GIST-AiTeR System for the Diarization Task of the VoxCeleb Speaker Recognition Challenge (VoxSRC) 2022

Dongkeon Park, Yechan Yu, Kyeong Wan Park, Ji Won Kim, Hong Kook Kim

Audio Intelligence Technology & Research Lab (AiTeR),
Gwangju Institute of Science and Technology (GIST)

22 September, 2022

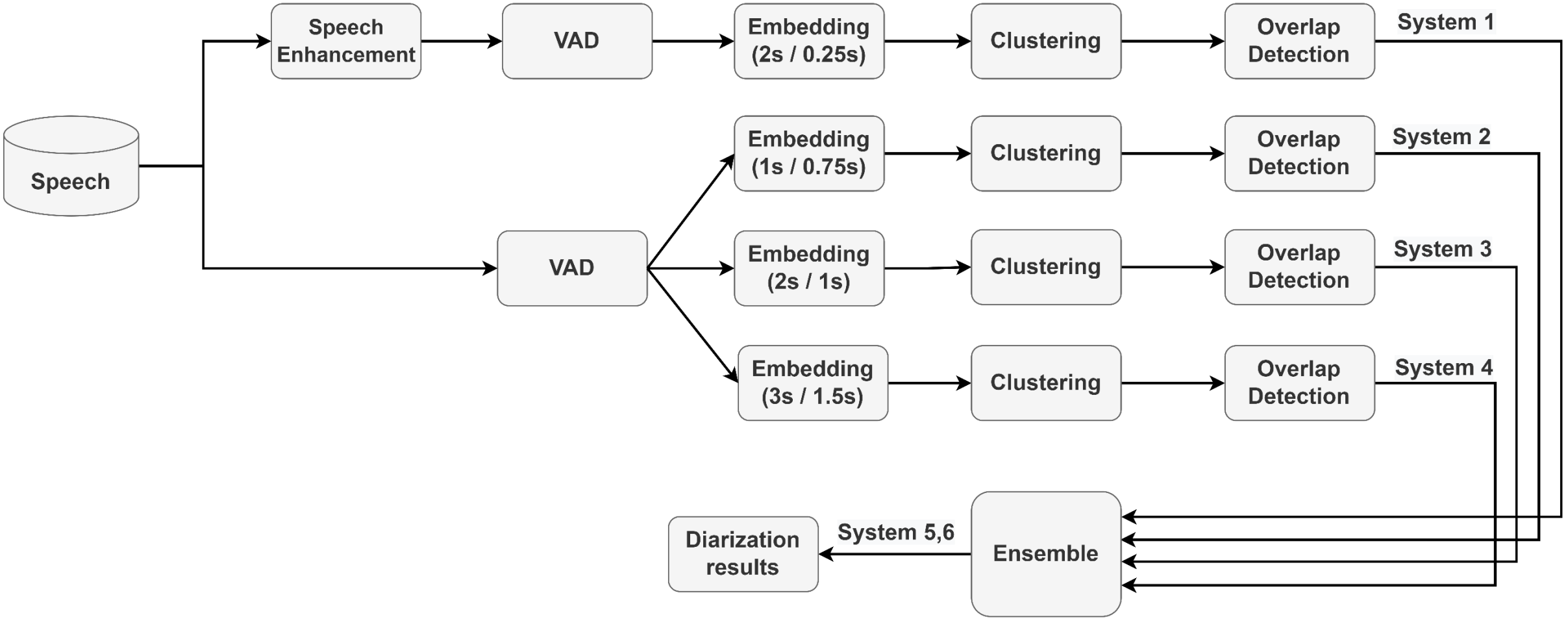
1. Dataset

- **DEV402**
  - Voxconverse 2020 dev set + first 186 recordings of voxconverse 2020 test set
- **VAL46**
  - Last 46 recordings of voxconverse 2020 test set

- **Mixed training set**
  - AMI, AISHELL-4, DIHARD I & II, CALLHOME
- **Other**
  - Voxceleb1&2, MUSAN, RIRs

- **We follow VoxSRC21 winner (Team DKU-DukeECE-Lenovo)**

2. System Overview



2.1 Speech Enhancement

Enhance Voxconverse set

- Pre-processing with pretrained speech enhancement model
  - FullSubNet[1]
    - Trained with DNS Challenge (INTERSPEECH 2020) dataset

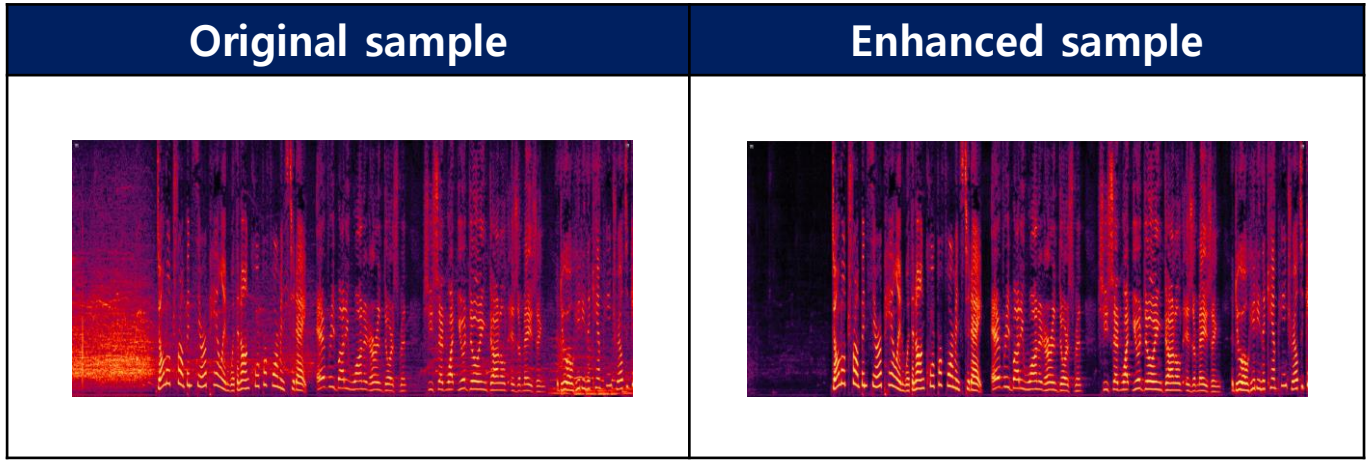


Figure 1. Spectrogram of original / enhanced voxconverse sample

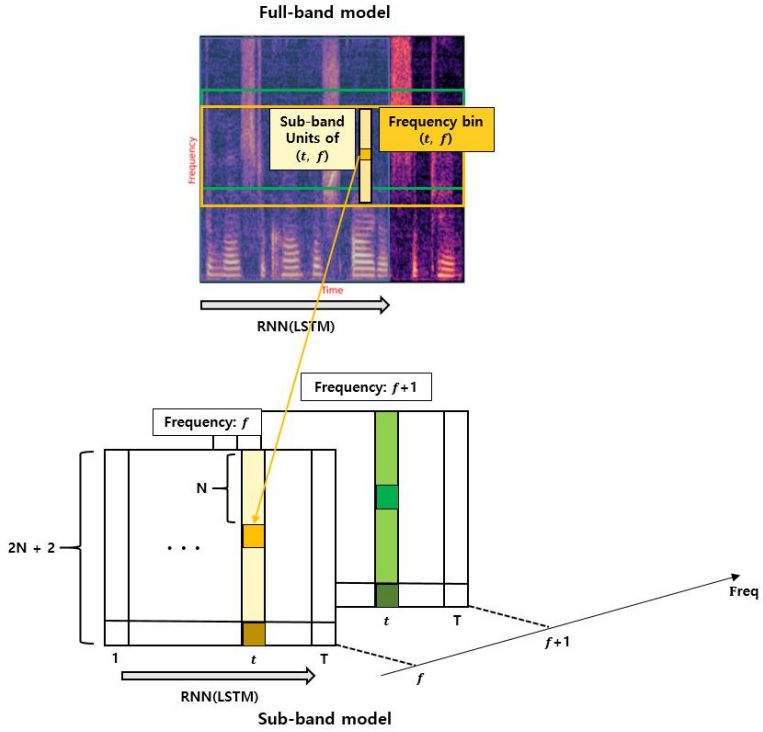


Figure 2. Speech Enhancement processing

2.2 Voice Activity Detection (VAD)

1. ResNet+LSTM

- Almost same as [2]
- Front-end: ResNetSE34[3]+ Statistical pooling ($S = \{1, 2\}$)
- Trained on the mixed training set and fine-tuned on enhanced DEV402, augmented by MUSAN and RIRs.

2. SincNet+LSTM (Pyannote 2.0)[4][5]

- It transferred from pre-trained using DEV402 without speech enhancement

2.2 Voice Activity Detection (VAD)

◆ Fusion

- Ensemble by averaging the posterior value from the ResNet+LSTM and SincNet+LSTM model

Model	FA [%]	Miss [%]	Acc [%]	F1 [%]
1. ResNet (S=1)	2.03	1.61	96.34	97.94
2. ResNet (S=2)	2.39	1.47	96.14	97.82
3. SincNet+LSTM	2.23	1.47	96.30	97.92
Fusion (1+2)	2.16	1.54	96.31	97.92
Fusion (1+2+3)	2.08	1.51	96.41	97.98

Table 1. Comparison of the false alarm (FA), miss detection (Miss), accuracy (Acc) and F1 score of three different VAD models and their fusions on VAL46

2.3 Speaker Embedding Extraction

- **Model**: MFA-Conformer[6]
- **Training set**: Voxceleb 1&2
- **Input**: Multi-scale input in each mini-batch[7]
- **Augmentation**: MUSAN noise or RIR, SpecAug
- **Evaluation**: Voxceleb1 test set

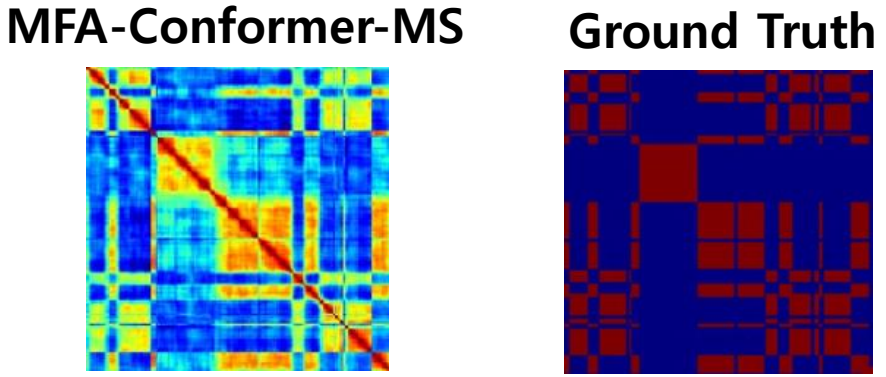


Figure 3. Our Speaker Embedding Extraction Cosine Similarity Score matrix

	Pooling	Multi-scale Input	EER(%)
MFA-Conformer	ASP	2.0 sec	0.697
MFA-Conformer-MS	ASP	[1.0, 2.0, 3.0] sec	0.867

Table 2. Our speaker embedding model results on VoxCeleb1 test set

2.4 Scoring + Clustering

Scoring PLDA model

- PLDA model is interpolated from VoxCeleb assigned a weight of 0.9 and DEV402 assigned a weight of 0.1

AHC with PLDA for initial assignment

- Short cluster identified using a duration threshold[8]
  - Merged into the closest long cluster or treated as a new cluster by SV threshold (=0.5)
- Higher SV threshold value caused slight underclustering

VB-HMM Clustering

- The parameters of VB-HMM[9] were tuned on VAL46 on each time-scale segments

2.5 Overlapped Speech Detection (OSD)

◆1. ResNet+LSTM

- Almost same as ResNst+LSTM VAD system
- Weighted Cross Entropy (WCE) Loss to deal with imbalanced dataset

◆2. SincNet+LSTM

- It transfered from pre-trained using DEV402

◆Fusion

- Ensemble by averaging the posterior value
- Threshold was intentionally set so that the precision became high

Model	Prec. [%]	F1 [%]
1. ResNet+LSTM (S=1)	68.55	68.22
2. ResNet+LSTM (S=2)	67.55	67.40
3. SincNet+LSTM	68.83	66.79
Fusion (1+2)	83.94	56.99
Fusion (1+2+3)	**88.81**	**52.45**

Table 3. Comparison of precision (Prec.) and F1 score of different OVD models on VAL46

3. Result

With Dover-Lap

- Significantly reduce DER

System	Time-scale (Segment / hop length)	Speech Enhancement	VAL46		VoxSRC22 test set	
			DER[%]	JER[%]	DER[%]	JER[%]
1	1s / 0.75s	No	4.41	27.47	-	-
2	2s / 1s	No	3.97	27.45	-	-
3	3s / 1.5s	No	4.02	26.92	-	-
4	2s / 0.25s	Yes	4.14	27.75	-	-
5	Fusion (1+2+3)		3.66	26.63	-	-
6	**Fusion (1+2+3+4)**		**3.56**	**27.63**	**5.12**	**30.815**

Table 4. Performance comparison of our different versions of speaker diarization systems

References

[1] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in Proc. ICASSP, 2021, pp. 6633–66

[2] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, "The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge," arXiv preprint arXiv:2109.02002, 2021

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. CVPR, 2018, pp. 7132–7141.

[4] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M. P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in ICASSP 2020-2020 IEEE International Conferenceon Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp.7124–7128.

[5] H. Bredin and A.Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in Proc. Interspeech 2021, Brno, Czech Republic, August 2021.

[6] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification," arXiv preprint arXiv:2203.15249, 2022.

[7] Y. Kwon, H.-S. Heo, J.-w. Jung, Y. J. Kim, B.-J. Lee, and J. S. Chung, "Multi-scale speaker embedding-based graph attention networks for speaker diarisation," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 8367–8371.

[8] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, S. Chen, Y. Zhao, G. Liu, Y. Wu, J. Wuetal., "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021,pp.5824–5828.

[9] F. Landini et al., "But System for the Second Dihard Speech Diarization Challenge," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6529-6533, doi: 10.1109/ICASSP40776.2020.9054251.