BUT SYSTEM FOR THE SECOND DIHARD SPEECH DIARIZATION CHALLENGE

Federico Landini[1]*, Shuai Wang*[1,2]*, Mireia Diez*[1]*, Lukáš Burget*[1]*, Pavel Matějka*[1]*, Kateřina Žmolíková*[1]*, Ladislav Mošner*[1]*, Anna Silnova*[1]*, Oldřich Plchot*[1]*, Ondřej Novotný*[1]*, Hossein Zeinali*[1]*, Johan Rohdin*[1]

[1]Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence, Czechia
[2] Speechlab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{landini,mireia}@fit.vutbr.cz

ABSTRACT

This paper describes the winning systems developed by the BUT team for the four tracks of the Second DIHARD Speech Diarization Challenge. For tracks 1 and 2 the systems were mainly based on performing agglomerative hierarchical clustering (AHC) of x-vectors, followed by another x-vector clustering based on Bayes hidden Markov model and variational Bayes inference. We provide a comparison of the improvement given by each step and share the implementation of the core of the system. For tracks 3 and 4 with recordings from the Fifth CHiME Challenge, we explored different approaches for doing multi-channel diarization and our best performance was obtained when applying AHC on the fusion of per channel probabilistic linear discriminant analysis scores.

Index Terms— Speaker Diarization, Variational Bayes, HMM, DIHARD, CHiME

1. INTRODUCTION

With the aim of bringing attention to diarization performed on challenging data, the Second DIHARD Diarization Challenge [1] proposed a common ground for comparison of different diarization systems on multiple domains. Alike the first DIHARD Diarization Challenge [2], tracks 1 and 2 consisted in performing diarization on single-channel recordings from different domains with and without oracle voice activity detection (VAD) labels, respectively. Tracks 3 and 4 focused on multi-channel data from the Fifth CHiME Challenge [3] also with and without VAD labels, respectively.

Our effort allowed us to obtain the first position on all four tracks. This paper describes those four winning systems but due to the lack of space some details are omitted. For specific parameter configurations we refer the reader to [4]. Together with this publication, the code of the most relevant modules of the winning system of track 1 has been made available [5].

The challenge proposed four tracks working with two different datasets and this paper will be structured accordingly. Section 2 describes the complete processing pipeline that we used for speaker diarization in tracks 1 and 2: We describe the used signal pre-processing, x-vector extraction and agglomerative hierarchical clustering (AHC) of x-vectors applied to obtain initial labels for the following step. In the next step, which is the core of our diarization pipeline, x-vectors are clustered using Bayesian hidden Markov model (BHMM). This method is often referred to as *VB diarization* [6]. Next, frame-level re-segmentation is performed based on another BHMM and finally, the overlapped speech post-processing is applied. Since we did not have oracle VAD labels for track 2, we describe how we trained a VAD system. Experiments related to each part of the system are described in each subsection followed by their corresponding discussions.

Section 3 focuses on processing the CHiME data used in tracks 3 and 4. We describe the clustering method used for diarization and the experiments we carried out to take advantage of the multi-channel data. We also show the performance when using a VAD system instead of the oracle labels.

Finally, we comment on the conclusions we reached to during and after the challenge. We also comment on the challenges that we see in the task of speaker diarization and the paths we believe we should follow.

2. SYSTEMS FOR TRACKS 1 AND 2, DIHARD DATA

In previous works we have shown that the BHMM-based diarization system can be tuned for different domains in order to achieve much better performance [7]. Still, this requires a module capable of classifying recordings with great accuracy. Since misclassification in this step can be quite harmful, we opted for focusing on a single diarization system for all domains. The different parts of the system are described and their results discussed in the following subsections. In this section, results in gray indicate "cheating" results where the test data is also used for training (i.e. the development set).

2.1. Signal Pre-processing

Due to the high levels of noise and reverberation in some of the recordings, we explored four different enhancing methods for improving the quality of the signals. We compared

The work was supported by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

the method provided by the organizers[1], one based on the Wave-U-Net [8], one based on neural network (NN) autoencoders [9] and the weighted prediction error (WPE) [10, 11] method which removes late reverberation. Alike the first edition of the challenge [12], we found WPE to be the most effective in this regard. Therefore, we pre-processed all training, development and evaluation recordings using this technique.

2.2. x-vectors

A very successful recent approach for speaker diarization is to cluster deep neural network (DNN) based speaker embeddings known as x-vectors [13, 6]. From the input recordings, x-vectors are typically extracted every 0.75s from 1.5s overlapping sub-segments. For this challenge, we use a higher x-vector frame-rate of 0.25s for tracks 1 and 2 as we found it to significantly improve the results [14]. In our diarization pipeline, x-vectors are clustered in two steps using AHC and BHMM as described in the following sections.

The x-vector extractor used is based on the SRE16 recipe [15] from the Kaldi toolkit [16] with some modifications: We use a larger and deeper neural network for x-vector extraction, which is trained on VoxCeleb 1 and 2 [17] with data augmentation. We use more training epochs than the original recipe and a different strategy to generate the input examples from the training speech recording. More details can be found in [4].

2.3. AHC Initial Clustering

The x-vectors extracted from an input recording are first clustered by means of AHC with similarity metric based on probabilistic linear discriminant analysis (PLDA) [18] log-likelihood ratio scores, as used for speaker verification. We train two PLDA models for this purpose: The first one is trained on x-vectors extracted from 3 seconds speech segments from VoxCeleb 1 and 2 which are mean centered, whitened to have identity covariance matrix and length-normalized [19]. The centering and whitening transformation are estimated on the joint set of DIHARD development and evaluation data to have a better estimate since unsupervised use of the evaluation data is allowed. To take advantage of the in-domain data, the second PLDA model is trained on x-vectors extracted in a similar way using the DIHARD development data. The centering and whitening transformation are also estimated on the joint set of DIHARD development and evaluation data. Note that these transformations are applied both to development and evaluation x-vectors when performing diarization.

The final "domain-adapted" PLDA model used for AHC-based clustering is obtained as an interpolation of the two PLDA models: means, within- and across-class covariance matrices from the two models are averaged. Table 1 shows that, on both the development and evaluation sets, the interpolated PLDA improves the diarization performance of AHC as compared to using PLDA trained only on out-of-domain VoxCeleb data. Still, the percentage of files where using the interpolated version worsens the results is 32% for the development set and 45% for evaluation. However, the result on development data with the interpolated PLDA is overoptimistic as the test data are also used for the PLDA training.

Note again that, in our diarization pipeline, the AHC is only used to obtain initial labels for the following BHMM based clustering. For more detailed analysis of the AHC-based diarization subsystem, we refer the reader to [14].

	VoxCeleb	Interpolated	Same	Improved
dev	20.46	19.74	9%	59%
eval	21.12	20.96	11%	45%

Table 1. DER on development and evaluation sets for AHC diarization using the PLDA trained on VoxCeleb and domain-adapted interpolated PLDA. Also percentage of files with equal or improved DER (when using the interpolated PLDA).

2.4. Bayesian HMM for x-vector clustering

BHMM is used to cluster x-vectors as described in detail in [6]. Before the BHMM-based clustering, the x-vectors as well as the parameters of the PLDA model are projected using linear discriminant analysis (LDA). The LDA projection is calculated directly from the parameters of the original (interpolated) PLDA model described above. More details in [4]. The resulting PLDA model is used in the BHMM to model speaker distributions as described in [6].

Variational Bayes (VB) inference for BHMM needs initial assignment of x-vectors to speaker clusters. This is taken from the previous AHC step, which needs to be run so as to under-cluster the x-vectors. This way the VB inference has more freedom to search for the optimal results and potentially remove redundant speakers[2]. A more thorough analysis on this matter is presented in [14].

Iterative VB inference is run until convergence to update the assignment of x-vectors to speaker clusters. Automatic relevance determination (ARD) [20] inherent in BHMM results in dropping redundant speaker clusters and allows us to properly estimate the number of speakers in each recording.

To optimize the diarization performance, the VB inference is controlled by a number of tunable parameters. In [6], we have newly introduced the *speaker regularization coefficient* F_B, which affects the model to be more or less aggressive when dropping the redundant speakers. *Acoustic scaling factor* F_A is introduced to compensate for the incorrect assumption of statistical independence between observations (i.e. x-vectors). P_{loop} is the probability of not changing speakers between observations, which serves as speaker turn duration model. Note that our system uses a higher frame-rate

[1] https://github.com/staplesinLA/denoising_DIHARD18

[2] Note that the inference in BHMM cannot converge to higher number of speakers than what is suggested by the AHC-based initialization.

for x-vector extraction compared to the ones used in former works, which requires using higher value of P_{loop} and lower value of F_A as compared to the optimal values reported in [6]. Details about the specific parameter values used for the challenge can be found in [4].

	VoxCeleb	Interpolated	Same	Improved
dev	18.34	17.90	14%	60%
eval	19.14	18.39	22%	56%

Table 2. DER on dev and eval sets for BHMM-based diarization using the PLDA trained on VoxCeleb and the domain-adapted interpolated PLDA. Also percentage of files with equal or improved DER (when using the interpolated PLDA).

Table 2 shows the improvement obtained by using the BHMM model for diarization. When using the PLDA model trained on VoxCeleb data, the BHMM approach improves around 2% absolute DER on both development and evaluation set with regard to the AHC (Table 1). Moreover, by using the PLDA interpolation only 26% of files in development and 22% in evaluation have worse results than when using the PLDA trained on VoxCeleb, showing even better improvement than with AHC-based diarization.

2.5. BHMM-based Re-segmentation

Given that the output from the previous step has a time resolution of 0.25s, we apply an additional frame-level VB re-segmentation step similar to the one used in our previous challenge submission [12]. VB re-segmentation is based on another BHMM which operates on mel-frequency cepstral coefficients (MFCC) with 10 ms frame rate as the input observations. In this case, state distributions are modeled by an i-vector extractor like model (i.e Gaussian mixture models with parameters constrained by eigenvoice priors) [21, 22], which is pre-trained on VoxCeleb 2 data. The initial assignments of MFCC frames to HMM states (speakers) is derived from the previous x-vector clustering step. Note that in the re-segmentation step only one VB iteration is performed [6] instead of running the algorithm until convergence.

	BHMM	+ reseg.
dev	17.90	18.23
eval	18.39	18.38

Table 3. DER obtained using only BHMM at x-vector level and adding BHMM at frame-level (re-segmentation) on the development and evaluation sets.

The effect of this extra re-segmentation step is shown in Table 3. Unlike in our previous participation [12], where the frame level re-segmentation gave great gains over simply doing AHC, this time this approach gives marginal gains. The reason for this is several-fold: the x-vectors are newly extracted every 0.25s, which provides 3 times better time resolution as compared to the typical 0.75s. Further, the BHMM-based x-vector clustering step produces a better diarization output than the AHC, leaving less margin for improvements. More importantly, BHMM-based x-vector clustering uses a PLDA model adapted to the target domain, whereas the re-segmentation step uses models solely trained on VoxCeleb data[3]. Adapting also the re-segmentation BHMM (i.e. the built-in i-vector model) would likely improve the re-segmentation results. See [14] for more details.

2.6. Overlapped Speech Post-processing

Given that none of our models accounts for overlapped speech (i.e. they all assume one speaker speaking at a time), we perform overlapped speech detection and apply a heuristic to label segments for more than one speaker.

For each recording, silence segments are removed and speech segments are concatenated. Then x-vectors are extracted from 1.5s sub-segments every 0.25s and classified using a logistic regression classifier as overlapped or non-overlapped speech. The classifier is trained on x-vectors extracted from the development set and labeled as overlapped speech if more than half of the original segment contains overlapped speech. Once overlap segments are detected, the heuristic consists in assigning for each frame in an overlapped speech segment the two closest speakers (in time) according to the diarization labels given by the previous step.

	No ov. proc.	With ov. proc.
dev	18.23	18.02
eval	18.38	18.21

Table 4. DER before and after doing overlapped speech post-processing on development and evaluation sets.

Table 4 shows the comparison of results with and without overlap post-processing. Although the x-vector extractor is trained with single speaker speech, the x-vectors still capture relevant information for overlapped speech detection. Note that 18.21 is slightly better than the result from our winning system[4]. In our submission, a more complex PLDA adaptation scheme was used: Directions with larger within- and across-class variability in the in-domain PLDA model (trained on DIHARD dev) than in the out-of-domain PLDA (trained on VoxCeleb) were identified and the extra variability was added to the corresponding covariance matrices in the out-of-domain PLDA so that they would have at least as much variability as in the in-domain PLDA model. Later, we found that the simple interpolation of the two PLDA models is sufficient and even slightly improves the results.

[3] This also explains the degradation obtained with re-segmentation step on dev data in Table 3

[4] 18.42 as shown in http://dihard.ldc.upenn.edu/competitions/73#results

2.7. Voice Activity Detection

Track 2 consisted in evaluating the same DIHARD set without using oracle VAD labels. Although we explored using the same VAD system we had used before [12], we found out that a DNN-based VAD system trained on the development set for binary, speech/non-speech, classification of speech frames provided better performance. More details in [4].

The model proposed for this track is essentially the same as for track 1 using the estimated (instead of oracle) VAD labels. However, due to a lack of time, PLDA interpolation is not done (the model trained on VoxCeleb data is used) and no overlapped speech post-processing is applied. With the DNN-based VAD, we obtain 23.81 DER on the development set ("cheating") and 27.11 DER on the evaluation set.

3. SYSTEMS FOR TRACKS 3 AND 4, CHIME DATA

Tracks 3 and 4 proposed for the first time a multi-channel diarization task using CHiME 5 data. This posed new challenges with respect to performing diarization on the DIHARD data as we had to devise a system capable of using data collected in 4 channels arranged in microphone arrays. For these tracks we propose a much simpler diarization system.

3.1. Multi-channel AHC Clustering

Our diarization system is based on performing AHC of x-vectors using information from all channels. We analyzed two ways of taking advantage of multiple channels. One possibility we explored is to apply beamforming [23] using the four signals in order obtain a single beamformed channel to use as input for the diarization system. The second approach consists in extracting x-vectors from each channel, computing the corresponding pairwise similarity PLDA score matrices, averaging them and then performing AHC on the resulting score matrix.

For all approaches, the AHC of x-vectors is performed in a similar manner as for track 1: recordings from all channels are processed with the WPE method, x-vectors are extracted as described in 2.2 and AHC is performed on the pairwise-similarity PLDA score matrices as in 2.3. In this case, the PLDA model trained on VoxCeleb segments is adapted in an unsupervised way to the train and development data from the CHiME corpus.

Table 5 shows results for the different approaches: performing x-vector AHC on each channel separately, on the beamformed signal and on the PLDA score matrix obtained from the fusion of the PLDA matrices of each channel. Even though we found that the optimal thresholds for the different channels are different, all values correspond to the same threshold for the sake of comparison. As can be seen, there are some differences on performance when different channels are evaluated (up to 0.83 DER).

Using the beamformed signal does not provide better results than simply any of the single channels; still, the parameters used for producing the beamformed signal might not be optimal. Tuning of parameters was not explored during the challenge and remains as future work.

However, fusing the score matrices does improve the results. This pattern was seen for every threshold; however, the best AHC threshold for the *Fusion* approach is always closer to 0, suggesting that the score fusion provides better calibrated scores. Note that the DER on the evaluation set for the *Fusion* approach shown in the table is worse than our winning system[5]. This is because a different AHC threshold was used for the submission: Since we saw different thresholds impacted considerably the performance in the dev+train set, we decided to submit diarization outputs obtained with different thresholds for evaluation expecting the performance would also vary.

	CH1	CH2	CH3	CH4	Beam.	Fusion
dev+train	55.43	55.34	55.78	54.95	55.75	53.58
eval	48.55	48.37	48.19	48.3	50.31	47.93

Table 5. DER of our system on each channel separately, on the beamformed channel and when using fusion of scores, on the development+training and evaluation sets.

3.2. Voice Activity Detection

For the track 4 of the challenge, the oracle VAD labels are not available and a VAD system [24] is run only on one of the channels to produce VAD labels.

With the NN-based VAD and the best aforementioned system, we obtain 58.92 DER which compares to 45.65 DER when oracle VAD labels are used. Training a tailored VAD system on the training set analogously as in track 2 remains as future work.

4. CONCLUSIONS

Another edition of the DIHARD challenge has taken place and it has proven to be too hard for McClane[6]. One year has passed, the quality of x-vector extractors has improved and they have become the cornerstone for top-performing diarization systems. However, we have shown that it is possible to take more advantage of them when using BHMM in comparison to simply doing AHC. Given the current performance of the systems, the overlapped speech gains more relevance accounting for more than 50% of the DER in our best systems in both sets. We believe this has to be addressed in the future as well as other approaches for adapting x-vectors to in-domain data. The challenge gave us the opportunity to work for the first time on multi-channel diarization.

[5] 45.65 as shown in http://dihard.ldc.upenn.edu/competitions/75#results

[6] One day before the deadline, putting our faith in our latest system allowing us to obtain the first position in the ranking, we named the submission "McClane". However, a literally last minute submission allowed us to even improve its performance http://dihard.ldc.upenn.edu/competitions/73#results

5. REFERENCES

[1] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," 2018.

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[4] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "BUT System Description for DIHARD Speech Diarization Challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[5] L. Burget, M. Diez, S. Wang, and F. Landini, "VBHMM x-vectors Diarization (aka VBx)." `https://speech.fit.vutbr.cz/software/vbhmm-x-vectors-diarization`.

[6] M. Diez, L. Burget, S. Wang, J. Rohdin, and H. Černocký, "Bayesian HMM based x-vector clustering for Speaker Diarization," in *Proceedings of Interspeech*, 2019.

[7] M. Diez, L. Burget, F. Landini, and H. Černocký, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[8] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[9] O. Plchot, L. Burget, H. Aronowitz, and P. Matějka, "Audio Enhancing With DNN Autoencoder For Speaker Recognition," in *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 2016*, pp. 5090–5094, IEEE Signal Processing Society, 2016.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1717–1731, Sept 2010.

[11] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation*, Oct 2018.

[12] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Žmolíková, O. Novotný, K. Veselý, O. Glembek, O. Plchot, *et al.*, "BUT System for DIHARD Speech Diarization Challenge 2018," in *Proceedings of Interspeech 2018*, pp. 2798–2802, 2018.

[13] 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," in *Interspeech*, pp. 2808–2812, ISCA, 2018.

[14] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černocký, "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[15] Kaldi, "SRE16 v2." https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[18] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *in Proceedings of Odyssey*, June 2010.

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech 2011*, 2011.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[21] M. Diez, L. Burget, and P. Matějka, "Speaker diarization based on bayesian hmm with eigenvoice priors," in *Proceedings of Odyssey 2018, The speaker and Language Recognition Workshop*, 2018.

[22] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," tech. rep., Montreal: CRIM, 2008.

[23] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," 2006.

[24] P. Matějka and et.al., "BUT-PT system description for nist lre," in *Proceedings of NIST Language Recognition Workshop 2017*, pp. 1–6, National Institute of Standards and Technology, 2017.