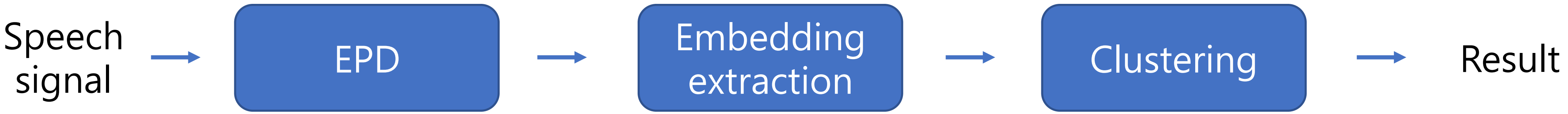
NAVER Clova Submission To The Third DIHARD Challenge

Hee-Soo Heo, Jee-weon Jung, Youngki Kwon, You Jin Kim,
Jaesung Huh, Joon Son Chung, Bong-Jin Lee

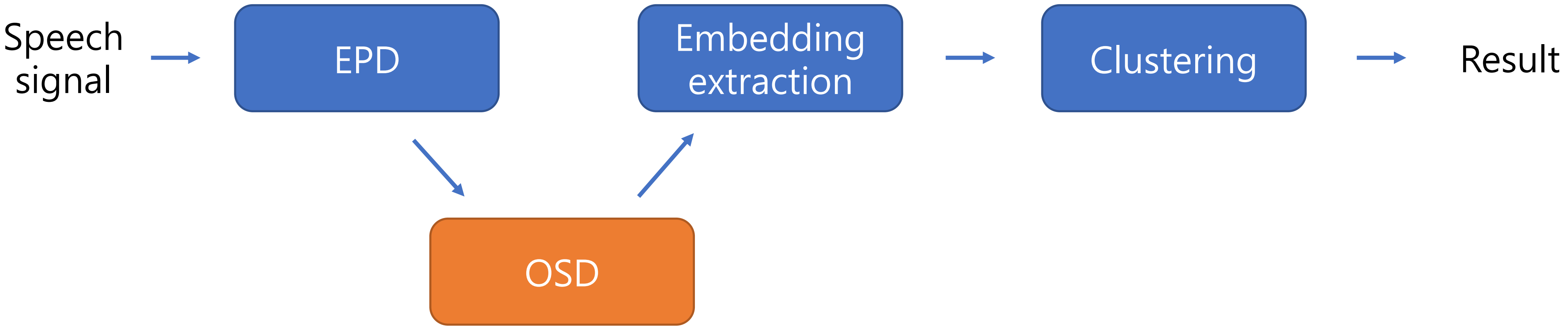
Pipeline

- Common step-wise pipeline
  - EPD: same as baseline
  - Embedding extractor: ResNet34
  - Clustering: spectral clustering
  - No sequential modeling


Speech signal → EPD → Embedding extraction → Clustering → Result

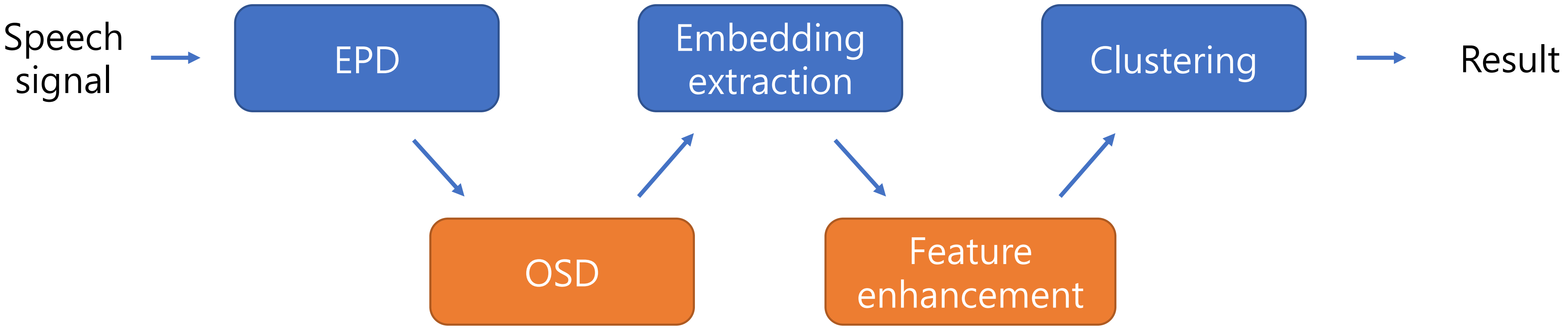
Contribution I

- Overlapped speech detection (OSD)
  - CRNN-based model ensemble



Contribution II

- Feature enhancement designed for diarization task
  - Session-level dimensionality reduction
  - Attention-based aggregation



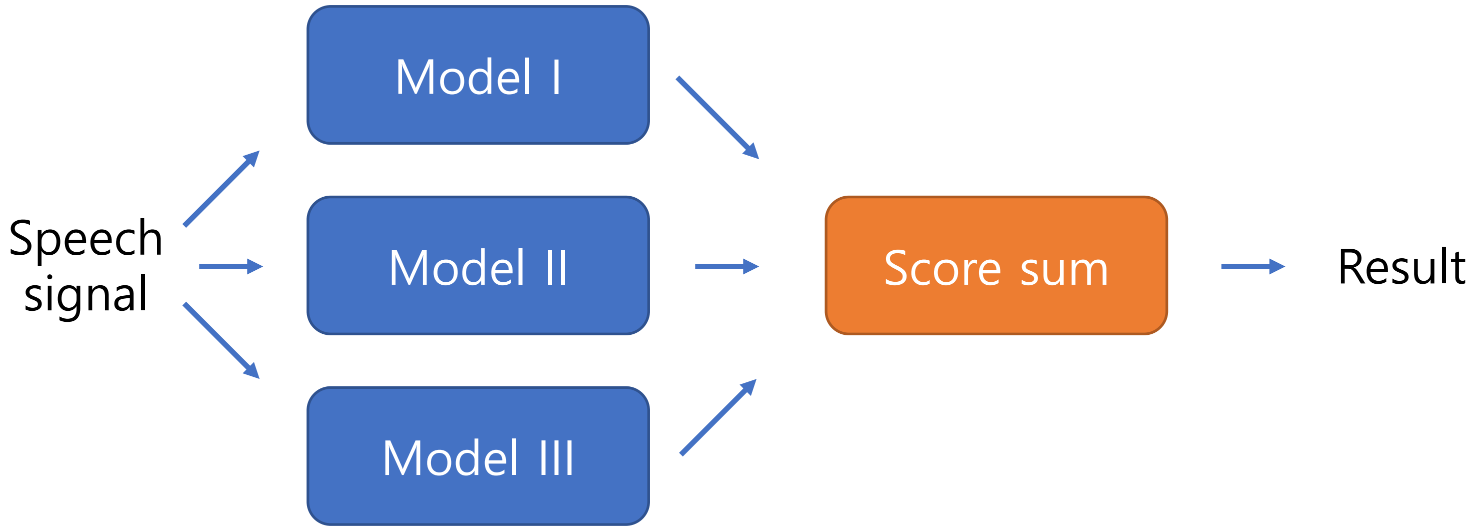
Overlapped speech detection

- Detects segments that includes multiple speakers
  - Outputs onset/offset of overlapped speech segment
- Key features
  1. Three class DNN classifier: non-speech, single speaker speech, overlapped speech
     - Test phase: use score of overlapped speech
  2. Within session overlapped speech augmentation
     - Label unbalanced: <10% overlapped speech in train dataset
     - Add another speaker’s segment into single speaker speech
  3. CRNN* architecture

* CRNN: convolutional recurrent neural network

Overlapped speech detection

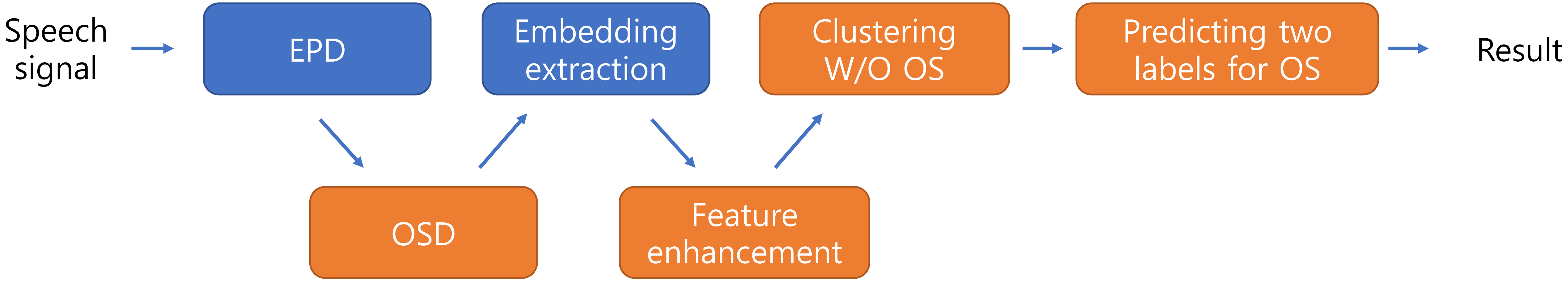
- Final system: score-level ensemble of three variants
  - Model I: 2D-CRNN w/ SE*
  - Model II: 2D-CRNN w/o SE
  - Model III: 1D-CRNN w/o SE



* SE: squeeze-excitation

Modification of pipeline

- To minimize clustering error caused by OS
  - Clustering without embeddings of OS
  - Predicting two labels based on cluster centroids



Feature enhancement

- Common in machine learning field,
  but **haven't been explored** for diarization task

- Characteristic of diarization
  - **Limited number of speakers to consider**
  - Within session comparison only

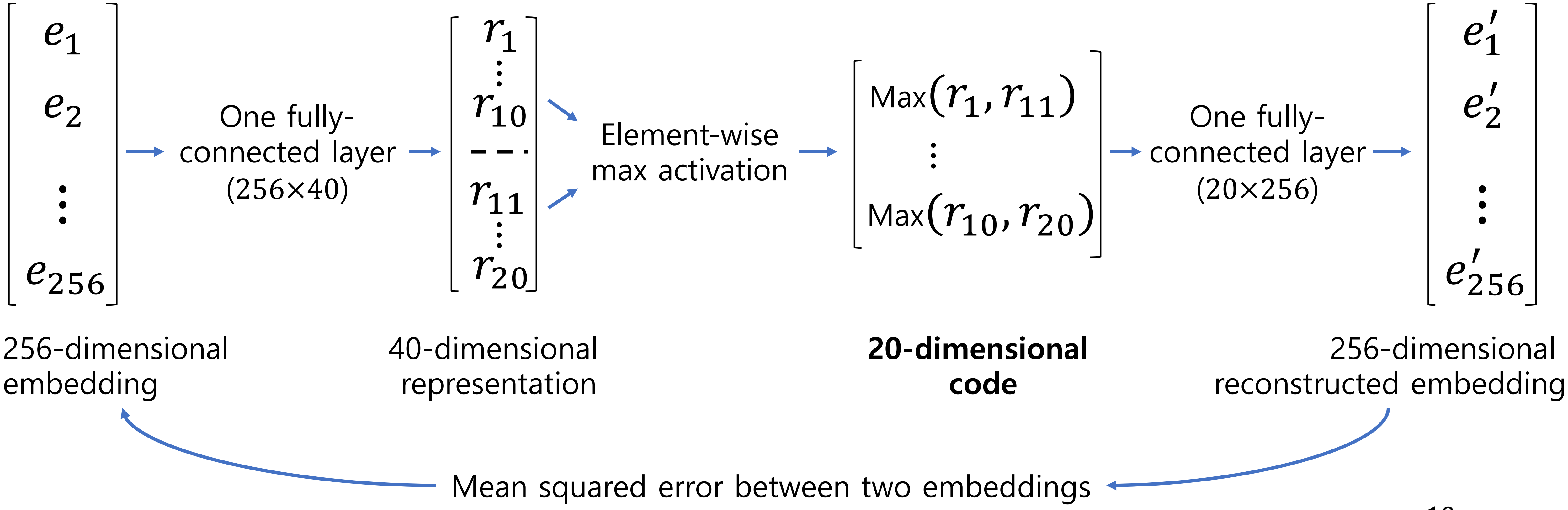
Dimensionality reduction

- Training auto-encoder for each session
  - Shallow architecture with max feature-map layer
  - From 256-dimensional embedding to 20-dimensional code

  - Training configuration
    - 200 epoch training for each session
    - Adam optimizer with 0.001 learning rate
    - No regularization

Dimensionality reduction

- Architecture of auto-encoder



$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{256} \end{bmatrix}$$

One fully-connected layer (256×40)

$$\begin{bmatrix} r_1 \\ \vdots \\ r_{10} \\ \hline r_{11} \\ \vdots \\ r_{20} \end{bmatrix}$$

Element-wise max activation

$$\begin{bmatrix} \mathrm{Max}(r_1, r_{11}) \\ \vdots \\ \mathrm{Max}(r_{10}, r_{20}) \end{bmatrix}$$

One fully-connected layer (20×256)

$$\begin{bmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_{256} \end{bmatrix}$$

256-dimensional embedding

40-dimensional representation

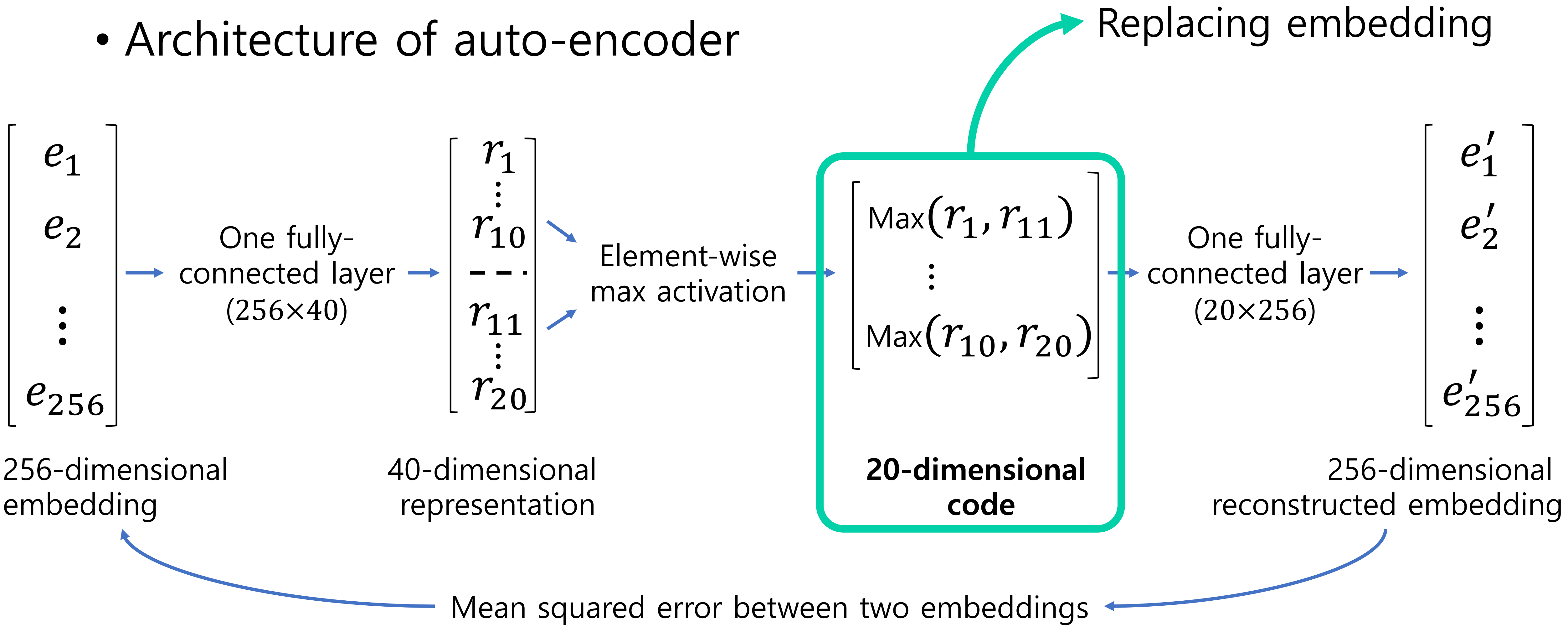
20-dimensional code

256-dimensional reconstructed embedding

Mean squared error between two embeddings

Dimensionality reduction

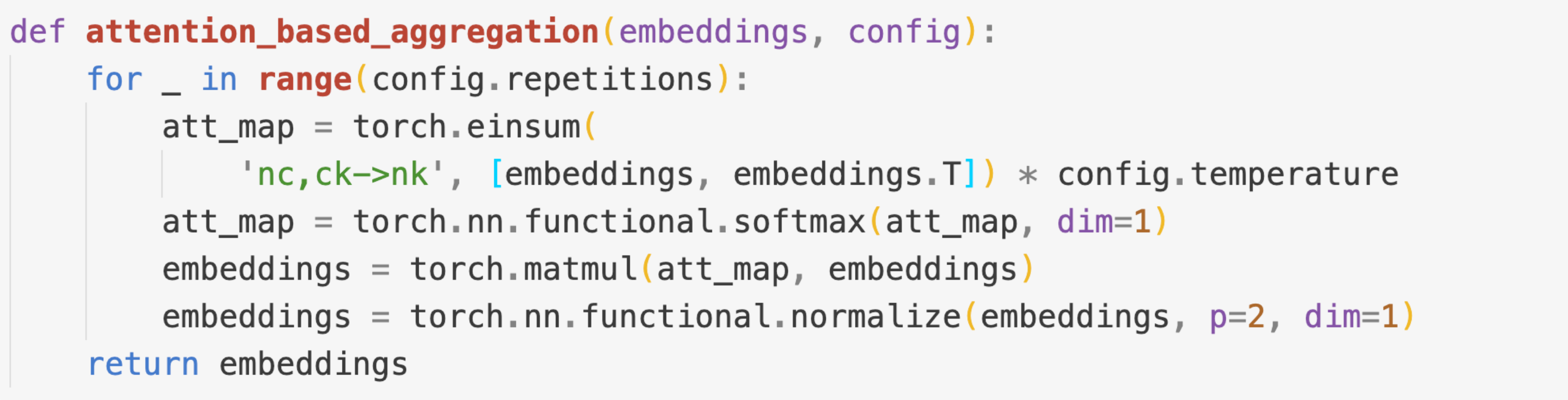
- Architecture of auto-encoder



Replacing embedding
$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{256} \end{bmatrix}$
One fully-connected layer (256×40)
$\begin{bmatrix} r_1 \\ \vdots \\ r_{10} \\ \hline r_{11} \\ \vdots \\ r_{20} \end{bmatrix}$
Element-wise max activation
$\begin{bmatrix} \text{Max}(r_1, r_{11}) \\ \vdots \\ \text{Max}(r_{10}, r_{20}) \end{bmatrix}$
One fully-connected layer (20×256)
$\begin{bmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_{256} \end{bmatrix}$
256-dimensional embedding
40-dimensional representation
20-dimensional code
256-dimensional reconstructed embedding
Mean squared error between two embeddings

Attention-based aggregation

- Soft version of clustering with two hyper-parameters
  - Number of repetitions: 5
  - Temperature value before softmax function: 15

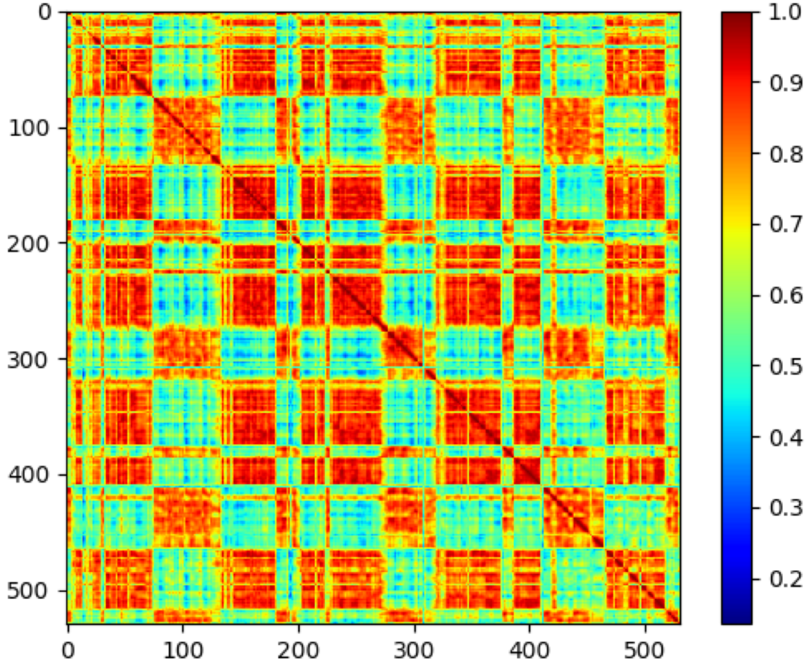


```
def attention_based_aggregation(embeddings, config):
    for _ in range(config.repetitions):
        att_map = torch.einsum(
            'nc,ck->nk', [embeddings, embeddings.T]) * config.temperature
        att_map = torch.nn.functional.softmax(att_map, dim=1)
        embeddings = torch.matmul(att_map, embeddings)
        embeddings = torch.nn.functional.normalize(embeddings, p=2, dim=1)
    return embeddings
```

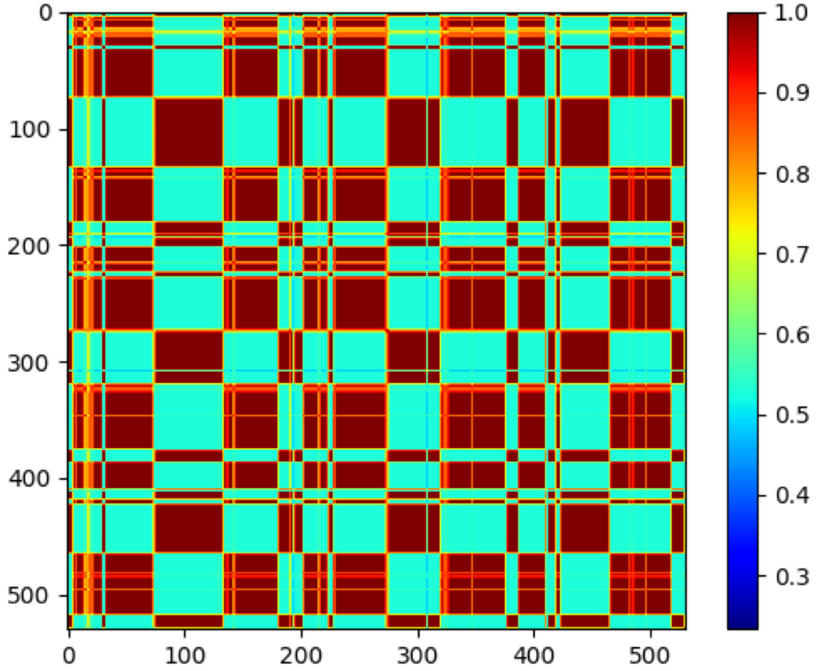
Attention-based aggregation

- Robust to outliers
- Refinement of affinity matrix

W/O AA

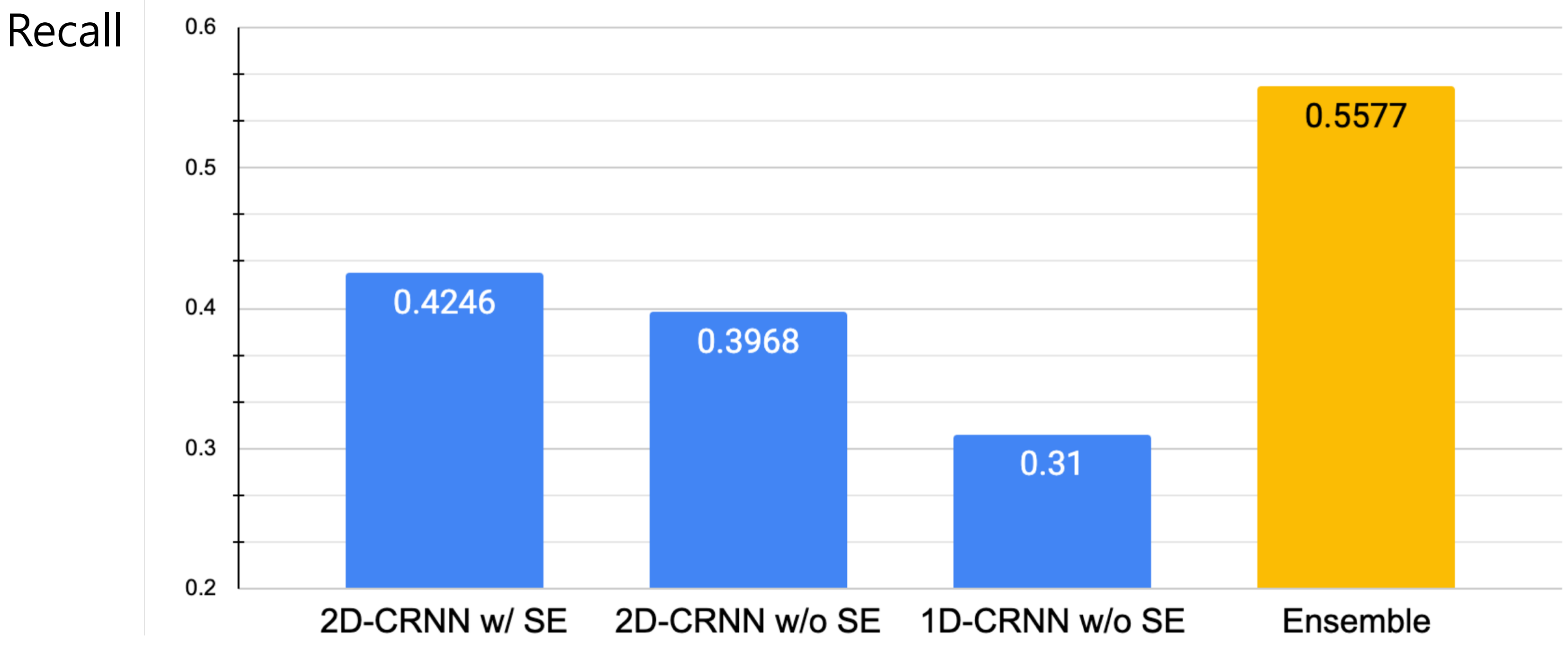


W/ AA



Experiments

- OSD configuration
  - Trainset: AMI corpus & VoxConverse & DIHARD 1&2 devset
  - Submitted model tuned using DIHARD 3 dev set
    - Set threshold that matches precision = 0.8



Embedding extractor

- Public ResNetSE34V2 architecture and training protocol[1)]

- Trainset: VoxCeleb1 & VoxCeleb2 dev set
- Frame-level feature: 64-dimensional mel-filterbank
- Number of filters in the first conv layer: 64
- Aggregation: average pooling
- Dimension of embedding vector: 256

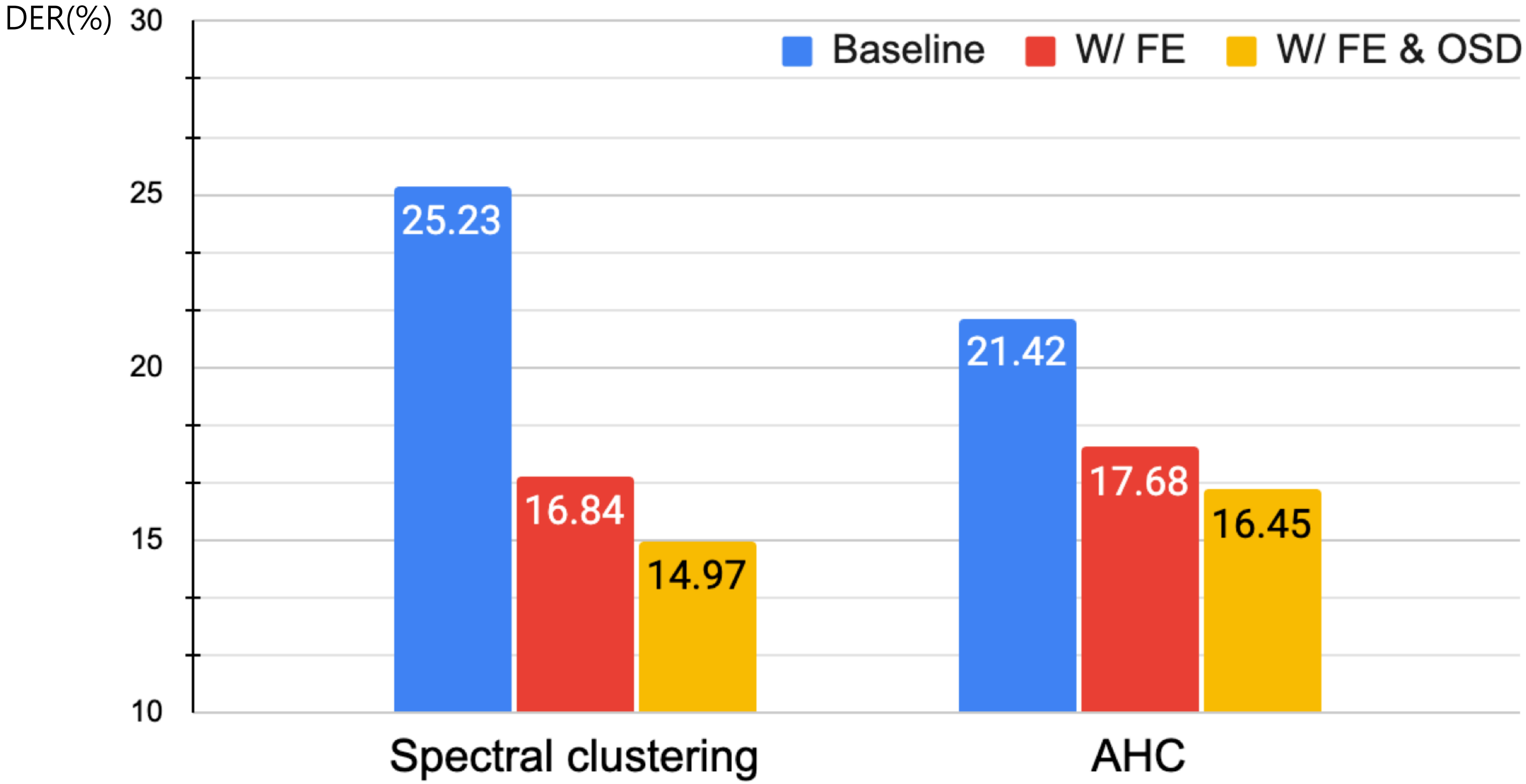
1) https://github.com/clovaai/voxceleb_trainer

Spectral clustering

- Cosine similarity-based affinity matrix
- **No additional refinement processes**
- Determining the number of clusters: eigenvalue 20 ↑
- K-means of spectral embeddings

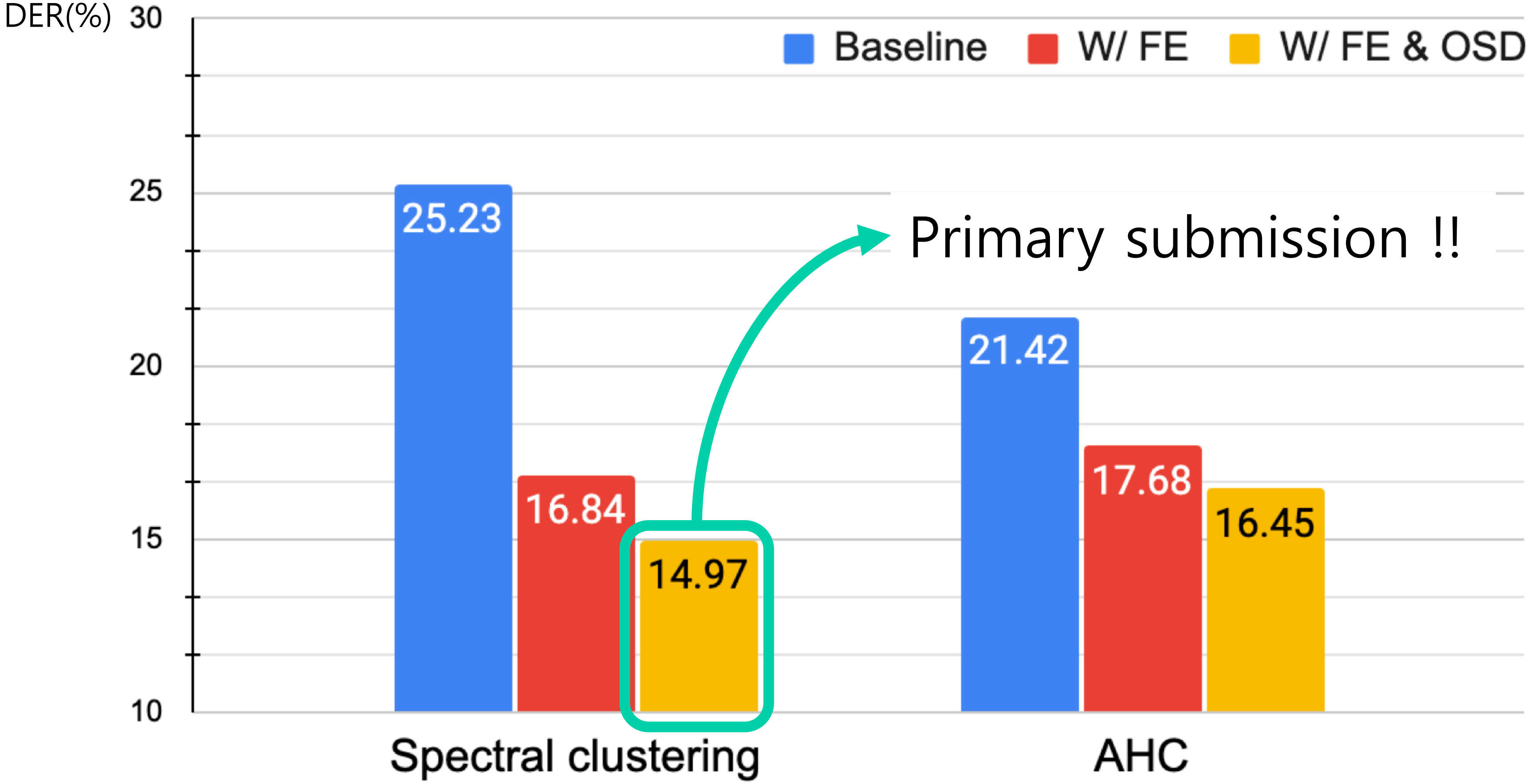
Experimental results

- DER on DIHARD III dev set (track1)



Experimental results

- DER on DIHARD III dev set (track1)



Results from leaderboard

- Performances on DIHARD III eval set (track1)
  - Core: 15.40% DER, 43.07% JER
    ranked 3rd
  - Full: 13.95% DER, 37.43% JER
    ranked 5th

Summary

- Two contribution to step-wise pipeline
  - Overlapped speech detection
    - Ensemble of CRNN-based models
  - Feature enhancement for speaker diarization
    - Dimensionality reduction & attention-based aggregation

- 15.40% DER on core evaluation set (track1)
- 13.95% DER on full evaluation set (track1)