System Description for Team DKU-Duke-Lenovo

Weiqing Wang, Qingjian Lin, Danwei Cai, Lin Yang, Ming Li

Content

- Dataset Description
- VAD
- Speaker Embedding
- Attention-based Scoring
- Target Speaker VAD
- Experimental Results

Dataset Description

- Statistics of the Dihard III Dataset

Table 1: *Statistics on the DIHARD III development set*

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

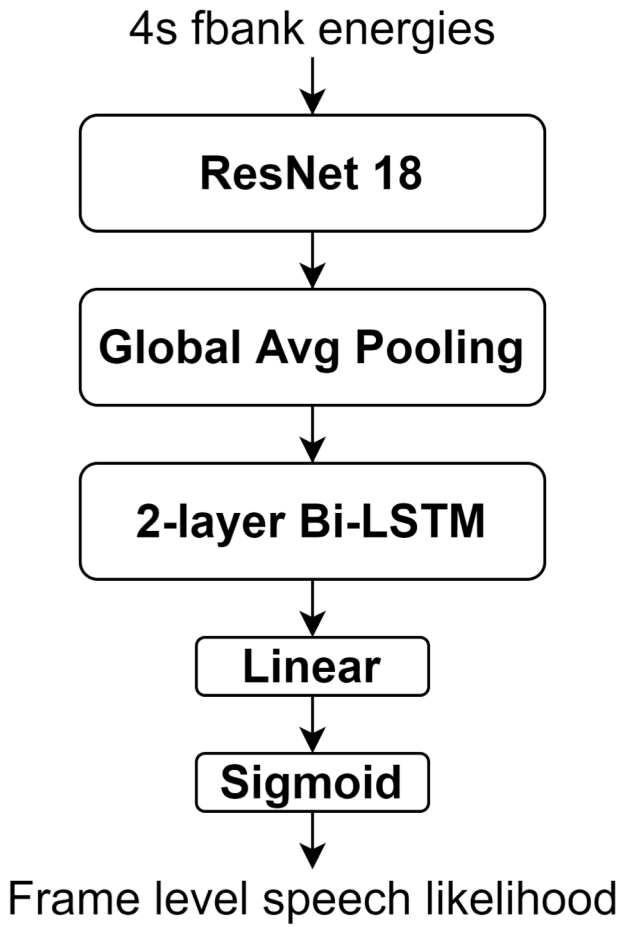
- Split dataset into CTS data (8kHz) and non-CTS (16kHz) data

VAD

- Training set: 90% of dev set
- Validation set: 10% of dev set
- Augmentation: MUSAN and RIRS

Table 3: *VAD accuracy on the development set*

	Training set	**Validation set**
Accuracy	96.8%	94.9%



Speaker Embedding[1]

- Architecture: ResNet34 + GSP + Linear (128-d) + ArcFace[2]
- Training set: Voxceleb 1 & 2 (8k for CTS data & 16k for non-CTS data)
- Augmentation: MUSAN and RIRS

[1] Qin, X., Li, M., Bu, H., Das, R. K., Rao, W., Narayanan, S., & Li, H. (2020). The FFSVC 2020 Evaluation Plan. arXiv preprint arXiv:2002.00387.
[2] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4690-4699).

Attention-based scoring for non-CTS data[2]

- Segmentation:
  - Training: 1.5s length / 0.75s shift
  - Infering: 1.5s length / 0.25s shift
- Training set: AMI, ICSI and Voxconverse dev
- Finetuning set: Dihard III dev set
- Post-processing:

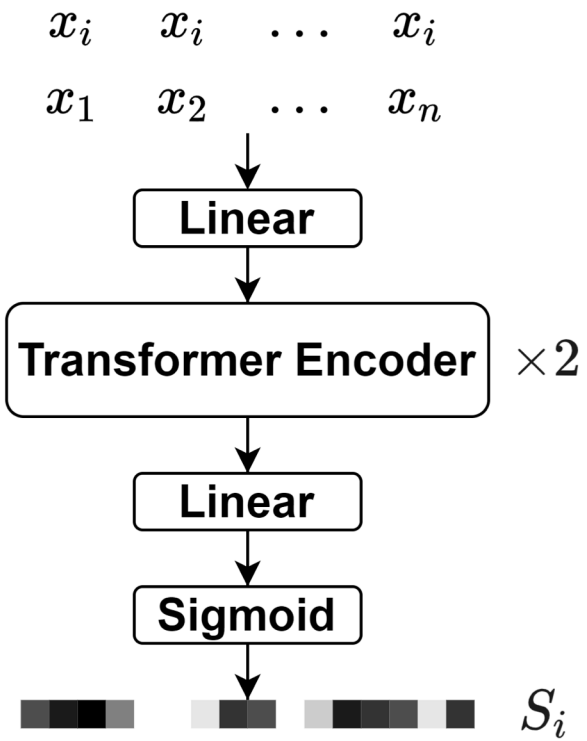
a) Symmetrization: $Y_{i,j} = \max(S_{ij}, S_{j,i})$
b) Diffusion: $Y \leftarrow YY^{\mathrm{T}}$
c) Row-wise max normalization: $S_{ij} = Y_{ij} / \max_k Y_{ik}$

- Spectral Clustering

$$\boldsymbol{S}_i = [S_{i1}, S_{i2}, ..., S_{in}] = f_{\mathrm{att}}(\boldsymbol{m}_i)$$

$$\boldsymbol{m}_i = \begin{bmatrix} \boldsymbol{x}_i & \boldsymbol{x}_i & ... & \boldsymbol{x}_i \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & ... & \boldsymbol{x}_n \end{bmatrix},$$

Speaker embedding sequence



[2] Lin, Q., Hou, Y., & Li, M. (2020). Self-attentive similarity measurement strategies in speaker diarization. In Proc. Interspeech (Vol. 2020, pp. 284-288).

Target Speaker VAD for CTS data

- AHC
  - Segmentation:
    - Uniform segmentation: 0.5s length / 0.25 shift
    - AHC-based segmentation[3]: threshold is 0.6
  - Only 2 speakers in CTS data
  - Center embedding: mean of all segments in the cluster
  - Stop threshold: 0.6 (for TSVAD)
  - Overlap threshold: 0.0

[3] Xiao, X., Kanda, N., Chen, Z., Zhou, T., Yoshioka, T., Zhao, Y., ... & Gong, Y. (2020). Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020. arXiv preprint arXiv:2010.11458.

Target Speaker VAD for CTS data

- TSVAD
  - Training set: Switchboard, SRE 04, 05, 06, 08
  - Finetuning set: 41 recordings in the CTS data
  - Validation set: 20 recordings in the CTS data
  - Post-processing
    - 11-tap median filtering
    - Threshold: 0.65
    - Correct non-speech frame

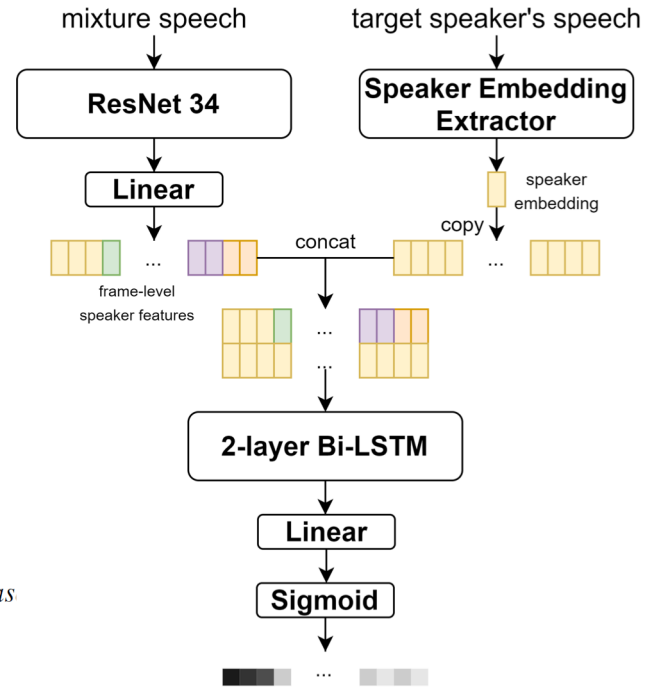


Table 4: *System performance (DER) on development datas (Track 1)*

Dataset	Method	DER (%)
NCTS	att-v2s + SC	16.05
CTS	Cosine + AHC	15.07
CTS	TSVAD	10.60
CTS (adapt)	TSVAD round 1	7.80
CTS (adapt)	TSVAD round 2	7.63

[4] Ding, S., Wang, Q., Chang, S. Y., Wan, L., & Moreno, I. L. (2019). Personal VAD: Speaker-Conditioned Voice Activity Detection. arXiv preprint arXiv:1908.04284.

[5] Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., ... & Romanenko, A. (2020). Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario. arXiv preprint arXiv:2005.07272.

Experimental Results

Table 5: *System performance (DER) on evaluation dataset (Track 1 & 2)*

	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

Thanks!