

SELF-SUPERVISED METRIC LEARNING WITH GRAPH CLUSTERING FOR SPEAKER DIARIZATION

Prachi Singh and Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Electrical Engineering, Indian Institute of Science, Bangalore, India.
prachisingh@iisc.ac.in, sriramg@iisc.ac.in

ABSTRACT

In this paper, we propose a novel algorithm for speaker diarization using metric learning for graph based clustering. The graph clustering algorithms use an adjacency matrix consisting of similarity scores. These scores are computed between speaker embeddings extracted from pairs of audio segments within the given recording. In this paper, we propose an approach that jointly learns the speaker embeddings and the similarity metric using principles of self-supervised learning. The metric learning network implements a neural model of the probabilistic linear discriminant analysis (PLDA). The self-supervision is derived from the pseudo labels obtained from a previous iteration of clustering. The entire model of representation learning and metric learning is trained with a binary cross entropy loss. By combining the self-supervision based metric learning along with the graph-based clustering algorithm, we achieve significant relative improvements of 60% and 7% over the x-vector PLDA agglomerative hierarchical clustering (AHC) approach on AMI and the DIHARD datasets respectively in terms of diarization error rates (DER).

Index Terms— Speaker diarization, x-vectors, path integral clustering, neural PLDA, self-supervised learning.

1. INTRODUCTION

Speaker diarization, the problem of unsupervised temporal sequence segmentation into speaker specific regions, is one of first processing steps in the conversational analysis of multi-talker audio. The performance of a speaker diarization system is adversely influenced by factors like short speaker turns, overlaps between multiple speakers, far-field effects in audio recording and environmental artifacts. The DIHARD evaluations explored various challenging environments for bench-marking diarization performance [1–3].

In the past decade, the dominant approach to speaker diarization involved a two step process. The first step consists of deriving embeddings from relatively short windowed segments of speech (typically 1 – 2s of audio) while the second step involves the clustering of the embeddings. The early breakthroughs were reported with the i-vector embedding extraction using unsupervised factor analysis modeling [4]. The availability of large amounts of speaker supervised speech recordings along with the advancements in deep learning have propelled the use of deep neural network (DNN) based embedding extractors like the x-vectors [5].

In terms of clustering, the common approach to speaker diarization is the bottom-up clustering approach where the goal is to suc-

cessively merge to achieve a one-to-one correspondence between the ground-truth speakers and clusters [6]. The most popular approach is the agglomerative hierarchical clustering (AHC) [7]. The inputs to the clustering algorithms commonly employ pre-processing techniques on the embeddings like length normalization [8], principal component analysis (PCA) [9] and PLDA based affinity matrix computation [4]. Another common approach to clustering is the spectral clustering approach [10]. In most of these approaches, the affinity matrix computation and clustering are performed as two independent steps with different cost functions.

A neural diarization approach (termed as end-to-end (EEND)) [11, 12], proposed recently, explores transformer models for speaker diarization. The key successes reported have been for recordings with 2 – 3 speakers. However, training the EEND system with more speakers is challenging because of permutation-invariant loss computation. Further, the performance on recordings with more than 4 speakers does not improve over the clustering based approaches [13].

The self-supervision in diarization can provide effective representations for downstream tasks without requiring the ground-truth labels [14]. In this paper, we extend our previous work, based on path integral clustering (PIC) [15, 16], with metric learning framework inspired by neural PLDA [17]. The previous work explored graph based clustering with self-supervised representation learning. The learned representations were used to derive cosine affinity score based adjacency matrix. This adjacency matrix, containing pair-wise similarity scores, was used in path integral clustering for speaker diarization [15]. In the proposed work, we further explore a learnable metric based on neural PLDA in the self-supervised learning framework. In particular, both the embeddings and the adjacency matrix for graph based clustering are jointly learned. Using this joint learning, we show significant performance improvements over baseline systems and previous models based on self-supervised graph clustering methods [15, 16].

2. RELATED WORK AND CONTRIBUTIONS

The most common clustering approach used in speaker diarization is based on agglomerative hierarchical clustering (AHC) [7]. The affinity measures explored for AHC consists of cosine similarity score [18] or PLDA [4, 19]. Other methods for clustering include k-means [20] and spectral clustering [21]. The long short term memory (LSTM) network used for affinity computation, proposed by Lin et. al. [22] and a fully supervised speaker diarization, using unbounded interleaved recurrent neural networks (RNN), proposed by Zhang et. al. [23], have been investigated recently.

The approaches based on re-segmentation [24,25] consists of using the clustering based results to initialize a hidden Markov model

(HMM) based clustering. Recently, the x-vector embedding based HMM variational Bayes, termed as VBx, has shown promising results [26].

For self-supervised clustering, the loss functions based on k-means [27], spectral clustering loss [28] and agglomerative clustering [29] have been investigated for image/text. A recent work based on neural modeling of discriminatively trained PLDA model [17] was proposed for speaker verification. The neural formulation of PLDA allows the learning of the parameters in a Siamese network.

This paper extends our previous works on self-supervised learning and graph-based clustering [15, 16]. The previous works proposed representation learning and graph based clustering in an iterative self-supervised learning framework. The triplet training loss based on cosine similarity was used in the representation learning. The final embeddings were used with the agglomerative hierarchical clustering (AHC) [16] or with a more robust graph based clustering called path integral clustering (PIC) [15]. The affinity measures in both the approaches was based on the cosine similarity score.

In this work, we include the PLDA parameters as a learnable metric in the self-supervised learning framework inspired by Ramoji et. al [17]. The PLDA parameters are learnable along with the neural network parameters for the embedding extraction. The entire model is trained using binary cross entropy loss (BCE). The advantage of the proposed approach over the previous work is the direct optimization of the adjacency matrix used in the graph based clustering. We call the proposed approach as self-supervised PLDA based metric learning with path integral clustering (SelfSup-PLDA-PIC).

3. BACKGROUND

This section describes the pre-processing steps, the probabilistic linear discriminant analysis (PLDA) model and the path integral clustering algorithm used in our approach.

3.1. Pre-processing steps

The x-vectors of each segment in a given recording, extracted using the TDNN network [5], are mean normalized and whitened. The whitened x-vector features are then processed using length normalization [8]. Further, a principal component analysis (PCA) based dimensionality reduction at the recording level is also applied [9].

3.2. PLDA model

The PLDA is a generative model which factorizes the input into the speaker factor and the channel factor. The simplified and widely used model is a generative linear-Gaussian model [19], where x-vector $\mathbf{x} \in \mathbb{R}^D$ represents segment embedding. The vector $\mathbf{y} \in \mathbb{R}^D$ represents the speaker factor. The distribution of \mathbf{x} given the speaker factor \mathbf{y} is assumed to be a Gaussian distribution,

$$p(\mathbf{x}|\mathbf{y}) = N(\mathbf{x}; \mathbf{y}, \Phi_w) \quad (1)$$

where $\Phi_w \in \mathbb{R}^{D \times D}$ is the within class covariance matrix. The latent vector \mathbf{y} is assumed to be distributed according to prior distribution:

$$p(\mathbf{y}) = N(\mathbf{y}; \mathbf{m}, \Phi_b) \quad (2)$$

where $\mathbf{m} \in \mathbb{R}^D$ and $\Phi_b \in \mathbb{R}^{D \times D}$ are global mean and between-speaker covariance matrix respectively. Further, Φ_w and Φ_b can be simultaneously diagonalized using diagonalizing transform $\mathbf{V} \in \mathbb{R}^{D \times D}$:

$$\mathbf{V}\Phi_w\mathbf{V}^T = \mathbf{I}, \quad \mathbf{V}\Phi_b\mathbf{V}^T = \Psi \quad (3)$$

Algorithm 1: Path Integral Clustering (Sec. 3.3)

Initialize: Construct a graph $G = (V, E)$ where, vertices V are the input data $\mathbf{X} = \{x_1, x_2, \dots, x_{N_r}\}$. The weighted adjacency matrix \mathbf{W} is computed and the transition probability matrix \mathbf{P} is obtained by normalizing \mathbf{W} ; Form n_c initial clusters $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_{n_c}\}$ by assigning each sample x_i to a cluster, using nearest neighbor merging; N^* = required number of speakers
while $n_c > N^*$ **do**
 1. Merge \mathcal{C}_a and \mathcal{C}_b , if $\{\mathcal{C}_a, \mathcal{C}_b\} = \operatorname{argmax}_{\mathcal{C}_a, \mathcal{C}_b \in \mathcal{C}} \mathcal{A}(\mathcal{C}_a, \mathcal{C}_b)$
 where, $\mathcal{A}(\mathcal{C}_a, \mathcal{C}_b)$ is given in Equation (9)
 2. $\mathcal{C}_c \leftarrow \{\mathcal{C}_c \setminus \{\mathcal{C}_a, \mathcal{C}_b\} \cup \{\mathcal{C}_a \cup \mathcal{C}_b\}\}$ and $n_c = n_c - 1$
 3. Recompute \mathcal{A}
end
Termination: \mathcal{C}_c

where Ψ is a diagonal covariance matrix. Given this model, the generative graphical model can be expressed in terms of \mathbf{u}, \mathbf{v} , where \mathbf{u}, \mathbf{v} denote the latent vectors representing speaker variable and the speaker embedding in the projected space respectively. If $\mathbf{A} = \mathbf{V}^{-1}$, the generative model is expressed as:

$$\begin{aligned} p(\mathbf{v}) &= N(\cdot | \mathbf{0}, \Psi) \\ p(\mathbf{u}|\mathbf{v}) &= N(\cdot | \mathbf{v}, \mathbf{I}), \\ \mathbf{x} &= \mathbf{m} + \mathbf{A}\mathbf{u} \end{aligned} \quad (4)$$

3.2.1. PLDA Score computation

The embeddings from the PLDA model are used to obtain the pairwise similarity score matrix \mathbf{S} which captures the similarity between embeddings in the speaker space. The similarity score between a pair of embeddings \mathbf{x}_i and \mathbf{x}_j , denoted as $s(i, j)$, is based on the log-likelihood ratio score between same-speaker hypothesis \mathcal{H}_s and different-speaker hypothesis \mathcal{H}_d . This can be computed using the PLDA model. We project the embeddings \mathbf{x} into latent space using Equation (4),

$$\mathbf{u} = \mathbf{A}^{-1}(\mathbf{x} - \mathbf{m}) = \mathbf{V}(\mathbf{x} - \mathbf{m}) = \mathbf{V}\mathbf{x} - \mathbf{b} \quad (5)$$

The similarity score can be computed as [19]:

$$\begin{aligned} s(i, j) &= -\frac{1}{2} \sum_{k=1}^d \log(\Psi[k] + \frac{1}{2}) + 2 \log(\Psi[k] + 1) \\ &+ \frac{(\bar{\mathbf{u}}[k])^2}{\Psi[k] + \frac{1}{2}} + \sum_{l \in \{i, j\}} \left((\mathbf{u}_l[k] - \bar{\mathbf{u}}[k])^2 + \frac{(\mathbf{u}_l[k])^2}{\Psi[k] + 1} \right) \end{aligned} \quad (6)$$

where $\Psi[k]$ is the diagonal element of the k-th row of Ψ , $\mathbf{u}_i[k]$ is the k-th dimension of \mathbf{u}_i , and $\bar{\mathbf{u}}[k] = \frac{\mathbf{u}_i[k] + \mathbf{u}_j[k]}{2}$

3.3. The path integral clustering

The path integral clustering (PIC) [30] is a graph-based agglomerative clustering algorithm, introduced in [15] for speaker diarization.

In PIC, a directed graph is created such that vertices represent the input features connected by the set of edges.

Let the x-vector embeddings from a recording r be denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_r}\} \in \mathbb{R}^D$, where N_r is the total number of embeddings present. We compute an adjacency matrix \mathbf{W} using the similarity scores (cosine or PLDA scores). From the adjacency

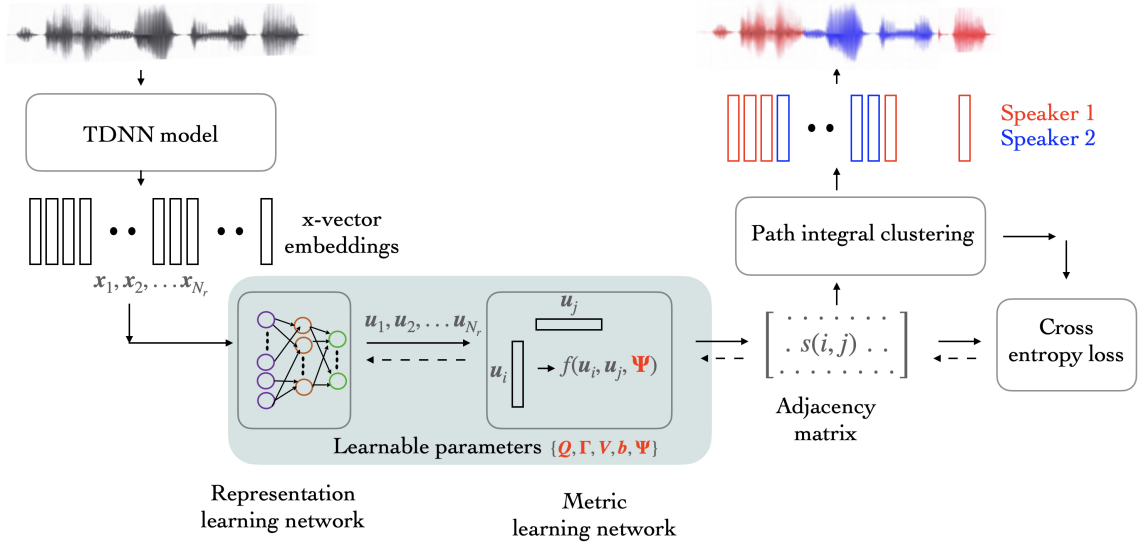


Fig. 1. Block schematic of the proposed self-supervised metric learning approach to speaker diarization.

matrix, we select only the K -nearest neighbors of x_i . The matrix \mathbf{W} is converted to transition probability matrix \mathbf{P} by dividing each row with its row sum.

The PIC involves computation of path integral \mathcal{S}_{C_a} and conditional path integral $\mathcal{S}_{C_a|C_a \cup C_b}$ for every cluster pair C_a and C_b at each step of merging as follows:

$$\mathcal{S}_{C_a} = \frac{1}{|C_a|^2} \mathbf{1}^T (\mathbf{I} - \sigma \mathbf{P}_{C_a})^{-1} \mathbf{1} \quad (7)$$

$$\mathcal{S}_{C_a|C_a \cup C_b} = \frac{1}{|C_a|^2} \mathbf{1}_{C_a}^T (\mathbf{I} - \sigma \mathbf{P}_{C_a \cup C_b})^{-1} \mathbf{1}_{C_a} \quad (8)$$

where, $|C_a|$, cardinality of C_a , is used for normalization of the path integrals. \mathbf{P}_{C_a} and $\mathbf{P}_{C_a \cup C_b}$ are the sub-matrices of the transition probability matrix \mathbf{P} , column vector $\mathbf{1}$ is a vector of all ones and of size $|C_a|$ and $\mathbf{1}_{C_a} \in \mathbb{R}^{|C_a \cup C_b|}$ is a binary column vector containing ones and zeros corresponding to the nodes of C_a and C_b respectively. The scalar $0 < \sigma < 1$ introduces discounting of longer paths. The cluster affinity measure for the PIC algorithm is computed as,

$$\mathcal{A}(C_a, C_b) = \mathcal{S}_{C_a|C_a \cup C_b} - \mathcal{S}_{C_a} + \mathcal{S}_{C_b|C_a \cup C_b} - \mathcal{S}_{C_b} \quad (9)$$

where, $\mathcal{S}_{C_a|C_a \cup C_b} - \mathcal{S}_{C_a}$ is the incremental path integral of C_a . It represents the sum of weighted paths between C_a and C_b such that the starting and ending vertices are in C_a . Thus, higher affinity shows denser connections between clusters. The clusters with maximum affinity are merged at each time step. The pseudocode is given in Algorithm 1.

4. PROPOSED APPROACH

The block schematic of the proposed self-supervised metric learning with graph based clustering algorithm (SelfSup-PLDA-PIC) is given in Figure 1. The model consists of a representation learning network and a metric learning network. The x-vectors extracted from short overlapping audio segments are used as inputs to the model. The model generates adjacency matrix which is used in PIC.

The SelfSup-PLDA-PIC¹ jointly performs representation learning and metric learning using the initial clustering results. The output of clustering generates speaker labels. These labels are used to form same speaker and different speaker score level targets for the adjacency matrix. The model training is performed using the binary cross entropy (BCE) loss using the target adjacency matrix. Since the model updates the parameters based on the unsupervised clustering labels, it is known as self-supervised training. Our previous work involved self-supervised representation learning using fixed cosine similarity scoring [16] whereas, here we have introduced metric learning block inspired by PLDA scoring (Sec. 3.2.1). The similarity score is computed using Equation (6), where, along with embeddings \mathbf{u} , the PLDA parameters Ψ is also learned. Therefore, it is also referred as neural PLDA [17]. In the following subsection, we discuss model architecture and the joint representation learning and metric learning approach.

4.1. Model architecture and training

The representation learning network is a three layer DNN with $\{D, d, d\}$ units where D is the x-vector dimension. It takes x-vectors of a recording r as input and generates d -dimensional embeddings $\mathbf{u} = \{\mathbf{u}_1, \dots, \mathbf{u}_{N_r}\}$. Let, $\{\mathbf{Q}, \mathbf{\Gamma}, \mathbf{V}\}$ denote the learnable weights of each layer respectively and let \mathbf{b} denote the bias of last layer. The embeddings \mathbf{u} are passed to the metric learning network, which performs pairwise neural PLDA scoring (discussed in Sec. 3.2.1) using learnable parameter $\Psi \in \mathbb{R}^{d \times d}$. In the forward pass, we generate adjacency matrix \mathbf{W} using pairwise PLDA scores and perform graph based path integral clustering (discussed in Sec. 3.3).

In the backward pass, we compute the ideal adjacency matrix using the clustering solution from the PIC step. The ideal adjacency matrix is a binary matrix consisting of target (label of 1) and non-target speaker similarity (label of 0) scores. Using the target and model output, a binary cross entropy (BCE) based loss function is used to update the learnable parameters $\{\mathbf{Q}, \mathbf{\Gamma}, \mathbf{V}, \mathbf{b}, \Psi\}$. A sig-

¹github code link: <https://github.com/iiscleap/SelfSup-PLDA.git>

moid non-linearity is also applied on the neural PLDA scores before the BCE loss computation.

5. EXPERIMENTS

5.1. Evaluation data

- **AMI:** The AMI dataset [31] contains meeting recordings from four different sites (Edinburgh, Idiap, TNO, Brno). The official speech recognition partition of AMI dataset comprises of development and evaluation sets consisting of 18 and 16 recordings sampled at 16kHz respectively. We use the single distant microphone (SDM) condition of AMI dataset for experiments. We also compare results on beamformed multi-distant microphone (MDM) recordings with other published results. The number of speakers and duration ranges from 3-5 and 20-60 mins respectively. For the AMI dataset experiments, we use the diarization error rate (DER) metric with a 250ms collar and by ignoring the overlap regions (as is the common practice on the AMI dataset).
- **DIHARD III:** The third DIHARD challenge dataset [3] was released as the third in series of DIHARD speech diarization challenges. It consists of development and evaluation sets of recordings with duration of 0.5-10 mins. These recordings are drawn from 11 domains including audio-books, telephone recordings, clinical interviews, restaurant conversations, web videos etc. The number of speakers varies from 1-10 with diverse regions of overlapping speech and speaker turn behavior. There are 254 and 259 recordings in the development and evaluation sets respectively. For the DIHARD dataset experiments, we use the DER metric with the overlaps and without providing a collar region.

5.2. Baseline Model

Our baseline model is based on DIHARD III baseline recipe [3]. It involves feature extraction followed by x-vector embedding extraction. The x-vectors are extracted using the extended-TDNN (ETDNN) [32] network. For training the ETDNN model, we use 40D mel-filterbank features using a 25ms window with 10ms shift.

The 13-layer ETDNN model follows the architecture described in [32, 33]. The ETDNN model is trained on the VoxCeleb1 [34] and VoxCeleb2 [35] datasets, for speaker identification task, to discriminate among the 7, 146 speakers. It has 4 TDNN layers which alternates with 4 fully connected layers of size 1024D. This is followed by 2 feed forward layers containing {1024, 2000} units. The segment pooling layer is of 4000D, containing mean and standard deviation computed at the segment level for the 1500D layer. From the segment level features, the 512 dimensional output of the affine component from the 11th layer is taken as the x-vector embedding. The pre-processing steps mentioned in section 3.1 are applied to x-vectors which include whitening transform obtained from the DIHARD development set, length normalization and recording level PCA. We preserve 30 dimensions for the AMI dataset. For the DIHARD dataset, we choose the number of dimensions which preserves 30% of total variance.

The PLDA model is trained using 3sec x-vectors extracted from the subset of Voxceleb-1 and 2. These x-vectors are whitened using a PCA transform learned from DIHARD development set. We use the same PLDA model for the DIHARD and AMI datasets. However, we extract the embeddings using a segment size of 1.5 s and a temporal shift of 0.75 s for the AMI dataset due to longer duration recordings while the segments of size 1.5 s are extracted with a shift of 0.25 s for the DIHARD dataset.

5.3. Model Initialization

The initialization is a critical step for self-supervised training to generate reliable labels. We initialize our model parameters $\{Q, \Gamma\}$ using the whitening transform and the recording level PCA from the baseline system respectively. The third layer's weight and bias $\{V, b\}$ of representation learning network are initialized with diagonalizing transform V and bias b from PLDA Equation (5), obtained after applying recording level PCA transform. Similarly, we initialize metric learning parameter Ψ using diagonal between class covariance matrix defined in Equation (3).

We perform initial clustering (AHC/PIC) till initial N^0 number of clusters. The value of N^0 is based on a threshold applied to the similarity scores for the AHC system. With the PIC, we use the stopping threshold applied on eigen-values of PIC affinity matrix (Equation (9)) [15] to estimate N^0 . We select a threshold higher than the optimal threshold to avoid over-clustering.

For the AMI dataset, the AHC threshold is set as $th = 0.0$ to obtain N^0 for SelfSup-PLDA-AHC training. For the SelfSup-PLDA-PIC system, the eigen-value based threshold is set at $th = 0.7$ for initial clustering in the self-supervised training. For the DIHARD dataset, the threshold for the SelfSup-PLDA-AHC system is set at $th = -0.7$. Since the DIHARD dataset has huge variation in number of speakers (1-10), the number of speakers estimated by AHC is used as N^0 in SelfSup-PLDA-PIC experiments.

5.4. Choice of hyper parameters

The hyper-parameters involved in our approach are selected based on the performance on the development set of both datasets. The nearest neighbour K and scaling factor σ are the hyper-parameters for PIC. The value of $K = 30$ and $\sigma = 0.1$ are the best values obtained for the AMI dataset. Similarly for DIHARD dataset, we found the best values of $K = 40$ and $\sigma = 0.5$. After model training, a temporal continuity of similarity scores can be incorporated [15]. This is done by multiplying the similarity score $s(i, j)$ with an exponential decay given by,

$$s'(i, j) = s(i, j)\beta^{\min(n_b, |i-j|)} \quad (10)$$

where, β is a positive decay factor < 1 , $|i-j|$ is the absolute segment index difference value of embeddings from the i th and j th segment, and n_b is the maximum value of the decay constant β .

6. RESULTS

6.1. AMI dataset

The results for various system configurations with the AMI development and evaluation datasets are reported in Table 1. These experiments use the ETDNN based x-vectors. We consider two cases in evaluation, with the known number of speakers N^* and with the unknown number of speakers. The baseline system is the x-vector with PLDA scoring and AHC. The use of graph based clustering with PIC improves the baseline system significantly.

The self-supervised clustering (SSC) with cosine based affinity matrix proposed previously [15] further improves over the PIC based system. The joint metric learning with the representation learning, proposed in this paper, denoted as SelfSup-PLDA, is shown to provide significant improvements over the previously proposed SSC-Cosine model. The relative DER improvement over the baseline system for the SelfSup-PLDA with temporal continuity is 66% and 60% for the AMI development and evaluation datasets for the condition with unknown number of speakers. Further, the application of

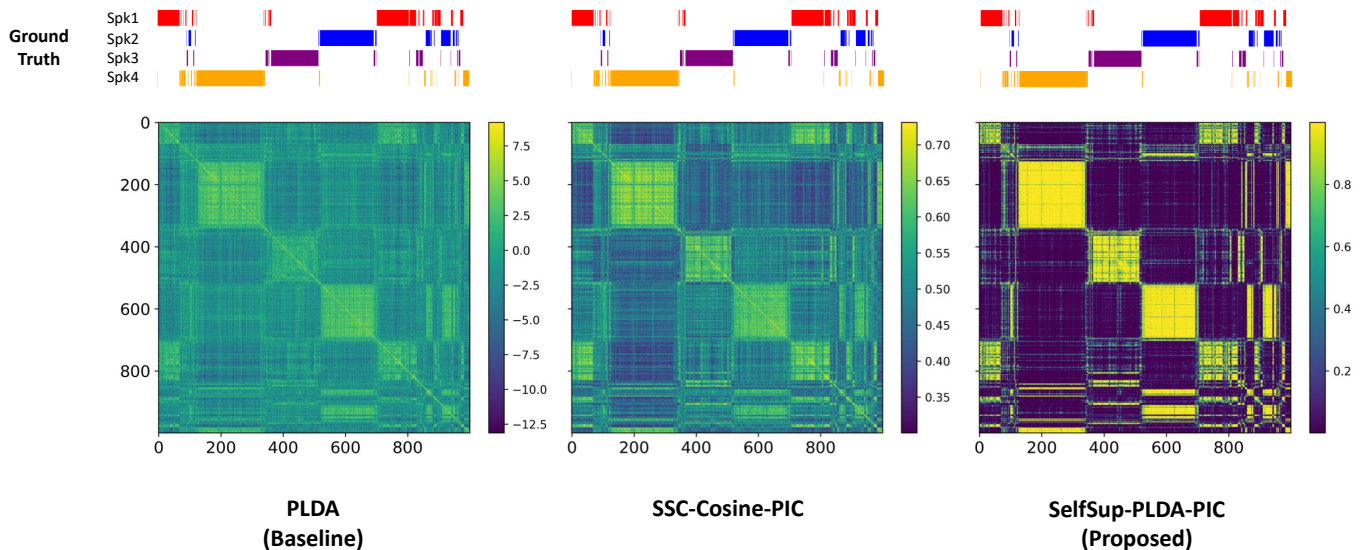


Fig. 2. Similarity score matrices using PLDA, SSC-Cosine-PIC [15] and SelfSup-PLDA-PIC (proposed) for a 4-speaker recording from AMI development set. The ground truth labels are plotted across time on top of affinity matrices for comparison.

VBx re-segmentation [36] on the outputs from the SelfSup-PLDA-PIC system provides final DER values of 2.9% and 4.2% on the development and evaluation datasets. Our VBx setup is based on baseline ETDNN x-vectors with PLDA adapted from DIHARD dev set. To the best of our knowledge, these results on the SDM data of the AMI corpus constitute the lowest DER reported in literature.

The results using residual network (ResNet101) based x-vector embeddings [26] are shown in Table 2. For these experiments, we use the ResNet 101 architecture. The first layer is a 2D convolutional layer. This is followed by four residual blocks. Each residual block consists of 3 residual convolutions. The output of the residual blocks is fed to the statistics pooling layer (from each of the 8 heads). The dense network layer following the pooling layer is used to extract the ResNet x-vectors. The training data and the cost function used to train the ResNet model is similar to the ETDNN framework.

Comparing the system with PLDA scoring and PIC for the ETDNN x-vectors (Table 1) and the ResNet x-vectors (Table 2), we find that the ResNet x-vectors improve significantly over the ETDNN based x-vectors although both models contain similar no. of parameters ($\sim 10^6$). The performance on the evaluation data for this system is 6.2% DER. Further, even with this improved baseline model, the proposed approach of SelfSup-PLDA with graph based clustering and the incorporation of the temporal continuity yields significant improvements. For the self supervised metric learning, it is seen that the final results from either of the x-vector models achieve similar DER results on the AMI evaluation dataset.

6.2. Adjacency matrix analysis

The similarity score matrix (adjacency matrix used in graph clustering) for the baseline x-vector PLDA system (left), self-supervised representation learning with cosine scoring [15] (middle) and the proposed self-supervised metric learning (right) are shown in Figure 2. The adjacency score matrix used in the proposed approach is processed with sigmoid non-linearity for training with BCE loss. The ground truth speaker activity for the four speakers in this record-

Table 1. DER (%) using ETDNN x-vectors on the AMI dataset.

System	Known N^*		Unknown N^*	
	Dev.	Eval.	Dev.	Eval.
x-vec + PLDA + AHC	15.9	12.2	13.1	12.3
x-vec + PLDA + PIC	5.1	10.2	5.8	11.4
SSC-Cosine-PIC [15]	5.3	6.2	6.5	8.4
SelfSup-PLDA-AHC	7.9	7.3	7.7	9.4
SelfSup-PLDA-PIC	4.2	6.2	4.4	6.9
+ Temporal continuity	4.2	4.2	4.4	4.9
SelfSup-PLDA-PIC + VBx [36]	-	-	2.9	4.2

Table 2. DER (%) using the Resnet x-vectors on the AMI dataset.

System	Unknown N^*	
	Dev.	Eval.
x-vec + PLDA + PIC	6.0	6.2
SelfSup-PLDA-PIC	4.6	6.0
+ Temporal continuity	4.4	4.3
SelfSup-PLDA-PIC + VBx [36]	3.4	4.5

ing is also shown in this figure. The same speaker regions of the similarity matrix with the baseline x-vector PLDA system are not well pronounced. The self-supervised embedding learning with cosine scoring improves the contrast between the scores from same speaker and across speaker segments. The proposed approach of metric learning with self-supervised principles is seen to provide the best contrast between the scores from same speaker and across speaker regions of the given audio recording. This increase in contrast partly explains the significantly improved DER results observed in Table 1 for the proposed SelfSup-PLDA+PIC model.

6.3. Comparison with prior literature

We attempt to compare the recent works proposed in Landini et. al. [26] and Dawalatabad et. al. [37] with the work proposed in this

Table 3. DER (%) on the MDM recordings of AMI dataset (without TNO recordings).

System	Unknown N^*	
	Dev.	Eval.
x-vec(ResNet101)+AHC+VBx [26]	2.78	3.09
ECAPA-TDNN [37]	3.66	3.01
x-vec(ETDNN)+ SelfSup-PLDA-PIC	5.38	4.63
- + VBx [36]	2.18	3.27

paper. These previous works report using the beamformed audio from the AMI corpus (multi-distant microphone or MDM), while all the previous results reported in this work used the more challenging single distant microphone (SDM) condition. In order to make the direct comparison, we did not perform any adaptation or fine-tuning on the MDM data. Rather, the same models used for the SDM evaluations are employed to perform diarization on the MDM recordings. Further, keeping in line with the prior works, we have also omitted the TNO recordings in the development and evaluation set in these results. The comparative analysis is shown in Table 3. The addition of VBx based re-segmentation to the proposed approach provides the best performance on the AMI development set compared to the prior works, with a final DER of 2.18%. Further, the result on the evaluation set (DER of 3.27%) is slightly inferior to the best reported result of 3.01%. This analysis highlights that, even without fine-tuning or adapting the self-supervised model parameters to the MDM condition, the techniques reported in this paper can match the best state-of-art results for the beamformed audio recordings.

6.4. DIHARD dataset

The results on the DIHARD dataset are reported in Table 4. The PIC approach improves slightly over the AHC approach used in the baseline system [38]. The self-supervised learning approach with cosine similarity [15] degraded the performance over the baseline system. This was analyzed to be partly due to the reduced duration of the recordings, large number of speakers within the given recording and the lack of robustness in the simple cosine similarity scoring.

The proposed approach of SelfSup-PLDA improves over both the AHC and PIC systems respectively. Without the VBx re-segmentation, the best results are achieved for the SelfSup-PLDA-AHC model. However, the VBx re-segmentation did not improve over the re-segmentation applied on the baseline model.

As seen here, the improvements in the DIHARD dataset are less significant compared to those observed in AMI dataset. The primary reason for this reduction in improvement is the shorter duration files (0.5-10min duration) in DIHARD dataset compared to the 20-60min duration of the AMI recordings. The self-supervised metric learning approaches proposed in this work rely on recording level labels to improve the adjacency matrix used in the graph based clustering. With a reduced number of embeddings, the training of the SelfSup-PLDA model is compromised. Secondly, the DIHARD datasets have diverse domains with some domains having a large number of speakers (more than 7 speakers) in the given recording.

The large number of speakers also decreases the quality of the pseudo cluster labels used in the self-supervised learning. In order to analyze the impact of the increased number of speakers, we split the results reported in Table 4 into two conditions - recordings having less than or equal to 7 speakers and those with more than 7 speakers. This analysis is reported in Table 5. As seen here, the model of self-supervised metric learning provides consistent performance

Table 4. DER (%) when number of speakers (N^*) are unknown for the DIHARD dataset.

System	Unknown N^*	
	Dev.	Eval.
x-vec + PLDA + AHC [38]	19.7	19.5
- + VBx [36]	17.0	16.6
x-vec + PLDA + PIC	19.7	18.9
- + VBx [36]	16.8	16.3
SSC-Cosine-PIC [15]	23.9	21.1
SelfSup-PLDA-AHC	18.9	18.2
SelfSup-PLDA-PIC	19.2	18.2
- + VBx [36]	17.5	17.2

Table 5. Average DER (%) on the DIHARD dataset for recordings with ≤ 7 speakers and > 7 speakers

System	≤ 7 speakers		> 7 speakers	
	Dev.	Eval.	Dev.	Eval.
x-vec + PLDA + AHC	18.0	19.3	36.6	27.1
x-vec + PLDA + PIC	17.7	17.8	36.5	24.0
SelfSup-PLDA-PIC	17.0	17.2	39.5	28.1

improvements for the recordings having less than 8 speakers. On the other hand, for recordings with more than 7 speakers, the self-supervised metric learning results in a performance degradation. As hypothesized earlier, the degradation is attributed to the errors in the pseudo-labels used in the self-supervised learning. In future, a confidence measure will also be explored in the self-supervised learning framework to avoid over-fitting to the noisy cluster labels.

7. SUMMARY

We have proposed an approach to perform metric learning and clustering jointly for the task of diarization. The metric learning is performed in a self-supervised manner by updating the neural PLDA model using cluster identities provided by graph based path integral clustering. Using an iterative procedure of metric learning and clustering, we show that the proposed algorithm provides improved similarity scores and precise speaker clusters. With challenging diarization datasets, we have illustrated the performance improvements obtained using the proposed approach. In particular, the self supervised metric learning algorithm provides the best results reported thus far for the AMI single distant microphone conditions. With the more challenging DIHARD dataset evaluations, the proposed approach did not show improvements when the number of speakers in the given recording was greater than 7. However, for the recordings with less than 8 speakers, the model showed consistent performance improvements over the baseline systems compared.

8. ACKNOWLEDGEMENTS

The authors would like to thank Abhishek Anand and Michael Free of British Telecom Research for their inputs in the model development. The authors would also like to acknowledge the efforts of Rajat Varma in DIHARD experiments as well as in the manuscript preparation.

9. REFERENCES

- [1] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, “First DIHARD challenge evaluation plan,” 2018, *tech. Rep.*, 2018.
- [2] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, and Sriram Ganapathy, “The second DIHARD diarization challenge: Dataset, task, and baselines,” in *Proc. of INTERSPEECH*, 2019, pp. 978–982.
- [3] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, “The Third DIHARD Diarization Challenge,” *arXiv preprint arXiv:2012.01477*, 2020.
- [4] Gregory Sell and Daniel Garcia-Romero, “Speaker diarization with plda i-vector scoring and unsupervised calibration,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.
- [5] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [6] Xavier Anguera, Chuck Wooters, and Jose M Pardo, “Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system,” in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 346–358.
- [7] William HE Day and Herbert Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.
- [8] Daniel Garcia-Romero and Carol Y Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Proc. of INTERSPEECH*, 2011, pp. 249–252.
- [9] Weizhong Zhu and Jason Pelecanos, “Online speaker diarization using adapted i-vector transforms,” in *IEEE ICASSP*, 2016, pp. 5045–5049.
- [10] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S Huang, “A spectral clustering approach to speaker diarization,” in *Proc. of INTERSPEECH*, 2006, pp. 2178–2181.
- [11] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, “End-to-End Neural Speaker Diarization with Self-attention,” in *ASRU*, 2019, pp. 296–303.
- [12] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, “End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors,” in *INTERSPEECH*, 2020.
- [13] Prachi Singh, Rajat Varma, Srikanth Raj, Venkat Krishnamohan, and Sriram Ganapathy, “LEAP diarization system for the third DIHARD challenge,” in *Proc. of INTERSPEECH*, 2021.
- [14] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song, “Using self-supervised learning can improve model robustness and uncertainty,” in *Advances in Neural Information Processing Systems*, 2019, pp. 15663–15674.
- [15] Prachi Singh and Sriram Ganapathy, “Self-supervised representation learning with path integral clustering for speaker diarization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1639–1649, 2021.
- [16] Prachi Singh and Sriram Ganapathy, “Deep Self-Supervised Hierarchical Clustering for Speaker Diarization,” in *Proc. of INTERSPEECH*, 2020, pp. 294–298.
- [17] Shreyas Ramoji, Prashant Krishnan, and Sriram Ganapathy, “NPLDA: A Deep Neural PLDA Model for Speaker Verification,” in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 202–209.
- [18] Jan Silovsky and Jan Prazak, “Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4193–4196.
- [19] Sergey Ioffe, “Probabilistic linear discriminant analysis,” in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.
- [20] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and James Glass, “Exploiting intra-conversation variability for speaker diarization,” in *Proc. of INTERSPEECH*, 2011, pp. 945–948.
- [21] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno, “Speaker diarization with LSTM,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5239–5243.
- [22] Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras, “LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization,” in *Proc. INTERSPEECH*, 2019, pp. 366–370.
- [23] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang, “Fully supervised speaker diarization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6301–6305.
- [24] Mireia Diez, Lukáš Burget, and Pavel Matejka, “Speaker diarization based on bayesian HMM with eigenvoice priors,” in *Proceedings of Odyssey*, 2018, pp. 147–154.
- [25] Prachi Singh, Harsha Vardhan, Sriram Ganapathy, and Ahilan Kanagasundaram, “LEAP diarization system for the second DIHARD challenge,” in *Proc. of INTERSPEECH*, 2019, pp. 983–987.
- [26] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget, “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks,” *arXiv preprint arXiv:2012.14952*, 2020.
- [27] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong, “Towards k-means-friendly spaces: Simultaneous deep learning and clustering,” in *international conference on machine learning*, 2017, pp. 3861–3870.
- [28] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger, “Spectralnet: Spectral clustering using deep neural networks,” *arXiv preprint arXiv:1801.01587*, 2018.
- [29] Jianwei Yang, Devi Parikh, and Dhruv Batra, “Joint unsupervised learning of deep representations and image clusters,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5147–5156.
- [30] Wei Zhang, Deli Zhao, and Xiaogang Wang, “Agglomerative clustering via maximum incremental path integral,” *Pattern Recognition*, vol. 46, no. 11, pp. 3056–3065, 2013.

- [31] 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 Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*. Noldus Information Technology, 2005, pp. 137–140.
- [32] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.
- [33] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, “BUT system description to voxceleb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.
- [34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: A large-scale speaker identification dataset,” *Proc. of INTERSPEECH*, pp. 2616–2620, 2017.
- [35] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” in *Proc. of INTERSPEECH*, 2018, pp. 1086–1090.
- [36] Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget, Pavel Matějka, Kateřina Žmolíková, Ladislav Mošner, Anna Silnova, Oldřich Plchot, Ondřej 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)*, 2020, pp. 6529–6533.
- [37] Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na, “ECAPA-TDNN embeddings for speaker diarization,” *arXiv preprint arXiv:2104.01466*, 2021.
- [38] Neville Ryant, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, “Third DIHARD challenge evaluation plan,” *arXiv preprint arXiv:2006.05815*, 2020.