BUT extended abstract for The Third DIHARD Speech Diarization Challenge Workshop

Federico Landini[1], *Alicia Lozano-Diez*[1], *Lukáš Burget*[1], *Mireia Diez*[1], *Anna Silnova*[1], *Kateřina Žmolíková*[1], *Ondřej Glembek*[1], *Pavel Matějka*[1], *Themos Stafylakis*[2], *Niko Brümmer*[2]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]Omilia - Conversational Intelligence, Greece

{landini, lozano, mireia}@fit.vutbr.cz

Abstract

This paper describes the systems developed by the BUT team for The Third DIHARD Speech Diarization Challenge. The systems for both tracks consist of a DOVERlap fusion of an end-to-end NN system with x-vector based clustering systems in the form of spectral clustering and VBx. Given that the x-vector clustering systems do not provide overlapping speakers, overlapped speech is detected by a TasNet-based detector before the final fusion with the end-to-end approach. This system allows us to obtain competitive results on Track 1. For Track 2 we simply use the VAD from the baseline system provided by the organizers, instead of the oracle VAD labels.

Index Terms: Speaker Diarization, DIHARD, VBx diarization, end-to-end diarization, overlapped speech detection

1. Introduction

The Third DIHARD Challenge [1] is the third of a series of yearly diarization challenges focusing on hard conditions. This edition brings also the attention to telephone conversations as a new domain and two evaluation conditions: core (all domains have roughly the same duration) and full (a superset of core with more recordings for telephone and clinical interviews).

Taking into account the differences in nature between telephone channel and the rest of the domains, we use two substantially different diarization approaches and combine them with the aid of a telephone channel detector. Telephone conversations (according to the detector) are processed only with an end-to-end (E2E) neural network (NN)-based system. The rest of the recordings are processed with the aforementioned system and with different systems based on x-vector clustering which are in turn fused and processed with an overlapped speech detector (OVD) to allow for simultaneous speakers.

This pipeline was used for both Track 1, where oracle voice activity detection (VAD) labels were available, and Track 2, where diarization had to be performed from scratch.

2. Systems for Track 1

Our final submission for Track 1 consists of the fusion of four different systems. At the upper level of the fusion we used a telephone channel detector based on analyzing the average energy level on the upper part of the spectrogram for each recording. Recordings with an average level below a threshold were identified as telephone conversations and processed by an end-to-end NN-based diarization system. This system follows the approach described in [2] based on self-attention and encoder-decoder long short-term memory based attractors. It allows overlapped speakers and also performs VAD. However, we output always the most likely speaker and other speakers simultaneously depending on a threshold, and combine the outputs with the given oracle VAD boundaries. The system is trained on artificially created telephone conversations and finetuned to real telephone conversations from CALLHOME [3]. All recordings had a sampling rate of 8 kHz, and, accordingly, the challenge files were also subsampled to such frequency for this system.

Recordings identified as non-telephone were also processed by the E2E model described above, and by three more systems based on x-vectors extracted on the speech segments according to the oracle VAD labels. We will refer to those systems as *VBx adapted PLDA*, *VBx HTPLDA* and *SC* and we describe them below. *VBx adapted PLDA* is based on applying Bayesian hidden Markov model (BHMM) diarization on x-vectors extracted with a 152-layer ResNet [4] exactly as described in [5]. However, the probabilistic linear discriminant analysis (PLDA) model used was obtained by adapting the original PLDA to a PLDA trained using speaker segments from DIHARD III development set following the approach described in [6]. *VBx HTPLDA* is based on applying BHMM on time-delay neural network x-vectors [7]. In this case, the PLDA model was replaced by a heavy-tailed PLDA and parameters of the system tuned for such model. *SC* consists in applying spectral clustering with k-means based on cosine similarities between each pair of ResNet152 x-vectors.

The four systems were fused using DOVERlap [8]. Then, given that only E2E accounts for overlapped speech, we applied on the resulting system an heuristic to assign a second speaker on segments. This heuristic uses the closest in time speaker [9] in segments marked by an overlapped speech detector (OVD).

The OVD is based on the Conv-TasNet architecture [10] and its implementation in Asteroid [11]. It uses the encoder and separator parts of Conv-TasNet followed by softmax to classify 2 ms frames into three classes: silence, single-speaker speech and overlapped speech. For training, we use DIHARD III dev set, VoxConverse [12] dev set and three meeting datasets: ICSI [13], ISL [14] and AMI [15] train set (both beamformed and Mix-Headset). At first, we sample from the datasets with ratio $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ for DIHARD:VoxConverse:meeting datasets and anneal towards 1:0:0 in each training iteration. In half of the samples, we use the real data directly and in the other half, we artificially mix two segments to create more overlap examples.

A comparison of the results obtained by each system detailed by the domains on the development set is presented in Table 1. We observe that the E2E approach trained on telephone speech obtains the best result in the cts domain (telephone) by far. However, it also has a performance on par with the others for audiobooks and sociolinguistic lab domains. The reason for this is still unclear and requires further experiments. On one side, these are among the cleanest domains in terms of background noise; however, the same could be said about the

Table 1: *DER (%) for the different systems for Track 1 on each of the domains of the development set, core condition.*

System	ALL	audiob.	broadc.	clinical	court	cts	maptask	meeting	restaurant	soc. field	soc. lab	webvideo
VBx adapted PLDA	16.66	3.83	2.11	10.32	2.73	17.24	4.92	26.13	40.54	13.36	7.88	36.36
VBx HTPLDA	16.33	2	2.41	10.04	2.9	16.52	4.89	26.52	39.89	12.82	8.13	35.12
SC	16.63	0.38	3.13	11.2	3.5	16.7	6.09	26.87	38.93	13.77	8.33	36.32
E2E	24.17	0.56	14.42	21.62	25.31	9.29	16.97	39.02	53.96	18.86	7.18	40.36
DOVERlap	15.86	0	2.42	9.43	3.01	16.29	4.63	25.94	39.59	12.28	6.99	35.45
+ ov. handling	15.03	0	2.32	9.17	2.77	13.78	3.36	24.59	39.16	11.95	6.33	34.33
Final fusion	14.56	0	2.32	9.17	2.77	9.29	3.36	24.59	39.16	11.95	6.33	34.33

Table 2: *Results (%) for the different systems for Track 1 on development and evaluation sets, core and full conditions.*

System	Development										Evaluation			
	Core					Full					Core		Full	
	DER	Miss	FA	SER	JER	DER	Miss	FA	SER	JER	DER	JER	DER	JER
VBx adapted PLDA	16.66	10.95	0	5.72	37.19	16.26	10.93	0	5.33	33.68	16.67	37.69	15.74	33.75
VBx HTPLDA	16.33	10.95	0	5.38	36.82	15.98	10.93	0	5.05	33.35	16.54	37.82	15.5	33.61
SC	16.63	10.95	0	5.69	38.67	16.51	10.93	0	5.58	34.97	16.56	38.72	15.79	34.46
E2E	24.17	8.89	1.69	13.59	56.68	20.59	7.82	1.88	10.89	49.76	23.51	53.45	19.06	45.87
DOVERlap	15.86	10.94	0.01	4.92	38.36	15.57	10.92	0	4.65	34.5	16.22	39.47	15.26	35.08
+ ov. handling	15.03	9.76	0.09	5.18	37.72	14.30	9.38	0.11	4.82	33.62	16.07	39.09	14.25	34.32
Final fusion	14.56	9.37	0.27	4.91	37.42	13.49	8.17	0.82	4.49	32.95	15.46	38.68	13.29	33.45

Table 3: *Results (%) for the different systems for Track 2 on development and evaluation sets, core and full conditions.*

System	Development										Evaluation			
	Core					Full					Core		Full	
	DER	Miss	FA	SER	JER	DER	Miss	FA	SER	JER	DER	JER	DER	JER
VBx adapted PLDA	19.49	12.6	0.91	5.98	39.83	19.14	12.59	0.96	5.58	36.43				
SC	19.58	12.61	0.91	6.06	41.78	19.52	12.6	0.96	5.95	38.18				
E2E	26.14	10.41	2.49	13.24	57.61	22.68	9.39	2.76	10.54	50.86				
DOVERlap	19.07	12.57	0.91	5.59	41.66	18.74	12.54	0.97	5.23	37.86				
+ ov. handling	17.89	10.32	1.35	6.22	40.99	16.89	9.84	1.4	5.65	36.76				
Final fusion	17.52	10.09	1.51	5.91	40.72	16.32	9.17	2.02	5.12	36.17	24.62	44.49	21.09	39.28

broadcast interviews or maptask and yet this approach shows notable performance degradation on those domains. As for the x-vector based approaches, although all of them have similar performance for the whole core set, we see differences of 1 point between the best and worst approach for almost every domain. With the fusion, we see that in all domains the resulting system has either better performance than the best of the four or the performance is close to that of the best system.

When applying the overlapped speech handling, we obtain over 5% relative improvement in terms of DER, with some gain on all of the domains. However, the OVD system was partially trained on the development set so the results are over-optimistic. Comparing the results for development and evaluation on Table 2, this is clear with only 0.15% DER improvement when applying the overlap handling on the core condition. The effect on the full condition is larger mainly because a large improvement is obtained on cts, a domain with a larger proportion in full than in core. However, the E2E approach has even better performance on such domain and, it is not over-optimistic; then, the gain on the evaluation set for full condition is even more.

3. Systems for Track 2

For Track 2 we proposed a similar, yet slightly simpler system. The same procedure as for Track 1 was followed but without using *VBx HTPLDA* for the fusion. Also, instead of the oracle VAD labels, we used those given by the baseline VAD [1]. Results for the different approaches are presented in Table 3. Although the performance of the final system shows 22% relative deterioration in Track 2 wrt Track 1 on the development set, the degradation is 55% relative on the evaluation set, showing that the results on the dev set are overoptimistic given that the VAD was trained on such set. When exploring other approaches for VAD, we saw that usually different domains require different parameters for voice detection, proving that producing a one-fits-all VAD system is indeed challenging.

4. Conclusions

Systems outperformed those in previous DIHARD challenge editions, but overlapped speech still remains as one of the main challenges. We used for the first time an end-to-end NN-based diarization approach for the challenge as it can deal with overlapped segments by default and without post-processings. However, these systems by themselves did not attain competitive results for the challenge and needed to be combined with more standard approaches. Yet, we believe that they will be a key in the future of diarization and we will explore them further.

5. Acknowledgements

The work was supported by Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, European Union's Horizon 2020 project No. 833635 ROXANNE and European Union's Marie Sklodowska-Curie grant agreement No. 843627. Some of the methods were implemented during the JSALT2020 workshop, hosted by JHU.

6. References

[1] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[2] 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 *Proc. Interspeech 2020*, 2020, pp. 269–273. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1022

[3] A. F. Martin and M. A. Przybocki, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *7th European Conference on Speech Communication and Technology, Eurospeech*, vol. 7, num. 2, September 2001, pp. 787–790.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] F. Landini, O. Glembek, P. Matějka, J. Rohdin, L. Burget, M. Diez, and A. Silnova, "Analysis of the BUT Diarization System for VoxConverse Challenge," *arXiv preprint arXiv:2010.11718*, 2020.

[6] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černockỳ, "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6519–6523.

[7] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotnỳ *et al.*, "BUT System for the Second DIHARD Speech Diarization Challenge," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6529–6533.

[8] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs," *arXiv preprint arXiv:2011.01997*, 2020.

[9] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 683–686.

[10] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[12] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[13] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[14] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Seventh International Conference on Spoken Language Processing*, 2002.

[15] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.