USTC-NELSLIP System Description for DIHARD-III Challenge

Yuxuan Wang
University of Science and Technology of China
Hefei, Anhui, China
yxwang1@mail.ustc.edu.cn

Maokui He
University of Science and Technology of China
Hefei, Anhui, China
hmk1754@mail.ustc.edu.cn

Shutong Niu
University of Science and Technology of China
Hefei, Anhui, China
niust@mail.ustc.edu.cn

Lei Sun
iFlytek Research
Hefei, Anhui, China
leisun8@iflytek.com

Tian Gao
iFlytek Research
Hefei, Anhui, China
tiangao5@iflytek.com

Xin Fang
iFlytek Research
Hefei, Anhui, China
xinfang@iflytek.com

Jia Pan
iFlytek Research
Hefei, Anhui, China
jiapan@iflytek.com

Jun Du
University of Science and Technology of China
Hefei, Anhui, China
jundu@mail.ustc.edu.cn

Chin-Hui Lee
Georgia Institute of Technology
Atlanta, Georgia, USA
chl@ece.gatech.edu

Abstract—This system description describes our submission system to the Third DIHARD Speech Diarization Challenge. Besides the traditional clustering based system, the innovation of our system lies in the combination of various front-end techniques to solve the diarization problem, including speech separation and target-speaker based voice activity detection (TS-VAD), combined with iterative data purification. We also adopted audio domain classification to design domain-dependent processing. Finally, we performed post processing to do system fusion and selection. Our best system achieved DERs of 11.30% in track 1 and 16.78% in track 2 on evaluation set, respectively.

Index Terms—speech diarization, speech separation, TS-VAD, Third DIHARD Challenge

I. Notable Highlights

1. Iterative speech separation (ISS) based diarization system.
2. Iterative TS-VAD (ITS-VAD) based diarization system.
3. Domain-dependent processing.

II. Data Resources

The overall framework of our system is shown in Figure 1. In this section, we introduce all data sources for submodules:

Audio Domain Classification: The whole DIHARD-III development set (LDC2020E12) was divided into 2 parts, 9/10 for training and 1/10 for testing.

Speech Enhancement: The clean speech data were from WSJ0 [1], AIShell-1 [2], THCHS-30 [3], and Librispeech [4], and the noise data included 115 types of noise [5] and MUSAN [6] corpus. The input noisy mixture was made at -5 dB, 0 dB and 5 dB, and finally we got a 1000-hour training set.

Clustering Based Diarization System: We directly referred to the system of BUT in [7]. For SAD, we used oracle SAD results for track 1, and scratched SAD results for track 2. The

pre-trained model was trained on the 600-hour home-made realistic speech data in iFlytek. The speech quality is not very stable due to the complicated acoustic environments. And the DIHARD-III development set was used for finetuning. For x-vector extraction, the training data was drawn from VoxCeleb 1 [8] and 2 [9] with data augmentation, amounting to 6 million utterances from 7146 speakers. For PLDA, the out-of-domain PLDA was trained with VoxCeleb 1 and 2, and the in-domain PLDA using DIHARD-II development set (LDC2019E31).
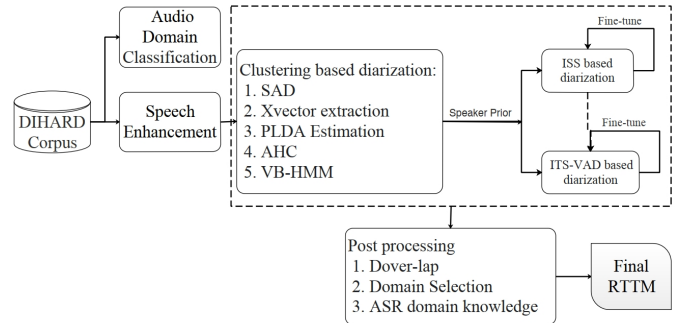

Fig. 1. An illustration of overall framework.

ISS Based Diarization System: The pre-trained model was trained on the 250-hour 2-speaker mixed data simulated by Librispeech.

ITS-VAD Based Diarization System: We used VoxCeleb 1 and 2 with augmented with 3-fold speed perturbation to train the i-vector extractor. The pre-trained model was trained on the 2500-hour data, in which Switchboard-2 (LDC97S62), AMI Meeting Corpus [10], Voxconverse [11] DEV set were

selected as realistic conversations and Librispeech as simulated meeting-style multi-speaker conversations.

ASR: The whole CHIME-6 training set [12] was used for ASR model training.

III. Detailed Description of Algorithm

A. Audio Domain Classification

We used a ResNet based network with 17 convolutional layers [13] [14] to classify the diverse 11 domains both in the development set and evaluation set. 64-dimensional log-mel filterbanks were used as input acoustic features, and 11 output-layer cells corresponded to 11 domains. During training, all sessions were truncated into 10-second segments, each segment was assgined with the corresponding domain label, the initial learning rate was set to 0.1, and the batch-size was set to 15. We trained it for 510 epochs. During testing, we used voting strategy to get session-level classifciation results. The proposed model achieved 100% accuracy on the development set, and we directly applied it on the evaluation set to predict the domain for each session.

B. Speech Enhancement

We employed the progressive multi-target network based speech enhancement model [15] on RESTAURANT domain for its loud background noise. 257-dimensional PELPS and PRM were selected simultaneously. 7-frame expansion was used for input, the number of LSTM memory cells in each layer was 1024, and the progressive increasing SNR between two adjacent targets was set to 10 dB. Besides, we applied speech enhancement on all domains before feeding into SAD module in Track2. Here $PELPS_1$ (Progressively Enhanced LPS at target layer 1) enhanced speech was found optimal for the following diarization results and SAD results.

C. Clustering Based Diarization System

The traditional clustering based diarization system was used as the basis system for the whole framework, and the diarization results were used as initialization for the subsequent submodules except the RESTAURANT domain. Experiment results showed that performing SS or TS-VAD based diarization on RESTAURANT domain data led to the performance degradation, so we collected the results of clustering based diarization system of RESTAURANT domain as the final submitted results directly.

The parameters finetuned on development set for RESTAURANT domain were set as followings: the interpolation value of the two PLDAs was 0.57, that is, the combination weight of the in-domain PLDA was 0.57, and the out-of-domain PLDA 0.43, the thresholds used as stopping criteria for the AHC were finetuned as $threshold\_bias$ = 0.5, $target\_energy$ = 0.3, the parameters for the BHMM was finetuned as $maxIters$ = 7, $smoothing\_factor$ = 4.0, lda_dim = 512. The rest parameters on RESTAURANT domain and all the parameters on other domains kept consistent with [7].

For SAD, We employed three different networks for frame-wise binary classification of speech and non-speech. The DNN model adopted a small and compact structure using 2 hidden layers with 256 and 128 hidden units in each layer and a final dual output layer. The input features were 39-dimensional PLP features and expanded an input context of 5 neighbouring frames (±2). The CLDNN model [16] adopted 2 CNN layers, 2 LSTM layers and 2 DNN layers. The input features were 40-dimensional filterbanks. The TDNN model was equipped with the same structure as the DIHARD-III SAD baseline [17]. As mentioned in III-B, all of those models were finetuned on enhanced DIHARD-III development set and tested on enhanced DIHARD-III evaluation set (LDC2020E13). Finally, we voted from the three systems and got a fusion SAD result.

D. ISS Based Diarization System

The SS framework simply contains two parts: separation and detection. In the separation part, we trained a fully convolutional time-domain audio separation network (Conv-TasNet) [18] as our pre-trained model. In the detection part, we directly used a DNN-based SAD to detect speaker presence. Combine all SAD results along the time axis, then speech diarization results were attained, and the overlap regions were automatically labeled.
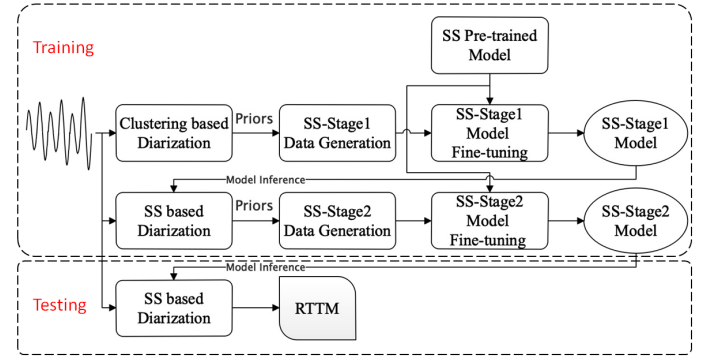

Fig. 2. The training and testing procedure for ISS Based Diarization System

To improve the generalization ability of the separation model, we adopted the iterative strategy, that is, we used the speaker priors drawn from the traditional clustering based diarization system to simulate data, about 5000 mixed audios for each session, and retrained to let the SS pre-trained model adapt current session. Then we got the SS-stage1 model. The same procedure was taken once more except the speaker priors were drawn from SS-stage1 model, which can significantly improve the separation performance by the more accurate priors. Finally, we used the SS-stage2 model to separate the entire session. To be mentioned that we finetuned the separation model for each session. The whole framework is shown in figure 2. Here we only applied this method on CTS domain in our submitted system.

The asteroid [19] was used as our speech separation toolkit to train the Conv-TasNet model. In the pre-training process, the learning rate was set to 0.001, and the batch-size was set to 6. We trained it for 75 epochs. In the finetuning process,

the learning rate was also set to 0.001, the batchsize was set to 4 and we trained it for 3 epochs.

E. ITS-VAD Based Diarization System

The TS-VAD model [20] takes log mel filterbanks as input, along with i-vectors corresponding to the speakers, and predicts per-frame speech activities for all the speakers simultaneously.

When the speaker number in a session was smaller than the number of output nodes N, we assigned the remain nodes to dummy speakers selected from the training set randomly, labeled them with silence when training and discarded them when testing. When the number of speakers was larger than N, we randomly selected N speakers for training and selected N speaker with longest voice for testing. All the speaker numbers above were estimated from traditional clsutering based diarization systems, and we set N to 8 here.

To improve the generalization ability of the TS-VAD model, we also applied iterative strategy here. First, we decoded the TS-VAD pre-trained model with i-vectors extracted from the clustering based diarization results. Then, we removed overlapping segments detected by TS-VAD from clustering based results and used the remaining non-overlapping segments to simulate multiple speaker dialogue data, about 4 hours for each session, for TS-VAD finetuning. Finally, we used the i-vectors extracted from the non-overlapping segments to decode the TS-VAD finetuned model and generated the iterative TS-VAD diarization results. Also we finetuned TS-VAD pre-trained model for each session. The whole framework is shown in figure 3. The number of iterations for iterative training depended on the different domains.

We implemented TS-VAD related experiments with kaldi and pytorch toolkit. In pre-training process, the training data was divided into 8 seconds each segment, the learning rate of the adam optimizer was set to 0.0001, and the batch-size was set to 32. And we trained it for 2 epochs. In finetuning process, the learning rate was also set to 0.0001, and the batch-size was set to 8.
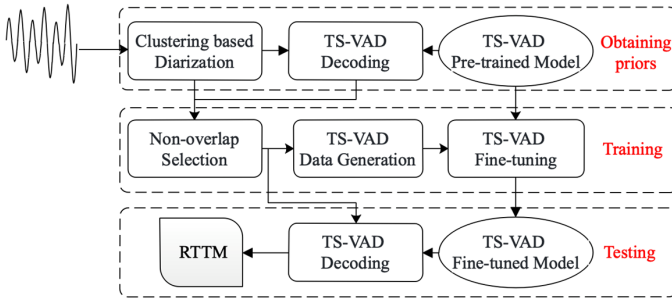

Fig. 3. The training and testing procedure for ITS-VAD Based Diarization System

F. Post-processing

As mentioned before, we got several diarziation systems including clustering based diarization, ISS based diarization and ITS-VAD based diarization with different priors. We first performed system fusion with Dover-lap [21] of above results, which can effectively improve overall performance. Dover-lap here was not only used for fusing different systems, but also fusing different epochs in the same system during the iterative processing.

Based on the results of audio domain classification, we also performed domain selection, that is, select the best system and the best set of parameters for each domain according to the performance on the development set.

Specially, we utilized domain knowledge from the speech recognition task, the recognized token [laugh] indicates where overlapping speech segments often occur, especially in multi-talker scenarios. We simply assigned laughter segments to each neighborhood speaker. This also gave a slight performance improvement.

IV. RESULTS

In this section, we introduce the performance of the sub-mission systems and other major system components.

TABLE I
PERFORMANCE COMPARISION ON FULL DEVELOPMENT SET CTS DATA
FOR TRACK 1 AMONG DIFFERENT SYSTEMS.

System	Miss(%)	FA(%)	SpkErr(%)	DER(%)
Clustering based diarization system	12.00	0.00	4.22	16.22
Clustering based diarization system +SS based diarization system	7.61	2.61	2.73	12.95
Clustering based diarization system +ISS based diarization system	5.12	1.74	1.90	8.76
Clustering based diarization system +ISS based diarization system +Dover-lap	5.40	1.21	1.70	8.31

Table I lists the performance comparison of different systems on CTS domain of full development set with oracle SAD. CTS data is collected in telephone environment, and has 2 speakers in a session with quite a bit of overlapping segments. We get a DER of 16.22% for the first clustering based diarization system, and the miss value is observed to be the largest part in the whole DER, which indicates that the traditional algorithm cannot well handle overlapped speech. To solve this problem, we add the second SS based diarization system (pre-trained model), the miss value achieves a large decrease of 4.39% (from 12.00% to 7.61%), and the increased FA value is due to the separation error, which may create additional sounds on the original speech segment. The corresponding DER also decreases from 16.22% to 12.95%. Next, when we add the third ISS based diarization system (SS-stage2 model), the performance is further improved, the DER changes from 12.95% to 8.76% with the decrease of 4.19%. The last system we add dover-lap, the dover-lap here is used to fuse different iterations during iterative training, and we finally get a DER of 8.31%, totally 48.7% relative improvement over the first traditional clustering based system.

Table II compares the DER performance on domain-wise full development set for track 1 except three domains: AUDIO-BOOKS, WEBVIDEO, and RESTAURANT. There is only one

TABLE II
DER(%) PERFORMANCE COMPARISION ON DOMAIN-WISE FULL DEVELOPMENT SET FOR TRACK 1 AMONG DIFFERENT SYSTEMS.

System	MAPTASK	BROADC.	COURT.	SOC. LAB	CTS	CLINICAL	SOC. FIELD	MEETING
Clustering based diarization system	5.02	2.60	2.95	7.97	16.22	10.97	11.87	26.41
Clustering based diarization system +TS-AVD based diarization system	6.71	2.94	3.15	8.81	10.21	16.48	13.79	24.72
Clustering based diarization system +ITS-VAD based diarization system	2.27	2.37	2.46	5.17	7.76	9.83	10.74	23.05

speaker for each session in AUDIOBOOKS, so we simply apply SAD processing, and assign each speech segment with the same speaker in one session. For RESTAURANT domain, as mentioned above, it has many speakers and overlapped speech with loud background noise, TS-VAD leads to performance degradation. So does WEBVIDEO domain. From the results we can learn that, TS-VAD based diarization system (pre-trained model) only performs better than clustering based diarzation on well matched domains like CTS and MEETING (a decrease of 6.01% and 1.69%, respectively) corresponding to Switchboard and AMI in training set, respectively. While when we perform ITS-VAD based diarization system here, considerable improvement is brought on all the eight domains, which shows its huge generalization abilities. Note that the iterative training adopts the priors from ISS based diarization system instead of clustering based one for CTS domain here.

TABLE III
THE OVERALL ERROR (%) OF DIFFERENT SAD SYSTEMS FOR THE CORE/FULL DEV AND EVAL SETS. THE PART. COLUMNS INDICATES WHETHER SCORING WAS PERFORMED USING THE FULL OR CORE DEV/EVAL SET.

system	part.	dev	eval
DNN	core	1.55	/
	full	1.51	6.18
CLDNN	core	1.60	/
	full	1.64	6.55
TDNN	core	1.44	/
	full	1.42	6.16
Fusion	core	1.32	/
	full	1.36	5.27
Baseline	core	2.30	7.26
	full	2.42	6.51

Table III shows the overall error of the three single and the fusion SAD models. Different structures complement each other in the detection performance, so we use a weighted combine to fuse the single models, and achieve an error improvement of 1.06% (from 2.42% to 1.36%) over baseline SAD on full development set. The same fusion strategy is applied on the full evaluation set, and we get an error improvement of 1.24% (from 6.51% to 5.27%).

Table IV and V lists the performance of baseline and our best submission systems in track 1 and track 2, respectively. For both tracks, both sets and both metrics, our systems all significantly outperform baseline, and finally we ranked 1st on the leaderboard among all participants.

TABLE IV
THE DER (%) AND JER(%) COMPARISION OF BASELINE AND USTC-NELSLIP BEST SUBMISSION SYSTEM FOR THE CORE/FULL DEV AND EVAL SETS IN TRACK 1.

System	part.	DER(%)		JER(%)	
		dev	eval	dev	eval
baseline	core	20.25	20.65	46.02	47.74
	full	19.41	19.25	41.66	42.45
USTC-NELSLIP	core	13.30	13.45	34.05	34.94
	full	11.07	11.30	29.48	29.94

TABLE V
THE DER (%) AND JER(%) COMPARISION OF BASELINE AND USTC-NELSLIP BEST SUBMISSION SYSTEM FOR THE CORE/FULL DEV AND EVAL SETS IN TRACK 2.

System	part.	DER(%)		JER(%)	
		dev	eval	dev	eval
baseline	core	22.28	27.34	47.75	51.91
	full	21.71	25.36	43.66	46.95
USTC-NELSLIP	core	14.49	19.37	35.19	39.22
	full	13.25	16.78	31.25	34.42

V. HARDWARE REQUIREMENTS

The infrastructure used to run the experiment was a GPU, TeslaV100-PCIE, with a total memory of 12GB unless specified otherwise.

For audio domain classification, the processing time for the full development set is about 3 minutes, among which about 2 minutes to get the segment-level results, and 1 minute to vote to get senssion-level results and save.

For speech enhancement, the processing time for 10 minutes of audio is about 20s, so the full development set cost about 200 minutes.

The clustering based diarization system processes the full development set using about 6 hours and 40 minutes, among which the time for AHC and BHMM clustering is about 4 hours and 20 minutes.

The ISS based diarization system processes 10 minutes of audio using about 20 seconds, and the total full development set CTS data using about 20 minutes.

The ITS-VAD based diarization system processes 10 minutes of audio using about 10 seconds.

The DNN-SAD, CLDNN-SAD, and TDNN-SAD model process the full development set using about 40 seconds, 45 seconds, and 45 seconds, respectively.

The post processing totally uses about 2 minutes to process the full development set.

REFERENCES

[1] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[2] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[3] D. Wang and X. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[5] L. Sun, J. Du, T. Gao, Y.-D. Lu, Y. Tsao, C.-H. Lee, and N. Ryant, "A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5234–5238.

[6] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[7] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný *et al.*, "But system for the second dihard speech diarization challenge," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6529–6533.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[10] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88. Citeseer, 2005, p. 100.

[11] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[12] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] M. D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 141–145.

[15] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, and C.-H. Lee, "Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7099–7103.

[16] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[17] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[18] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[19] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[20] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," *arXiv preprint arXiv:2005.07272*, 2020.

[21] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," *arXiv preprint arXiv:2011.01997*, 2020.