# USC-SAIL System for DIHARD III: Domain Adaptive Diarization System

*Tae Jin Park, Raghuveer Peri, Arindam Jati and Shrikanth Narayanan*

University of Southern California, Signal Analysis and Interpretation Laboratory (SAIL)

taejinpa@usc.edu, rperi@usc.edu, jati@usc.edu, shri@sipi.usc.edu

## Abstract

DIHARD challenge focuses on the hard diarization problem and the DIHARD dataset includes a number of challenging domains that are very challenging to obtain low diarization error rates. We propose a novel approach to deal with domain mismatch problems by estimating the domain of the given input session. We take advantage of three different embedding extractors trained on different datasets. Based on these multiple embedding extractors, our domain adaptive speaker diarization system employs two different approaches: Hard decision and soft decision. In the hard decision method, we estimate the given session into one of the three categories and select an embedding extractor suited to that category. On the other hand, in the soft decision method, we train a neural affinity score fusion network that estimates the desirable weights between affinity scores we obtain from the three embedding extractors. We show the performance gain of each method and how our domain estimator models are trained. In addition, we introduce the auto-tuning spectral clustering method to enable a parameter-free speaker diarization system.

**Index Terms**: Speaker Diarization, Domain Adaptation, DIHARD 3

## 1. Introduction

Speaker diarization often suffers from sparse training dataset since training dataset for speaker diarization is not as abundant as training dataset for other applications such as automatic speech recognition (ASR). To tackle this issue, we propose a domain adaptive speaker diarization system that estimates the most suitable speaker embedding extractor or estimates the weights between the speaker embedding extractors that are trained on different dataset. We show the performance gains based on the methods we propose for track 1.

## 2. System Description

### 2.1. Dataset and Speaker Embedding Extractor

In USC-SAIL DIHARD3 speaker diarization system, we employ three different x-vector models [1] trained on different datasets with a few modifications in the x-vector model architecture. x-vector CH is trained on SRE challenge dataset, x-vector VOX is trained on Voxceleb dataset and x-vector CHIME is trained on Voxceleb dataset and then adapted on CHIME-6 dataset. We use window length of 1.5 s, hop-length of 0.25 s and minimum window length of 0.5 s. We employ cosine similarity to measure the similarity between two speaker embedding vectors. For the hard decision method, we only use affinity matrix that is obtained from a single x-vector model while we use weighted sum of three affinity matrices in the soft decision method. Based on the affinity matrix we get from domain adaptation process, we employ the same clustering method for all the given sessions.

### 2.2. Domain Classifier

#### *2.2.1. Hard Decision Method*

As previously mentioned, we employ three different speaker embeddings, as we found different embeddings to be optimal for different domains. Motivated by the findings from Table 1, we developed a classifier that can predict the speaker embedding to use for each session. We classify each session into one of the three categories, and employ the corresponding speaker embedding to perform speaker diarization. Our approach can be thought of as a crude domain classifier, where our goal is not to accurately predict the domain, but to broadly assign each session to its optimal speaker embedding. This kind of domain grouping has been previously explored in [2], but in this work we grouped them based on the optimal speaker embedding to use. Hereafter we will refer to the classifier we developed as domain classifier.

We employed a deep neural network as our domain classifier. It takes the concatenation of the three speaker embeddings as input. It consists of two 1-D convolutional layers followed by an average pool layer resulting in fixed dimensional embeddings. The embeddings are then passed through a fully-connected layer which assigns probabilities to each class. The network was trained on the Development Set using cross-entropy loss.

#### *2.2.2. Soft Decision Method*

For the soft decision method, we employ a neural network architecture that is similar to Siamese-network [3] and we refer to this neural network model as Neural Affinity Score Fusion (NASF) module. We use weight sharing network and feed three different set of x-vector embeddings (CH, VOX and CHIME6) from two different segments. Thus, there are six different embedding vectors (two sets of three) that are fed to NASF module. The NASF module outputs weight between these three x-vector inputs and we use this weight to calculate the weighted sum of three affinity matrices. The final weighted affinity matrix is then fed to clustering module to obtain speaker labels.

### 2.3. Clustering

We use the auto-tuning clustering method appeared in [4] which does not require development set for tuning the clustering algorithm. We employ sparse-search where we only allow maximum of 20 threshold values to be searched.

## 3. Experimental results

### 3.1. Development Set

The following table shows the DER achieved by each x-vector type for each domain. Note that this result is obtained from the FULL set, Track 1.

Table 1: *Dev Set DER for each domain and x-vector type.*

| Track 1 | x-vector Type | | |
|---|---|---|---|
| **Domain** | **CH** | **VOX** | **CHIME6** |
| Audiobooks | 0.46 | 1.74 | 1.21 |
| Broadcast Interview | 4.19 | 4.05 | 5.88 |
| Clinical | 23.66 | 20.96 | 16.63 |
| Court | 4.63 | 9.9 | 5.09 |
| CTS | 16.3 | 18.91 | 19.13 |
| Maptask | 6.69 | 6.37 | 6.75 |
| Meeting | 31.44 | 30.85 | 28.3 |
| Restaurant | 57.71 | 55.27 | 52.64 |
| Socio Field | 16.21 | 14.24 | 13.78 |
| Socio Lab | 8.45 | 8.52 | 9.61 |
| Webvideo | 41.32 | 39.24 | 40.66 |
| Total | 19.89 | 20.52 | 19.44 |

### 3.2. Evaluation Set

The following table shows the results we obtain from the hard decision model and soft decision model.

Table 2: *Evaluation set results for Track 1.*

| | CORE | | FULL | |
|---|---|---|---|---|
| Type | DER | JER | DER | JER |
| x-vector CHIME6 | 22.140 | 48.850 | 20.310 | 43.700 |
| Hard Decision | 22.250 | 48.370 | 19.660 | 42.520 |
| Soft Decision | 19.760 | 43.030 | 18.190 | 38.330 |

## 4. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 5329–5333.

[2] M. Sahidullah, J. Patino, S. Cornell, R. Yin, S. Sivasankaran, H. Bredin, P. Korshunov, A. Brutti, R. Serizel, E. Vincent *et al.*, “The speed submission to dihard ii: Contributions & lessons learned,” *arXiv preprint arXiv:1911.02388*, 2019.

[3] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *Proceedings of the Workshop on Deep Learning in International Conference on Machine Learning, ICML*, 2015.

[4] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.