Domain-Dependent Speaker Diarization 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

Abstract

This report presents the system developed by the ABSP Laboratory team for the third DIHARD speech diarization challenge. Our main contribution in this work is to develop a simple and efficient solution for acoustic domain dependent speech diarization. We explore speaker embeddings for *acoustic domain identification* (ADI) task. Our study reveals that i-vector based method achieves considerably better performance than x-vector based approach in the third DIHARD challenge dataset. Next, we integrate the ADI module with the diarization framework. The performance substantially improved over that of the baseline when we optimized the thresholds for agglomerative hierarchical clustering and the parameters for dimensionality reduction during scoring for individual acoustic domains. We achieved a relative improvement of 9.63% and 10.64% in DER for core and full conditions, respectively, for Track 1 of the DI-HARD III evaluation set.

Index Terms: acoustic domain identification (ADI), agglomerative hierarchical clustering (AHC), diarization error rate (DER), Jaccard error rate (JER), speaker diarization.

1. Introduction

Speaker diarization is the task of generating time-stamps with respect to the speaker labels [1]. For *rich transcription* (RT) of spoken documents involving multiple speakers, besides "what is being spoken" and "by whom", it is also imperative to know "who spoke when". In earlier studies, RT researchers mainly considered audio recordings of broadcast news, telephone conversations and meetings/conferences [1]. With the rise in multimedia content over the years, more variety is observed in the recording environments of the spoken documents[1]. Thus, instead of a *one size fits all* method, a domain-dependent speaker diarization approach might cater to the diverse and challenging recording conditions in a better way.

In our submission to the **Track 1** of the third DIHARD challenge [2], we propose a simple but efficient method for acoustic domain identification (ADI) using speaker embeddings. Next, we show that domain-dependent threshold for speaker clustering helps to improve the diarization performance only when probabilistic linear discriminant analysis (PLDA) adaptation uses audio-data from all the domains. Additionally, experimental optimization of the principal component analysis (PCA) parameters in a domain-specific way further improves the diarization performance.

2. The third DIHARD challenge

The third edition of the DIHARD challenge series, DIHARD III, has two tracks: **Track 1**: Diarization using reference SAD
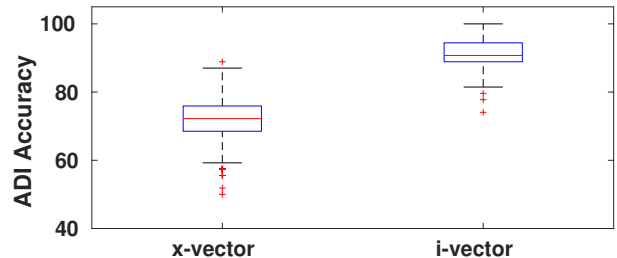

Figure 1: *Acoustic domain identification performance using x-vector and i-vector embeddings.*

and **Track 2**: Diarization from scratch. The challenge data set is taken from eleven diverse domains. This brings a lot of variability in data conditions like background noise, language, source, number of speakers, and speaker demographics. The database is divided into development and evaluation sets which consists of 5 - 10 minute duration samples. Evaluation has to be done on two partitions of the evaluation data: Core evaluation set - a balanced set where total duration of the each domain is approximately equal, and Full evaluation set - where all the samples are considered for each domain. Primary evaluation metric of the challenge is *diarization error rate* (DER), which is the sum of all the errors associated with the diarization task. *Jaccard error rate* (JER) is the secondary evaluation metric which is based on the Jaccard similarity index. The details of the challenge with rules are available in the challenge evaluation plan [2].

3. Acoustic domain identification system

Our ADI system is based on the speaker embeddings as sentence-level feature and nearest neighbor classifier. Though the speaker embeddings are principally developed for speaker characterization, they also capture information related to acoustic scene [3], recording session [4], and channel [5]. In this work, we study two frequently used speaker embeddings: discriminatively trained x-vectors and generative i-vectors. In an earlier study, these speaker embeddings were investigated for the second DIHARD dataset on related tasks [6, 7].

The experiments were performed on the development set consisting of 254 speech utterances from 11 different domains. We randomly selected 200 utterances for training and used the remaining 54 for test. For similarity measurement between training and test data, *cosine similarity* was employed. We repeated the experiments 1000 times and obtained average accuracy of 71.39% and 90.81% for x-vector and i-vector system, respectively. The comparison of ADI performance of the two systems, shown in Fig. 1, confirms that the i-vector system is substantially better than the x-vector system for ADI on DI-HARD III dataset. In the subsequent experiments, we have used this system (but trained with all 254 sentences) for identifying the domains of evaluation data samples.

Table 1: *Results showing the impact of domain-dependent processing on speaker diarization performance (DER in %/JER in %) for for Clinical and Court subsets of the development set of third DIHARD challenge.*

Domain	Method	First Pass		Re-segmentation with VB-HMM	
		Full	Core	Full	Core
Clinical	Baseline	17.55 / 28.88	16.08 / 26.38	17.69 / 28.46	16.72 / 27.21
	Domain-dependent threshold and PLDA adaptation	20.06 / 29.92	18.88 / 28.11	16.71 / 25.50	15.21 / 23.66
	Domain-dependent threshold and PLDA adaptation with full-data	15.81 / 23.69	14.61 / 22.37	14.69 / 22.07	13.78 / 21.59
	Same as above + domain-dependent parameter for PCA	14.67 / 22.66	12.91 / 20.83	13.79 / 20.65	12.66 / 20.05
Court	Baseline	10.81 / 38.75	10.81 / 38.75	10.17 / 37.63	10.17 / 37.63
	Domain-dependent threshold and PLDA adaptation	12.19 / 43.99	12.19 / 43.99	9.03 / 37.97	9.03 / 37.97
	Domain-dependent threshold and PLDA adaptation with full-data	5.82 / 23.91	5.82 / 23.91	4.77 / 22.22	4.77 / 22.22
	Same as above + domain-dependent parameter for PCA	5.03 / 17.30	5.03 / 17.30	3.82 / 16.04	3.82 / 16.04

4. Experimental setup

Our experimental setup is based on the baseline system created by the organizers [8]. We have used the toolkit[2] with the same frame-level acoustic features, embedding extractor, scoring method, etc. In order to extract utterance-level embeddings for ADI task, we used pre-trained x-vector and i-vector model trained on VoxCeleb audio-data[3].

5. Results on the challenge dataset

5.1. Impact of domain-dependent processing

In the first experiment with the baseline system, we analyzed the diarization performance for each domain and noticed that the performance varied with domains, as is expected. Next, the diarization experiment was separately performed with each domain. Interestingly, this degraded the performance compared to the baseline. This is more likely due to limited speaker and acoustic variability in domain-specific subset which is also used as in-domain target data for PLDA adaptation. Hence, we propose to apply domain-specific threshold for speaker clustering but PLDA adaptation is performed with audio-data from all the eleven subsets. The baseline system uses recording-dependent PCA during PLDA scoring where 30% of the total energy is preserved during dimensionality reduction. We experimentally optimized this value for all the domains separately which further improved the diarization performance. In Table 1, we show the results for two subsets clinical and court. Table 2 summarizes the results for the entire development set.

Table 2: *Results showing the speaker diarization performance using baseline (B) and proposed methods (M1 and M2) on **development set** of third DIHARD challenge. M1: domain-dependent threshold, M2: domain-dependent threshold and domain-dependent parameter for PCA.*

Method	Full		Core	
	DER	JER	DER	JER
B	19.59	43.01	20.17	47.28
M1	17.97	40.33	18.73	44.77
M2	17.40	38.08	17.95	42.12

5.2. Results on evaluation set

We performed domain-dependent diarization on the evaluation set by first predicting the domain for every utterance, followed by grouping of the utterances according to the predicted domains. We used domain-specific threshold for speaker cluster-

---

²https://github.com/dihardchallenge/dihard3_baseline
³https://kaldi-asr.org/models/m7

Table 3: *Same as Table 2 but for **evaluation set**.*

Method	Full		Core	
	DER	JER	DER	JER
B (Submission ID: 1044)	19.19	43.28	20.39	48.61
M1 (Submission ID: 1218)	17.56	38.60	19.23	43.74
M2 (Submission ID: 1373)	17.20	37.30	18.66	42.23

ing and percentage of total energy for dimensionality reduction optimized on corresponding domain of the development set. The results summarized in Table 3 indicate substantial improvement over the baseline.

6. Conclusion

In this work, we explored domain-dependent speaker diarization on the third DIHARD dataset by integrating an ADI system with the baseline system. Our ADI system is based on i-vector embeddings and nearest neighbour classifier. We applied domain-specific thresholds for speaker clustering and used domain-dependent PCA parameters for dimensionality reduction during PLDA scoring. The PLDA adaptation was performed with audio-data from all the domains. This way we achieved about ten percent relative improvement with respect to the baseline system for both the conditions in Track 1 of the challenge. The work can be extended with advanced embedding extractor based on ResNet, Extended-TDNN, etc.

7. Acknowledgements

8. References

[1] X. Anguera *et al.*, "Speaker diarization: A review of recent research," vol. 20, no. 2, pp. 356–370, 2012.

[2] N. Ryant *et al.*, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[3] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge," in *Proc. DCASE2018 Workshop*, 2018.

[4] D. Raj *et al.*, "Probing the information encoded in x-vectors," in *Proc. IEEE ASRU*, 2019, pp. 726–733.

[5] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Proc. INTERSPEECH*, 2017, pp. 1497–1501.

[6] M. Sahidullah *et al.*, "The Speed submission to DIHARD II: Contributions & lessons learned," *arXiv preprint arXiv:1911.02388*, 2019.

[7] T. Fennir, F. Habib, and C. Macaire, "Acoustic scene classification for speaker diarization," Université de Lorraine, Tech. Rep., 2020.

[8] N. Ryant *et al.*, "The third DIHARD diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.