CRIM's System Description for the Third Edition of DIHARD Challenge 2020

Jahangir Alam, Vishwa Gupta

Computer Research Institute of Montreal, Canada

jahangir.alam,vishwa.gupta@crim.ca

Abstract

In this work, we describe the systems developed for tackling speaker diarization problem for the 3rd edition of DIHARD 2020 challenge. We submit systems only for track 1 task of this challenge. For this task, our developed systems employ the well-known x-vector/PLDA/AHC framework followed by the Bayesian Hidden Markov Model (HMM) with eigenvoice priors applied at the x-vector embeddings domain. For the extraction of x-vector embeddings we adopt three deep learning architectures, namely, TDNN with statistics pooling, TDNN-LSTM and TDNN with multi-level (i.e., from more than one layer) statistics pooling. PLDA model is trained on the out-of-domain data and then adapted to the DIHARD 2020 development data. 30-dimensional Mel-filterbank features are used as frontend. As a pre-processing step, we dereverberate the development and evaluation data of DIHARD 2020 using weighted prediction error (WPE) dereverberation algorithm.

Index Terms: Speaker diarization, WPE, agglomerative hierarchical clustering, variational Bayes HMM, speech activity detection.

1. Introduction

Automatic determination of speaker turns in a conversational audio recordings is denoted as Speaker Diarization (SD). It is often referred as ”who spoke when”. In general, a diarization system partitions multi-speaker speech recordings into short segments and clusters them according to speaker identities [1, 2, 3, 4].

Speaker diarization has a wide range of applications in three primary domains namely, broadcast news audio, recorded meeting and telephone conversation. Applications of diarization include audio and speaker indexing, content structuring, audio information retrieval, speaker verification in the presence of multiple-speaker recordings, speech-to-text transcription, and video processing [1, 2, 3, 4, 5].

Speaker diarization has received much attention in the recent years. On some specific task or dataset, such as CallHome [1, 3, 5, 6], researchers were able to attain state-of-the-art speaker diarization performances but the performances do not generalize to more challenging and realistic data including web videos, speech in the wild, child language recordings, Video Annotation for Speech Technology (VAST) etc [7, 8, 9].

In order to draw researcher's attention on these more challenging scenarios the first DIHARD challenge [7] was launched in 2018 focusing on 'hard' conditions. Following the success of first DIHARD challenge the second and third editions of the challenge have been launched in 2019 and 2020, respectively [8, 9]. The main purpose of DIHARD series of challenges was to provide a common framework with standardized data, tasks, and metrics for facilitating comparison of current and future research works as well as to promote research works on building robust diarization systems [2, 3, 4, 7, 8, 10, 9, 11].

In this work, we provide a description of the systems developed for third edition of the DIHARD challenge 2020 [9, 11].

2. Data Resource

We employ following source of corpora for all our experiments for the DIHARD challenge 2020:

- Voxceleb 1 & 2 [12].
- NIST SRE Mixer data (excluding the MIXER6 Speech (LDC2013S03), NIST SRE10 & SRE12 evaluation data) [13].
- NIST SRE 2016 [14] and 2018 [15] Evaluation data.
- DIHARD II Development data [8].

Following the NIST SRE 2016 kaldi recipe [13] data Augmentation is performed to the above mentioned data with additive noise and room impulse response (RIRs) taken from MUSAN [16] and SLR28 [17] datasets. MUSAN and SLR28 datasets are available in the following link https://www.openslr.org/.

Augmented data is used for training the speaker embeddings extractor and the original training data excluding DIHARD II development data is used for training out-of-domain PLDA parameters. DIHARD III development data is used for tuning the system parameters and results are reported on the evaluation set (from leaderboard). In-domain data (DIHARD II development and DIHARD III development) is used for training in-domain PLDA parameters.

3. Description of our Developed Systems

Our developed systems consists of following steps:

3.1. Dereverberation using Weighted Prediction Error Algorithm

As pre-processing step for all systems we apply weighted prediction error (WPE) - based single channel dereverberation for enhancing the 3rd DIHARD challenge 2020 development and evaluation data. The WPE performs dereverberation using a linear time invariant filter and produces M-channel outputs from M-channel inputs [18]. Here, the number of channel is $M = 1$.

3.2. Speech Activity Detection (SAD)

Since we participate only on track 1 of the DIHARD challenge 2020 we use reference SAD segmentations provided by the organizer for the challenge development and evaluation datasets. Speech segmentations generated by an energy-based SAD are used for other corpora.

3.3. Features Extraction

We extract 30-dimensional Mel-filterbank (MelFB) features that cover the frequency regions 20-7600 Hz. Mean normalization is applied with a sliding windowing of 3 sec. For features extraction an analysis frame length of 25 msec is used with a frame shift of 10 msec.

3.4. Extraction of Speaker Embeddings

One of the main components for speaker recognition and speaker diarization is the extraction of speaker embeddings. This is normally done by training an embedding extractor in unsupervised fashion (such as i-vector extractor [19]) or in supervised way (such as neural network-based embeddings extractor [20, 21, 22]) on the top of frame level acoustic features.

For speaker diarization, extraction of speaker embeddings using a deep learning architectures, such as TDNN, in supervised fashion is proved to be very effective. In this work, we employ following three supervised embeddings extractors:

- TDNN: This embeddings extractor is based on time delay neural network (TDNN) with a single statistics pooling layer and is similar to the one used in DIHARD III baseline [9] and in [20]. Our baseline system use this for the extraction of speaker embeddings.

- Extended TDNN: This supervised extractor is based on a extend version of TDNN architecture with multi-level statistic pooling (i.e., statistics are pooled from more than one layer) [10]. For our contrastive system we use this supervised extractor to extract speaker embeddings.

- Extended TDNN-LSTM: This extractor is similar to the extended TDNN speaker extractor but in this case a LSTM layer is employed between 1st and 2nd TDNN layers. Multi-level statistics pooling is used including one from the LSTM layer. For our primary system speaker embeddings are extracted using this supervised extended TDNN-LSTM extractor.

Once training is done speaker embeddings are extracted from 3s segments from the out-of-domain and in-domain PLDA training data as mentioned in section 2.

Speaker embeddings from the DIHARD III development and evaluation data are extracted from 1.5 sec segments with a 0.25 sec shift. Extracted embeddings are centered and whitened using statistics estimated from the DIHARD III in-domain data, followed by length normalization [10].

3.5. Similarity Scoring

Probabilistic linear discriminant analysis (PLDA) is employed to perform similarity scoring between any two speaker embeddings in the same audio which are then used in the clustering step by the clustering algorithm.

3.6. Clustering

Speaker embeddings are then clustered using agglomerative hierarchical clustering (AHC) and a similarity matrix generated by scoring with a PLDA model.

This initial clustering is then refined either using frame-level Variational Bayes Hidden Markov Model (VB-HMM) resegmentation [23] or by performing another clustering based on Bayes hidden Markov model and variational Bayes inference [10].

4. Primary and Contrastive Systems

4.1. Primary System

In this system diarization is performed by segmenting each observed recording into short overlapping segments after performing dereverberation with WPE algorithm, extracting speaker embeddings using extended TDNN-LSTM extractor, performing similarity scoring with adapted probabilistic linear discriminant analysis (PLDA) model, carrying out Bayesian HMM clustering with the LDA followed by a initial clustering with the agglomerative hierarchical clustering (AHC) [10] algorithm. The adapted PLDA model is obtained by the interpolation of out-of-domain PLDA and in-domain PLDA models.

4.2. Contrastive System

This system conduct diarization by dividing each recording into short overlapping segments, extracting speaker embeddings, performing similarity scoring with out-of-domain PLDA model, clustering using agglomerative hierarchical clustering (AHC) and then refining the AHC output using Variational Bayes Hidden Markov Model (VBHMM) with posterior scaling (denoted as VB resegmentation) [9]. This system is similar to DIHARD 3 baseline but with following differences - (i) Extended TDNN architecture with multi-level pooling is used instead of TDNN, (ii) development and evaluation sets are dereverberated employing WPE dereverberation technique.

4.3. Baseline System

Baseline system is similar to DIHARD III baseline as described in [9].

5. Results on the Evaluation Set

In this section, we evaluate the the performances of our developed systems. Official evaluation metrics, Diarization Error Rate (DER) and Jaccard Error Rate (JER), are used for reporting results on the evaluation set of task 1 for the 3rd DIHARD challenge 2020.

Table 1 provides a comparison of performances on the evaluation set between our developed (primary & contrastive) and the baseline systems on the core condition of task 1. We can see that both primary and contrastive systems outperform the baseline system in both official evaluation metrics. Primary system provides the best performance in this condition.

Speaker diarization results reported in terms of DER and JER metrics in table 2 demonstrate performance comparison of our developed system with that of the baseline system. It is observed from this table that with our developed systems we achieved better performance than the baseline with primary system yielding the best results.

Use of WPE-based dereverberation, extended TDNN-LSTM - based embeddings extractor, adapted PLDA by interpolation of out-of-domain PLDA and in-domain PLDA as well as use of additioanal clustering of speaker embeddings based on Bayes hidden Markov model and variational Bayes inference led to better performance with our primary system.

6. Conclusions

In this work, we presented the systems developed for tackling speaker diarization problem of task 1 for the 3rd edition of DIHARD challenge 2020. We adopted the widely used x-vector/PLDA/AHC framework followed re-clustering by the

Table 1: *Speaker Diarization performances on the evaluation set of 3rd DIHARD challenge 2020, Task 1 Core condition.*

System	DER (%)	JER (%)
Primary	16.440	37.400
Contrastive	19.580	46.990
Baseline [9]	20.650	47.740

Table 2: *Speaker Diarization performances on the evaluation set of 3rd DIHARD challenge 2020, Task 1 Full condition.*

System	DER (%)	JER (%)
Primary	15.500	33.350
Contrastive	18.330	41.390
Baseline [9]	19.250	42.450

Bayesian Hidden Markov Model (HMM) with eigenvoice priors applied at the x-vector embeddings domain. We employed three supervised speaker embeddings extractor based on TDNN with statistics pooling, extended TDNN-LSTM and extended TDNN with multi-level (i.e., from more than one layer) statistics pooling. For our primary system, the PLDA model was trained on the out-of-domain data and then adapted to the DIHARD 2020 in-domain data data. On the core and full conditions of task 1 our primary submission yielded better performance than the DIHARD III baseline system providing a ranking of 7th and 9th on the leaderboard, respectively.

7. Acknowledgements

The authors would like to thank Ministry of Economy and Innovation (MEI) of the Government of Quebec for the continued support. Authors also wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2019-05381. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the NSERC.

8. References

[1] P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of telephone conversations using factor analysis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[2] V. Gupta and J. Alam, “Crim's speaker diarization system for the dihard diarization challenge,” 2018. [Online]. Available: https://dihardchallenge.github.io/dihard1/system_descriptions/crim_systems.pdf

[3] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, “Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge,” in *Proc. Interspeech 2018*, 2018, pp. 2808–2812. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1893

[4] Q. Lin, W. Cai, L. Yang, J. Wang, J. Zhang, and M. Li, “Dihard ii is still hard: Experimental results and discussions from the dku-lenovo team,” 2020.

[5] S. Hernawan, “Speaker diarization: Its developments, applications, and challenges,” 2012.

[6] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with lstm,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5239–5243.

[7] N. Ryanta, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta, M. Liberman, Y. Lu, M. Maciejewski, F. Metze, J. Profant, L. Sun, Y. Tsao, and Z. Yu, “Enhancement and analysis of conversational speech: Jsalt 2017,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5154–5158.

[8] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, “The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines,” in *Proc. Interspeech*, 2019, pp. 978–982. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1268

[9] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, “The third dihard diarization challenge,” 2021.

[10] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, “But system description for dihard speech diarization challenge 2019,” 2019.

[11] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, “Third dihard challenge evaluation plan,” 2020. [Online]. Available: https://dihardchallenge.github.io/dihard3/docs/third_dihard_eval_plan_v1.1.pdf

[12] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018. [Online]. Available: http://www.danielpovey.com/files/2018_icassp_xvectors.pdf

[14] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, “The 2016 NIST Speaker Recognition Evaluation,” in *Proc. of Interspeech*, 2017, pp. 1353–1357. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-458

[15] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, “The 2018 NIST Speaker Recognition Evaluation,” in *Proc. Interspeech 2019*, 2019, pp. 1483–1487. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1351

[16] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[17] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[18] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge,” in *REVERB Challenge Workshop*, Florence, Italy, May 2014.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” pp. 5329–5333, 2018.

[21] G. Bhattacharya, M. J. Alam, and P. Kenny, “Deep speaker recognition: Modular or monolithic?” in *INTERSPEECH*, 2019.

[22] J. Monteiro, M. J. Alam, and T. H. Falk, “Combining speaker recognition and metric learning for speaker-dependent representation learning,” in *INTERSPEECH*, 2019.

[23] M. Diez, L. Burget, and P. Matejka, “Speaker diarization based on bayesian hmm with eigenvoice priors,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 147–154. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-21