based diarization

SAD extracts speech segments from recordings

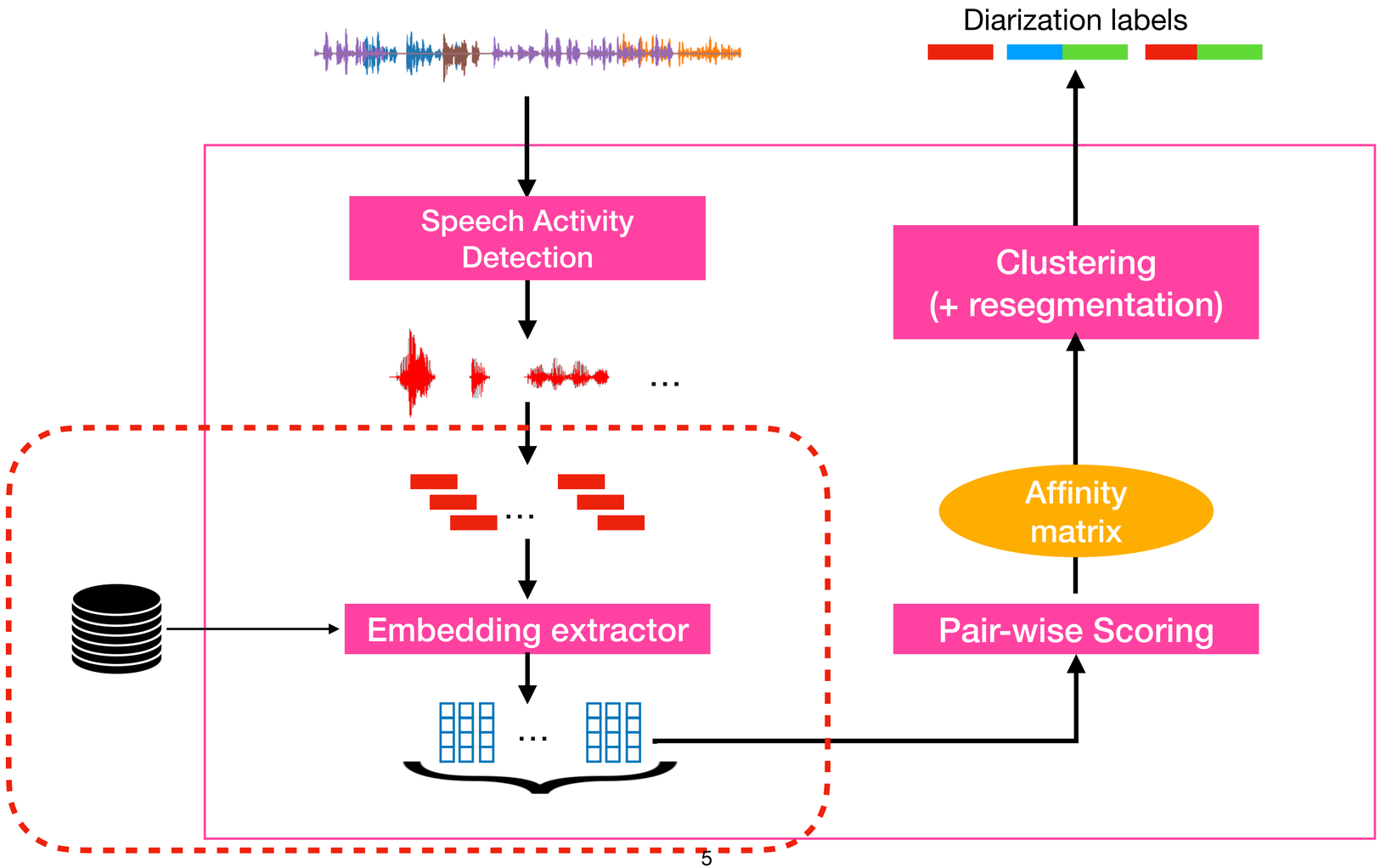


SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Traditional clustering-based diarization

Embeddings extracted for small subsegments

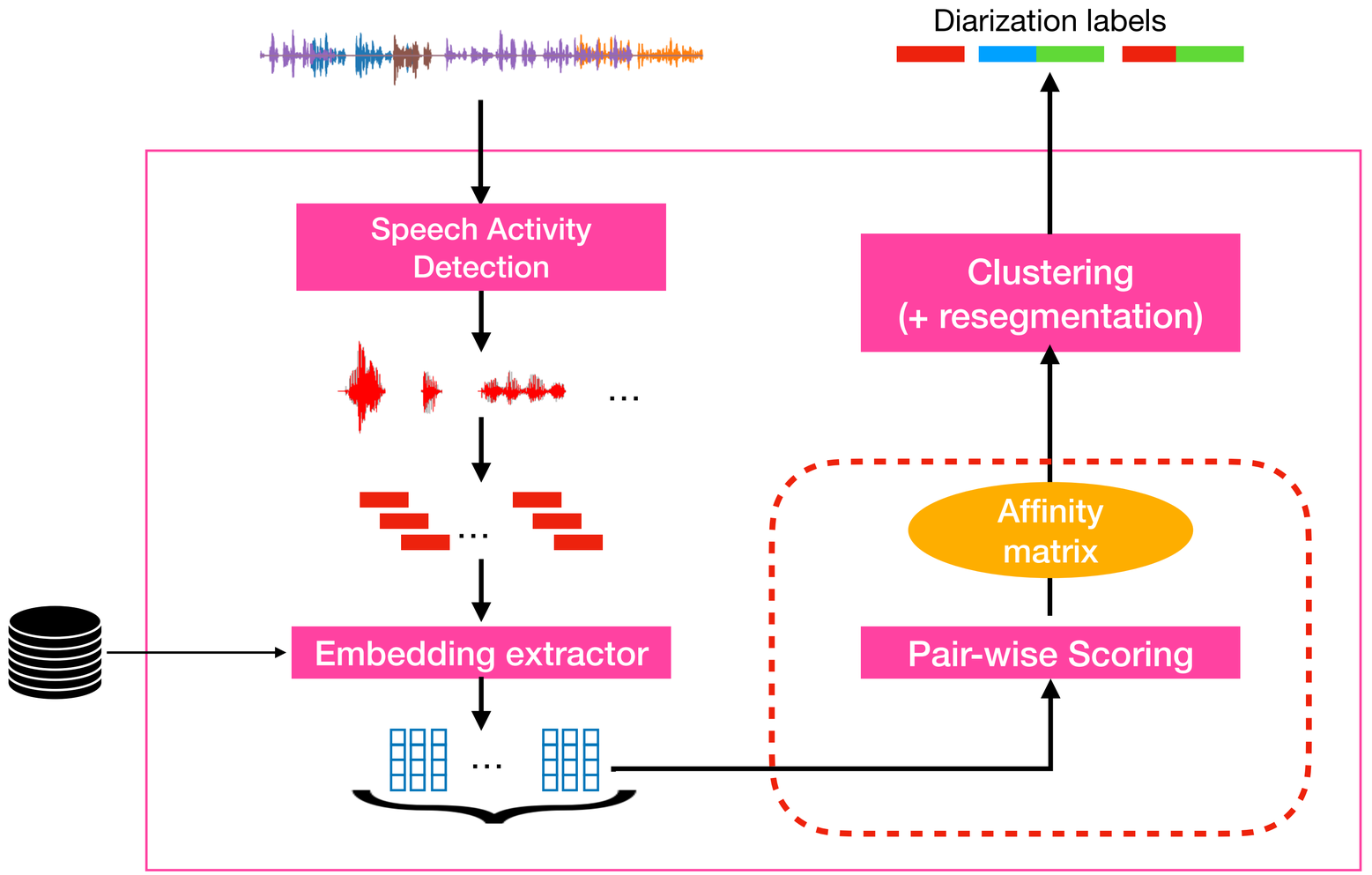


SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Traditional clustering-based diarization

Pair-wise scoring of subsegments



PLDA scoring
Cosine scoring

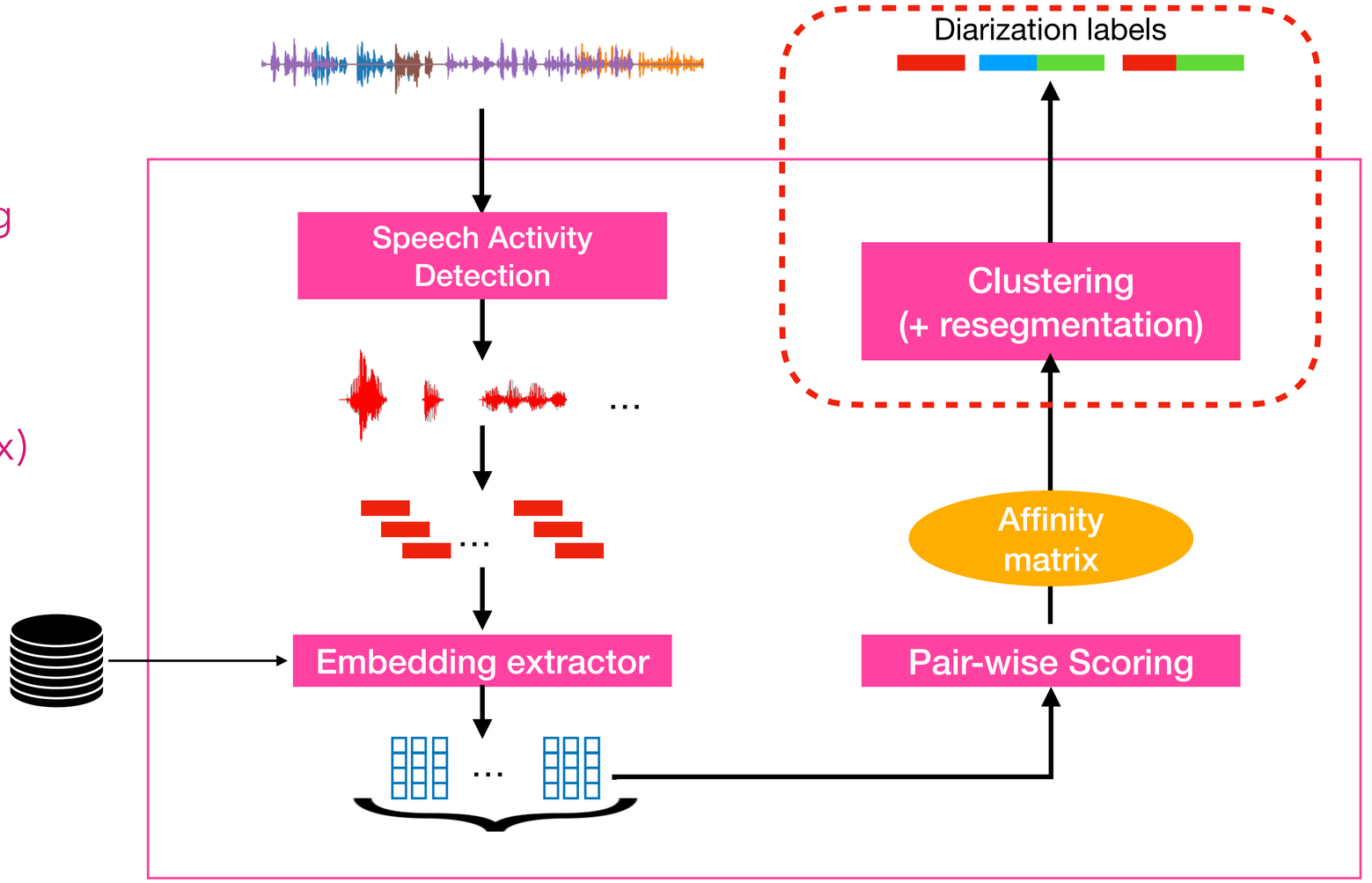
Traditional clustering-based diarization

Clustering based on the affinity matrix, followed by optional resegmentation

Agglomerative hierarchical clustering

Spectral clustering

Variational Bayes (VBx)



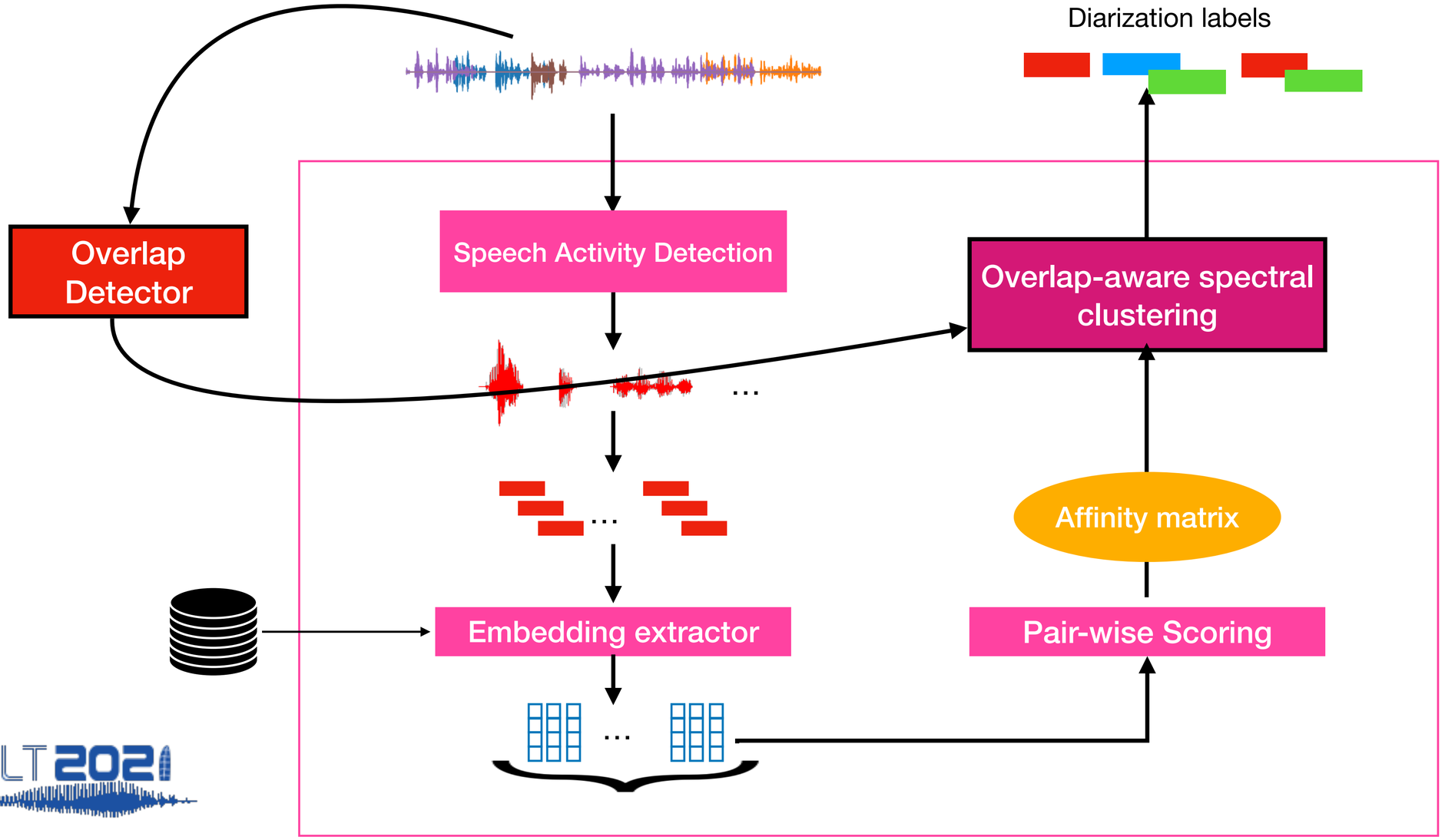
SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Clustering paradigm assumes single-speaker segments

So overlapping speakers are completely ignored!

Overlap-aware spectral clustering



SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results on AMI Mix-Headset eval

12.0% relative improvement over spectral clustering baseline

System	DER
Spectral clustering	26.9
AHC	28.3
VBx	26.2
Overlap-aware SC	**24.0**

Park et al., “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” IEEE Signal Processing Letters, 2020.

Garcia-Romero et al., “Speaker diarization using deep neural network embeddings,” ICASSP 2017.

Dîez et al., “Speaker diarization based on Bayesian HMM with eigenvoice priors,” Odyssey 2018.

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results on AMI Mix-Headset eval

Comparable with other overlap-aware diarization methods

System	DER
VB-based overlap assignment	23.8
Region proposal networks	25.5
Overlap-aware SC	**24.0**

Bullock, et al., “Overlap-aware diarization: resegmentation using neural end-toend overlapped speech detection,” ICASSP 2020.

Huang et al., “Speaker diarization with region proposal network,” ICASSP 2020.

Does not require **matching training data** or **initialization** with other diarization systems.

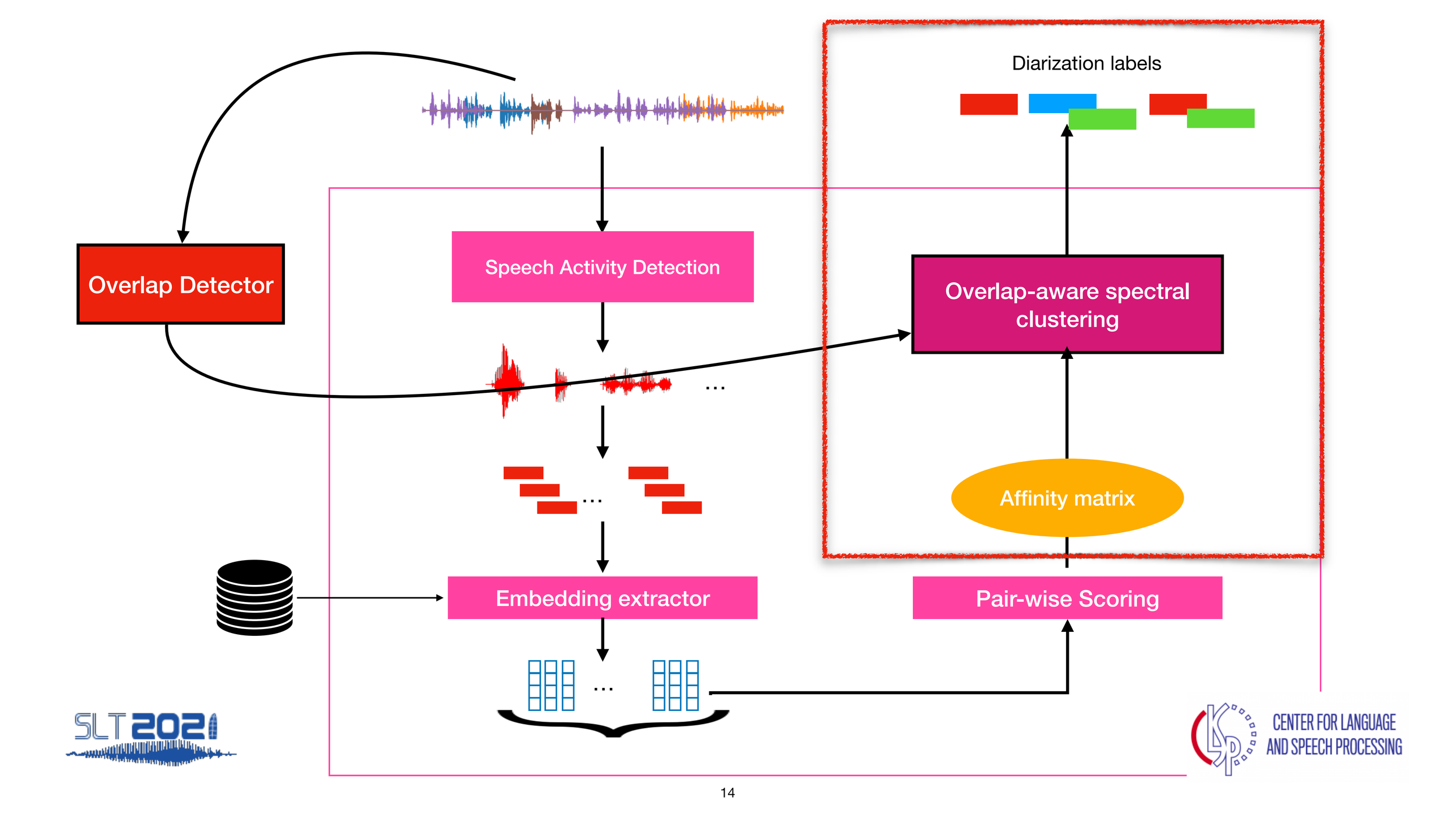
SLT 2021

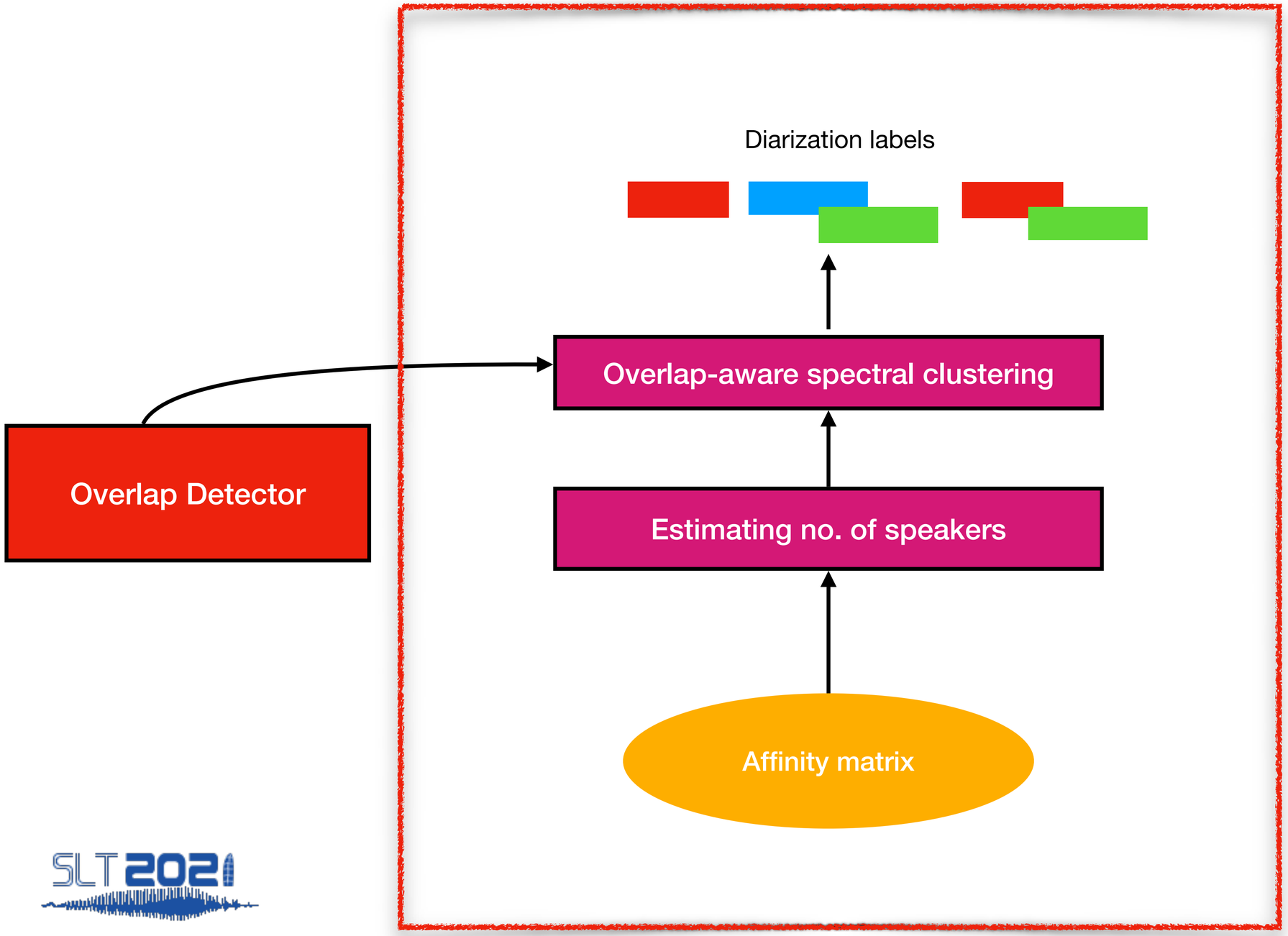
CENTER FOR LANGUAGE AND SPEECH PROCESSING

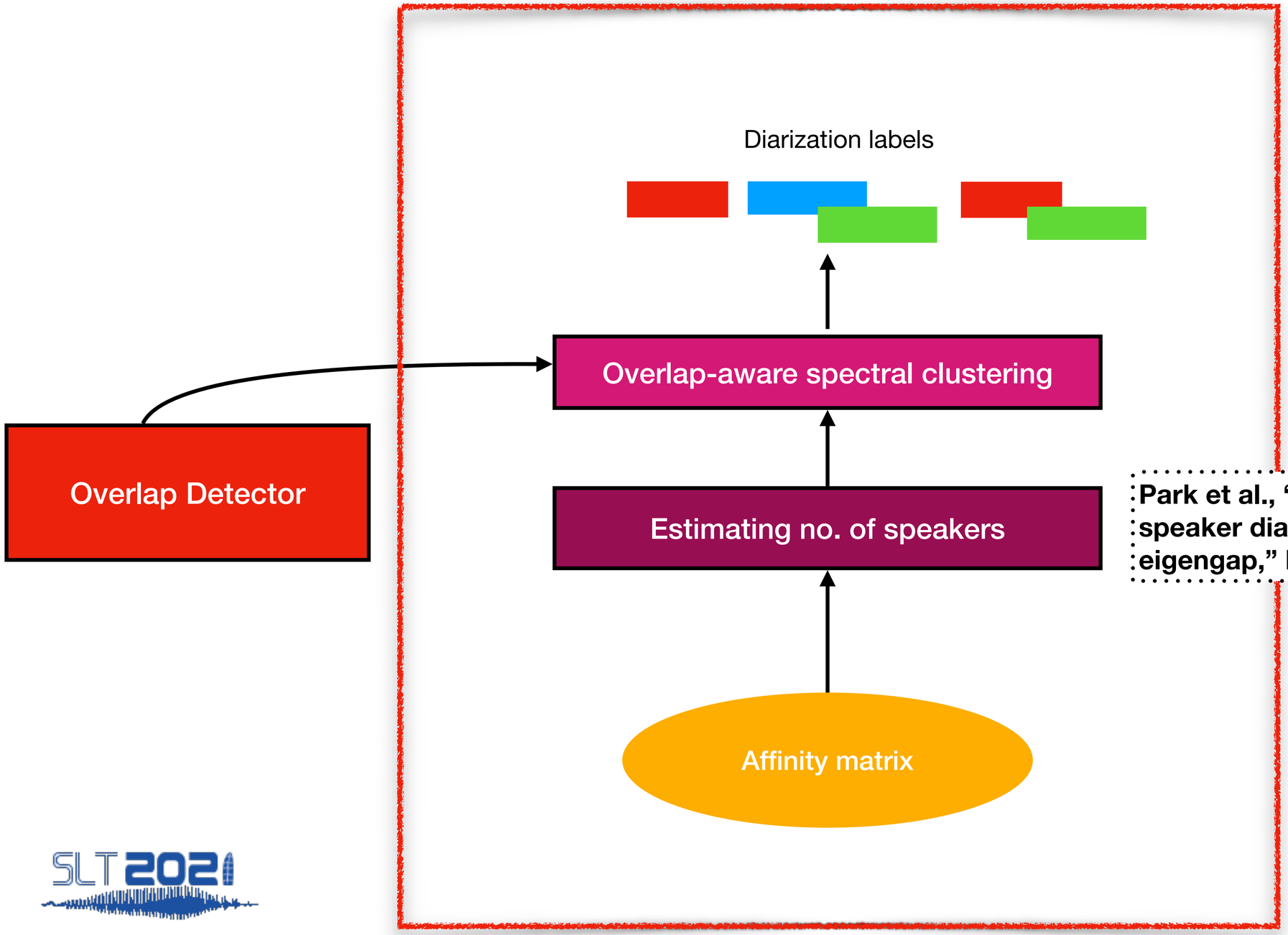
End of Highlight

Overview

- Overlap-aware spectral clustering
  - Estimating number of speakers
  - Reformulation of the clustering problem
  - Incorporating overlap detector output
- HMM-DNN overlap detector: Overview
- More Results
  - Error analysis on AMI


Diarization labels
Overlap Detector
Speech Activity Detection
Overlap-aware spectral clustering
Affinity matrix
Embedding extractor
Pair-wise Scoring
SLT 2021
CENTER FOR LANGUAGE AND SPEECH PROCESSING


Diarization labels
Overlap-aware spectral clustering
Overlap Detector
Estimating no. of speakers
Affinity matrix



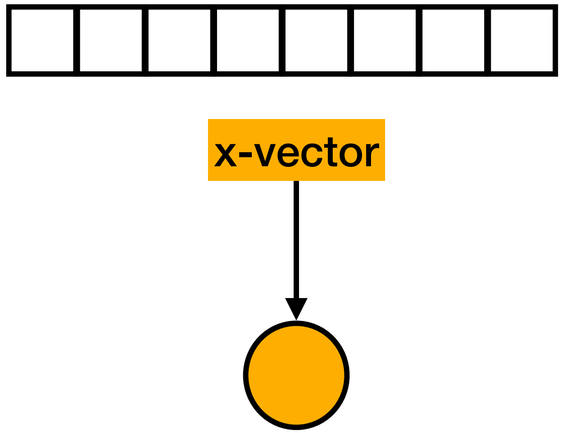
Park et al., “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” IEEE Signal Processing Letters, 2020.

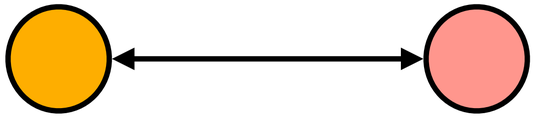
SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

New formulation for spectral clustering

The basic clustering problem: a graph view

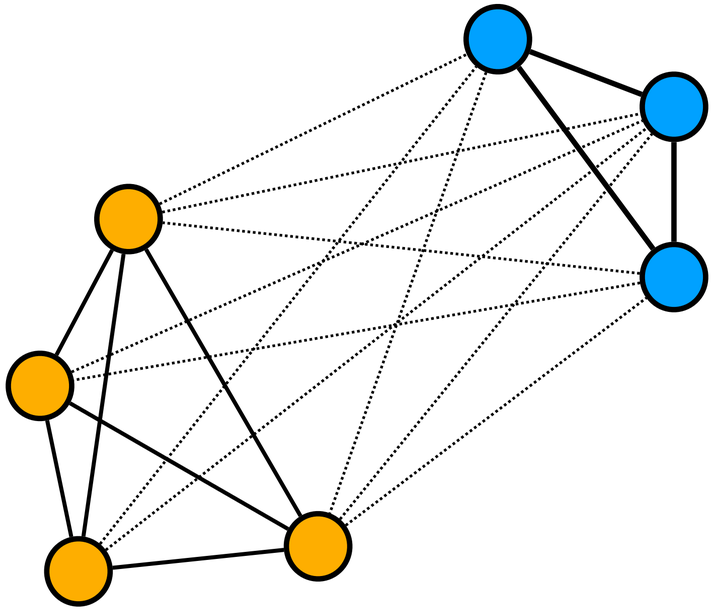




Cosine similarity

New formulation for spectral clustering

The basic clustering problem: a graph view



Speaker B

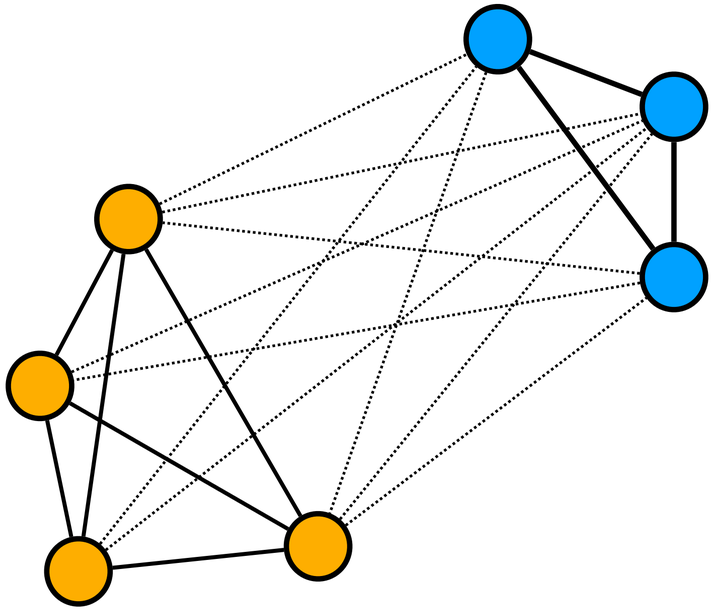
Speaker A

Edge weights within a group

Edge weights across groups

New formulation for spectral clustering

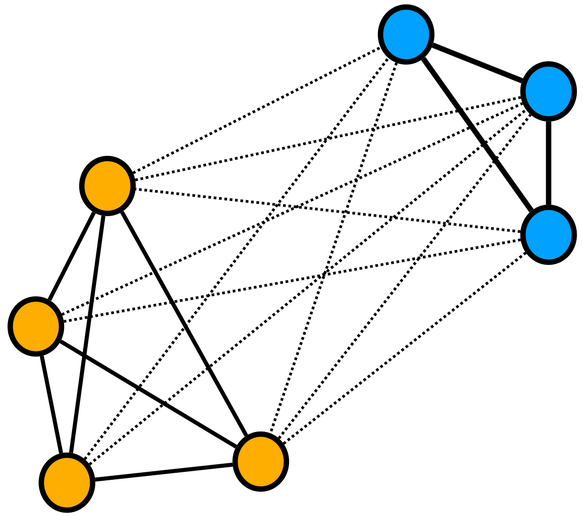
The basic clustering problem: a graph view



$$\textit{maximize} \quad \frac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$$

New formulation for spectral clustering

The basic clustering problem: a graph view



maximize $\dfrac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$

$$\text{maximize} \quad \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T\mathbf{A}X_k}{X_k^T\mathbf{D}X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N\times K},$$

$$X\mathbf{1}_K = \mathbf{1}_N\,.$$

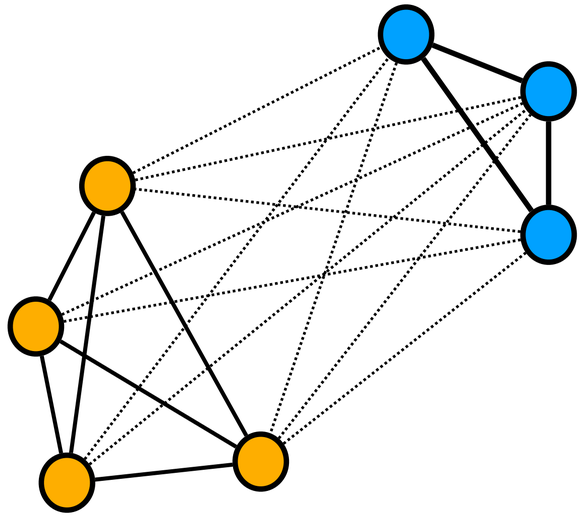
K speakers, **N** segments

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

New formulation for spectral clustering

The basic clustering problem: a graph view



maximize $\dfrac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$

$$\text{maximize} \quad \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N\times K},$$

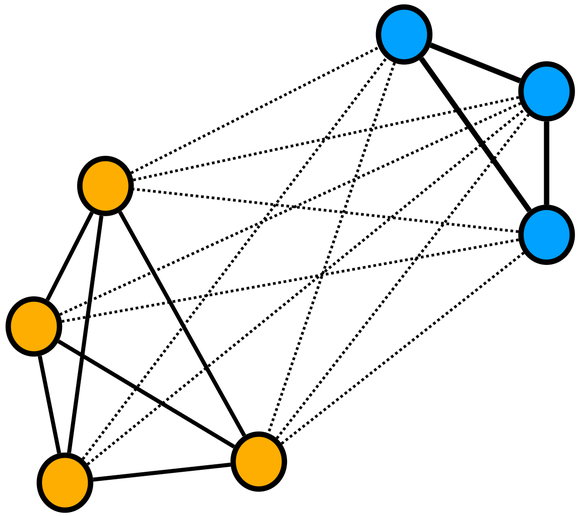
$$X\mathbf{1}_K = \mathbf{1}_N .$$

Affinity matrix

Diagonal matrix containing degree of nodes

New formulation for spectral clustering

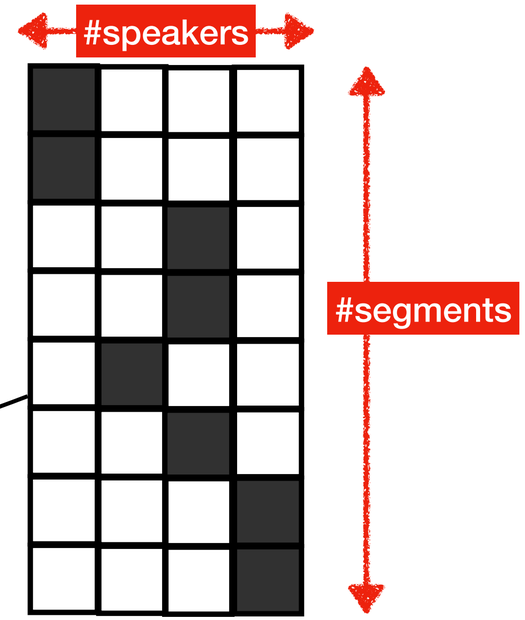
The basic clustering problem: a graph view



$$\text{maximize} \quad \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T\mathbf{A}X_k}{X_k^T\mathbf{D}X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N\times K},$$

$$X\mathbf{1}_K = \mathbf{1}_N .$$



Final cluster assignment matrix

New formulation for spectral clustering

This problem is NP-hard!

$$\text{maximize} \quad \epsilon(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N \times K},$$

$$X\mathbf{1}_K = \mathbf{1}_N.$$

Remove the discrete constraints to make the problem solvable

New formulation for spectral clustering

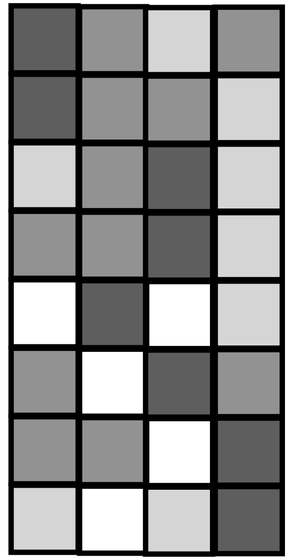
Relaxed problem has a set of solutions

$$\text{maximize} \quad \epsilon(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N \times K},$$

$$X\mathbf{1}_K = \mathbf{1}_N.$$

Taking the Eigen-decomposition of $D^{-1}A$

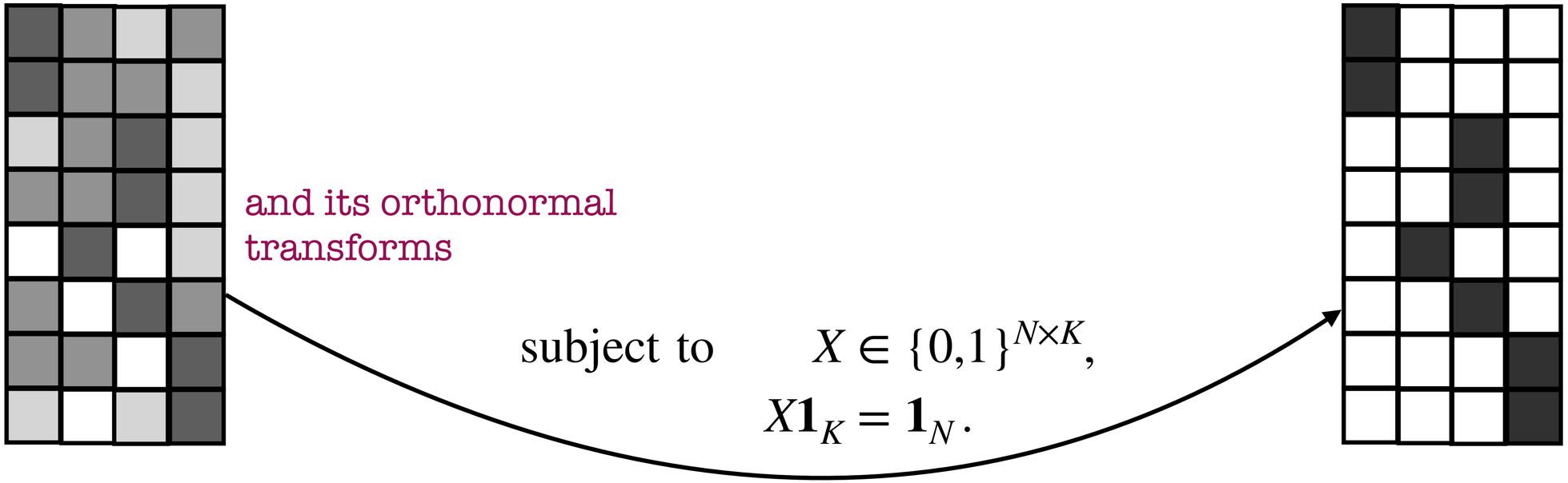


and its orthonormal transforms

Set of solutions to the **relaxed** problem

New formulation for spectral clustering

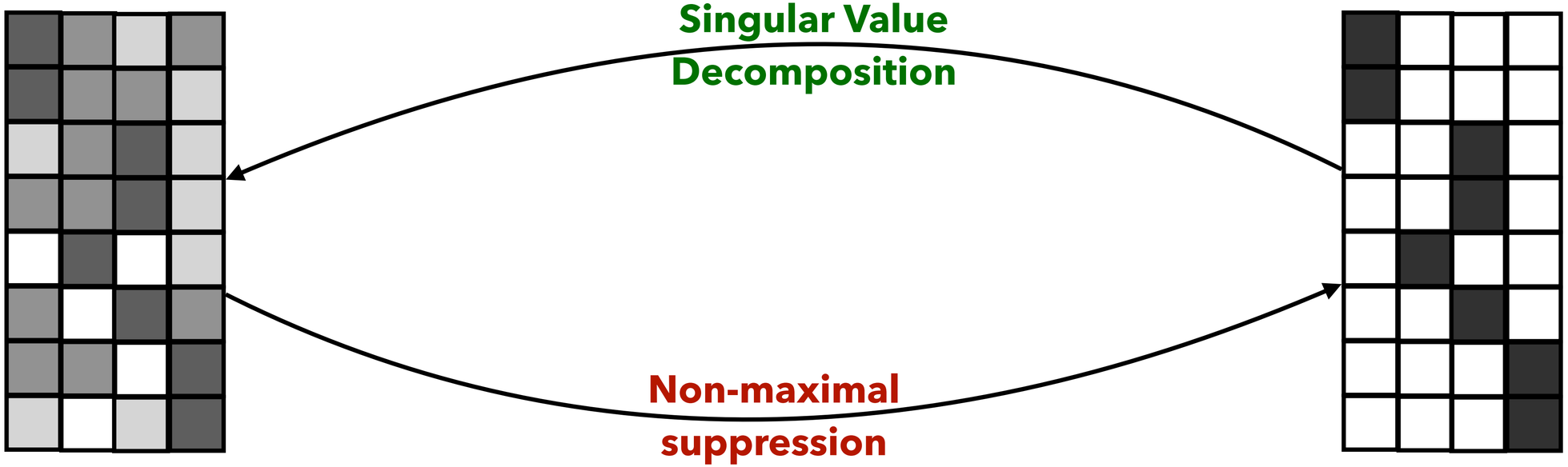
Now we need to discretize this solution!



Find a matrix which is **discrete** and also close to any one of the **orthonormal transformations** of the relaxed solution

New formulation for spectral clustering

Now we need to discretize this solution!

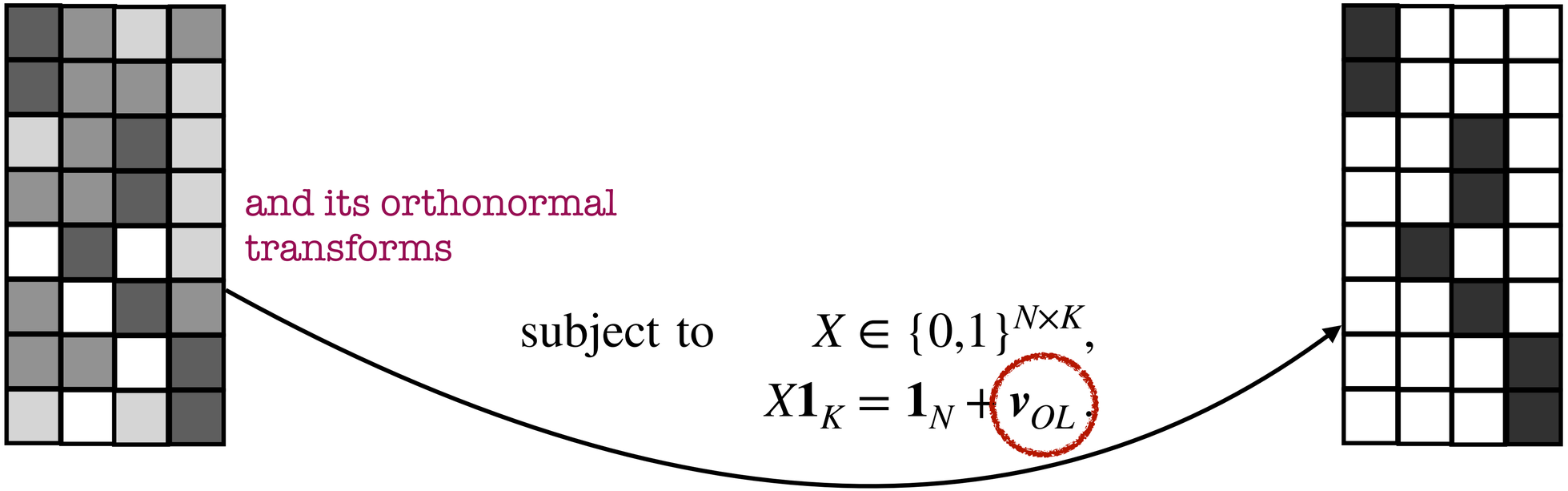


and its orthonormal transforms

Iterate until convergence

Let us now make it overlap-aware

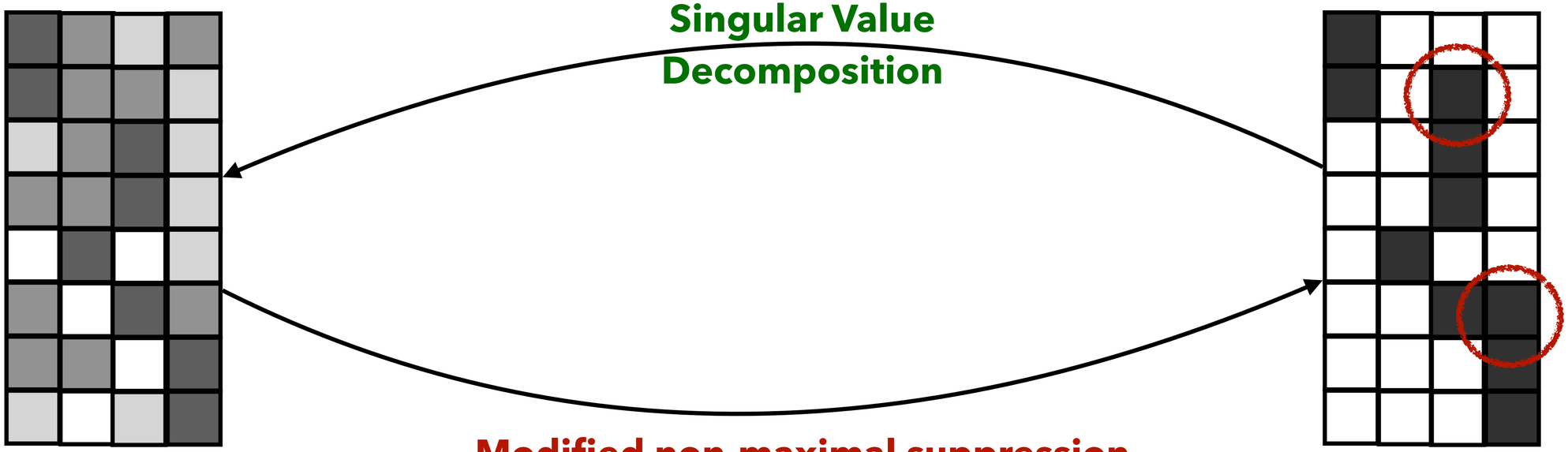
Suppose we have $\mathbf{v}_{OL}$



Discrete constraint is modified to include overlap detector output

Let us now make it overlap-aware

Modify non-maximal suppression to pick top 2 speakers



Iterate until convergence

How does it compare with …

… VB-based overlap assignment

Bullock, et al., "Overlap-aware diarization: resegmentation using neural end-toend overlapped speech detection," ICASSP 2020.

	Overlap-aware SC	VB based overlap assignment	
✓	No initialization required	Needs initialization, e.g. from AHC system	✗
✗	Does segment-level assignment (coarse)	Does frame-level assignment (fine-grained)	✓

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

How does it compare with …

… EEND, RPN, TS-VAD

	Overlap-aware SC	RPN, EEND, TS-VAD	
✗	Uses external overlap detector	Includes overlap detection/assignment	✓
✓	Matched training data not required	Requires matched (simulated) training data	✗

Huang et al., “Speaker diarization with region proposal network,” ICASSP 2020.

Fujita et al., “End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification,” ArXiv.

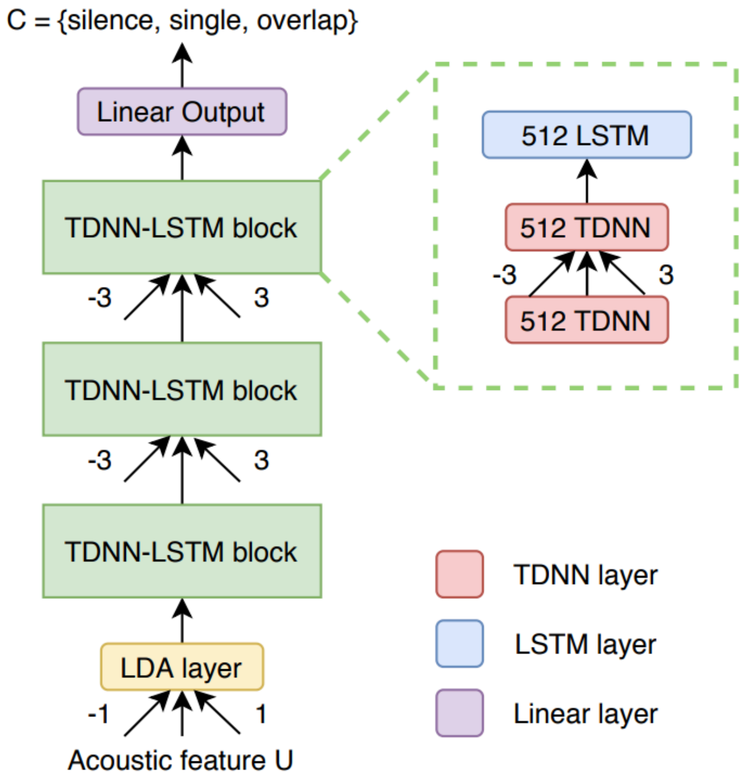
Medennikov, et al., “Target speaker voice activity detection: a novel approach for multispeaker diarization in a dinner party scenario,” Interspeech 2020.

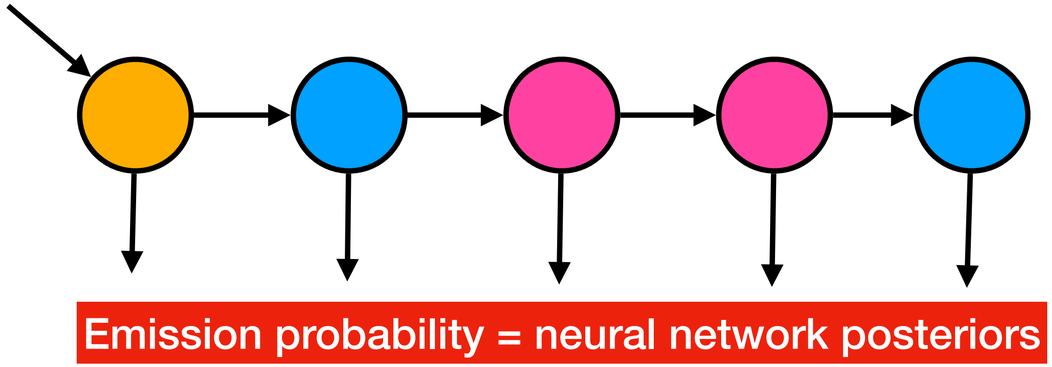
SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Hybrid HMM-DNN overlap detector

(Now merged as a KALDI recipe)





Viterbi decoding used for inference

Results: DER breakdown on AMI eval

System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results: DER breakdown on AMI eval

Missed speech decreases significantly

System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results: DER breakdown on AMI eval

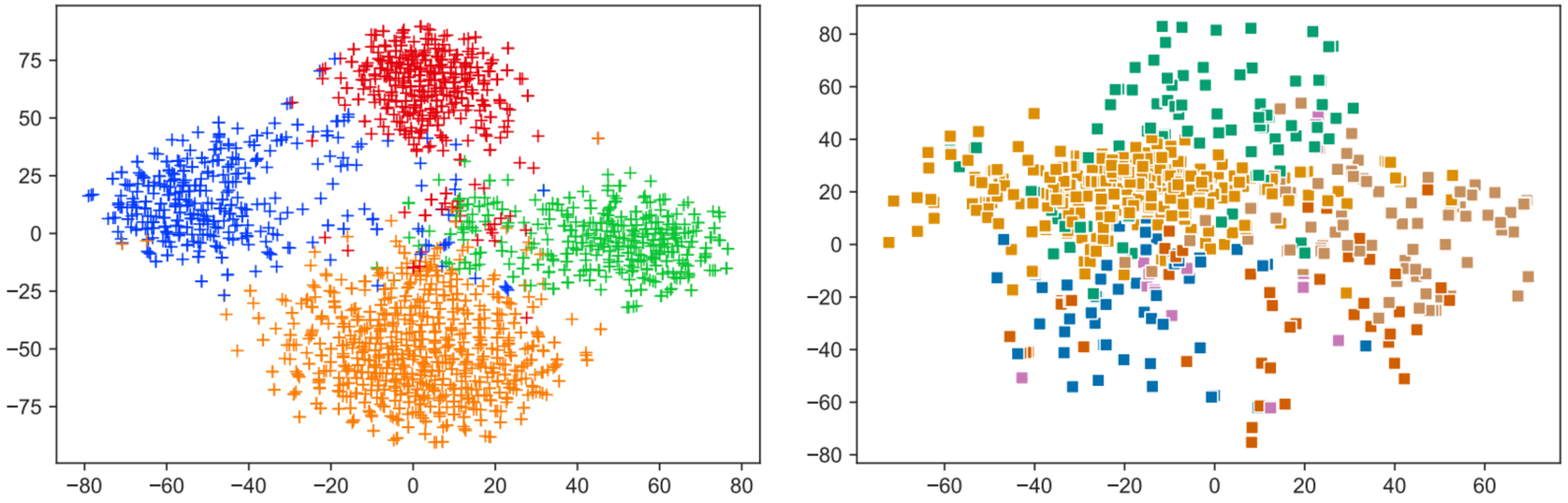
Speaker confusion increases

System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Future work: train a more robust x-vector extractor



Non-overlapping segments

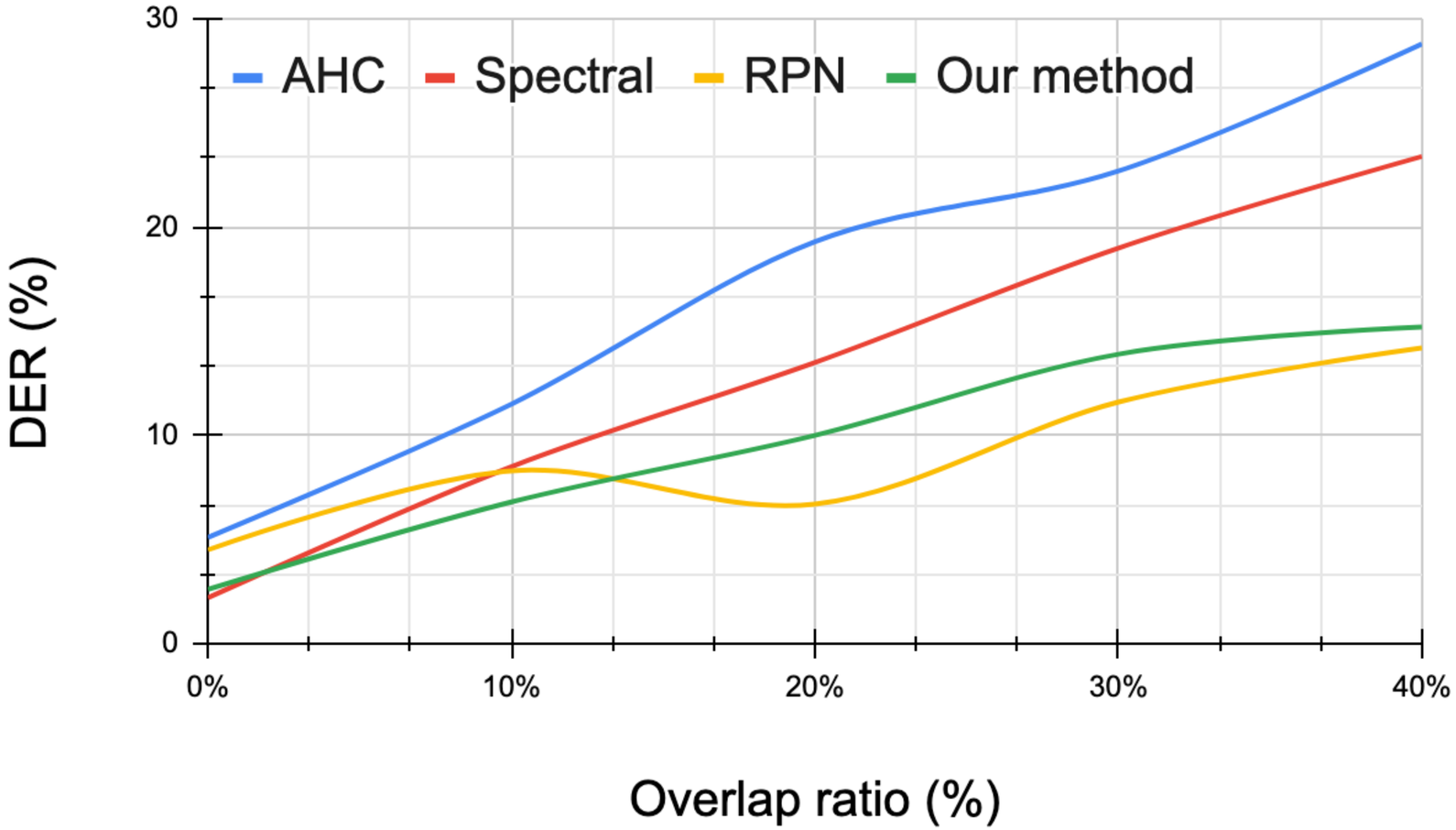
Overlapping segments

T-SNE plot of x-vector embeddings

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

More results: DER on LibriCSS



SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Try out the code!

Acknowledgments:

Paola Garcia, for helpful discussions and insights.

Takuya Yoshioka, for simulation script for generating LibriCSS training data.

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING