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

李明
语音及多模态智能信息处理实验室
昆山杜克大学大数据研究中心
Speech and Multimodal Intelligent Information Processing Lab (SMIIP)
Data Science Research Center
Duke Kunshan University

July 10th 2021

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

DUKE KUNSHAN UNIVERSITY

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

Duke UNIVERSITY + WUHAN UNIVERSITY 1893 + Kunshan Government

DUKE KUNSHAN UNIVERSITY

Outlines

- Introduction of Speaker Diarization
- Deep learning based speaker diarization with supervised training
  - Modular system
  - End-to-end system
- Self-supervised speaker representation learning

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Introduction of Speaker Diarization
a Who-Spoke-When problem

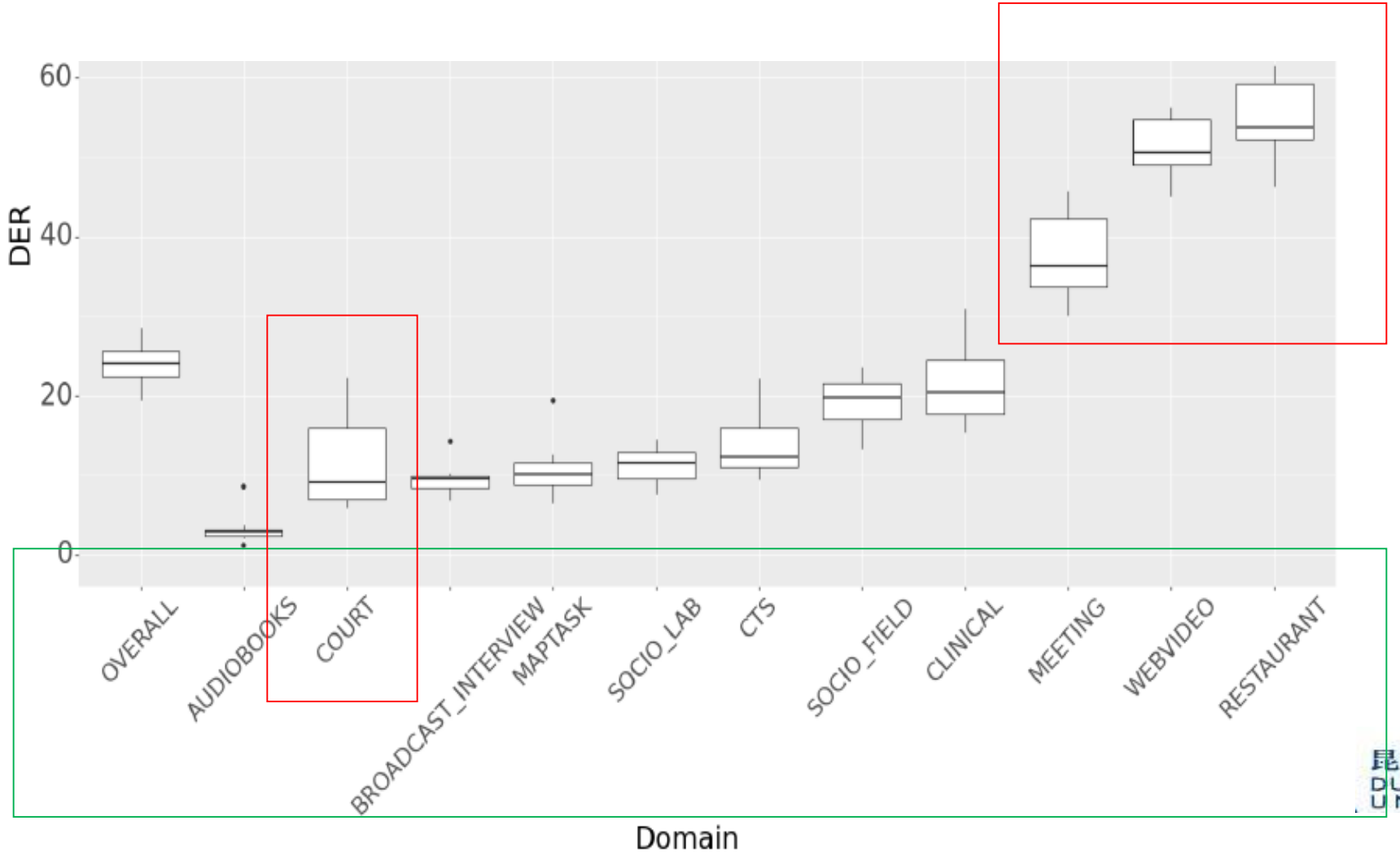
Speakers : A B C

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 ~ 5	12	2.06	2.06	1.2
Clinical	2	48	2.06	4.27	4.8
Courtroom	5 ~ 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 ~ 10	14	2.45	2.45	28.9
Restaurant	5 ~ 8	12	2.03	2.03	33.7
socio_field	2 ~ 6	12	2.01	2.01	8.1
socio_lab	2	16	2.67	2.67	5.0
Web video	1 ~ 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

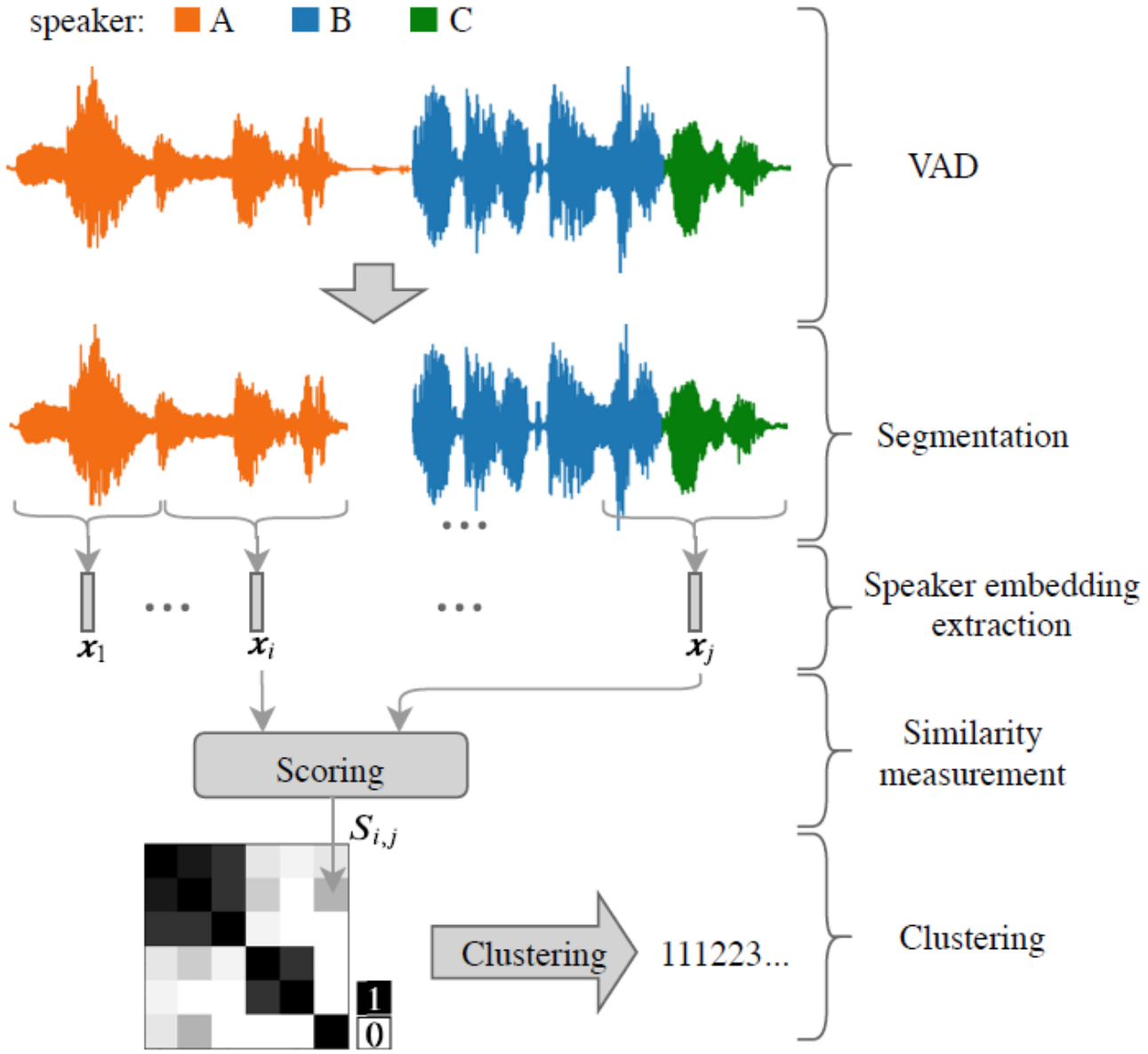
Potential applications / Dihard3 track 2 results



昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Deep learning based speaker diarization with supervised training

Pipeline style modular system



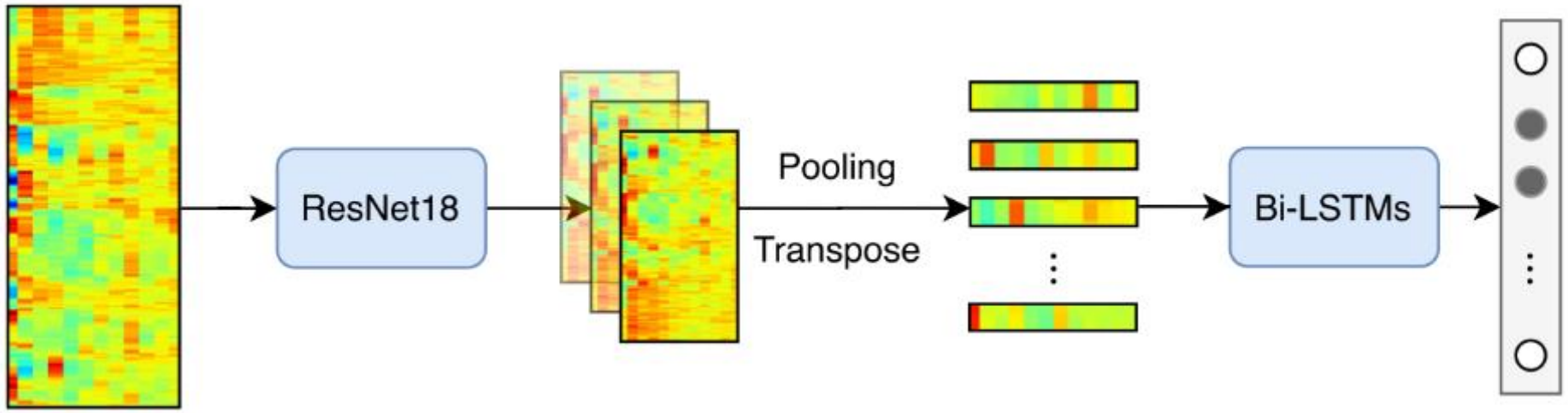
VAD, Segmentation (speaker change point detection), Speaker embedding extraction (e2e SV), similarity measurement (PLDA modeling)

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

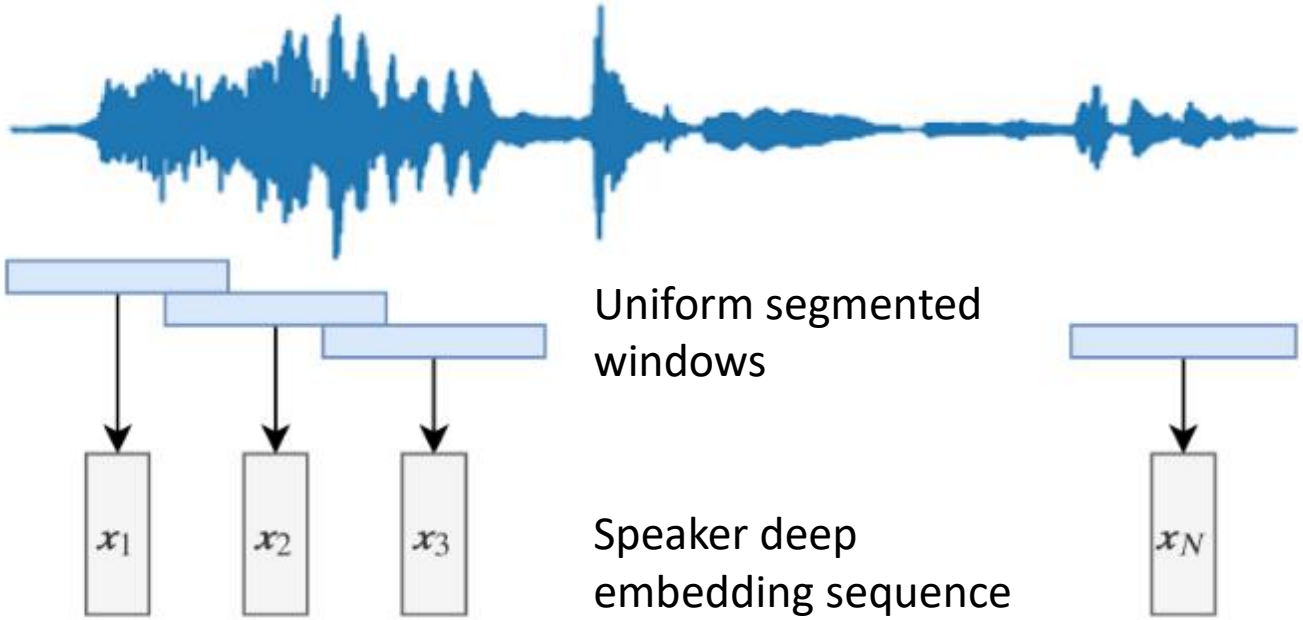
昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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

VAD and Segmentation



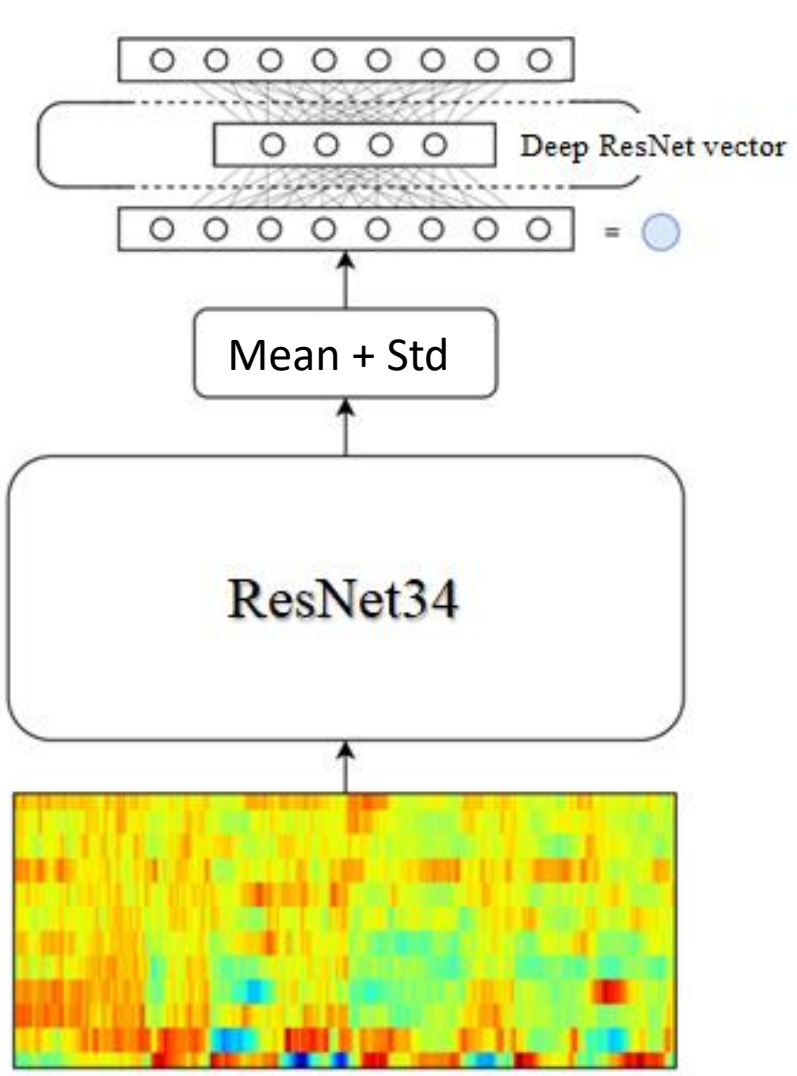
Lin, Qingjian, and Ming Li Tingle Li. "The DKU speech activity detection and speaker identification systems for fearless steps challenge phase-02." Proc. of Interspeech, 2020.

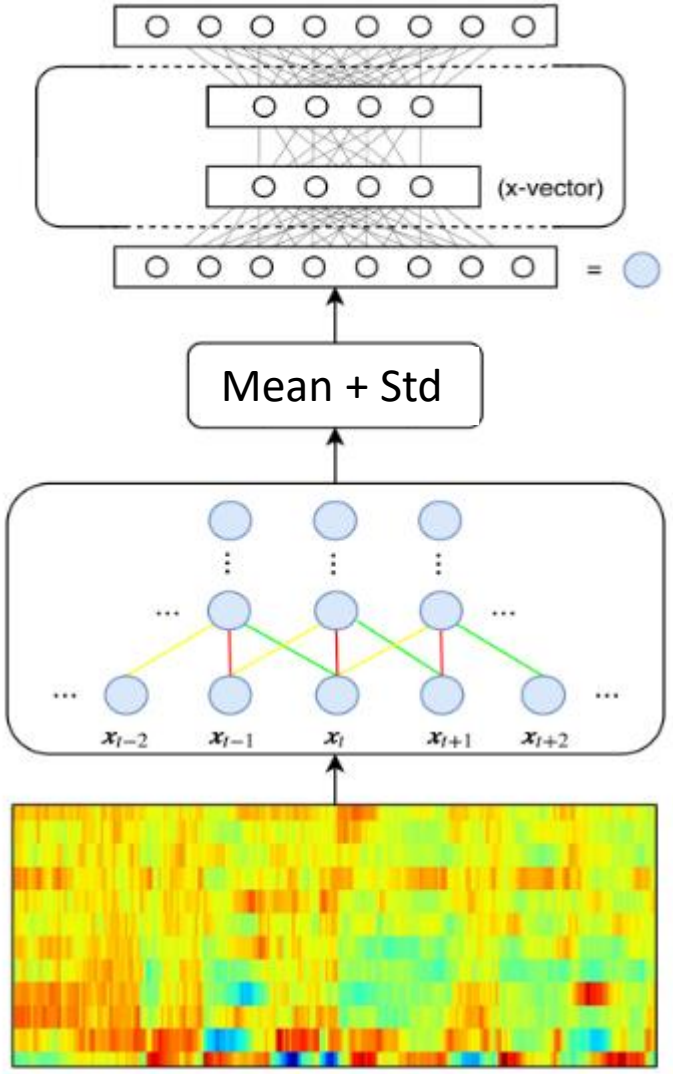


昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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





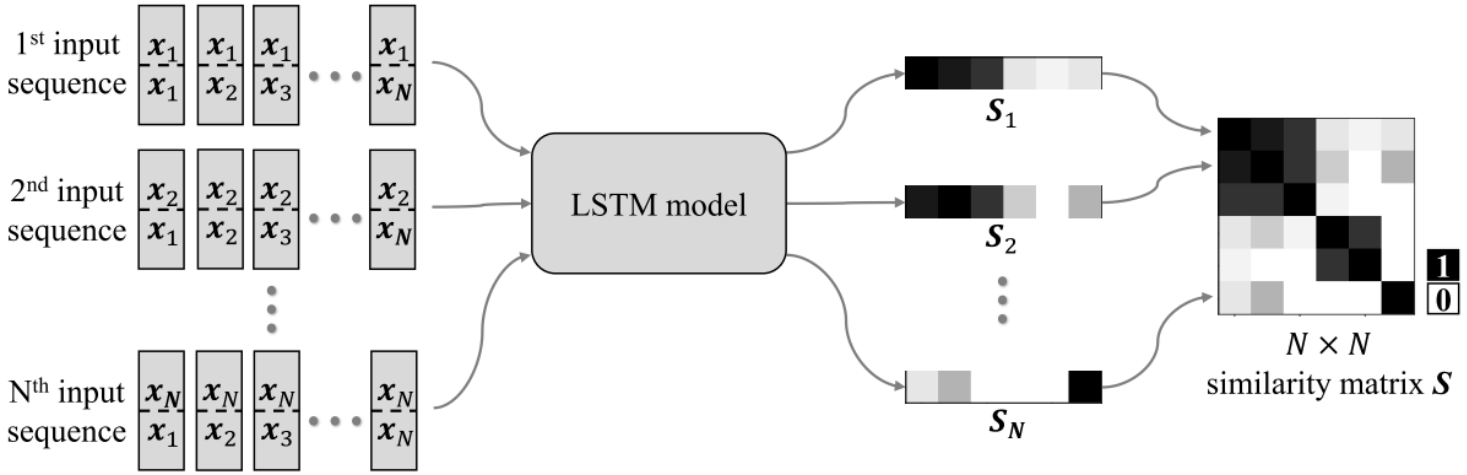
昆山杜克大学
DUKE KUNSHAN UNIVERSITY

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

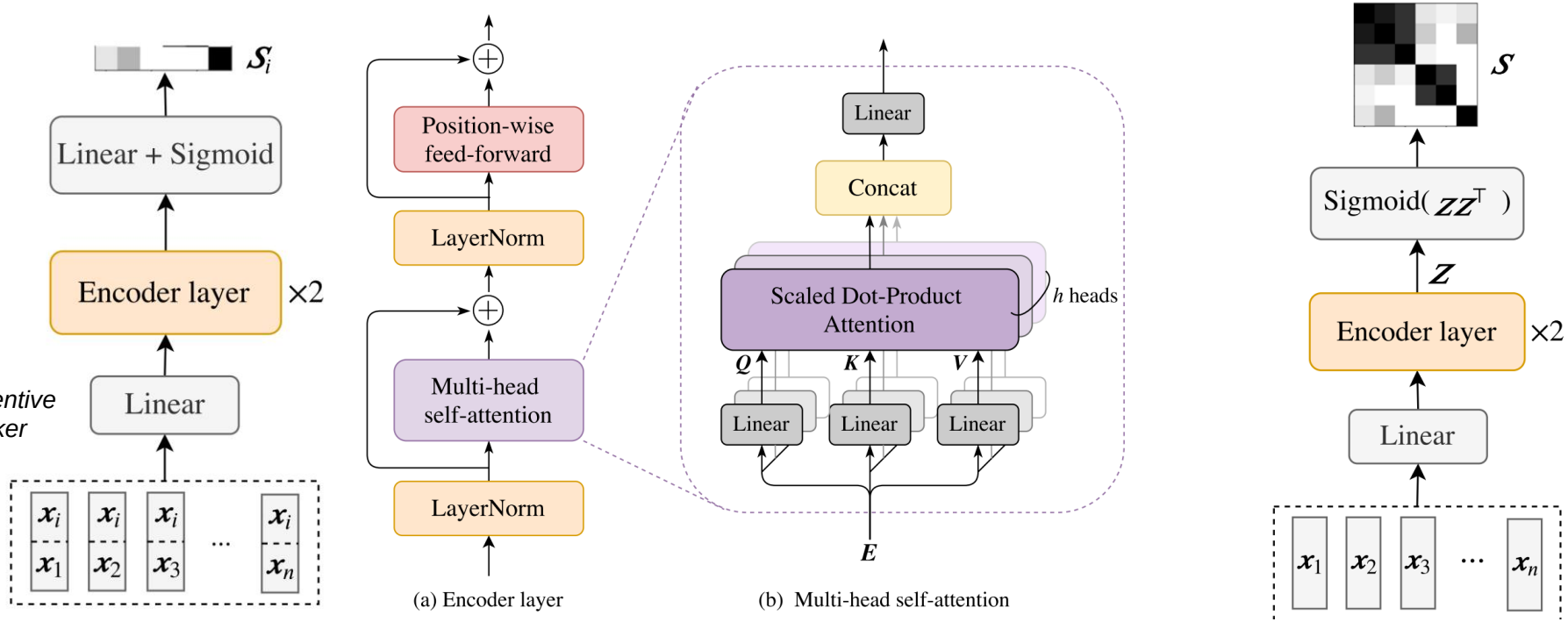
LSTM based scoring

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

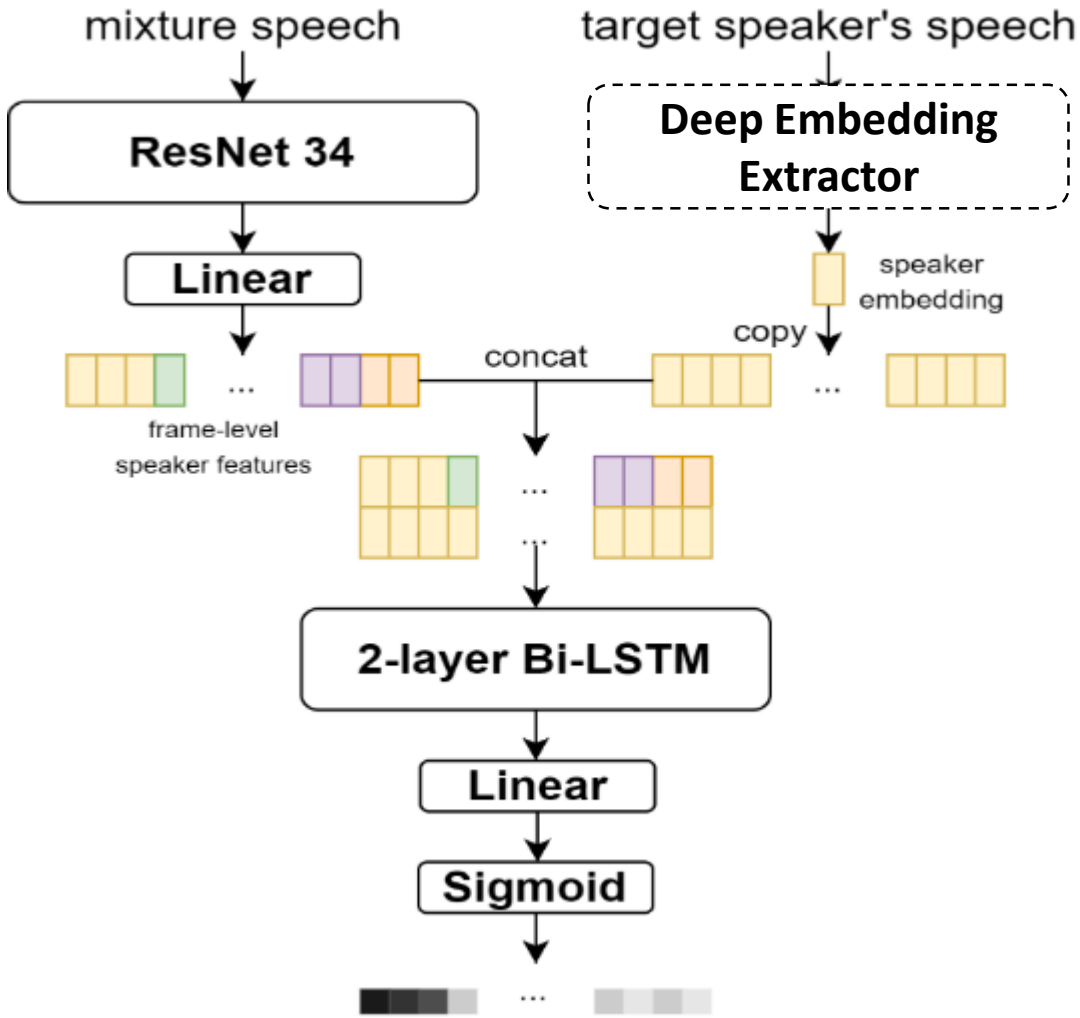
Results on Dihard II task 1

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	

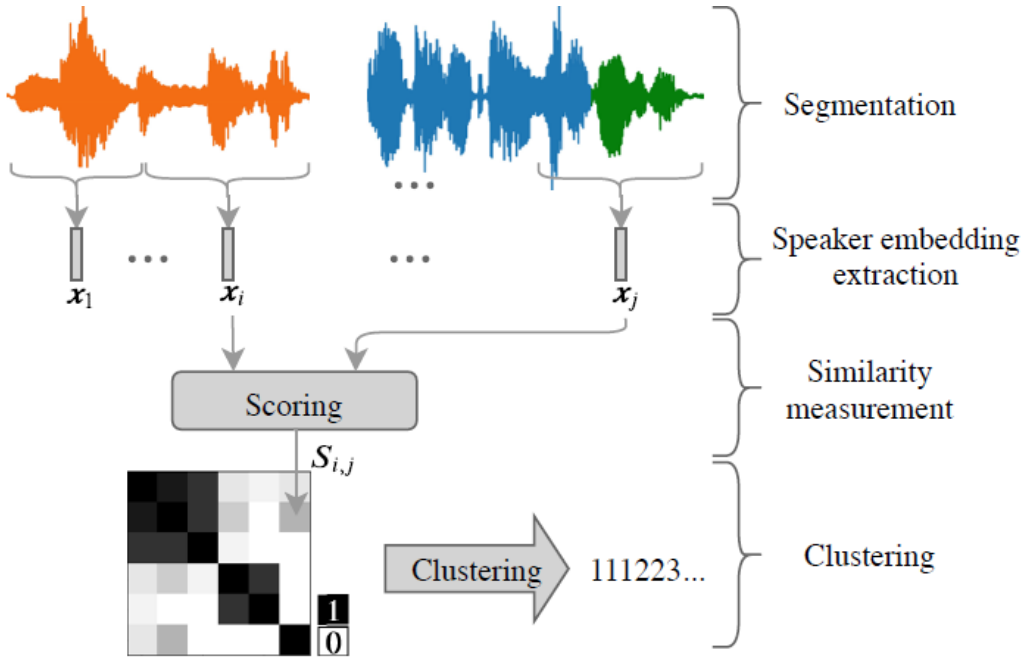
Lin, et.al, “Self-Attentive Similarity Measurement Strategies in Speaker Diarization”, Interspeech 2020.

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



Target speaker VAD round 1 → Target speaker VAD round 2

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

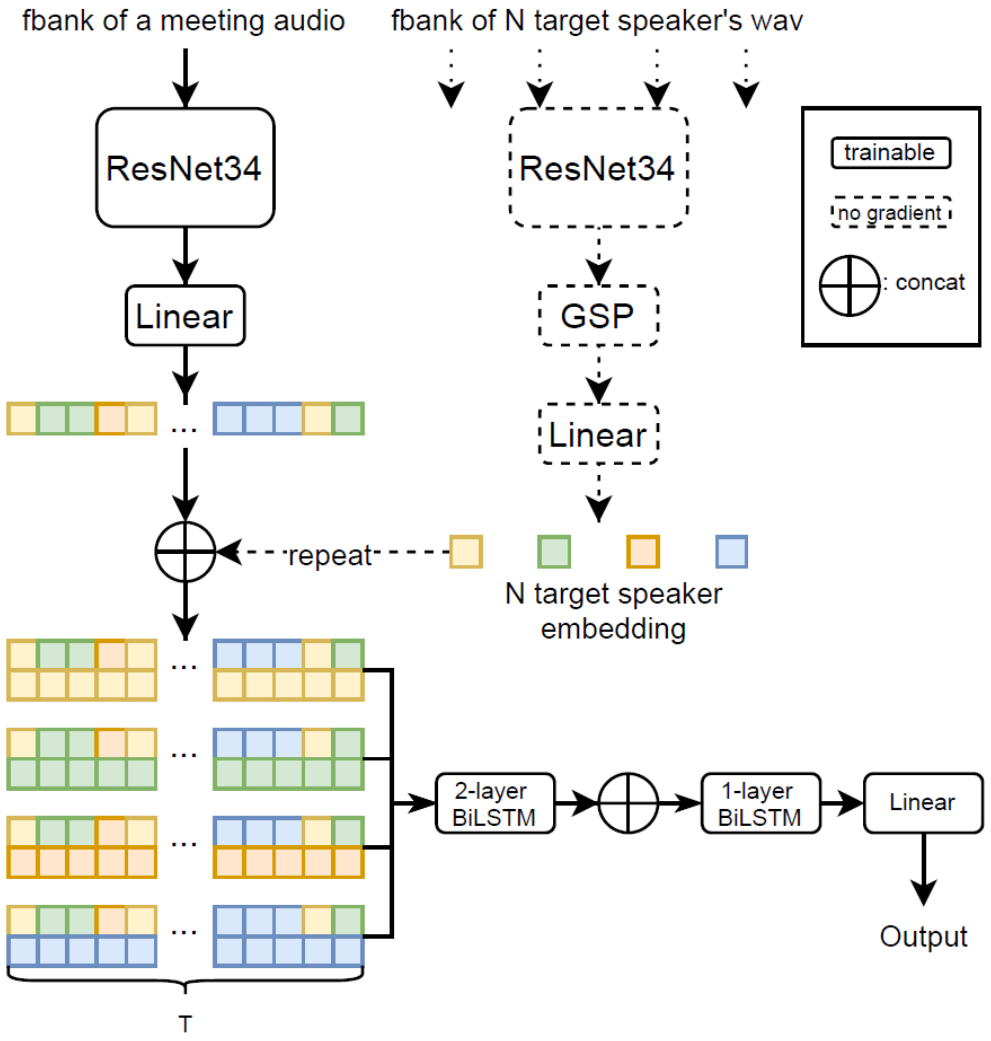
Results on DIHARD3 full test data

	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

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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



昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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

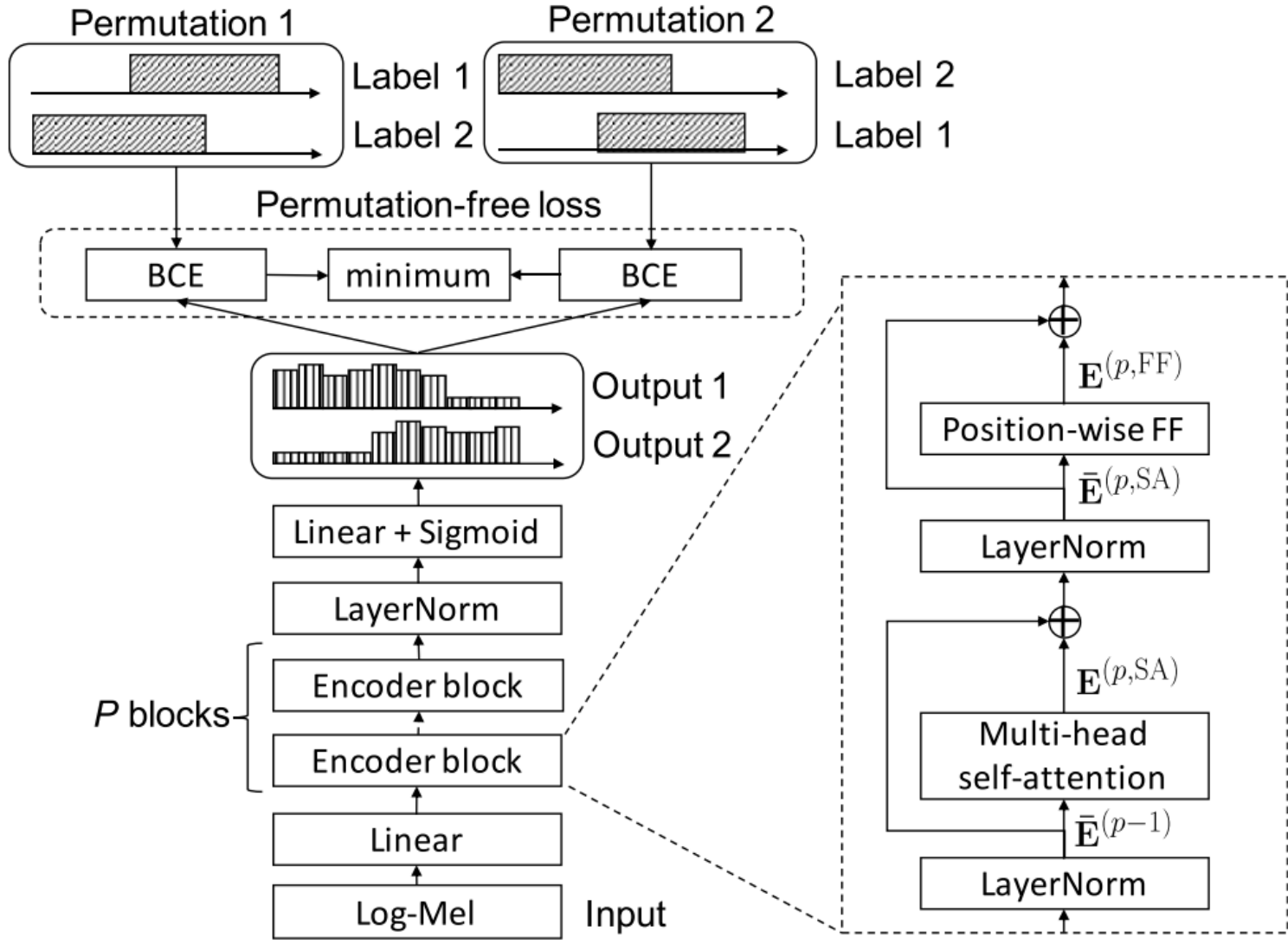
Model	Dev		Eval	
	Track 1	Track 2	Track 1	Track 2
1 LSTM	21.48	13.56	-	-
2 Att-v2s	22.57	15.11	-	-
3 AHC (uni-seg)	20.83	13.33	-	-
4 AHC (ahc-seg)	21.39	14.21	-	-
5 TSVAD (round 0)	20.75	11.88	43.99	13.85
6 TSVAD (round 1)	20.94	11.99	-	-
Fusion (1+2+3+4)	20.39	12.70	44.56	14.63
Fusion (1+2+3+4+5)	-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

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

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

End-to-end Speaker Diarization

Self-Attentitive End-to-End Neural Diarization (SA-EEND)



[1] Y. Fujita, et,al ,End-to-end neural speaker diarization with self-attention, in ASRU, 2019.

End-to-end Speaker Diarization

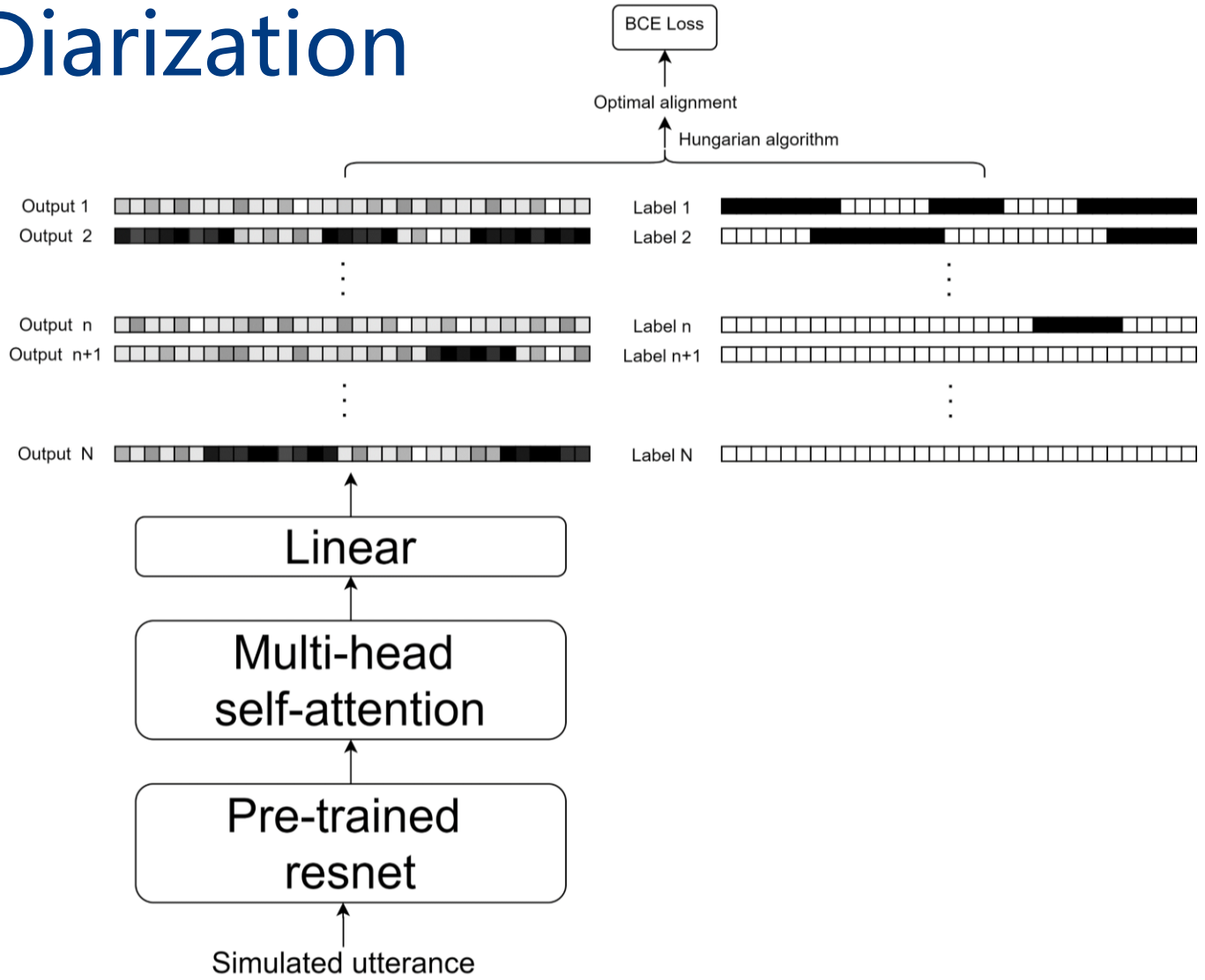
Data augmentation

using voxceleb data to create mixing data

Loss function

Hungarian algorithm can find

the optimal alignment in polynomial time [5]



Results on Dihard II task 2

[1]G. Sell, et.al, “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD chal- lenge,” in INTERSPEECH, 2018.
[2] S. Novoselov, et.al, “Speaker diarization with deep speaker embeddings for DIHARD Challenge II,” in INTERSPEECH, 2019.
[3] F. Landini, et.al, “BUT system for the Second DIHARD Speech Diarization Challenge,” in ICASSP, 2020,.
[4] Horiguchi, et.al, “End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors”, in arXiv:2005.09921. 2020.
[5] qingjian Lin, et.al, “Optimal Mapping Loss: A Faster Loss for End-to-End Speaker Diarization”, Odyssey 2020.

DIHARD II task 2 Baseline[1]	Best pre-is2019-deadline[2]	Best post-is2019-deadline[3]	SA-EEND + EDA[4]	Resnet + self-attention [5] (our method)
40.86	35.10	27.11	32.59	33.69

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Self-Supervised Speaker Representation Learning

Introduction

Why self-supervised learning?

- Collecting a dataset with clean labels is expensive
- Availability of vast numbers of unlabeled speeches/videos/images
  - Facebook: 350 million photos uploaded everyday[1]
  - Youtube: 500+ hours of content uploaded every minute[2]

What is self-supervised learning?

- A form of unsupervised learning where the data provides supervision
- Learning model trains itself and generates labels accurately
  - Data labeling is automated
  - Human interaction is eliminated

1. https://www.socialreport.com/insights/article/360000094166-The-Latest-Facebook-Statistics-2018
2. https://blog.youtube/press/

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Introduction

Two methods in self-supervised speaker representation learning

Generative

- Spectral feature reconstruction via self-supervised speaker representation + phone decoder [1]
- Autoencoder [2]

Discriminative

- Contrastive learning: each data sample defines its own class [3, 4, 5]

Audio-visual based methods [6, 7]

1. T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget, “Self-Supervised Speaker Embeddings,” in Interspeech, 2019, pp. 2863–2867.
2. U. Khan and J. Hernando , “The UPC Speaker Verification System Submitted to VoxCeleb Speaker Recognition Challenge 2020 (VoxSRC-20),” arXiv:2010.10937, 2020.
3. N. Inoue and K. Goto, “Semi-Supervised Contrastive Learning with Generalized Contrastive Loss and Its Application to Speaker Recognition,” arXiv:2006.04326, 2020.
4. J. Huh, H. S. Heo, J. Kang, S. Watanabe, and Joon S. Chung, “Augmentation Adversarial Training for Unsupervised Speaker Recognition,” arXiv:2007.12085, 2020.
5. M. Ravanelli and Y. Bengio, “Learning Speaker Representations with Mutual Information,” in Interspeech, 2019, pp. 1153–1157.
6. A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, “Disentangled Speech Embeddings Using Cross-Modal Self-Supervision,” in ICASSP, 2020, pp. 6829–6833.
7. S. W. Chung, H. G. Kang, and J. S. Chung, “Seeing Voices and Hearing Voices: Learning Discriminative Embeddings Using Cross-Modal Self-Supervision,” arXiv:2004.14326, 2020.

Contrastive Self-Supervised Learning

General framework for speaker representation learning

Data sampling

- Segments from different utterances belong to different speakers
- Two segments $x_{i,1}, x_{i,2}$ are randomly sampled from utterance x_i

Data augmentation

- Key to learn good self-supervised representation [1, 2]
- Reverberation, additive noise, etc.

$$x'_{i,1} = \mathrm{aug}(x_{i,1}) \qquad x'_{i,2} = \mathrm{aug}(x_{i,2})$$

Representation extraction

Loss function

1. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv:2002.05709*, 2020.
2. J. Huh, H. S. Heo, J. Kang, S. Watanabe, and Joon S. Chung, "Augmentation Adversarial Training for Unsupervised Speaker Recognition," arXiv:2007.12085, 2020.

Contrastive Self-Supervised Learning

General framework for speaker representation learning

- Data sampling
- Data augmentation
- Representation extraction
  - Encoding network f_θ: can be any type of speaker embedding network
  - Representations: $z_{i,1} = f_\theta(x'_{i,1})$ $z_{i,2} = f_\theta(x'_{i,2})$
- Loss function
  - Negative contrastive estimation (NCE) loss [1, 2]
    - Cross entropy to classify the the 'positive' sample from the 'negative' samples

      $\mathcal{L}(\theta) = -\log \frac{\exp\left(\Phi(z_{i,1}, z_{i,2})\right)}{\exp\left(\Phi(z_{i,1}, z_{i,2})\right) + \sum_{j \neq i} \exp\left(\Phi(z_{i,1}, z_{j,2})\right)}$ $\Phi(\cdot)$ is similarity measurement

      minimize the distance between segments from same utterance and maximize the distance between different utterances

  - ...

1. O. J Hénaff, A. Razavi, C. Doersch, S. Eslami, and A. van den Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding," *arXiv:1905.09272*, 2019.
2. A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv:1807.03748*, 2019.

Iterative Training with Pseudo-Labels

Drawbacks in contrastive self-supervised learning

- Assumption: segments from different utterances belong to different speakers
- Naturally introduce label error when training speaker embedding network

To avoid this label error

- Iterative, self-evolving framework [1, 2]
  - Start with contrastive self-supervised learning and generate pseudo labels
  - Iteratively train the speaker network with the pseudo labels
- Idea
  - Take advantage of the DNN’s ability to learn from data with label noise
  - Bootstrap the DNN’s discriminative power

1. D. Cai, W. Wang, and M. Li, “An Iterative Framework for Self-Supervised Deep Speaker Representation Learning,” arXiv:2010.14751, 2020
2. J. Thienpondt, B. Desplanques, and K. DemuynckThe, “IDLAB VoxCeleb Speaker Recognition Challenge 2020 System Description,” arXiv:2010.12468, 2020

Iterative Training with Pseudo-Labels

Relative works

- Iterative PLDA for unsupervised speaker verification [1]
  - Based on the traditional i-vector/PLDA pipeline
  - Iteratively optimize the PLDA scoring backend with the fixed speaker representation of i-vectors
- Deep clustering [2]
  - Jointly learn the parameters of a neural network and the cluster assignments (using statistical clustering method) of the resulting features
- Pseudo-label approaches for semi-supervised learning [3, 4]

1. W. Liu, Z. Yu, and M. Li, “An Iterative Framework for Unsupervised Learning in the PLDA based Speaker Verification,” in ISCSLP, 2014, pp. 78–82.
2. M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep Clustering for Unsupervised Learning of Visual Features,” in ECCV, 2018.
3. D. Yarowsky, “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods,” in ACL, 1995, pp. 189-196.
4. Z. Zhou, and M. Li, “Tri-Training: Exploiting Unlabled Data Using Three Classifier,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 11, pp. 1529–1541, 2005.

Iterative Training with Pseudo-Labels

Step 1 (initial round)

- Train a speaker embedding network with contrastive self-supervised learning

Step 2

- Extract speaker embeddings for the training data
- Perform a clustering algorithm to generate pseudo labels

Step 3

- Train the speaker embedding network with cross-entropy loss using the generated pseudo labels

Repeat step 2 and step 3 with limited rounds

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Iterative Training with Pseudo-Labels

- Generating pseudo labels by clustering: k-means
  - Pseudo labels $\{y_i\}$ and K centroids $\{c_1, c_2, \cdots, c_K\}$
- Purifying pseudo labels
  - Pseudo labels: massive label noise
  - Two simple steps to purify
    - Filter out p portion of the data with least clustering confidence
    - Keep the pseudo clusters with at least S samples: to reduce the possibility that one actual speaker appears in several pseudo clusters
  - Label noise level
    - Trade-off between the data size and the purifying intensity

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Iterative Training with Pseudo-Labels

Experiments

Speaker verification performance (minDCF and EER[%])

Model	#Utterances	#Clusters	NMI	Voxceleb 1 test		Voxceleb 1-E		Voxceleb 1-H		VoxSRC20 dev	
Supervised	1,092,009	5,994	1	0.097	1.51	0.102	1.59	0.178	3.00	-	-
CSL	1,092,009	-	-	0.508	8.86	0.570	10.15	0.710	16.20	0.857	20.11
Round 1	347,625	2,839	0.9381	0.429	6.96	0.433	7.91	0.561	11.73	0.700	14.59
Round 2	631,408	4,776	0.9404	0.341	5.42	0.358	6.22	0.479	9.60	0.606	12.05
Round 3	644,692	4,708	0.9603	0.300	4.73	0.316	5.29	0.433	8.17	0.546	10.42
Round 4	733,865	5,018	0.9638	0.263	4.16	0.278	4.55	0.391	7.39	0.503	9.48
Round 5	843,770	5,407	0.9618	0.241	3.45	0.246	4.02	0.363	6.57	0.464	8.60

NMI: normalized mutual information

- Measurement of clustering quality
- Two label assignment are largely independent: NMI → 0
- Two label assignment are in significant agreement: NMI → 1

昆山杜克大学 DUKE KUNSHAN UNIVERSITY

Challenges & Opportunities

- Large portion of overlapping
- long duration of recordings
- Large number of speakers
- Noisy and far-field condition
- Online diarization
- Multi-channel diarization
- Multimodal joint analysis
- Large scale database in complex scenarios
- etc.

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Thank you very much!

ming.li369@duke.edu

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

Qingjian Lin
SYSU Master student
2018-2020
Now at Lenovo

Weiqing Wang
Duke Ph.D. student
2019-2024

Work reported represents collaborative efforts with many students, colleagues and collaborators!

昆山杜克大学
DUKE KUNSHAN UNIVERSITY