# The USTC-NELSLIP Systems for DIHARD-III Challenge

*Maokui He*[1]*, Yuxuan Wang*[1]*, ShuTong Niu*[1]*, Lei Sun*[2]*, Tian Gao*[2]*, Xin Fang*[2]*, Jia Pan*[2]*, Jun Du*[1]*, Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]iFlytek Research, Hefei, Anhui, P.R.China
[3]Georgia Institute of Technology, Atlanta, Georgia, USA

hmk1754@mail.ustc.edu.cn, jundu@ustc.edu.cn

## Abstract

This technical report describes our submission system to the Third DIHARD Speech Diarization Challenge. Besides the traditional x-vector based system, the innovation of our system lies in the combination of various front-end techniques to solve the diarization problem, including speech separation, target-speaker based voice activity detection (TS-VAD) and iterative system optimizing. We also adopted audio domain classification to design different processing procedures under corresponding data domains. Finally, we used DOVER-Lap to do system fusion. 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

## 1. System overview

The overall framework of our system is shown in Figure 1. We will introduce the details of each component in the next chapters.

## 2. Track 1: Diarization form reference SAD

### 2.1. Audio Domain Classification

The Third DIHARD corpus mainly consists of 11 different domains [1], which differ on background noise, number of speakers, recording equipment from each other. It can be found that domain-dependent processing can bring additional benefits for final performance. Thus we used a ResNet based network with 17 convolutional layers as the audio domain classification module. The whole development set was divided into 2 parts, 9/10 for training and 1/10 for testing. All utterances were truncated into 10-second segments, each segment was assigned with the corresponding domain label. 64-dimentional log-mel filterbanks were used as acoustic features. During testing, we used voting strategy to get utterance-level classification results. The proposed model achieved 100% accuracy on the development set, and we directly applied it on the evaluation set. If there is no dominant category when testing, then we use a common diarization system without domain-dependent processing.

### 2.2. Speech Enhancement

We employed the progressive multi-target network based speech enhancement model which was introduced in [2]. Here we used the enhanced speech of target1 as the inputs of the following modules.

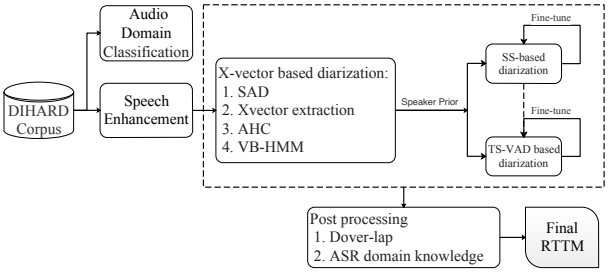


Figure 1: *An illustration of overall framework.*

### 2.3. X-vector based diarization system

In this section, we introduce the traditional speech diarization system which is mainly based on x-vector, AHC, and VB. We used it as the basis system which can yield diarization results as initialization for our proposed systems.

**X-vector Extractor:** We trained a time-delay neural network (TDNN) for extracting x-vectors [3]. The training data were drawn from VoxCeleb 1 and 2, and we used data augmentation using MUSAN corpus. 40-dimentional FBANK features were used for training, and 512-dimentional x-vectors were generated to respresent the speaker's characteristics. 1.5s window length and 0.25s window shift were selected for better performance.

**PLDA Scoring:** We used one PLDA model using the VoxCeleb 1 and 2 corpora, and one in-domain PLDA model using the DIHARD development set [4]. Both models are estimated from centered, whitened and length-normalized x-vectors extracted on 3s segments. On the development set, we combined the two models with a weight value between 0 and 1 for a specific domain selected according to the domain classification results.

**AHC Clustering:** We performed the AHC clustering as our first-stage clustering. The AHC results can be regarded as the initialization for the following VBHMM clustering. So we kept the AHC under-clustering. The threshold used as stopping criteria for the AHC was tuned on the development set according to the domain classification results.

**VBHMM based Clustering:** The VBHMM clustering took the AHC output and was regarded as our second-stage clustering [4]. The segments were assigned to different speakers again based on the first-stage clustering. The parameters $F_a$, $F_b$, $P_{loop}$ and iteration number for different domains were adjusted on the development set according to the domain classification results.

### 2.4. Speech separation based diarization system

By comparing the definitions of speech diarization and speech separation, it's found that these two tasks are very similar. Hence, we proposed a novel speech separation based framework for the speech diarization task, which can inherently cope with overlapping regions. The framework simply contains two parts: separation and detection, and achieved much better performance than traditional diarization methods, especially for telephone data like Fisher. Firstly, we used the Librispeech dataset to simulate 250 hours training data, and trained a fully convolutional time-domain audio separation network (Conv-TasNet) as our pretrained model. Secondly, to eliminate the instability of the model, we added a fine-tune procedure using speaker priors which could be drawn from the traditional diarization system. Simulated utterances were generated to let the pretrained model adapt current utterance. Finally, we used the adapted model to separate the entire utterance. For each separated channel, we directly used a DNN-based SAD to detect speaker presence. After combining all SAD results along the time axis, speech diarization results were attained where overlapped regions were automatically labeled. Because of time limitation, we only used this proposed system in telephone speech domain in our submitted system.

### 2.5. TS-VAD based diarization system

The TS-VAD [5] model takes FBANKs as input, along with i-vectors corresponding to the speakers, and predicts per-frame speech activities for all the speakers simultaneously. We used Voxceleb1 and Voxceleb2 with data augmentation techniques to train the i-vector extractor. We used Librispeech corpus to simulate multi-talker conversations as training data, and we also collected realistic conversations from Swithboard, AMI and Voxconverse. When training, if speaker number in an utterance is smaller than the number of output nodes $N$, we assign the remain nodes to dummy speakers selected from the training set and labeled them with silence. If the number of speakers is larger than $N$, top $N$ speakers with longer non-overlapping time are selected as training targets and the other speakers are discarded. The same strategy was adopted when testing, where speaker number was estimated from traditional VB-HMM results.

To eliminate the instability of the TS-VAD model, we also applied fine-tune strategy here. We used non-overlap segments as speaker priors which were generated from VB-HMM, and simulated multi-talker conversations to fine-tune the model. Finally, the fine-tuned TS-VAD model directly estimate presence probability of each speaker. In addition, one advantage of our system is that it can iteratively update the priors to further enhance the model performance.

Table 1: *Track 1 diarization results for the DEV and EVAL sets*

| System | Part. | DER (%) | | JER (%) | |
|---|---|---|---|---|---|
| | | Dev | Eval | Dev | Eval |
| DIHARD baseline | core | 20.25 | 20.65 | 46.02 | 47.74 |
| | full | 19.41 | 19.25 | 41.66 | 42.45 |
| Best Single System | core | 14.07 | 14.86 | 35.63 | 36.75 |
| | full | 12.22 | 12.41 | 30.96 | 31.54 |
| Post-processing | core | 13.30 | 13.45 | 34.05 | 34.94 |
| | full | 11.07 | 11.30 | 29.48 | 29.94 |

### 2.6. Post-processing

We designed a post-processing module to synthesize the results of all subsystems. Dover-lap was used to do system fusion [6], it can effectively improve overall performance. Specially, we utilized domain knowledge from the speech recognition task, which brings two kinds of information. One is the segmentation results of the recognizer can be used as an ASR-based VAD. The other one is that the recognized token [laugh] indicates where overlapping speech segments often occur, especially in multi-talker scenarios. The acoustic model was trained on CHiME-6 corpus, and we simply assigned the closest speakers if the token appeared.

## 3. Track 2: Diarization form scratch

Most modules are the same with Track1, we put more efforts on the SAD part.

### 3.1. SAD models

We employed three different networks for SAD training. The FCN model adopts 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 CLDNN model adopts CNN, LSTM and FCN. The TDNN model was equipped with the same structure as the DIHARD III SAD baseline. All of those models were fine-tuned on enhanced DIHARD III development sets and decoded on enhanced DIHARD III evaluation sets. The ASR-based VAD was also used as a supplement. Finally, we fused those systems through a voting strategy.

Table 2: *Track 2 diarization results for the EVAL sets*

| System | Part. | DER (%) | JER (%) |
|---|---|---|---|
| DIHARD baseline | core | 27.34 | 51.91 |
| | full | 25.36 | 46.95 |
| Best Single System | core | 20.66 | 40.73 |
| | full | 17.70 | 35.76 |
| Post-processing | core | 19.37 | 39.22 |
| | full | 16.78 | 34.42 |

## 4. References


[1] 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.

[2] 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*. IEEE, pp. 7099–7103.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018.

[4] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černockỳ, "Optimizing bayesian hmm based x-vector clustering for the second dihard speech diarization challenge," in *ICASSP 2020*. IEEE, pp. 6519–6523.

[5] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Y. Khokhlov, and etc., "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," *ArXiv*, vol. abs/2005.07272, 2020.

[6] 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.