Reviewing the DIHARD Challenges

Vladyslav Bondarenko

I. DIHARD-I AND DIHARD-II

A. Diarisation

Both DIHARD challenges aim at providing a comparative evaluation for systems performing diarisation task. Diarisation, also sometimes referred to as 'who spoke when', consists of segmenting and labelling a continues recording by a speaker. In modern systems, this task is usually split into several concrete steps. First speaker representations are extracted, usually in a form of i-vector [1] or x-vector [2]. Then, speech activity detection (SAD) is performed on each individual frame and frames are segmented according to pauses. Finally, in situations where there is an unknown number of speakers, clustering is usually performed to assign each of the segments to a speaker.

B. Challenges

Before DIHARD challenges, there was a lack of common diarisation tasks which resulted in fragmentation of efforts. Most teams worked on a specific domain, causing further concerns with lack of system generalisation to other domains. DIHARD challenges [3], [4] have addressed those issues with an introduction of a common task and evaluation data drawn from 11 different domains.

Both challenges split into several tracks. The first challenge provides with two tracks. For the first track, perfect segmentation has already been provided, which removes the need to perform sound activity detection (SAD). As will be seen later, it was found that error in SAD is usually one of the significant contributors to the final performance of the system. For the second track, diarisation had to be performed from scratch. Second DIHARD challenge kept the first two tracks and additionally provided with two for multi-channel audio. Submission for each of the tracks is evaluated separately.

C. Datsets

A key aspect of DIHARD is the variety of challenging data on which the systems are evaluated. Data is drawn from 11 different domains, including child language recording, clinical interviews, meetings, YouTube videos, audio-books and others. Over 19 hours of raw audio files drawn from those domains are provided to participants as training/development data. Participants are also free to use any external data sources to train their systems, with a few exceptions of sources from which evaluation data has been drawn. DIHARD-II provided with additional 6 hours of development data as well as provided even more accurate segmentation for track 1 on some of the domains. As well as that, multi-party dinner recordings are provided as a source of multi-channel data for track 3 and 4.

D. Evaluation

Key metric in both of the tasks is diarisation error rate (DER) which is the sum of all the different errors produced by a diarisation system. Formula as presented in first DIHARD challenge is shown below:

$$DER = \frac{FA + MISS + ERROR}{TOTAL} \tag{1}$$

Where *FA* is the total time that is incorrectly attributed to any of the speakers, *MISS* is the total time that is incorrectly not attributed to any of the speakers, and *ERROR* is total time attributed to incorrect speaker. The first two errors arise as a result of errors in SAD.

DIHARD-I additionally used mutual information evaluation and DIHARD-II instead of mutual information used Jackard error rate (JER). Nevertheless, some participants [5] reported that those errors are correlated with DER and so only reported later. Thus, only DER errors will be discussed as part of this report.

The final evaluation is done with a separate dataset that is not distributed to participants until the end of the competition. In the end, participants receive unlabeled evaluation data with some restriction on the investigation of the data that can be done. Participants then run the data through their system producing predictions which are sent back to the organisers and compared with true labels to produce the final scores. This process ensures fair competition and prevents participants from over-fitting their systems to the evaluation set.

Second DIHARD challenge additionally provided a baseline system, based on the winner of DIHARD-I submission. Baseline allows participants to compare their implementations and report any improvements with regards to the baseline. It also provides a good starting point for teams new to diarisation task, lowering the boundary to entry. Along with the success of the first challenge, this introduction has brought a lot more registration for DIHARD-II, as will be seen in the following section.

E. Participants

First DIHARD challenge attracted registrations from 20 teams of which 13 submitted a system. DIHARD-II more than doubled the registers with 48 teams from 17 different countries joining the challenge. Most successful submissions [5]–[8] that published their results came from relatively large teams from Universities. There was also a few teams with commercial background [9], [10].

II. Best Approaches

A. DIHARD-I

Top-performing Track 1 system for DIHARD-I was developed by a team from Johns Hopkins University (JHU) [6] with results demonstrated in the top row of Tab I. Their top-performing system used an x-vector representation extracted with Kaldi [11] toolkit. The Neural Net (NN) for feature extraction has been trained on VoxCeleb [12] dataset and demonstrate superior performance over the i-vector with more training data. It was found that the 16kHz sampling rate was more superior to 8kHz, which could be explained by the sampling rate collected data. The final step before clustering was to score the features using probabilistic linear discriminant analysis (PLDA) [13].

Once the feature vectors are obtained, they were clustered with an agglomerative hierarchical clustering (AHC) algorithm. The stopping threshold has been determined using a supervised approach. An attempt was made to train the domain-specific threshold but was found to over-fit. They have also experimented with fusing the PLDA scores from different systems but found little improvement when using x-vector. It seems like the most significant contribution came from applying a refined Variational Bayes (VB) algorithm to ensure more accurate segmentation. An interesting observation was made that VB was most effective with only one pass over the system, suggesting that underlying optimisation criteria could be less connected to improved diarisation.

TABLE I: Top evaluation results in DIHARD-I

Team	DER	
	Track 1	Track 2
JHU [6]	23.73	37.19
BUT [7]	25.07	35.51
USTC [8]	24.56	36.05
EUROCOME [9]	29.33	-

For the Track 2 the best performing team was from Brno University of Technology (BUT) [7] as demonstrated in Tab.I. Since the only difference between Track 1 and Track 2 is the availability of gold segmentation, the superior performance of BUT system on second track could be mostly attributed to a better SAD algorithm. Similarly to the JHU system, their SAD is a NN but trained on a different dataset. BUT have also chosen a minimal threshold to ensure that almost every voiced segment is recognised. Better results could likely be explained by better toleration of false alarms during the training of the diarisation system. Other than that their system was similar to JHU one in many ways - also using AHC with VB. An interesting investigation was made into overlapped speech detection, which was further developed for DIHARD-II and was the key contributor to their victory.

B. DIHARD-II

BUT team [5] came back for the second DIHARD challenge with a refined system and came out as a clear winner on all 4 tracks. Similarly to their previous submission, for track 1 and 2, an x-vector representation is used extracted using a deeper architecture with the higher frame rate. AHC clustering algorithm is used with a similarity metric based on PLDA. Clustering returns the initial set of labels which are then processed by Bayesian Hidden Markov Model based clustering algorithm [14] to obtain the final model of speaker distribution. The final model is trained using Variation Bayes inference similarly to the JHU system [6] described above. For the overlapped speaker detection, their system performs some post-processing where each frame is classified as either overlapped or not (binary-classification). If the segment is classified as overlapped, the second most probable speaker is added to the final prediction.

TABLE II: Top evaluation results in DIHARD-II

Team	DER			
	Track 1	Track 2	Track 3	Track 4
Baseline [3]	25.99	40.86	50.85	77.34
BUT [5]	18.21	27.11	47.93	58.92
LEAP [9]	21.90	-	-	-
NTIS [15]	23.47	-	-	-

As for the multi-channel system, they proposed a much simpler architecture with x-vector extraction and AHC performed per channel. Final results for each of the tracks are demonstrated in Tab.II, where significant improvement over the baseline is demonstrated for each of the tracks.

III. Alternative Approaches

The following section explores some of the alternative approaches that were investigated by teams for different steps of the diarisation pipeline.

A. Denoising

JHU team has found that any form of denoising or signal preprocessing prior SAD, has resulted in reduced performance. In contrast, a team from the University of Science and Technology of China (USTC) [8] demonstrated that a sufficiently advanced denoising model could contribute to better performance. Their main contribution was a Long-Short Memory Network (LSTM) architecture for predicting clean Linear Predictor Coefficients (LPS). Their final system demonstrated excellent performance achieving a second place in both track 1 and track 2 sub-challenges of DIHARD-I.

B. Representation

Although the majority of teams including the winners have focused on using i-vectors and x-vector as their front-end; a team from EUROCOME [9] have demonstrated that a relatively effective approach can be developed with an alternative

representation. They use an infinite impulse response, constant Q transform Mel-frequency cepstral coefficients (ICMC). They justify it as a more appropriate model of the human perception system.

C. Re-segmentation

It appears like the final re-segmentation plays the crucial part in final system performances; in-particular when gold segmentation is provided. As such, few teams [7], [15] that used less advance approaches such as Gaussian Mixture Models (GMM) achieved worth performance. NTIS [15] team, for example, used a very similar set up as BUT for the second DIHARD challenge - x-vectors with AHC clustering. However, their final re-segmentation was performed using a GMM which likely played a key role in almost 30% degradation in final performance over BUT winning system.

IV. Conclusion

A. Progress

First DIHARD challenge stood up to its name and demonstrated that diarisation is far from being a solved problem. Even top systems with lowest DER values performed worth than previous state-of-art [16]. For some domains error rates even exceed 50%, highlighting the problem of generalisation of current systems to a range of domains. Nevertheless, BUT team demonstrated great improvement in the performance of almost 30% from their first challenge system. It is clear that they are the leading research team in the area of diarisation and demonstrate promising improvements. However, such a drop in DER from DIHAD-I could also be partially attributed to better segmentation for track-1 data, and actual improvement in full diarisation pipeline calculated from track-2 is around 24%. Still, this could also be partially due to better SAD system, which remains the critical performance bottleneck.

Overall, there are some definite attributes of top-performing diarisation systems. Feature extraction using Deep Learning (x-vectors) shows promising improvements; particularly as datasets and computing power grow in size. AHC clustering is consistently chosen as a primary clustering algorithm due to its unique ability to produce a variable number of clusters. Finally, the key step in reducing the error and what seems to differentiate system performance the most is the choice of the re-segmentation algorithm. Variation of HMM trained using VB seem to perform the best. Some exploration has been done into performing partial domain-specific training of key parameters, but overall those demonstrate over-fitting [5], [6], [15] which could be attributed to lack of data for individual domains.

B. Challenges

Although there is some improvement, the DER is still too high for any operational uses. As been discussed previously and as demonstrated by different in errors between track 1 and track 2, one of the critical problems is errors in SAD systems. Gold segmentation, as provided for track 1 would not be available during a realistic use-case scenario, and SAD, is expected to be performed accurately in a live setting. Poor performance could be attributed to an extent to inferior denoising approaches and could be improved as demonstrated by [8].

Recognising multiple speakers within a single segment remains a big problem. BUT teams addressed this issue to an extent with their binary classification algorithm improving final results very slightly. Unfortunately, the performance of their classification algorithm is not reported, and it is hard to evaluate how much it contributes. Furthermore, their approach could only recognise if there are two speakers at a time, and potential extensions could be considered, such as multi-class classification.

It appears like diarisation performance on multi-channel audio is much worse than with a single channel as demonstrated by BUT team. Nevertheless, this could be attributed to a lack of data and evaluation on a single challenging domain. Moreover, there hasn't been a proper investigation done on best approaches for multi-channel predictions as this is the first year when track 3 and 4 have been introduced.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," tech. rep.

[2] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "DEEP NEURAL NETWORK-BASED SPEAKER EMBEDDINGS FOR END-TO-END SPEAKER VERIFICATION," tech. rep.

[3] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines," tech. rep., 2019.

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD Challenge Evaluation Plan," tech. rep., 2018.

[5] F. Landini1, . , Shuai Wang1, and Mireia Diez1, "BUT SYSTEM FOR THE SECOND DIHARD SPEECH DIARIZATION CHALLENGE."

[6] 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," tech. rep.

[7] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. K. Kateřinažmolíková, N. Novotn´y, K. V. Vesel´y, O. Glembek, O. Plchot, L. Mošner, and P. Matějka, "BUT system for DIHARD Speech Diarization Challenge 2018,"

[8] L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin, and C.-H. Lee, "Speaker Diarization with Enhancing Speech for the First DIHARD Challenge," tech. rep.

[9] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD Challenge,"

[10] A. Kanagasundaram, "LEAP Diarization System for the Second DIHARD Challenge," *Interspeech 2019*.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. S. Silovsk´y, G. Stemmer, and K. V. Vesel´y, "The Kaldi Speech Recognition Toolkit," tech. rep.

[12] A. Nagrani, J. Son Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," tech. rep.

[13] S. Ioffe, "Probabilistic Linear Discriminant Analysis," tech. rep., 2006.

[14] M. Diez, L. Burget, S. Wang, and J. Rohdin, "Bayesian HMM based x-vector clustering for Speaker Diarization," 2019.

[15] Z. Zajíc, M. Kunešová, M. Hrúz, and J. Vaněk, "UWB-NTIS Speaker Diarization System for the DIHARD II 2019 Challenge," tech. rep.

[16] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields," tech. rep.