基于深度编码的说话人日志
Deep Embedding based Speaker Diarization

李明
昆山杜克大学大数据研究中心SMIIP实验室
Speech and Multimodal Intelligent Information Processing Lab (SMIIP)
Data Science Research Center
Duke Kunshan University
Oct 27th 2021

SMIIP

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

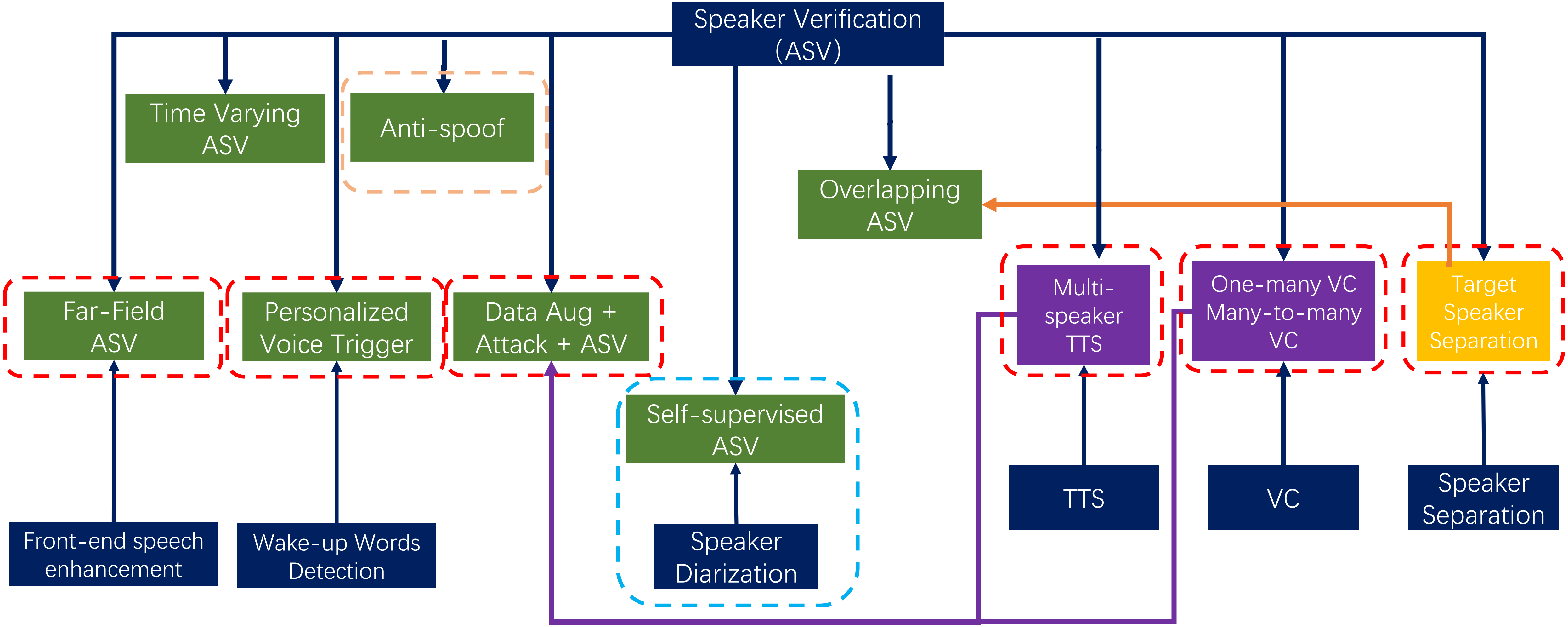
- **DUKE KUNSHAN UNIVERSITY**

- Sino-US Joint Venture University with independent legal status
- Duke-standard education and research
- Comprehensive and small

+

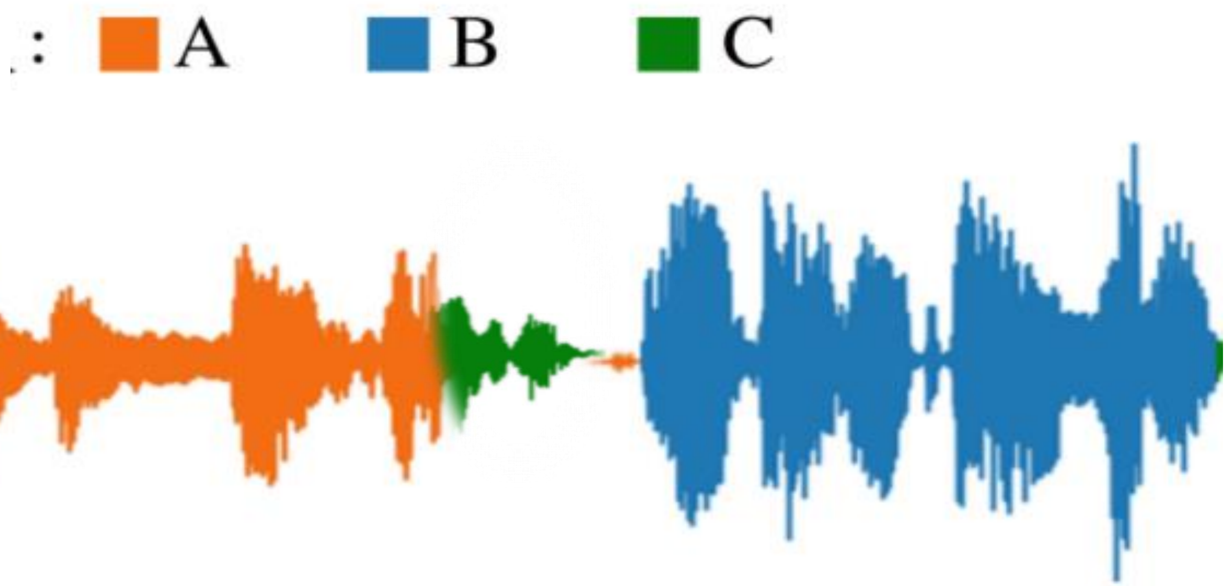
+

Speaker Verification Related Research Topics



Introduction of Speaker Diarization
a Who-Spoke-When problem

Speakers

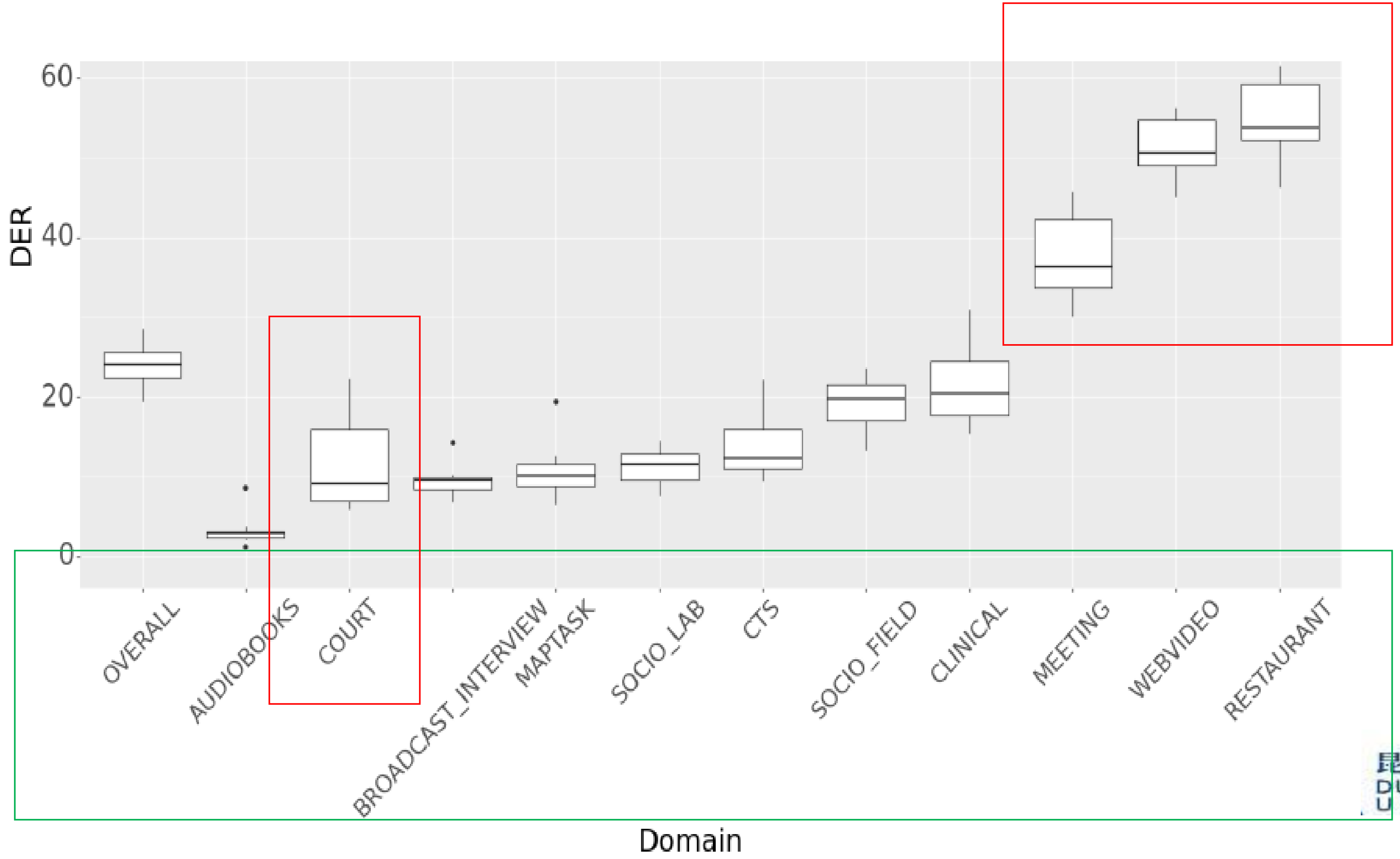


Potential applications / Dihard3 test data

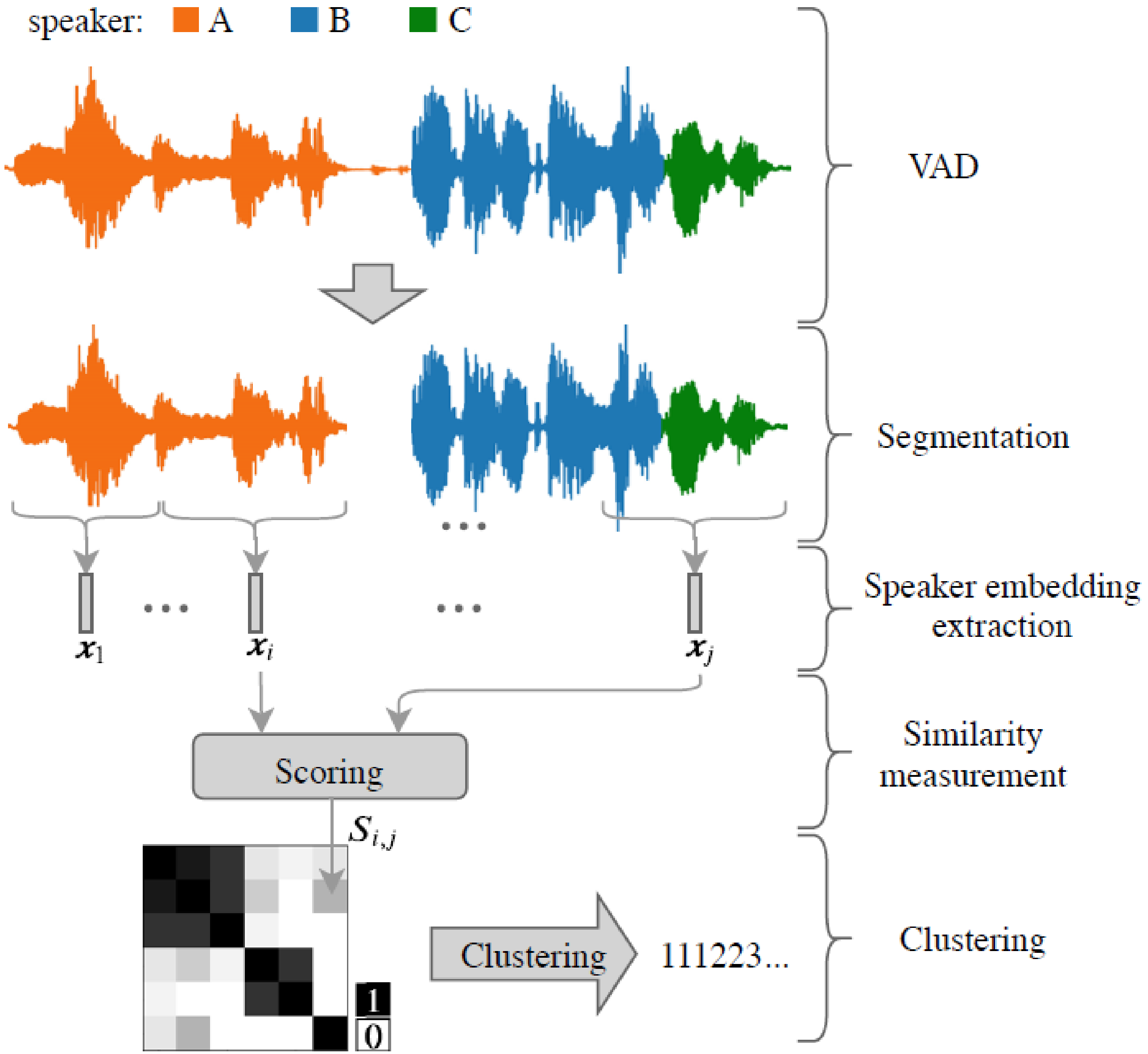
Domain	#Speakers	#Recordings	Duration of full set (h)	Duration of core set (h)	Overlap ratio (%)
Audiobooks	1	12	2.01	2.01	0
Broadcast interview	$3 \sim 5$	12	2.06	2.06	1.2
Clinical	2	48	2.06	4.27	4.8
Courtroom	$5 \sim 10$	12	2.08	2.08	1.9
CTS	2	61	2.17	10.17	13.6
Map task	2	23	2.53	2.53	2.9
Meeting	$3 \sim 10$	14	2.45	2.45	28.9
Restaurant	$5 \sim 8$	12	2.03	2.03	33.7
socio_field	$2 \sim 6$	12	2.01	2.01	8.1
socio_lab	2	16	2.67	2.67	5.0
Web video	$1 \sim 9$	32	1.89	1.89	27.7
Total	-	254	23.94	34.15	12.2

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Potential applications / Dihard3 track 2 results



Pipeline style modular system

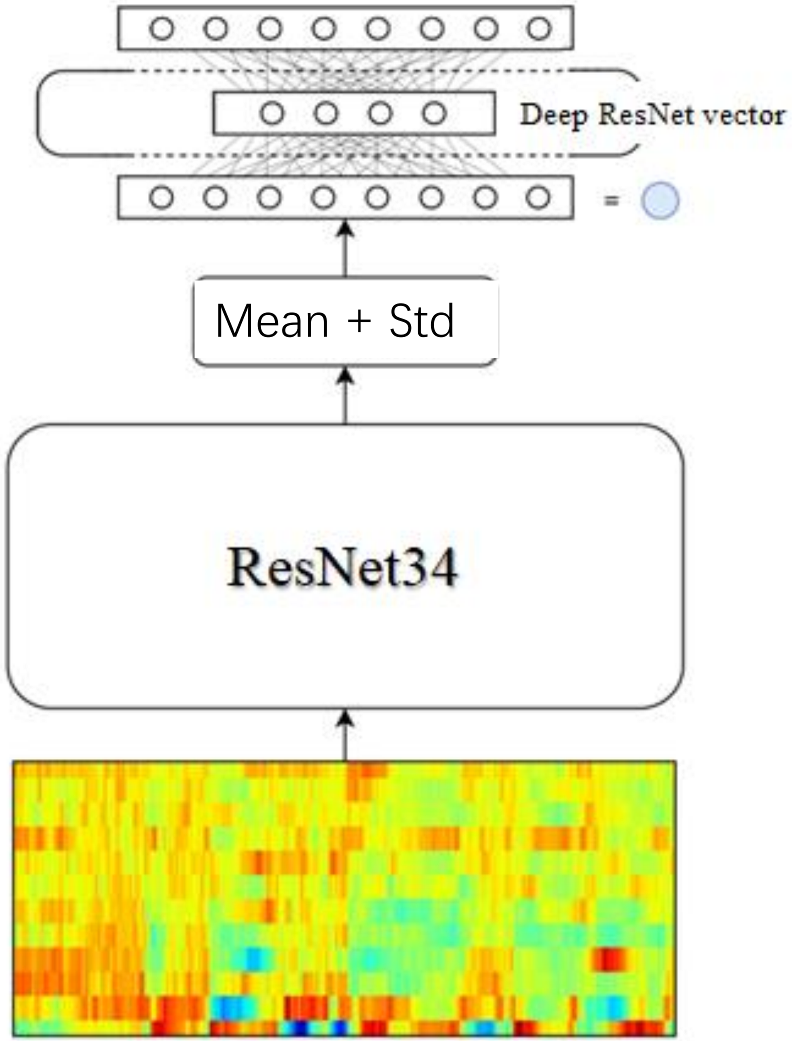


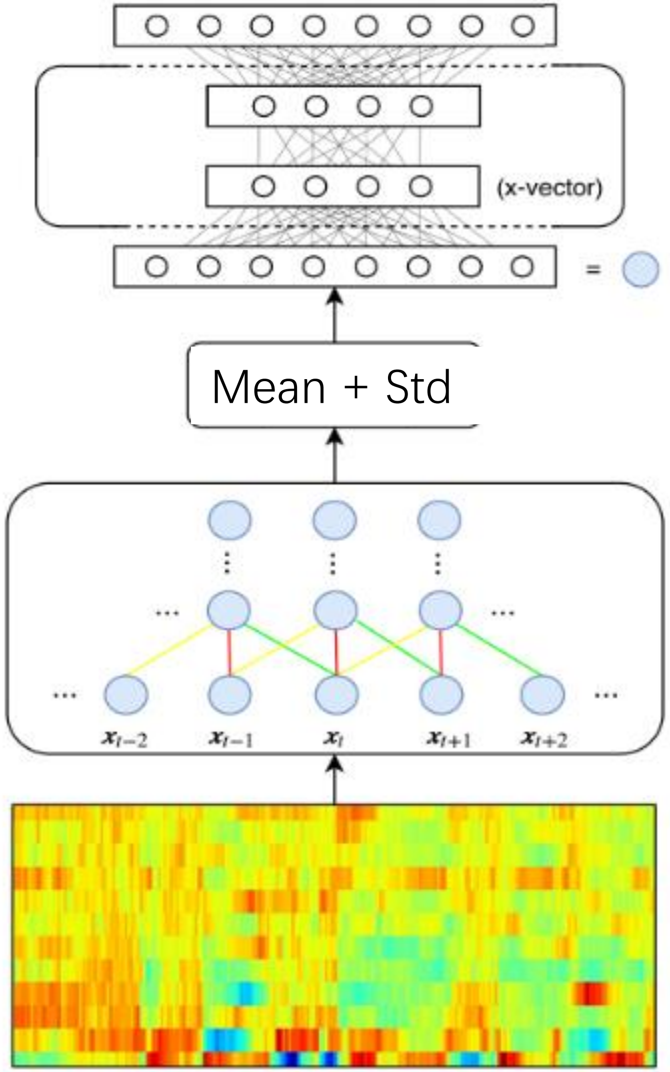
VAD, Segmentation (speaker change point detection), Speaker embedding extraction (e2e SV), similarity measurement (PLDA modeling)

Are these models are trained in the supervised manner (except the clustering)

Qingjian Lin, Weicheng Cai, Lin Yang, Junjie Wang, Jun Zhang, Ming Li, “DIHARD II is Still Hard: Experimental Results and Discussions from the DKU-LENOVO Team”, Odyssey 2020

Speaker Embedding Extraction





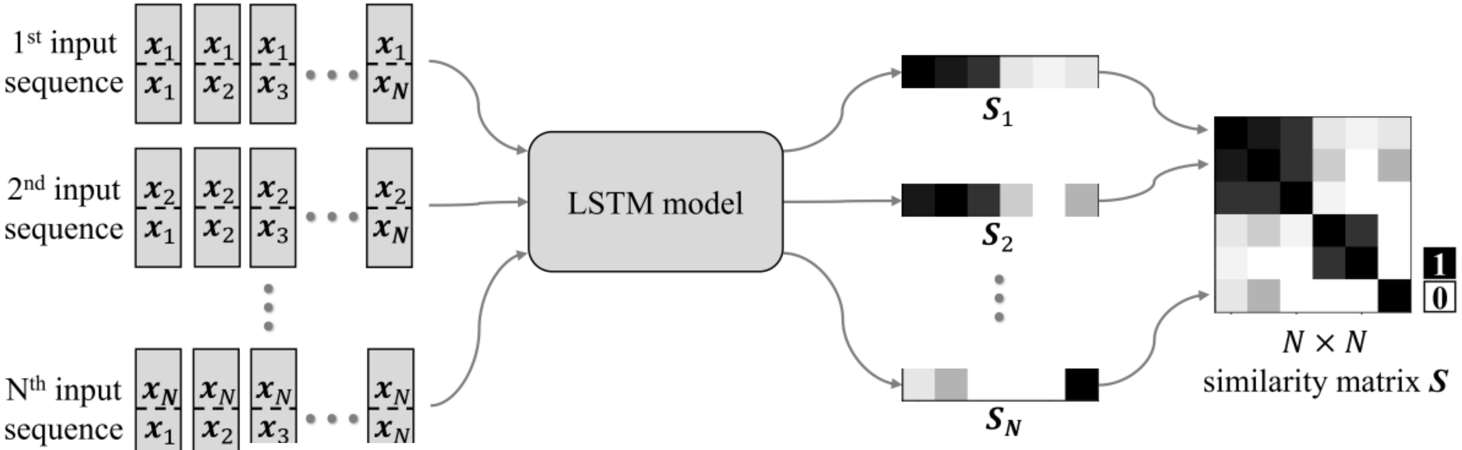
Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "X-vectors: Robust dnn embeddings for speaker recognition." Proc. of ICASSP, pp. 5329-5333, 2018.

Cai, Weicheng, Jinkun Chen, and Ming Li. "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System." In Proc. Odyssey, pp. 74-81. 2018.

Estimating the similarity matrix

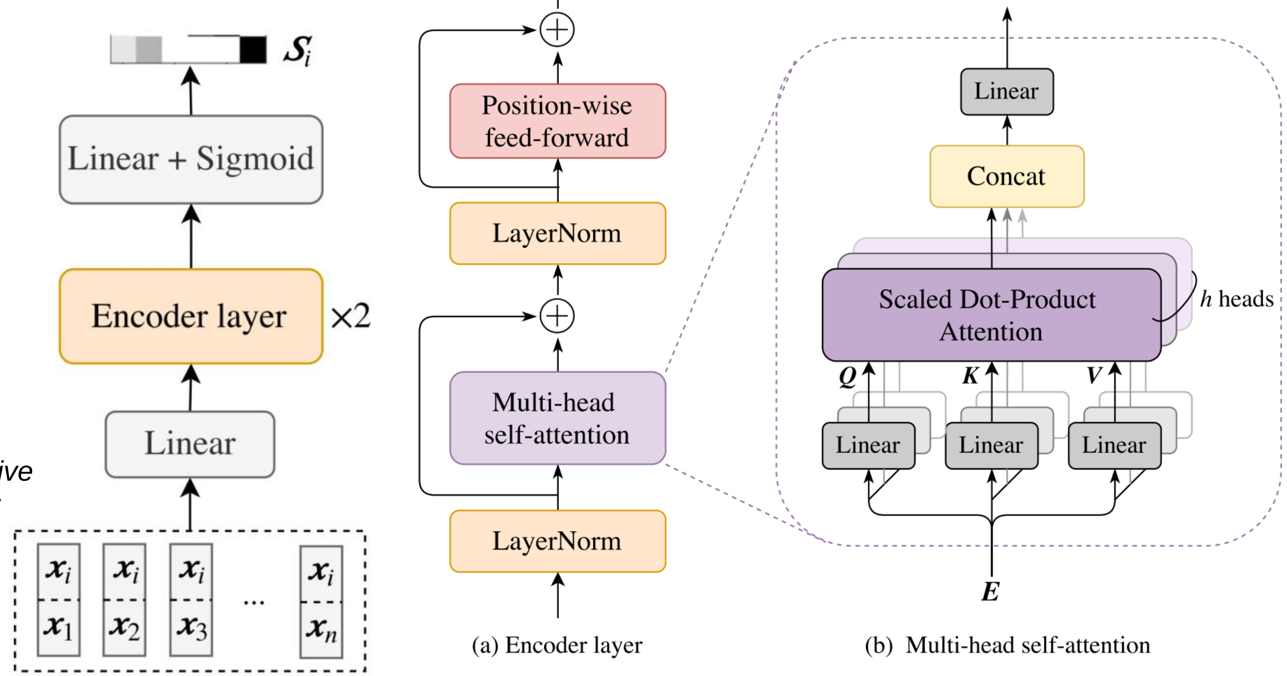
LSTM based scoring

Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin and Claude Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization", Interspeech 2019.

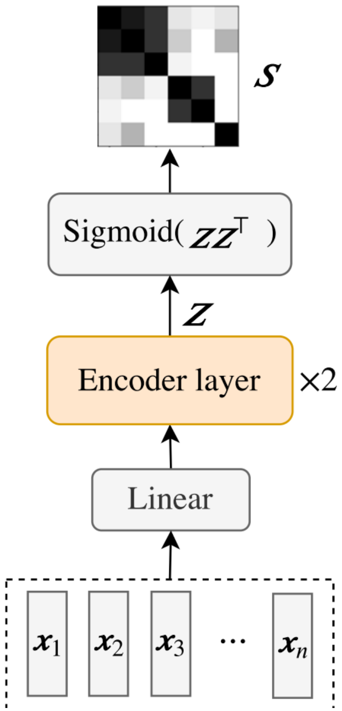


Attention based scoring

Qingjian Lin, Yu Hou and Ming Li, "Self-Attentive Similarity Measurement Strategies in Speaker Diarization", Interspeech 2020.



Attention vector-to-sequence



Attention sequence-to-sequence

Estimating the similarity matrix

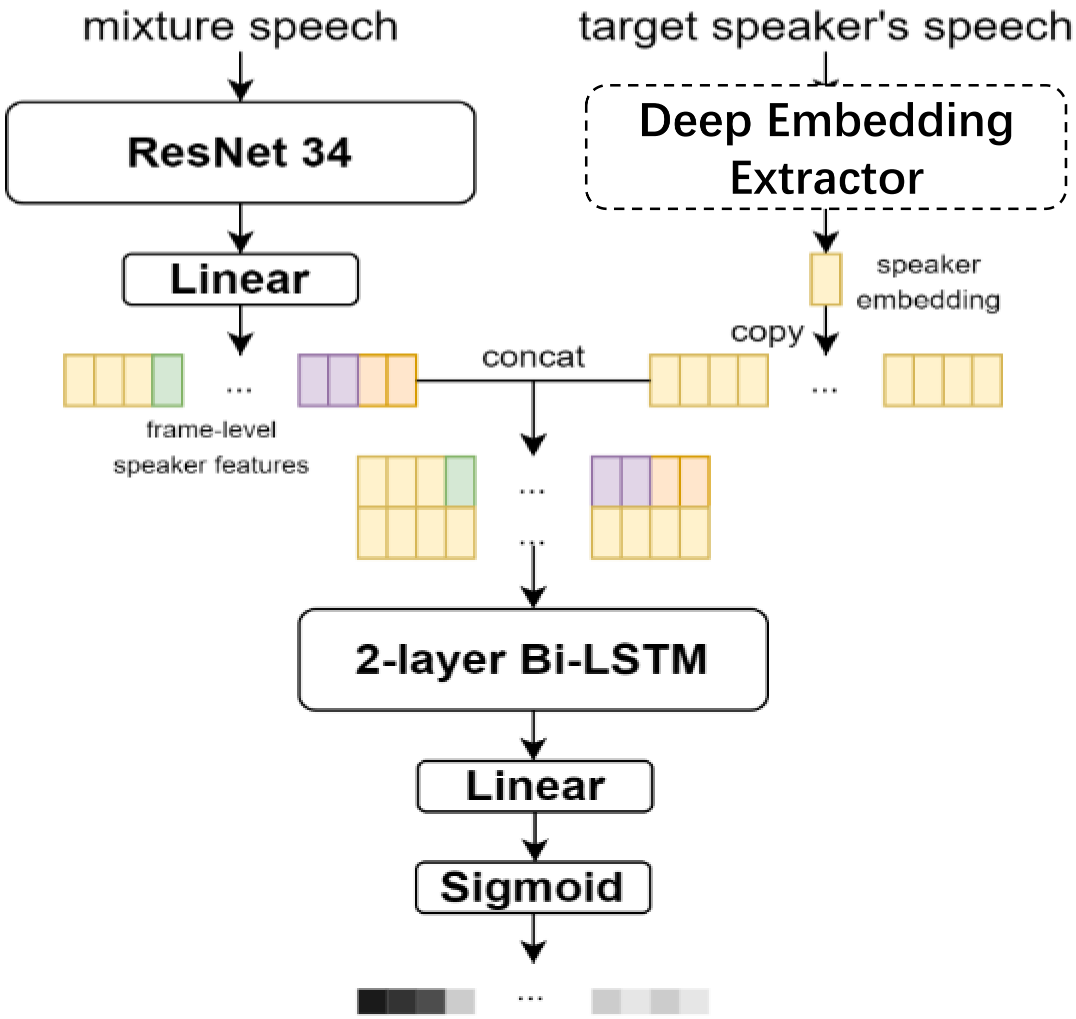
Results on Dihard II task 1

Table 2: *Evaluation on DIHARD II corpus. Results are reported with and without domain adaptation by the Dev Set.*

Model	+VB	Dev		Eval		Eval + adaptation		Time cost (Eval)
		DER(%)	JER(%)	DER(%)	JER(%)	DER(%)	JER(%)	
LSTM	×	19.65	49.60	20.57	50.25	19.72	46.49	67 min
	√	19.48	49.21	19.98	49.42	19.26	45.91	-
Att-v2s	×	**19.07**	**47.43**	**20.15**	**47.84**	**18.98**	43.20	148 min
	√	**18.76**	**46.77**	**19.46**	**47.01**	**18.44**	42.52	-
Att-s2s	×	19.39	48.42	21.46	48.71	21.45	**43.19**	24 s
	√	19.16	47.99	20.78	47.92	20.12	**41.73**	-
PLDA	×	23.48	57.17	-	-	23.73	56.84	51 s
DIHARD II winner system [27]						18.42	44.58	
DIHARD II official baseline [28]						25.99	59.51	

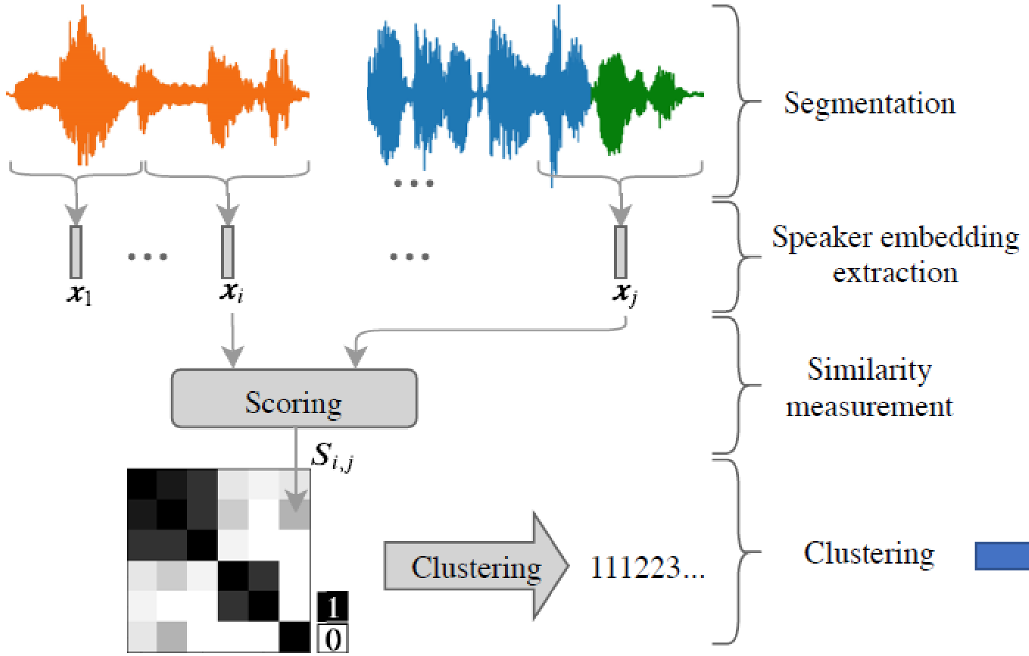
Lin, et.al, “Self-Attentive Similarity Measurement Strategies in Speaker Diarization”, Interspeech 2020.

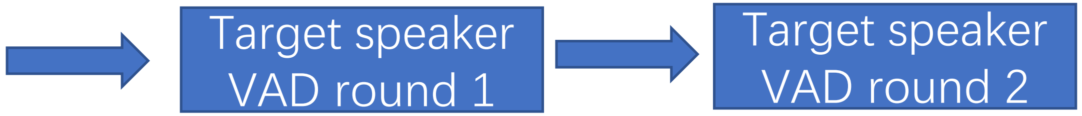
Target-speaker VAD based post processing



Weiqing Wang, Qingjian Lin, Danwei Cai, Lin Yang, Ming Li, "The DKU-Duke-Lenovo System Description for the Third DIHARD Speech Diarization Challenge", the Third DIHARD Speech Diarization Challenge Workshop, 2021

Target-speaker VAD based post processing





Results on DIHARD3 CTS data

Training data	Finetune data	Testing data	Methods	DER
N/A	CTS-dev-41	CTS-dev-20	X-vector + Spectral Cluster	15.07%
SRE+SWBD	N/A	CTS-dev-20	+ target speaker vad round 1	10.60%
SRE+SWBD	CTS-dev-41	CTS-dev-20	+ target speaker vad round 1	7.80%
SRE+SWBD	CTS-dev-41	CTS-dev-20	+ target speaker vad round 2	7.63%

Weiqing Wang, Qingjian Lin, Danwei Cai, Lin Yang, Ming Li, "The DKU-Duke-Lenovo System Description for the Third DIHARD Speech Diarization Challenge", the Third DIHARD Speech Diarization Challenge Workshop, 2021

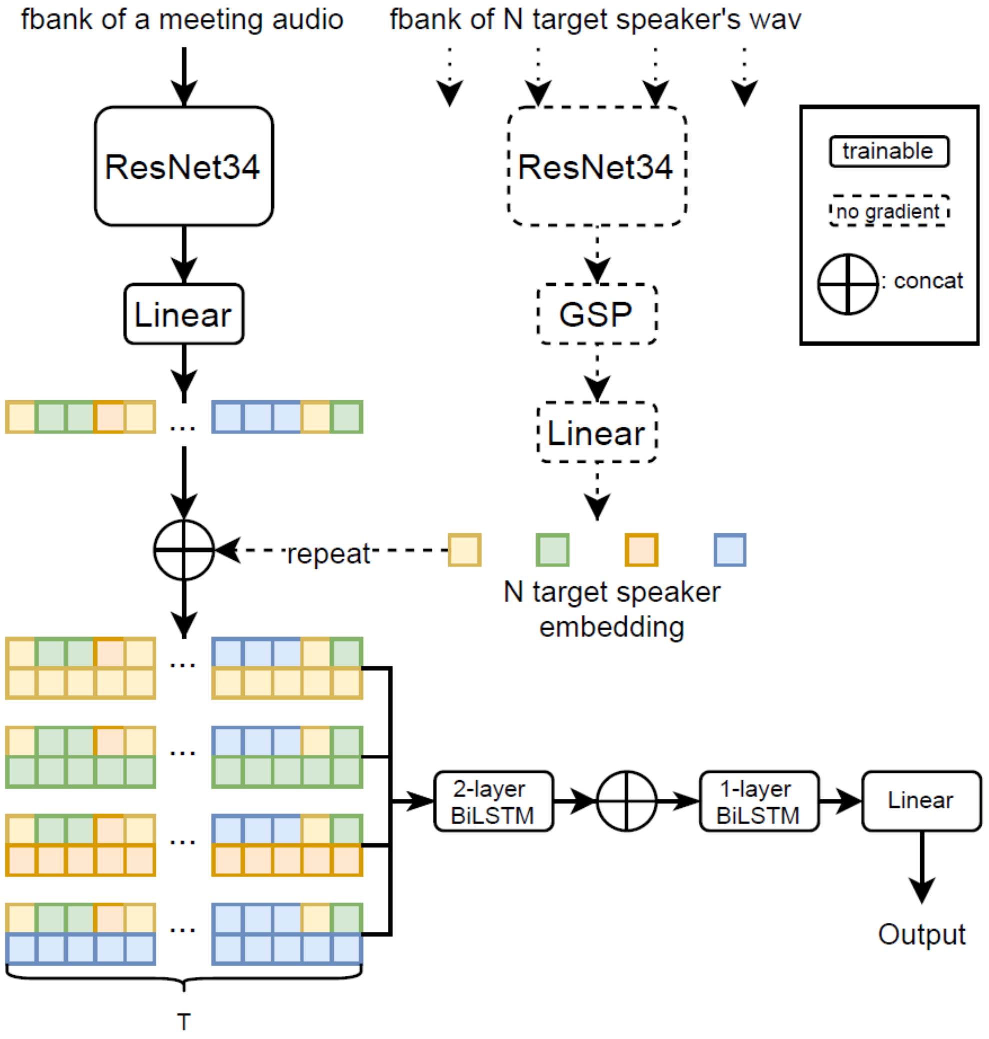
Target-speaker VAD based post processing

Results on DIHARD3 full test data

	Dataset	Method	DER on full set (%)	DER on core set (%)
Track1	NCTS (adapt) & CTS	att-v2s + SC & Cosine + AHC	16.34	17.03
	NCTS (adapt) & CTS (adapt)	att-v2s + SC & TSVAD round 2	13.39	15.43
Track2	NCTS (adapt) & CTS	att-v2s + SC & Cosine + AHC	-	-
	NCTS (adapt) & CTS (adapt)	att-v2s + SC & TSVAD round 2	18.90	21.63

Weiqing Wang, Qingjian Lin, Danwei Cai, Lin Yang, Ming Li, "The DKU-Duke-Lenovo System Description for the Third DIHARD Speech Diarization Challenge", the Third DIHARD Speech Diarization Challenge Workshop, 2021

Another target-speaker VAD based post processing



Weiqing Wang, Danwei Cai, Jin Wang, Qingjian Lin, Xuyang Wang, Mi Hong, Ming Li, "The DKU-Duke-Lenovo System Description for the Fearless Steps Challenge Phase III", INTERSPEECH, 2021

Target-speaker VAD based post processing

Results on the fearless step challenge phase III dataset

Model	Dev		Eval	
	Track 1	Track 2	Track 1	Track 2
1 LSTM	21.48	13.56	-	-
2 Att-v2s	22.57	15.11	-	-
3 AHC (uni-seg)	20.83	13.33	-	-
4 AHC (ahc-seg)	21.39	14.21	-	-
5 TSVAD (round 0)	20.75	11.88	43.99	13.85
6 TSVAD (round 1)	20.94	11.99	-	-
Fusion (1+2+3+4)	20.39	12.70	44.56	14.63
Fusion (1+2+3+4+5)	-	11.81	-	12.83
Fusion (3+4+5)	**19.19**	**11.40**	**42.21**	**12.32**

Weiqing Wang, Qingjian Lin, Danwei Cai, Lin Yang, Ming Li, "The DKU-Duke-Lenovo System Description for the Third DIHARD Speech Diarization Challenge", the Third DIHARD Speech Diarization Challenge Workshop, 2021

New results

DER (%) OF DIFFERENT SIMILARITY MEASUREMENT MODELS. EMBD-AUG IS THE DATA AUGMENTATION PERFORMED ON THE SPEAKER EMBEDDING, SP IS SEGMENTAL POOLING, AND JT DENOTES JOINTLY TRAINING.

Model	DIHARD II			DIHARD III			VoxConverse		
	Dev	Eval	Eval (+dev adapt)	Dev	Eval	Eval (+dev adapt)	Dev	Eval	Eval (+dev adapt)
BiLSTM	24.15	25.59	19.92	21.03	20.10	17.03	13.29	17.88	12.45
+ embd aug	17.44	18.25	18.12	16.15	15.85	15.62	4.50	6.91	6.70
+ SP	17.34	17.81	17.80	16.11	15.61	15.45	4.47	6.02	**4.57**
+ JT	-	-	**17.76**	-	-	**15.18**	-	-	4.63
Self-att	20.97	22.48	19.99	19.99	19.20	16.84	10.29	14.08	10.21
+ embd aug	18.00	18.71	18.41	16.47	16.04	15.85	7.10	9.26	7.25
+ SP	17.97	18.76	**18.00**	16.52	16.00	**15.65**	6.06	7.95	**5.67**
Official baseline	-	-	25.99 [61]	-	-	19.25 [62]	-	-	-
Winner system (Clustering)	-	-	18.42 [20]	-	-	15.47 [65]	-	-	-

Weiqing Wang, Qingjian Lin, Danwei Cai, **Ming Li** (*), “Segment-level Speaker Embedding Similarity Measurement in Speaker Diarization”, submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing

New results

Winner of VoxSRC 2021 speaker diarization track

TABLE III

DER (%) OF THE SEGMENT-LEVEL TS-VAD MODELS ON EVALUATION DATASET (N=8, FULLY ASSIGNED)

Pooling Size	DIHARD II				DIHARD III				VoxConverse			
	MISS(%)	FA(%)	SpkErr(%)	DER(%)	MISS(%)	FA(%)	SpkErr(%)	DER(%)	MISS(%)	FA(%)	SpkErr(%)	DER(%)
s=1 (80ms)	8.2	0.7	7.6	**16.48**	6.6	0.7	4.4	**11.62**	1.0	0.2	3.7	4.95
s=2 (160ms)	8.1	1.4	7.7	17.20	6.1	1.6	4.4	12.09	1.0	0.2	3.8	5.04
s=4 (320ms)	8.1	1.4	8.3	17.83	6.3	1.8	4.9	12.97	1.0	0.3	3.5	**4.72**
s=8 (640ms)	8.2	1.7	9.2	19.06	6.7	2.1	5.9	14.67	1.0	0.3	3.6	4.78
Clustering	9.7	0.0	8.1	17.76	9.5	0.0	5.7	15.18	1.6	0.0	3.0	4.57
Winner System (TS-VAD)	-	-	-	-	-	-	-	12.30 [65]	-	-	-	-

TABLE IV

DER (%) OF THE SEGMENT-LEVEL TS-VAD AS OVERLAP DETECTION ON EVALUATION DATASET (N=2, PARTIALLY ASSIGNED OVERLAPPED REGION)

Pooling Size	DIHARD II				DIHARD III				VoxConverse			
	MISS(%)	FA(%)	SpkErr(%)	DER(%)	MISS(%)	FA(%)	SpkErr(%)	DER(%)	MISS(%)	FA(%)	SpkErr(%)	DER(%)
s=1 (80ms)	8.0	1.0	8.1	**17.19**	5.7	1.5	5.7	**12.89**	1.1	0.3	3.1	**4.39**
s=2 (160ms)	7.9	2.1	7.9	17.94	5.4	2.8	5.4	13.57	1.0	0.5	3.0	4.49
s=4 (320ms)	8.2	1.9	7.9	17.98	5.7	2.8	5.3	13.77	1.1	0.3	3.0	4.40
s=8 (640ms)	8.3	1.7	8.0	18.00	6.2	3.1	5.3	14.49	1.0	0.5	3.0	4.52
Clustering	9.7	0.0	8.1	17.76	9.5	0.0	5.7	15.18	1.6	0.0	3.0	4.57

Weiqing Wang, Qingjian Lin, Danwei Cai, **Ming Li** (*), “Segment-level Speaker Embedding Similarity Measurement in Speaker Diarization”, submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing

Modularized Online Speaker Diarization

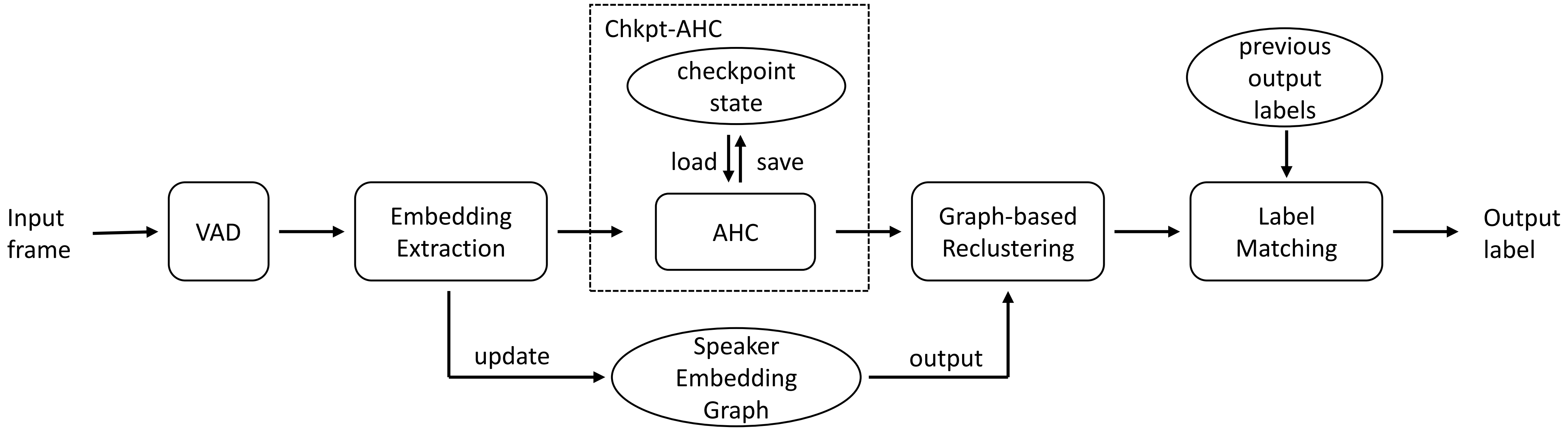


Fig 1 The pipeline of the purposed system

Yucong Zhang, Qinjian Lin, Weiqing Wang, Lin Yang, Xuyang Wang, Junjie Wang, Ming Li, "Online speaker diarization with graph-based label generation", submitted to ICASSP 2022.

Chkpt-AHC

Problem:

- Agglomerative hierarchy clustering causes high time complexity

Solution:

- Save the intermediate state of AHC
- Starting from limited number of clusters

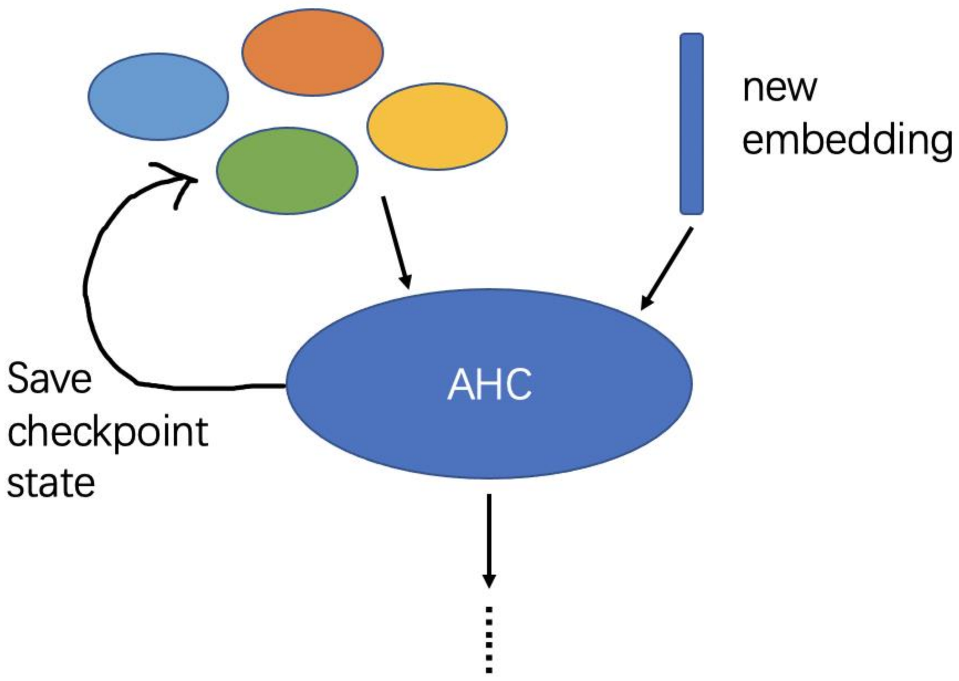


Fig 2 chkpt-AHC

Graph-based Reclustering

Speaker embedding graph

- Nodes represent speaker embeddings, **N** = {n_A,n_B,…,}
- Node n_K and n_L has edge e_{KL} if similarity between embedding K and L greater than pre-defined threshold θ, weight equals the similarity
- Graph pruning to reduce the time complexity

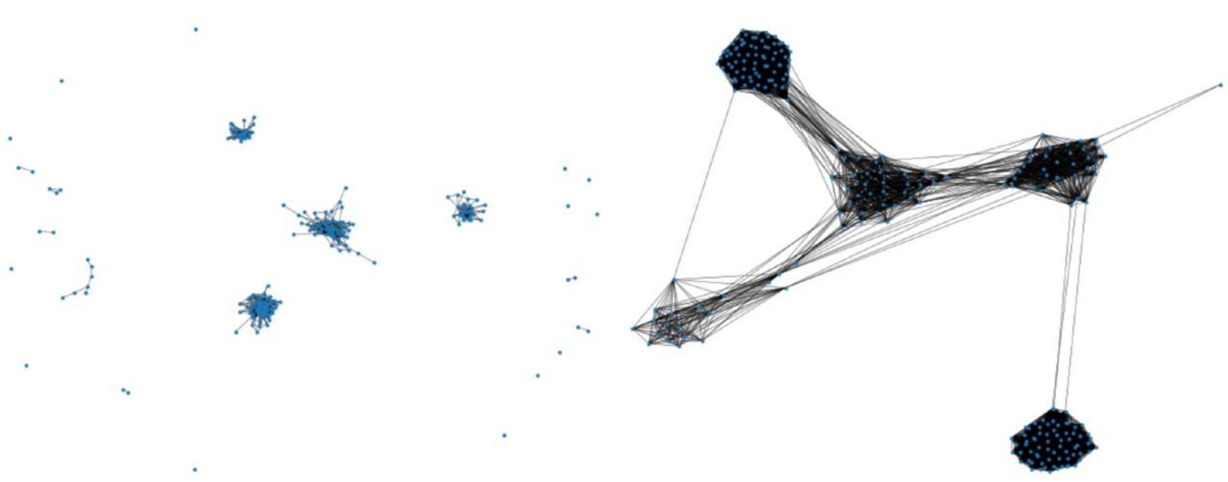


Fig 3 speaker embedding graph (threshold: left 0.6, right 0.3)

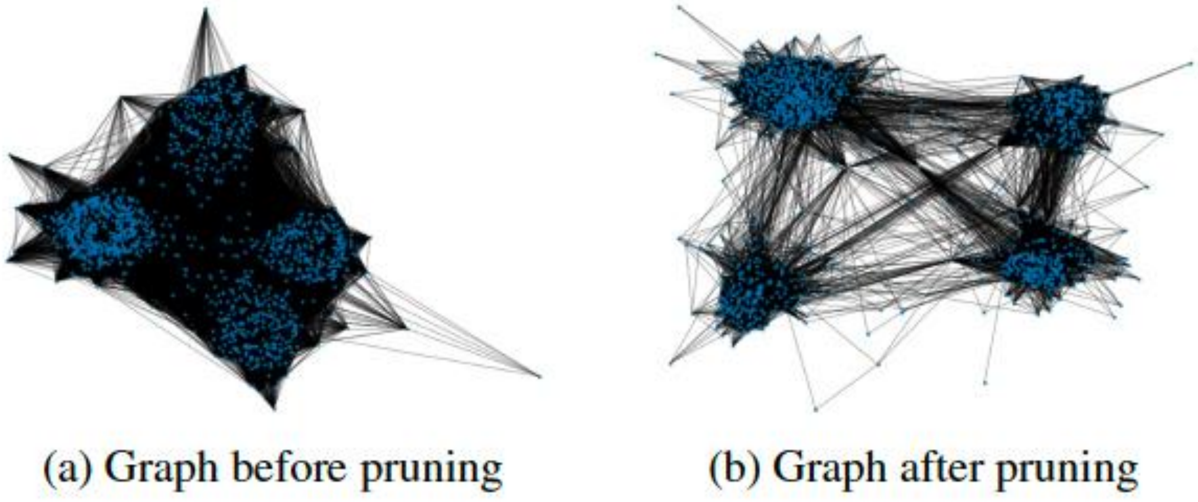


Fig 4 Effects of Graph pruning with threshold=0.3

Graph-based Reclustering

Problem:

- After chkpt-AHC, due to high stopping criteria, lots of small clusters left behind

Solution:

- Use smaller threshold to build speaker embedding graph
- Assign remaining embeddings to speaker clusters based on cluster likelihood

$$\mathcal{L}_{C_j}^{(i)} = \frac{\sum_{n_k \in C_j} w_{ik}}{|C_j|}$$

where C_j represents jth speaker cluster,
w_{ik} represents the weight of edge $e_{ik} \in E$

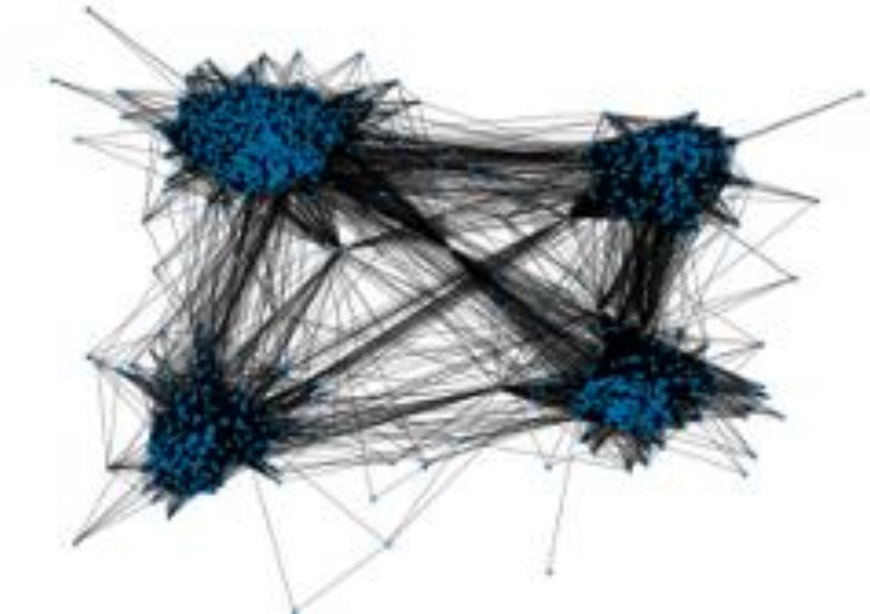


Fig 5 speaker embedding graph

Label Matching

Problem:

- Label consistency

Solution:

- Construct bipartite graph
- Hungarian Algorithm

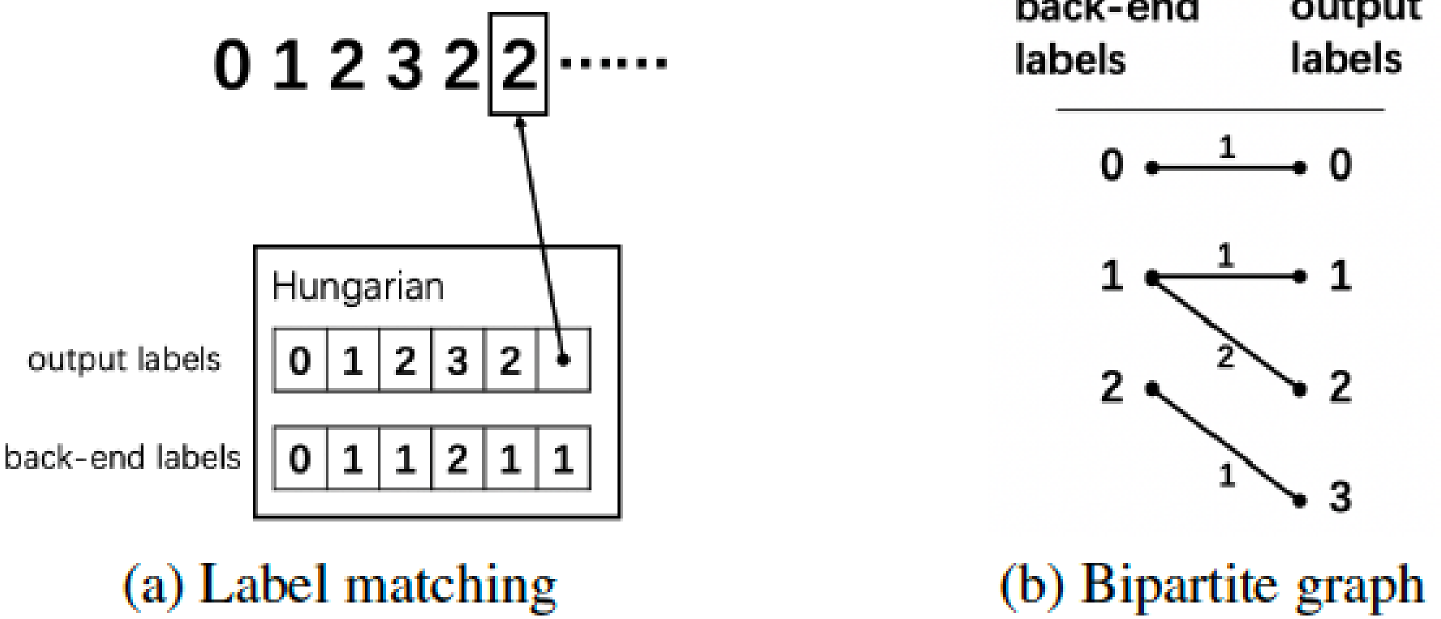


Fig 4 Label matching with Hungarian Algorithm

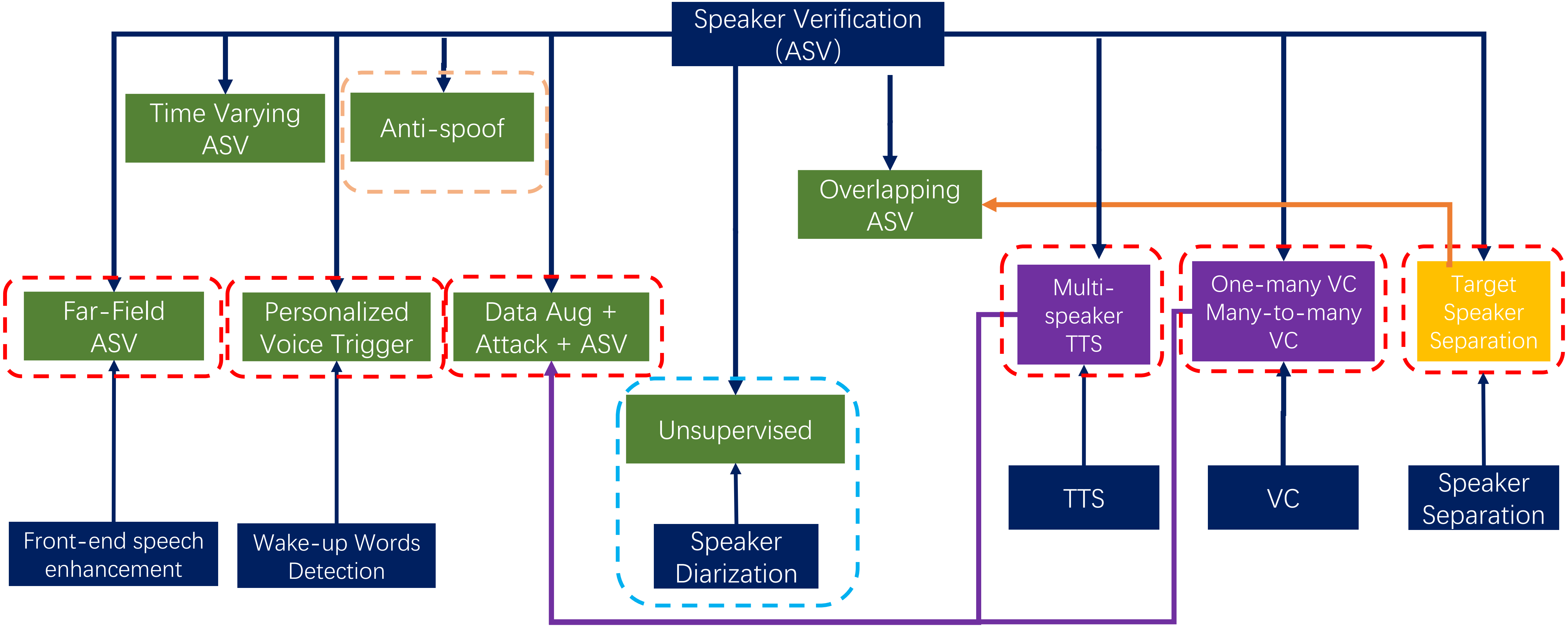
Demo

Results

Table 1. The DER (%) of the proposed speaker diarization system. The baseline system is introduced by [23] for DIHARD3 competition without VB-HMM resegmentation. System 1 is the offline version of our proposed diarization system.

System	Offline	Online	AHC	Chkpt-AHC	Naive Reclustering	Graph-based Reclustering	DIHARD3		VoxConverse	
							Dev	Eval	Dev	Eval
Baseline	√	-	-	-	-	-	20.71	20.75	-	-
1	√	-	√	-	√	-	17.63	16.82	3.94	4.68
2	-	√	√	-	√	-	20.17	19.68	5.20	6.28
3	-	√	-	√	√	-	20.78	20.05	5.91	6.71
4	-	√	-	√	-	√	**20.28**	**19.57**	**5.80**	**6.60**

Review


Speaker Verification (ASV)
Time Varying ASV
Anti-spoof
Overlapping ASV
Far-Field ASV
Personalized Voice Trigger
Data Aug + Attack + ASV
Multi-speaker TTS
One-many VC Many-to-many VC
Target Speaker Separation
Unsupervised
Speaker Diarization
Front-end speech enhancement
Wake-up Words Detection
TTS
VC
Speaker Separation

Thank you very much!

ming.li369@duke.edu

https://scholars.duke.edu/person/MingLi