# The DKU-DukeECE Diarization System for VoxSRC 2022

Weiqing Wang[2], Xiaoyi Qin[1], Ming Cheng[1], Yucong Zhang[1], Kangyue Wang[1], Ming Li[1,2]

[1] Data Science Research Center, Duke Kunshan University, Kunshan, China
[2] Department of Electrical and Computer Engineering, Duke University, Durham, USA

- Voice activity detection (VAD)
- Speaker embedding extraction
- Clustering-based method
  - Cosine + Agglomerative Hierarchical Clustering (AHC)
  - LSTM-based similarity measurement + Spectral Clustering (SC)
- Overlap speech detection

- **Voice activity detection (VAD)**
- Speaker embedding extraction
- Clustering-based method
  - Cosine + Agglomerative Hierarchical Clustering (AHC)
  - LSTM-based similarity measurement + Spectral Clustering (SC)
- Overlap Speech Detection (OSD)

- Model 1: ResNet34 + statistical pooling + transformer enc + linear
- Model 2: ResNet50 + convolution subsample + conformer enc + transformer dec
- Model 3: Pretrained *pyannote 2.0*[1].
- Model 4: *Kaldi*[2] ASR

**Table 1:** False alarm (FA), miss detection (MISS) and accuracy of the VAD model on Voxconverse test set

| #Model | FA [%] | MISS [%] | ERROR [%] |
|---|---|---|---|
| 1 | 2.94 | 1.33 | 4.27 |
| 2 | 2.70 | 1.77 | 4.47 |
| 3 | 2.25 | 2.10 | 4.35 |
| 4 | 0.81 | 11.87 | 12.68 |
| Fusion | 2.60 | 1.37 | 3.97 |

[1] https://github.com/pyannote/pyannote-audio/tree/develop.

[2] https://kaldi-asr.org/models/m13.

- Voice activity detection (VAD)
- **Speaker embedding extraction**
- Clustering-based method
  - Cosine + Agglomerative Hierarchical Clustering (AHC)
  - LSTM-based similarity measurement + Spectral Clustering (SC)
- Overlap Speech Detection (OSD)

- SimAM-ResNet34[3] + attentive statistic pooling + Linear + ArcFace
- Trained on Voxceleb2 dev set.
- Finetuned on VoxConverse dev set with pseudo labels.

**Table 2:** The performance of speaker embedding system.

| Model | Vox-O | | VoxSRC22 task4val | |
|---|---|---|---|---|
| | EER[%] | mDCF | EER[%] | mDCF |
| SimAM-ResNet | 0.726 | 0.036 | 5.84 | 0.220 |
| + fine-tune | - | - | 5.08 | 0.335 |

[3] **X. Qin, N. Li, C. Weng, D. Su, and M. Li, Simple attention module based speaker verification with iterative noisy label detection, in ICASSP 2022.**

- Voice activity detection (VAD)
- Speaker embedding extraction
- **Clustering-based method**
  - Cosine + Agglomerative Hierarchical Clustering (AHC)
  - LSTM-based similarity measurement + Spectral Clustering (SC)
- Overlap Speech Detection (OSD)

- Similar to Microsoft system[4] in VoxSRC 2020 without speech separation.
- AHC for segmentation:
  - Uniformly segment speech with a length of 1.28s and shift of 0.32s
  - Iteratively merge two closest consecutive segments with the largest cosine similarity until the preset threshold is reached
- AHC for clustering:
  - Perform a plain AHC on the segments with a relatively high threshold to get the clusters with high confidence
  - Split clusters into "long clusters" and "short clusters" by the total duration in each cluster
  - Assign each short cluster to the closest long cluster, and some short clusters are treated as new speakers if not matching any long clusters.

[4] **X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, S. Chen, Y. Zhao, G. Liu, Y. Wu, J. Wu et al., "Microsoft speaker diariza-tion system for the voxceleb speaker recognition challenge 2020, " in ICASSP, 2021.**

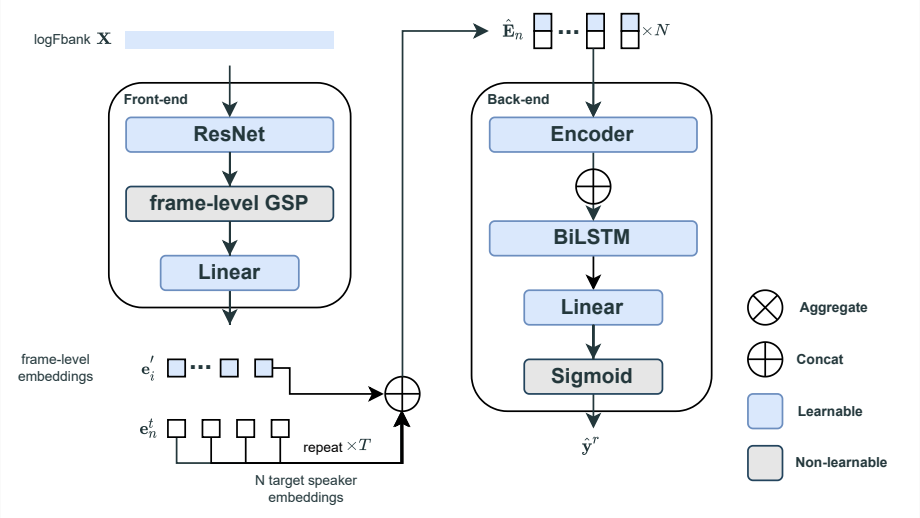
- BiLSTM + Linear + Sigmoid
- Uniformly segmented speech with a length of 1.28s and shift of 0.64s.
- Trained on the mixed training set, fine-tuned on voxconverse dev set, and validated on voxconverse test set
- After obtaining the affinity matrix S, perform spectral clustering on it to get the diarization output

$$\mathbf{S}_i = [\mathbf{S}_{i,1}, \mathbf{S}_{i,2}, ..., \mathbf{S}_{i,n}] = f(\begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_2 \end{bmatrix}, ..., \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_n \end{bmatrix}), \quad (1)$$

[5] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization, " in INTERSPEECH, 2019.

- Voice activity detection (VAD)
- Speaker embedding extraction
- Clustering-based method
  - Cosine + Agglomerative Hierarchical Clustering (AHC)
  - LSTM-based similarity measurement + Spectral Clustering (SC)
- **Overlap Speech Detection (OSD)**

- Plain overlap detection
  - Model architecture is the same as VAD model 1.
- Target-speaker Voice Activity Detection
  - Training
  - Inference


logFbank $\mathbf{X}$
Front-end
ResNet
frame-level GSP
Linear
frame-level embeddings
$\mathbf{e}_i^t$
$\mathbf{e}_n^t$
repeat $\times T$
N target speaker embeddings
$\hat{\mathbf{E}}_n$ $\times N$
Back-end
Encoder
BiLSTM
Linear
Sigmoid
$\hat{\mathbf{y}}^r$
Aggregate
Concat
Learnable
Non-learnable

- Training
  - Initialize the ResNet34 with the parameters from pre-trained speaker embedding model.
  - Pre-trained on Simulated Librispeech with front-end frozen and then unfrozen.
  - Finetuned on voxconverse dev set.
  - Validated on voxconverse test set.
  - Data augmentation is performed with MUSAN and RIRs.
- Inference
  - Fully assigning:
    - The TS-VAD output is the final results.
    - Keep the AHC results of speakers with short speech.
  - Partially assigning:
    - Only replace the overlap regions detected by TS-VAD.

**Table 3:** The performance of different speaker diarization systems in terms of DER (%) and JER (%).

| Model | Test (Oracle VAD) | | Test (System VAD) | | VoxSRC-22 Test | |
|---|---|---|---|---|---|---|
| | DER[%] | JER[%] | DER[%] | JER[%] | DER[%] | JER[%] |
| Baseline | - | - | - | - | 19.60 | 41.43 |
| AHC | 3.36 | 21.67 | 5.35 | 27.99 | - | - |
| + *OD* | 3.03 | 21.43 | 5.02 | 27.72 | - | - |
| + *TS-VAD (fully assigned)* | 3.60 | 22.21 | 5.61 | 28.08 | - | - |
| + *TS-VAD (partially assigned)* | **2.96** | 21.77 | **4.86** | 27.69 | 4.85 | 28.05 |
| LSTM-SC | 4.91 | 32.74 | 6.36 | 34.82 | - | - |
| + *OD* | 4.39 | 32.02 | 6.04 | 34.53 | - | - |
| + *TS-VAD (fully assigned)* | 4.12 | 31.70 | 5.68 | 33.92 | - | - |
| + *TS-VAD (partially assigned)* | 4.31 | 32.14 | 5.85 | 34.30 | - | - |
| Fusion | 3.09 | 23.14 | 4.94 | 28.79 | **4.74** | 27.84 |

Duke

**Table 4:** The performance of different speaker diarization systems in terms of DER (%) and JER (%).

| Model | Test (Oracle VAD) | | Test (System VAD) | | VoxSRC-22 Test | |
|---|---|---|---|---|---|---|
| | DER[%] | JER[%] | DER[%] | JER[%] | DER[%] | JER[%] |
| Baseline | - | - | - | - | 19.60 | 41.43 |
| AHC | 3.36 | 21.67 | 5.35 | 27.99 | - | - |
| + *OD* | 3.03 | 21.43 | 5.02 | 27.72 | - | - |
| + *TS-VAD (fully assigned)* | 3.60 | 22.21 | 5.61 | 28.08 | - | - |
| + *TS-VAD (partially assigned)* | **2.96** | 21.77 | **4.86** | 27.69 | 4.85 | 28.05 |
| LSTM-SC | 4.91 | 32.74 | 6.36 | 34.82 | - | - |
| + *OD* | 4.39 | 32.02 | 6.04 | 34.53 | - | - |
| + *TS-VAD (fully assigned)* | 4.12 | 31.70 | 5.68 | 33.92 | - | - |
| + *TS-VAD (partially assigned)* | 4.31 | 32.14 | 5.85 | 34.30 | - | - |
| Fusion | 3.09 | 23.14 | 4.94 | 28.79 | **4.74** | 27.84 |

- DER reduction compared with last year:
  - VAD: about 0.3%
  - AHC with better embedding: about 0.2%
  - TS-VAD: about 0.1%
- Fusion:
  - The most difficult part to be tuned.
  - DER mismatch between voxconverse test set and VoxSRC-22 test set.