LEAP Submission for the Third DIHARD Diarization Challenge

Prachi Singh, Rajat Varma, Venkat Krishnamohan, Srikanth Raj Chetupalli, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Electrical Engineering, Indian Institute of Science, Bangalore, India.

{prachisingh, rajatvarma, venkatk, sraj, sriramg}@iisc.ac.in

Abstract

The LEAP submission for DIHARD-III challenge is described in this paper. The proposed system is composed of a speech bandwidth classifier, and diarization systems fine-tuned for narrowband and wideband speech separately. We use an end-to-end speaker diarization system for the narrowband conversational telephone speech recordings. For the wideband multi-speaker recordings, we use a neural embedding based clustering approach, similar to the baseline system. The embeddings are extracted from a time-delay neural network (called x-vectors) followed by the graph based path integral clustering (PIC) approach. The LEAP system showed 24% and 18% relative improvements for Track-1 and Track-2 respectively over the baseline system provided by the organizers. This paper describes the challenge submission, the post-evaluation analysis and improvements observed on the DIHARD-III dataset.

Index Terms: speaker diarization, end-to-end system, x-vectors, path integral clustering

1. Introduction

The task of identifying "who spoke when" in a speech recording, referred to as "Speaker diarization", is important for speech analytics applications such as rich transcription, content retrieval and indexing, etc. Speaker diarization is challenging due to the source related factors, such as the number of speakers in the recording, nature of the conversation, turn taking behavior, and the channel related factors such as the conversation medium, presence of noise and reverberation etc. The DIHARD series of challenges [1, 2, 3] are designed to benchmark the performance of speaker diarization systems on a common set of speech recordings collected from diverse domains. In the current paper, we describe the LEAP lab submission to the DIHARD-III challenge [3].

The popular approach, including the baseline system for the DIHARD-III challenge [3, 4], comprises of (i) speech activity detection and segmentation of audio, (ii) extraction of speaker embeddings, (iii) pair-wise similarity scoring, (iv) clustering, and (v) resegmentation steps to generate the diarization output for a given recording. An example state-of-the-art system [3] uses embeddings extracted from a neural network (x-vectors) [5], probabilistic linear discriminant analysis (PLDA) model [6] for scoring [7], agglomerative hierarchical clustering (AHC) [8], and a variational-Bayes hidden Markov model (VB-HMM) for resegmentation [9]. However, a single speaker is assumed in each segment, which limits the performance of the systems designed in this approach on recordings with overlap. In other approaches [10], a separate overlap detection system is employed to assign more than one speaker label to the overlapping regions during resegmentation.

The end-to-end neural diarization (EEND) systems handle the speaker overlap naturally, and have gained interest in
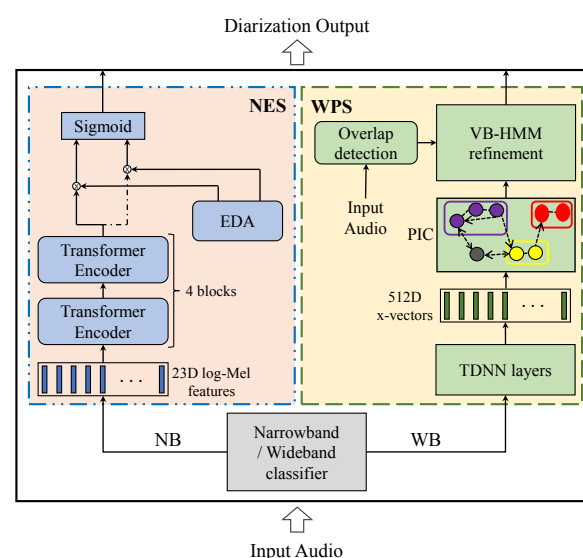

Figure 1: *Overall Block Diagram of LEAP system submitted to DIHARD III.*

the recent past [11, 12]. Short-term features with context and frame sub-sampling are input to a neural network, trained with a permutation invariant binary cross-entropy objective for each speaker activity. The presence of unknown number of speakers is handled using encoder-decoder based attractor [12] modeling.

In the LEAP submission, we explore different systems for the narrowband and wideband speech recordings, overlap processing using an off-the-shelf overlap detector. We have trained an end-to-end system for the telephone recordings in narrowband. As number of speakers varies from 1-10 in the wideband recordings in DIHARD, end-to-end training becomes computationally tedious and time consuming due to the number of permutations involved in the loss computation. Therefore, we have explored alternative graph based clustering approach known as path integral clustering (PIC) for diarization inspired by [13].

2. Detailed description of the systems

The overall block schematic of the LEAP system is given in Figure 1. The input audio is passed through a narrowband/wideband classifier. If the classifier output is narrowband (NB), the audio is passed through the narrowband end-to-end system (NES), otherwise a wideband PIC system (WPS) is used for diarization. The details of each independent system are discussed in the following sections.

2.1. Wideband-Narrowband classifier

A two-layer fully-connected feed-forward neural network with x-vectors as input features is used as the wideband-narrowband classifier. X-vectors are extracted using baseline TDNN model [3] containing 5 TDNN layers followed by statistics pooling layer which is followed by 2 fully connected layers as described in [5]. The model is trained using Voxceleb1 and 2 datasets .The input x-vectors are 512-dimensional and extracted every 5s using segments of duration 10s. During inference, a majority voting of the classifier neural network's prediction for all the segments is used to make a file-level decision on the recording bandwidth.

2.2. Wideband PIC system (WPS)

The diarization system for wideband speech recordings is inspired by the multi-stage baseline [3]. The WPS consists of neural embedding extraction followed by pair-wise similarity scoring and clustering of short speech segments and VB-HMM re-segmentation with overlap processing. In our setup, we have explored two models for embedding extraction.

2.2.1. Embedding extraction

We use x-vectors as the embeddings, which are extracted using two architecture variants, (i) extended-TDNN (ETDNN) and (ii) factorized-TDNN (FTDNN). We used 30-dimensional mel frequency cepstral coefficient (MFCC) features, extracted every 10 ms using a 25 ms window, as inputs to both the models. Both models are trained on the VoxCeleb1 [14] and VoxCeleb2 [15] datasets using speaker identification as the objective to discriminate among $7,320$ speakers. We extract the embeddings using a segment size of 1.5 s and a temporal shift of 0.25 s for both the models.

ETDNN: The 14-layer ETDNN model follows the architecture described in [5]. The model has 4 TDNN layers with 1 fully connected layer and ReLu activation after each TDNN layer of $512D$, with a total temporal context of ± 11 neighboring frames. Following the TDNN layers, the model has two feed-forward layers containing $\{512, 1500\}$ units followed by mean and standard deviation pooling layer of 3000D. The 512 dimensional output of the affine component of the 12th layer before non-linearity is taken as the x-vector embedding.

FTDNN: The architecture of the 14-layer FTDNN model is similar to that of ETDNN, with factorized TDNN layers [16] in place of the TDNN layers. The model has a reduced number of parameters due to the factorization of each layer's weight matrix M as a product of two low-rank matrices. The model enforces second factor in $M = AB$ to be semi-orthogonal ($MM^T = I$). The factorized TDNN layers are each of size 1024, and perform a sequence of convolutions that maps the input to 256 dimensions and back to the hidden layer size of 1024. The network has a longer temporal context of ± 16 frames. We extract 512-dimensional output from the 12th affine layer of the model as the x-vector embedding

2.2.2. Similarity Scoring

We consider (i) cosine score and (ii) PLDA score to compute the similarity between segment level x-vector features. To compute the cosine score, the x-vectors are first projected into a 30-dimensional space using principle component analysis (PCA), computed using the development dataset and the score is computed in the projected space. The probabilistic linear discriminant analysis (PLDA) models are trained separately for the

ETDNN and FTDNN x-vectors, and a binary hypothesis testing framework is used to obtain the PLDA similarity score between a pair of segments. The x-vectors extracted from subset of Voxceleb1 and Voxceleb2 containing 128k utterances and 7,200 speakers are used for training the PLDA model. We apply a recording level PCA on the x-vectors reducing the dimension based on 30% of total energy.

2.2.3. Path integral clustering

We explored a graph-structural-based agglomerative clustering algorithm, known as path integral clustering (PIC) [17], using the segment similarity scores to obtain the diarization output. The use of PIC for clustering is an extension of our previous work on self-supervised learning for speaker diarization [18]. The clustering process involves the creation of a directed graph $G = (V, E)$, where input features are the vertices (V) and E is the set of edges connecting the vertices. PIC merges clusters in a bottom-up manner using affinities computed using the across cluster edge/path weights defined as incremental path integral. Higher affinity indicates dense connections across clusters.

Let $\boldsymbol{X} = \{x_1, x_2, ..., x_{N_r}\}$ be the set of N_r x-vector embeddings extracted from a recording r. We first compute a sparse adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{N_r \times N_r}$ of graph G based on the K-nearest neighbours of each x_i. We use sigmoid of similarity scores (cosine/PLDA) as edge weights in $\boldsymbol{W}$. For initialization, embeddings are clustered based on their nearest neighbour. The cluster affinity measure between two clusters for the PIC algorithm is computed as,

$$\mathcal{A}\left(\mathcal{C}_a, \mathcal{C}_b\right) = \left[\mathcal{S}_{\mathcal{C}_a | \mathcal{C}_a \cup \mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_a}\right] + \left[\mathcal{S}_{\mathcal{C}_b | \mathcal{C}_a \cup \mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_b}\right] \quad (1)$$

where, terms in each square bracket denote the incremental path integral of $\mathcal{C}_a$ and $\mathcal{C}_b$ respectively. $\mathcal{S}_{\mathcal{C}_a}$ is the path integral of the cluster $\mathcal{C}_a$ defined as the weighted sum of transition probabilities of all possible paths from any vertex i to any other vertex j. Here i, j vertices belong to the cluster $\mathcal{C}_a$ and all the vertices along the path also belong to cluster $\mathcal{C}_a$. $\mathcal{S}_{\mathcal{C}_a | \mathcal{C}_a \cup \mathcal{C}_b}$ is the conditional path integral defined as the path integral of all paths in $\mathcal{C}_a \cup \mathcal{C}_b$ such that the paths start and end with vertices belonging to $\mathcal{C}_a$. The transition probability matrix $\boldsymbol{P}$ is obtained by normalizing each row of $\boldsymbol{W}$ by its sum.

Similar to AHC, we perform merging of two clusters at each time step based on the maximum affinity across all pairs till we reach the required number of speakers. Further details about the algorithm and implementation can be found in [13].

2.2.4. VB resegmentation and Overlap detection (VB-overlap)

For further refinement of segment boundaries, we apply VB-HMM resegmentation [9] with posterior scaling [4] as described in the baseline [3]. For overlap detection, we use the overlap detection module available in the pyannote toolkit [19]. The neural overlap detection model PyanNet architecture is described in [10]; it consists of SincNet filter layers followed by recurrent and fully-connected layers. We use the pre-trained network, trained on the DIHARD-I dataset, to compute the frame-level overlap scores. The overlapping speaker segments identified by the detector are then used to refine the VB-HMM resegmentation output, similar to the approach described in [10].

2.3. Narrowband End-to-End system (NES)

The architecture of the model is similar to the self-attentive end-to-end neural speaker diarization (SA-EEND) [11] combined with encoder-decoder based attractors (EDA) [12]. The model

Table 1: *DER(JER) performance for wideband system configurations using ETDNN x-vector model and for narrowband using SA-EEND-EDA model indicating the improvements from the proposed approaches for Track-1.* * *indicates baseline with oracle number of speakers.*

Wideband System config.	Dev DER(JER)
PLDA+AHC (S1)	20.09 (43.86)
PLDA + PIC (S2)	**18.91 (42.06)**
Cosine+ PIC	19.78 (43.61)
S1+VB-overlap	17.70 (42.93)
S2+VB-overlap	**17.01 (41.77)**
Narrowband System config.	**Dev DER(JER)**
Baseline w Oracle*	16.03 (20.21)
SA-EEND V1	**9.84 (12.00)**
SA-EEND V2	**9.34 (11.19)**

uses 4 stacked Transformer encoder blocks; each encoder consists of 256 attention units with 4 attention heads.

We use 23-dimensional log-Mel filterbank features, extracted every 10 ms using a frame length of 25 ms, with an added context of ± 7 frames. The resulting 345 dimensional vectors are sub-sampled by a factor of 10 and used as input to the SA-EEND+EDA model.

Training: To train the narrowband end-to-end system, we simulated $100k$ two-speaker mixtures from Switchboard-2 (Phase I, II, III), Switchboard Cellular (Part 1, Part 2) and NIST SRE datasets 2004-08, using the algorithm proposed in [20]. The model is trained on the simulated mixtures for 100 epochs using utterance-level permutation-invariant training (PIT) [21] criterion. This is followed by model adaptation on two-speaker recordings from CALLHOME dataset.

Evaluation: For the narrowband speech recordings, we fix the number of attractors to be generated to 2 and obtain the frame-wise posteriors. A threshold is applied to the posteriors to detect the presence of the speakers. We assign the speaker with the maximum posterior in the frames where model does not predict any speaker. The silence frames are then removed based on the SAD. To avoid abrupt speaker change and to make the model more memory efficient, we subsample the frame-level input features with different factors. For SA-EEND V1 and V2, we employ a subsampling factor of 10 and 5 respectively.

3. Data Resources

DIHARD III dataset [3] comprises of 0.5-10 minutes of recordings drawn from 11 different domains ranging from audiobooks to web videos. It consists of variations in number of speakers (1-10), overlapping speech (0-80%) and diverse speaker turn behavior, making the task of speaker diarization more challenging. The dataset is divided into dev and eval set for training and evaluation respectively, containing about 60 hours of audio. Each set is further divided into full and core for evaluation. The full set considers entire set for evaluation, whereas the core set excludes some files from domains with more hours of audio so that all domains contribute equally to the overall performance. Training datasets are listed here.

3.1. Wideband system datasets

- VoxCeleb1 [14] and VoxCeleb2 [15]: The two datasets respectively contains over 0.1 and 1 million utterances
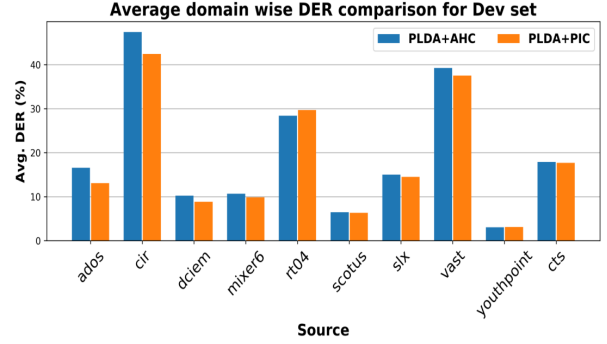

Figure 2: *Performance comparison of AHC and PIC algorithms across domains in DIHARD dev set.*

from thousands of celebrities, extracted from videos uploaded to YouTube. Both the dataset are fairly gender balanced with 55% and 61% of the speakers belonging to the male gender respectively. The video recordings included in the datasets are recorded in a large number of challenging visual and auditory environments.

3.2. Narrowband system datasets

Telephone recordings: We use Switchboard-2 (Phase I, II, III), Switchboard Cellular (Part 1, Part 2) and NIST SRE datasets 2004-2008.

CALLHOME (CH) : It is a collection of multi-lingual telephone call recordings sampled at 8 kHz, containing 500 recordings. The duration of each recording ranges from 2-5 minutes. The number of speakers in each recording varies from 2 to 7, and majority of the recordings have two speakers. The CH dataset is divided equally into two different sets, CH1 and CH2, with similar distribution of number of speakers.

4. Experiments & Results

The system performance is evaluated using the diarization error rate (DER) and Jaccard error rate (JER) measures, computed using the *dscore* toolkit [22].

We have applied different strategies for wideband and narrowband subset of DIHARD dev set. For wideband, we experiment with different scoring and clustering techniques. Table 1 shows DER(JER) performance of wideband system using ETDNN x-vectors. The PLDA+AHC is the baseline approach [3]. As discussed in Section 2.2.3, we have implemented PLDA and cosine with PIC. From Table 1, we observe that PLDA with PIC algorithm gives 1% absolute improvement compared to PLDA with AHC algorithm. The stopping criteria of PIC is based on number of speakers predicted by PLDA+AHC threshold obtained after fine tuning on dev set. The comparison of PIC and AHC is elaborated in Figure 2. It shows domain wise average DER for PLDA-AHC and PLDA-PIC using the ETDNN embedding extractor. The plot shows improvements for PIC algorithm over AHC in 7 out of 9 domains in wideband. Further refinement using VB-HMM and overlap detection gave an improvement of 3% absolute improvement over the baseline system. The miss-rate and false-alarm of the overlap-detection module on DIHARD dev set are 17.1% and 5.0 % respectively and on DIHARD eval set are 56.68% and 2.91% respectively.

For narrowband system, we use SA-EEND system with attractor. Table 1 also shows results on narrowband recordings

Table 2: *Performance of individual systems on Track-1 .*

Individual System	Set	Dev DER(JER)	Eval DER(JER)
Baseline[3]	full	19.10 (41.10)	19.68 (44.32)
	core	19.97 (45.52)	21.35 (48.89)
WPS (ETDNN)+NES	full	14.45 (**37.09**)	14.93 (37.09)
	core	16.43 (**42.45**)	18.2 (43.28)
WPS (FTDNN)+NES	full	**14.34** (37.31)	**14.88 (36.73)**
	core	**16.26** (42.75)	**18.07 (42.82)**

Table 3: *Performance of individual systems on Track-2 .*

Individual System	Set	Dev DER (JER)	Eval DER (JER)
Baseline[3]	full	21.35 (42.97)	25.76 (47.64)
	core	22.31 (47.28)	28.31 (52.44)
WPS(ETDNN)+NES	full	16.77 (**37.15**)	21.04 (39.68)
	core	18.64 (**41.93**)	24.92 (45.32)
WPS(FTDNN)+NES	full	**16.53** (38.50)	**21.09 (39.54)**
	core	**18.34** (43.62)	**24.99 (45.13)**

from CTS domain. We see a significant improvement of 40% (relative) for SA-EEND model over the baseline. As we increase the resolution by reducing subsampling factor, we only observe a marginal improvement.

Tables 2 and 3 show results of our systems and the baseline for Track-1 and Track-2 respectively. We have combined WPS with ETDNN/FTDNN for wideband domain files and NES for narrowband (CTS domain) files. The classification accuracy of wideband-narrowband classifier for eval set is 99.67%. In Track-1, we obtain 24% and 15% relative improvements in DER for full and core sets respectively using ETDNN model over the baseline. A similar performance gain is obtained using the FTDNN model also. In Track-2, we obtain around 18% and 11% relative improvements in DER for both ETDNN and FTDNN model for full and core sets over the baseline.

5. Post Evaluation Analysis

Here, we discuss details and results of our post evaluation analysis to improve performance of wideband system for Track-1.

5.1. VBx resegmentation

We perform hidden Markov model (HMM) based clustering on x-vectors using variational Bayes (VB) approach with VBx clustering [23]. It is a simplified version of VB-HMM model in which the state distributions are Gaussian and the observations are the x-vectors of the recording. Each state represents a speaker and transition probabilities across states represents speaker turn. The loop probability $Ploop = 0.8$ is the probabilty to remain in the same speaker state. The parameters of each Gaussian are obtained from a pre-trained PLDA model. As per [23], we train two PLDA models using VoxCeleb dataset and DIHARD dev set using x-vectors from 3 s segments. The pre-processing steps include whitening using 1.5 s DIHARD dev x-vectors with 0.25 s shift followed by length normalization. The x-vectors are then projected from 512D to 220D using Linear Discriminant Analysis (LDA). We perform interpolation of both PLDAs regulated by $\alpha = 0.5$. The AHC is used to initialize the HMM resegmentation. We have experimented with ETDNN model based on MFCC and mel filterbank features. The best results are obtained using mel filterbank features.

Table 4: *Post DIHARD analysis: The DER (JER) performance of individual systems on Track-1 .*

Individual System	Set	Dev DER(JER)	Eval DER(JER)
LEAP Eval Sub.	full	14.34 (37.31)	14.88 (36.73)
	core	16.26 (42.75)	18.07 (42.82)
WB VBx-ovp + NES	full	**12.87 (33.11)**	**14.53 (35.62)**
	core	**14.26 (37.64)**	**17.45 (40.90)**
WPS (Mel)+VBx + NES	full	13.22 (34.05)	15.04 (38.17)
	core	14.71 (38.68)	18.20 (44.08)
WB EEND + NES	full	19.19 (47.17)	18.66 (44.78)
	core	23.43 (55.08)	23.55 (52.59)

5.2. End-to-End Wideband system

We have explored the use of end-to-end system for other domains in DIHARD III excluding CTS. The architecture of the model is same as the one discussed in Section 2.3.

Training: We simulated $400k$ mixtures for 1-4 speakers from VoxCeleb1 and 2 and Librispeech [24] datasets. We trained the model on 2 speaker mixtures for 12 epochs followed by fine tuning on 1-,2-,3-,4- speaker mixtures for 4 epochs. Finally, the model is adapted on DIHARD III dev set.

Evaluation: We followed the same procedure for evaluation as described in section 2.3, except that instead of fixing the number of attractors to be generated to 2, we let the model infer the required number based on the attractor existence probabilities.

5.3. Results

Table 4 shows our contributions post DIHARD evaluation. VBx system combined with overlap detection gave best results for wideband on dev set. We combined wideband VBx with Narrowband end-to-end system (NES) to get best performance. We get 26% and 18% relative improvements over baseline in full and core eval sets respectively. The third row in the Table 4 shows results of WPS system with ETDNN trained with mel filterbank and interpolated PLDA where we obtain the number of speakers from VBx system. Owing to large diversity in the domains and the training complexity, the wideband end-to-end system (last row of Table 4) generated suboptimal results.

6. Conclusions

We have presented the detailed description of the LEAP submission to DIHARD-III diarization challenge. The key novel components of the system include, (i) a band classifier based approach to separately process conversational telephone data and the other wide-band domains, (ii) graph based clustering approach to diarization using path integral clustering and (iii) incorporation of overlap detector with the final re-segmentation step to improve the diarization performance. The paper also includes our analysis further to the conclusion of the challenge where we observe additional performance improvements with the use of re-segmentation based on the x-vector embeddings instead of frame-level features.

7. Acknowledgement

8. References

[1] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," *2018, tech. Rep.*, 2018.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, and S. Ganapathy, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Proc. of Interspeech*, 2019, pp. 978–982.

[3] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," *Proc of Interspeech 2021*, 2021. [Online]. Available: https://arxiv.org/abs/2012.01477

[4] P. Singh, H. Vardhan, S. Ganapathy, and A. Kanagasundaram, "LEAP diarization system for the second DIHARD challenge," in *Proc. of Interspeech*, 2019, pp. 983–987.

[5] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *IEEE ICASSP*, 2019, pp. 5796–5800.

[6] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[7] J. Silovsky, J. Prazak, P. Cerva, J. Zdansky, and J. Nouza, "PLDA-based clustering for speaker diarization of broadcast streams," in *Proc. Interspeech*, 2011.

[8] J. Luque and J. Hernando, "On the use of agglomerative and spectral clustering in speaker diarization of meetings," *Odyssey*, 2012.

[9] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on bayesian HMM with eigenvoice priors," in *Proceedings of Odyssey*, 2018, pp. 147–154.

[10] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection," in *IEEE ICASSP*, Barcelona, Spain, May 2020.

[11] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Self-attention," in *ASRU*, 2019, pp. 296–303.

[12] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. of Interspeech*, 2020.

[13] P. Singh and S. 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.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. of Interspeech*, pp. 2616–2620, 2017.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. of Interspeech*, 2018, pp. 1086–1090.

[16] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of Interspeech*, 2018, pp. 3743–3747.

[17] W. Zhang, D. Zhao, and X. Wang, "Agglomerative clustering via maximum incremental path integral," *Pattern Recognition*, vol. 46, no. 11, pp. 3056–3065, 2013.

[18] P. Singh and S. Ganapathy, "Deep self-supervised hierarchical clustering for speaker diarization," *Proc. of Interspeech 2020*, pp. 294–298.

[19] 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 *IEEE ICASSP*, 2020.

[20] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-free Objectives," in *Proc. of Interspeech*, 2019, pp. 4300–4304.

[21] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks," in *IEEE/ACM Trans. on ASLP*, vol. 25, no. 10, 2017, pp. 1901–1913.

[22] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *Proc. of Interspeech*, 2021. [Online]. Available: arXivpreprintarXiv: 2006.05815

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

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.