LEAP Submission for Third DIHARD Diarization Challenge

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

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science.

prachisingh@iisc.ac.in

Abstract

The LEAP submission for DIHARD-III challenge is described in this report. The LEAP system involves the use of End-to-end speaker diarization systems for the two-speaker conversational telephone speech recordings. For the wide-band multi-speaker recordings, the proposed approach for diarization uses embeddings from time-delay neural network (called x-vectors) followed by a graph based clustering approach called the path integral clustering. The LEAP system showed 24% and 18% relative improvements for track1 and track2 respectively over the baseline system provided by the organizers. This report provides details of the model and the experimental results on the DIHARD-III.

Index Terms: speaker diarization, End-to-End system, x-vectors

1. Description of LEAP Systems

The DIHARD dataset has a mix of narrowband and wide-band speech recordings. In the development set, 24% of the recordings are narrowband, and the remaining have wideband speech. The narrowband recordings have only two speakers in each recording. We use a combination of models optimized separately for narrowband and wideband speech, in combination with a bandwidth classifier, to design the diarization system. For wideband speech, the pipeline consists of an embedding extractor based on extended time delay neural network (E-TDNN), a graph based clustering scheme called path integral clustering (PIC) and VB-HMM re-segmentation. In addition, we also use an overlap detection based on a separate overlap detection model which is combined with the VB-HMM diarization system. For narrowband speech, we explore the supervised end-to-end model architecture [1], with known number of speakers (two).

1.1. Baseline system

The baseline setups for both tracks are implemented as described in [2]. We have used the best configuration obtained from track1 and used that for track2. We have used the pre-trained baseline SAD model for track2.

1.2. 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. The 512-dimensional input x-vectors are extracted every 5 s using segments of duration 10 s. During evaluation, majority voting of the segment-wise prediction of the neural network is used to decide on the bandwidth of the recording.

1.3. Wideband PIC system

This system is inspired by multi-stage baseline system, which consists of neural embedding extraction, followed by pair-wise similarity scoring and clustering of short speech segments, and VB-HMM based resegmentation. In our setup, we have explored two different models for embedding extraction as described below.

1.3.1. Embedding extraction and Scoring

We use x-vectors as the embeddings. The embeddings are extracted using two variants (i) extended-TDNN (ETDNN), and (ii) factorized-TDNN (FTDNN).

ETDNN: The 12-layer ETDNN model follows the architecture described in [3]. ETDNN model is trained on the VoxCeleb1 and VoxCeleb2 datasets, for speaker identification task, to discriminate among 7232 speakers. The 512 dimensional output of the affine component of the 11^{th} layer are taken as the x-vector embeddings. We extract the embeddings using a segment size of 1.5 s and a temporal shift of 0.25 s.

FTDNN: The architecture of the FTDNN model is similar to that of ETDNN, with factorized TDNN layers [4] in place of the TDNN layers. The model is trained in a similar manner to ETDNN. We extract 512-dimensional output from the 12th affine layer of the model as the x-vector embeddings, using the same resolution and segment-size as x-vectors from ETDNN.

We consider (i) cosine score, and (ii) PLDA score, to compute the similarity between segments. A separate PLDA model is trained for both ETDNN and FTDNN x-vectors. The similarity score between two segments is then obtained using a binary hypothesis testing framework. To compute the cosine score, the x-vectors are projected to a 30 dimensional space using PCA, computed using the development dataset, and the score is computed in the projected space.

1.3.2. Path integral clustering

We perform clustering of PLDA/Cosine scores to get the diarization output. We have implemented a graph-structural based agglomerative clustering algorithm known as path integral clustering (PIC) [5].The clustering process involves 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. The transition probabilities, computed from the similarity score (PLDA/Cosine) matrix, are used as the edge weights. Similar to AHC, PIC also merges two clusters at each time step based on maximum affinity, but the affinity is computed using the path integral as defined in [5]

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

For further refinement of segment boundaries, we apply Variational Bayes Hidden Markov Model (VB-HMM) resegmentation as described in baseline.

We use the overlap detection module available in the pyannote.audio python toolkit [6] to identify the segments with speaker overlap. The architecture of the neural overlap detection model is described in [7]; it consists of sincNet filter layers followed by recurrent and fully-connected layers. We use the pre-trained network, trained on DIHARD-I dataset to compute

Table 1: *Track1 DER(JER) of individual and fused systems. WPS is wideband PIC system and NES is narrowband end-to-end system.*

Individual System	Set	Dev DER(JER)	Eval DER(JER)
Baseline[2][1.1]	full	19.10 (41.10)	19.68 (44.32)
	core	19.97 (45.52)	21.35 (48.89)
WPS (ETDNN)[1.3]+NES[1.4]	full	14.45 (37.09)	14.93 (37.09)
	core	16.43 (42.45)	18.2 (43.28)
WPS (FTDNN)[1.3]+NES[1.4]	full	14.34 (37.31)	14.88 (36.73)
	core	16.26 (42.75)	18.07 (42.82)

the frame-level overlap scores. The segments identified as overlap by the detector are then used to refine the segments obtained after VB-HMM re-segmentation, similar to approach described in [7].

1.4. Narrowband End-to-End system

The architecture of the model is similar to the SA-EEND [8] combined with the encoder-decoder based attractor calculation (EDA) [1]. The model uses 4 stacked Transformer encoder blocks; each block consists of 256 attention units with 4 attention heads.

23-dimensional log-Mel-filterbank features, extracted every 10 ms using a frame length of 25 ms are used as input features, similar to [1]. A context of ± 7 frames is applied, and the resulting 345 dimensional vectors are subsampled by a factor of 10 and fed to the SA-EEND+EDA model.

Training: We simulated 1,00,000 two-speaker mixtures to train the Narrowband End-to-End system from Switchboard-2 (Phase I, II, III), Switchboard Cellular (Part 1, Part 2) and NIST SRE datasets 2004-2008, using the algorithm proposed in [9]. The model was trained on the simulated mixtures for 100 epochs using utterance-level permutation-invariant training (PIT) [10] criterion. This was followed by model adaptation on CALLHOME dataset two-speaker files.

Evaluation: For the narrowband audio files we hard-set the number of attractors to be generated to 2 and obtained the frame wise posteriors.

A threshold was applied on the posteriors to detect the presence of the speakers. If for any frame the model failed to detect any speaker based on our set threshold τ, we assigned the speaker with the maximum posterior in that frame. The silence frames were then removed based on ground truth SAD for track1 and pre-trained model SAD for track2.

1.5. Systems fusion

We experimented with weighted combination of PLDA scores from ETDNN and FTDNN models for wideband PIC system (WPS) to improve the overall performance. But results improved only marginally hence these are not mentioned in the table.

2. Experiments & Results

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 3 shows DER(JER) performance of Wideband system using ETDNN x-vectors. PLDA+AHC is the baseline approach. As discussed in Section 1.3.2, we have implemented PLDA and Cosine with path integral clustering (PIC). The stopping criteria of PIC is based on number of speakers predicted by PLDA+AHC threshold obtained after fine tuning on Dev set.

Table 2: *Track2 DER(JER) of individual and fused systems. WPS is wideband PIC system and NES is narrowband end-to-end system.*

Individual System	Set	Dev DER(JER)	Eval DER(JER)
Baseline[2][1.1]	full	21.35 (42.97)	25.76 (47.64)
	core	22.31 (47.28)	28.31 (52.44)
WPS(ETDNN)[1.3]+NES[1.4]	full	16.77 (37.15)	21.04 (39.68)
	core	18.64 (41.93)	24.92 (45.32)
WPS(FTDNN)[1.3]+NES[1.4]	full	16.53 (38.50)	21.09 (39.54)
	core	18.34 (43.62)	24.99 (45.13)

Table 3: *DER(JER) performances for wideband system configurations using ETDNN x-vector model and for narrowband using SA-EEND model indicating the improvements from the proposed approaches for track1.* indicates baseline with oracle number of speakers.*

Wideband System config.	Dev DER(JER)
PLDA+AHC (S1)	20.09 (43.86)
PLDA + PIC (S2)	**19.06** (42.44)
Cosine+ PIC	19.78 (43.61)
S1+VB-overlap	17.70 (42.93)
S2+VB-overlap	17.03 (41.92)
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)

For narrowband system, we use SA-EEND system with attractor. For evaluation, we subsample the frame-level features by different factor to avoid abrupt speaker change and to make it more memory efficient. Table 3 also shows results on narrowband recordings from cts domain. SA-EEND V1 involves subsampling by 10 and whereas SA-EEND V2 involves subsampling by 5.

3. References

[1] 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 *Interspeech*, 2020.

[2] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," 2020.

[3] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[4] 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. Interspeech 2018*, 2018, pp. 3743–3747. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1417

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

[6] 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 *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[7] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end over-

lapped speech detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7114–7118.

[8] 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.

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

[10] 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.