Domain-Dependent Speaker Diarization for the Third DIHARD Challenge

A Kishore Kumar, Shefali Waldekar, Goutam Saha
ABSP Laboratory, Department of E & ECE, Indian Institute of Technology Kharagpur

Md Sahidullah
MULTISPEECH, Inria, France

Introduction

- **Speaker diarization (SD)** is the task of generating **timestamps** with respect to the **speaker labels** in a s**poken document** [1].
- **More variety in the recording environments** due to the increase in multimedia content over the years,
- A **domain-dependent SD approach** might be better than a *one-size-fits-all* method for the **diverse and challenging recording conditions**.

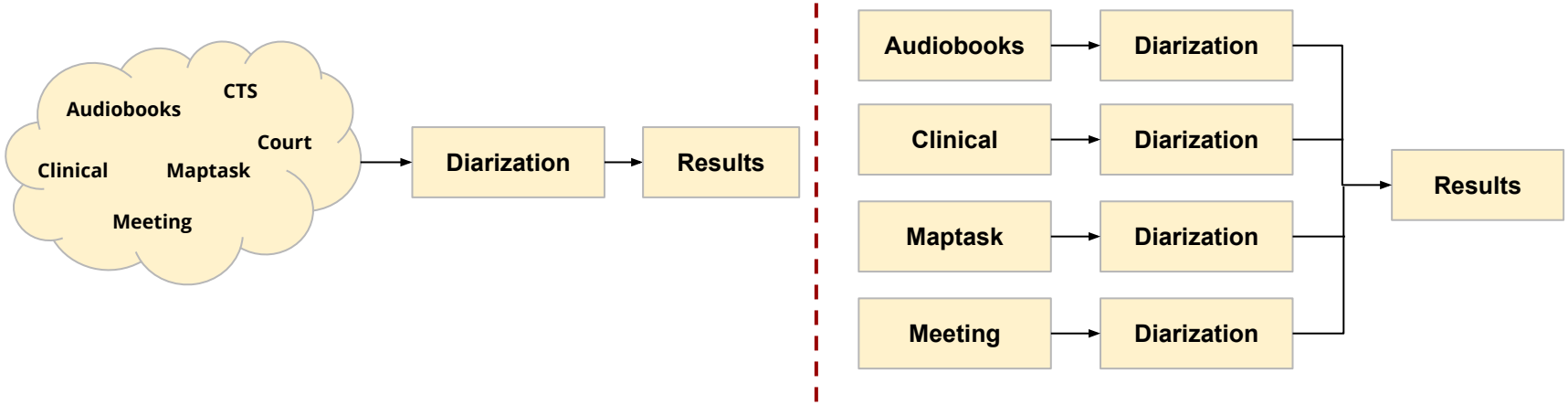

Figure 1: Illustration explaining difference between all-domain diarization and domain-wise diarization

[1] X. Anguera et al., vol. 20, no. 2, pp. 356–370, 2012.

Our contributions

- Development of acoustic domain identification (ADI) system:
  - Study of speaker embeddings

- Domain-dependent processing
  - Domain-dependent threshold for speaker clustering
  - Domain-dependent adaptation
  - Domain-dependent PCA parameters

Acoustic domain identification system

- 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 [2], recording session [3], and channel [4].
- We study two frequently used speaker embeddings: discriminatively trained x-vectors and generative i-vectors.

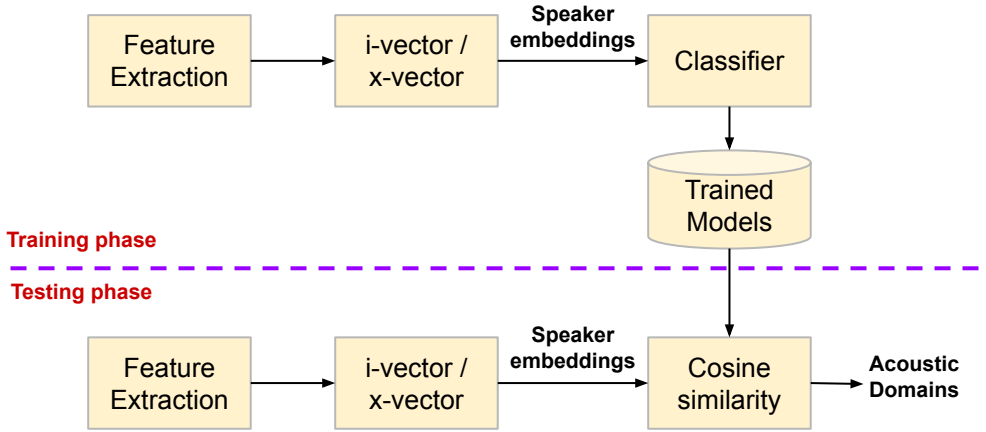

Figure 2:Block diagram of the proposed acoustic domain identification system

[2] H. Zeinali, L. Burget, and J. Cernocky, DCASE2018 Workshop, 2018. [3] D. Raj et al., Proc. IEEE ASRU, 2019, pp. 726–733. [4] S. Wang, Y. Qian, and K. Yu, INTERSPEECH, 2017, pp. 1497–1501.

Acoustic domain identification system

- 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.

- We repeated the experiments 1000 times and obtained average accuracy of 71.39% and 90.81% for x-vector and i-vector system, respectively.

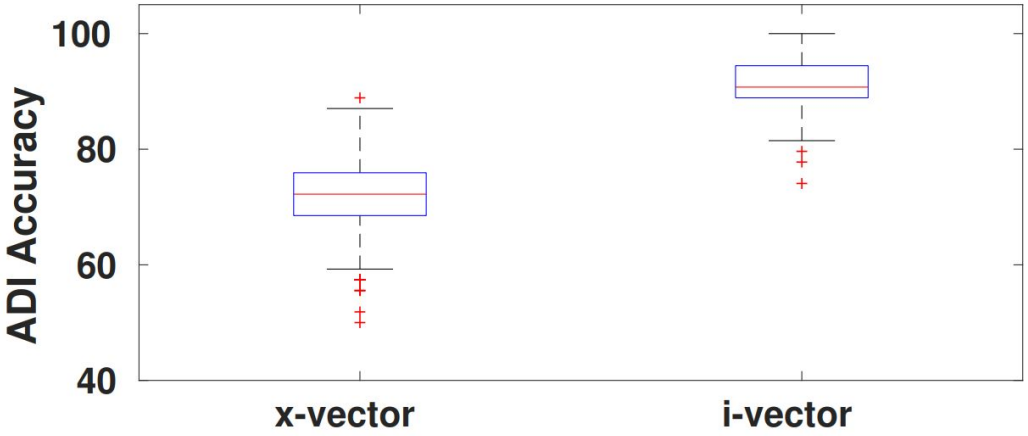

Figure 3:Acoustic domain identification performance using x-vector and i-vector embeddings

Experimental Setup

- For ADI system to extract utterance-level embeddings, we used pre-trained x-vector and i-vector model trained on VoxCeleb audio-data[1].
- Our experimental setup for speaker diarization is based on the baseline system created by the organizers[2].

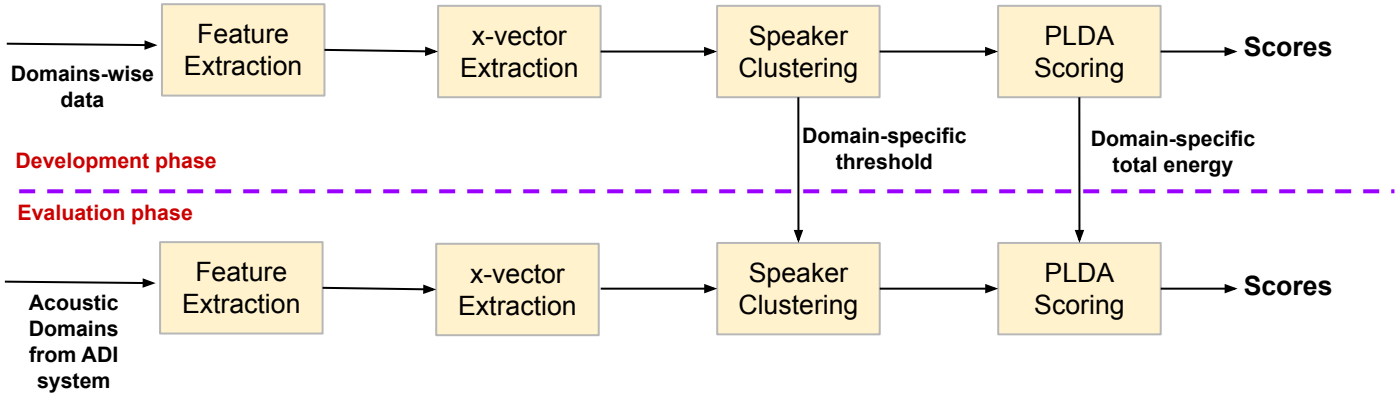

Figure 4: Block diagram representation of experimental setup

[1] *https://kaldi-asr.org/models/m7*
[2] *https://github.com/dihardchallenge/dihard3_ baseline*

Experimental Observations

- Domain-wise analysis shows that performance varied with domains
- Degradation was observed when each domain data was diarized separately
    - Limited speaker and acoustic variability could be the cause
- Domain-specific threshold for speaker clustering but PLDA adaptation with audio-data from all the eleven subsets gave better results
- Dimensionality-reduction in baseline using PCA: 30% of total energy is preserved
    - Better results when this was optimized for each domain separately

Experimental Results: Development Data

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

Domain	Method	Phase I		Phase II	
		Full (DER / JER)	Core (DER / JER)	Full (DER / JER)	Core (DER / JER)
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**

Experimental Results: Development Data

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**

Experimental Results: Evaluation Data

- Predict the domain for every utterance from ADI system
- Group the utterances according to predicted domains
- Use domain-specific parameters obtained from development set for clustering and dimensionality reduction

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**

Conclusion

- We explored **domain-dependent speaker diarization** on the third DIHARD dataset by integrating an **ADI system based on i-vector embeddings and nearest neighbour classifier** with the baseline system.
- 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.
- We achieved **about** t**en 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.