ABSP System for The Third DIHARD Challenge

Kishore A. Kumar, Shefali Waldekar, Goutam Saha, Md Sahidullah

ABSP System for The Third DIHARD Challenge

A Kishore Kumar[1], Shefali Waldekar[1], Goutam Saha[1], Md Sahidullah[2]
[1]Dept. of Electronics and ECE, Indian Institute of Technology Kharagpur, Kharagpur, India
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
kishore@iitkgp.ac.in

***Abstract*—This report describes the speaker diarization system developed by the ABSP Laboratory team for the third DIHARD speech diarization challenge. Our primary contribution is to develop acoustic domain identification (ADI) system for speaker diarization. We investigate speaker embeddings based ADI system. We apply a domain-dependent threshold for agglomerative hierarchical clustering. Besides, we optimize the parameters for PCA-based dimensionality reduction in a domain-dependent way. Our method of integrating domain-based processing schemes in the baseline system of the challenge achieved a relative improvement of 9.63% and 10.64% in DER for core and full conditions, respectively, for Track 1 of the DIHARD III evaluation set.**

I. Notable highlights

We participated in the **Track 1** of the third DIHARD challenge [1]. Our main focus was to apply domain-dependent processing which was found promising in preliminary studies with the second DIHARD dataset [2], [3]. We propose a simple modification of the baseline system of the challenge which results considerable reduction of the error rates compared to the baseline performance. The notable features of our submission to the challenge are as follows:

- We propose a simple but efficient method for acoustic domain identification (ADI) using speaker embeddings of the full-recording. We observed that i-vector-based speaker embeddings are considerably better than x-vector-based speaker embeddings for ADI task.
- We have found that that full domain-dependent processing with domain-dependent clustering and domain-dependent probabilistic linear discriminant analysis (PLDA) adaptation does not improve the diarization performance. However, this helps when the clustering is done in a domain-dependent way, but PLDA adaptation during scoring is made with audio-data from all the eleven domains.
- We also found that experimental optimization of the parameters for principal component analysis (PCA) in a domain-specific way further improves the diarization performance.
- The proposed system does not introduce much computational overhead over the baseline system for the diarization. Though this approach requires more time for empirical optimization of the parameters on the development set, the additional computational cost is negligible for the evaluation data.
- The proposed system does not have any fusion or system combination from evaluation perspective. Considering the fact that most of the top systems in this challenge are combination of two or more sub-systems, our algorithm is remarkably faster than other competitive systems.

II. Data resources

The ABSP system has two major components: ADI and speaker diarization. The ADI system uses i-vector speaker embeddings extracted with models trained on VoxCeleb 1[1] and 2[2] corpora.

On the other hand, the diarization system uses an embedding extractor trained on a combination of VoxCeleb 1 and VoxCeleb 2 augmented with additive noise and reverberation from MUSAN[3] and RIR[4] database, respectively.

III. Detailed description of algorithm

Our diarization system is primarily based on the baseline system created by the organizers [4]. We have used the toolkit[5] with the same frame-level acoustic features, embedding extractor, scoring method, etc. The ADI system is based on the speaker embeddings as sentence-level feature and nearest neighbor classifier [5]. In order to extract utterance-level embeddings for ADI task, we used pre-trained i-vector [6] model trained on VoxCeleb audio-data[6].

We can summarize the steps for the speaker diarization as follows:

1) *ADI task*: First, the ADI system was developed from the development set. We have used nearest neighbour classifier with cosine similarity. The full development set with all 254 files was used for training the final ADI system. More details about this system are reported in [5].
2) *Domain-dependent threshold selection*: The baseline system for the challenge finds the optimum threshold by computing diarization error rates (DERs) on full development set at different thresholds ranging from -1.5 to 0.0[7]. We follow the same process but for different acoustic-domains, independently. At the end of this step, the optimum thresholds for each domain are stored in a lookup table.

[1]https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html
[2]https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html
[3]https://www.openslr.org/17/
[4]https://www.openslr.org/28/
[5]https://github.com/dihardchallenge/dihard3_baseline
[6]https://kaldi-asr.org/models/m7
[7]https://github.com/dihardchallenge/dihard3_baseline/blob/master/recipes/track1/local/diarize.sh

3) *Domain-dependent dimensionality reduction*: The PLDA scoring involves dimensionality reduction of the embedding using PCA. The baseline system preserves 30% of the total energy during dimensionality reduction. Instead of applying fixed value of 0.3 for all the recording, we optimized this for each domain separately by varying it between 0.1 to 0.9 with a step of 0.1. Similar to the previous step, the optimum parameters for each domain are preserved in another lookup table.
4) *Diarization on the evaluation set*: Finally, during the diarization on the evaluation set, we first computed the i-vector of the full-recording to the be processed. Then, we predicted the corresponding acoustic domain using the ADI system. This is followed by the selection of clustering threshold and dimensionality reduction parameters corresponding to the predicted labels.

IV. Results on the development set

The speaker diarization results on development set are shown in Table I. We have also shown the results for the evaluaiton set in Table II. Both these results confirm considerable improvement over baseline system.

TABLE I
RESULTS SHOWING THE SPEAKER DIARIZATION PERFORMANCE USING BASELINE AND PROPOSED METHODS ON **DEVELOPMENT SET**.

Method	Full		Core	
	DER	JER	DER	JER
Baseline	19.59	43.01	20.17	47.28
Proposed	17.40	38.08	17.95	42.12

TABLE II
SAME AS TABLE I BUT FOR **EVALUATION SET**.

Method	Full		Core	
	DER	JER	DER	JER
Baseline	19.19	43.28	20.39	48.61
Proposed	17.20	37.30	18.66	42.23

V. Hardware requirements

The codes were run on Dell PowerEdge R730 server[8]. It has Intel Xeon (R) CPU E5-2695 v4 @ 2.10GHz processor and 128 GB memory with 72 CPUs, 2 sockets, 18 cores/socket, and 2 threads/core. We used Ubuntu 16.04 LTS 64-bit operating system.

We used 32 cores for the experiments. The system execution times to process the entire development set is $\approx$ 8 hours as this involves experimental optimization. However, for the evaluation set, the time required is almost identical to the baseline system ($\approx$ 1 hour). The additional time is just about 5 seconds per audio recording which includes i-vector extraction, domain-prediction and looktable search.

[8]https://www.dell.com/en-us/work/shop/povw/poweredge-r730

VI. Conclusion

Our study is a step towards advancing the baseline diarization system with domain-dependent processing. Our system showed substantially reduced error rates as we optimized the clustering threshold and the dimensionality reduction parameters for each domain separately. The future work involves investigating advanced embedding extractors and exploring more domain-dependent processing, e.g., domain-dependent acoustic front-end, embedding extractor, re-segmentation, etc.

VII. Acknowledgements

Experiments presented in this paper were partially carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

[1] N. Ryant *et al.*, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.
[2] M. Sahidullah *et al.*, "The Speed submission to DIHARD II: Contributions & lessons learned," *arXiv preprint arXiv:1911.02388*, 2019.
[3] T. Fennir, F. Habib, and C. Macaire, "Acoustic scene classification for speaker diarization," Université de Lorraine, Tech. Rep., 2020.
[4] N. Ryant *et al.*, "The third DIHARD diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.
[5] A. K. Kumar, S. Waldekar, G. Saha, and M. Sahidullah, "Domain-dependent speaker diarization for the third DIHARD challenge," *arXiv preprint arXiv:2101.09884*, 2021.
[6] N. Dehak *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.