Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
Johns Hopkins University, USA

gsell@jhu.edu

Abstract

We describe in this paper the experiences of the Johns Hopkins University team during the inaugural DIHARD diarization evaluation. This new task provided microphone recordings in a variety of difficult conditions and challenged researchers to fully consider all speaker activity, without the currently typical practices of unscored collars or ignored overlapping speaker segments. This paper includes details about research decisions made in the process of the evaluation, including explorations of training data, feature bandwidth, speech activity detection, and i-vector versus x-vector representations. The utility of system fusions and domain-specific processing are also considered. In all cases, deeper analyses are presented using the development data, for which truth labels were given during the evaluation, while evaluation scores are only presented for a subset of the systems. Finally, we discuss lessons learned and remaining challenges left for future work within the lens of this new approach to diarization performance measurement.

Index Terms: speaker diarization

1. Introduction

Speaker diarization is the problem of organizing a conversation into the segments spoken by the same speaker (often referred colloquially to "who spoke when"). While measurements of diarization performance have continued to improve, in recent years, individual research projects have tended to focus on specific datasets (such as Callhome, AMI, or broadcast). This effect, at the very least, makes it difficult to compare performance, and, more problematic, could be leading to divergent solutions that overfit to the characteristics of particular corpora.

In response to this research rut, the inaugural DIHARD challenge was intended to provide a standard set of data drawn from diverse and challenging conditions to evaluate current system performance and provide a standard set for future diarization research. The development (dev) data released with labels during the challenge included data from ten diverse domains ranging from monologues to interviews with children to meetings to internet videos. An additional three domains were included in the evaluation (eval) data as well, though truth marks for this set were not available at time of writing, and so the effects of those additional domains remain unknown. This resulted in a highly diverse and challenging dataset for diarization.

Additionally, teams were invited to participate on two tracks. In the first, oracle marks for speech activity were provided, following the standard often followed for Callhome diarization research. In Track 2, however, speech marks needed to be estimated with a speech activity detection (SAD) algorithm. This direction provides an additional pathway for challenge and for research opportunity.

This paper describes the submissions for the challenge from the Johns Hopkins University (JHU) team, as well as the series of experiments that shaped the final system built from the initial system designed with Callhome diarization. The discussion also includes possible directions for future work, as the relatively short duration of the challenge meant many paths were necessarily left unexplored.

2. DIHARD Challenge Experiments

Through the course of this challenge, we explored a number of system modifications. This section outlines many of these experiments, after describing the initial system used as a starting point. For simplicity, the results are discussed using diarization error rate (DER) with no unscored collars and including overlapping speech, which was one of the two official metrics of the evaluation. The second metric, mutual information (MI), is not included here because the high-level conclusions are essentially the same as with DER, and it is comparatively less familiar for diarization.

2.1. Initial System

Our initial baseline system divided the speech signal into short (1.5-2 second) segments with a hop of 1 second. An i-vector was extracted for each of the these segments, and the collection of segment i-vectors [1] was then scored with probabilistic linear discriminant analysis (PLDA) [2, 3] and clustered using agglomerative hierarchical clustering (AHC). System components were trained for this initial system for improved performance on Callhome [4], and the speech was processed at 8kHz.

The initial SAD system used for Track 2 was a bidirectional long short-term memory (BLSTM) DNN trained on a limited set of Switchboard speech with augmentation via added noise, reverberation, and low-bitrate speech coding. Like the initial clustering sytem, the initial SAD system processed all audio at 8kHz.

Scores for the initial system for the dev and eval set on both tracks are shown in Table 2. These numbers served as our starting point for the challenge.

It is worth noting that previous Callhome diarization research found refining the clustering marks with a frame-level diarization system utilizing subspace models via Variational Bayes (VB) [1] to be a highly effective technique [5]. However, in our initial experiments, this addition was found to be detrimental to performance, and several modifications were required. As

[1] Code available at http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors

a result, the VB refinement is left out of our initial system, and the process for improving it is left for Section 2.9.

2.2. Speaker Representation

The initial system utilized i-vectors as the speaker representation, but recent work has also shown that DNN-based representations called x-vectors [6] can also be effective for diarization [7]. Since that initial report, x-vectors have improved performance significantly with multiclass discriminative training [8] and augmented training data [9], and so utilizing x-vectors for the DIHARD challenge was a clear goal.

All JHU x-vector systems for this challenge utilized Kaldi [10]. 8kHz i-vector systems used features of 20 mel-frequency cepstral coefficients (MFCCs) along with the first-order differences (deltas) computed over 9 frames, while 8kHz x-vector systems used the log-compressed magnitudes of 23 mel-frequency spectral bands. For 16kHz systems, the feature dimensions were expanded to 30 mel filters for x-vectors and 24 MFCCs for i-vectors. X-vector systems were built in accordance with previous speaker recognition work [9], except the intermediate layers were limited to 128 dimensions instead of 512 (unless otherwise stated), while i-vectors were 64-dimensional, projected from the statistics of a 1024-component UBM.

Performance for x-vectors and i-vectors was measured for a number of different training lists built from combinations of VoxCeleb [11], various broadcast news corpora distributed by LDC (primarily in English, Arabic, or Chinese), audio from European Parliament videos, Librispeech, or Mixer 4/5.

As was shown in prior work [9], x-vectors are more able to capitalize on larger quantities of training data, as their performance continues to improve with additional data, while the i-vector system is never able to improve beyond the initial microphone-trained version with VoxCeleb only. PLDA training, similarly, was best for the i-vector system when only using VoxCeleb, while the x-vectors benefited from more data sources. Additionally, unlike in previous Callhome work [4], we found PLDA to be more effective when trained with all same-speaker audio in the same class, as opposed to training to the combination of speaker and channel. For whitening lists, however, the best data combination was found to include VoxCeleb, Mixer 4/5, and the provided DIHARD data, and this combination was best for all systems.

2.3. Speech Activity Detection

Track 2 of the DIHARD challenge required systems to estimate their own SAD marks. As described above, the initial SAD system utilized a BLSTM-DNN trained with the CURRENNT[2] toolkit on 8kHz telephony with synthetic variations. Input features were 13 MFCCs with deltas and double-deltas appended. This system was found to perform at a miss rate of 10.2% at a false alarm of 4.6%. The threshold for separating speech and non-speech can be modified in order to shift the balance of misses and false alarms, but we found that misses are a preferable error in SAD systems for segment-clustering diarization. This is presumably because including segments of non-speech can corrupt the clustering, creating additional errors beyond the SAD mistakes. Additionally, it is also the case that every second of corrected false alarms results in a second of overall error reduction, while corrected misses still need to be clustered properly to reduce final error, so it may simply be that lower false alarm rates more reliably map to improved final error.

[2]https://sourceforge.net/projects/currennt/

System	Track 1 DER	Track 2 DER
NB i-vector	24.81	35.84
WB i-vector	21.74	33.72
NB x-vector	23.42	34.69
WB x-vector	21.42	33.17

Table 1: *DER scores on the dev data comparing narrowband and wideband performance for both x-vectors and i-vectors. The additional bandwidth in wideband processing is clearly helpful in the task*

In order to improve SAD performance on this data, we first trained a 5-layer TDNN [12, 13] with 16kHz microphone data of audio from European Parliament videos, again with various synthetic variations to add more diversity to the training data. The benefit of a TDNN for this task is that it can incorporate a wider input context without exploding the number of parameters. In this case, each layer doubles the width, resulting in 150ms of input context for each decision. However, this system performed resulted in worse performance than the initial SAD system, with a much higher miss rate of 17.4% at a similar false alarm rate of 4.8%.

But, if we instead retrained only the final classifier layer with the provided dev data (or, when testing on dev, with a two-fold split of the dev data), the performance improved significantly (7.3% miss and 4.1% false alarm). A similar strategy using MFCC input features was also effective (7.3% miss and 6.0% false alarm), but the retraining the final layer of the TDNN was even better. Taking this strategy an extra step and training separate final layers for each domain was able to provide small additional gains for dev (6.1% miss and 4.2% false alarm), but reduced performance on eval, as will be discussed in more detail in Section 2.7.

2.4. Signal Bandwidth

The initial diarization system utilized by the JHU team was built for the telephone recordings of Callhome, and therefore was only trained for 8kHz (narrowband) data. However, the DIHARD challenge data is sampled at 16kHz (wideband), and so half of the bandwidth would be ignored without retraining with 16kHz data. Results comparing performance on narrowband or wideband data can be seen in Table 1 for both x-vectors and i-vectors. In both cases, the systems were only trained with VoxCeleb, though for the x-vector training, data augmentation was also employed.

The results in Table 1 show clearly that wideband processing provides an advantage for these systems, especially for Track 1. Based on these experiments, our research focused on utilizing wideband processing in both i-vector and x-vector systems.

2.5. Signal Enhancement

Given the presence of microphone recordings from multiple noisy environments, we tested the effects of signal enhancement via mask-based speech separation. The specific algorithm used to estimate the spectral masks utilized a BLSTM-DNN trained with CHiME-3 data [14]. The masking was tested in front of the test (dev) data only, as well as in front of the training data for PLDA or the i-vector extraction.

Overall the effects of mask-based separation were detri-

mental to diarization, dropping performance in the system used from 22.8% DER to 24.1% in total performance. The degradation was to a lesser degree when all elements of the pipeline were trained with enhanced data (with unmatched conditions, the DER dropped as far as 28.7%), which would typically be expected to be the best strategy, but even then, it is better to simply run the unprocessed data. However, for the SEEDLINGS domain, the enhancement led to consistent gains (from 44.7% DER to 38.7% in the matched case). More analysis is required to fully understand this effect, but it might indicate that some noisier files benefitted from the estimated masks, and that a mask estimation algorithm trained with a more diverse set of data sources could be more broadly effective as a diarization front-end.

2.6. AHC Threshold

An important parameter for speaker diarization with AHC determines when the AHC merges should stop. In past work, unsupervised estimation of this threshold proved to be effective for Callhome diarization [4], but the threshold can also be determined in a supervised fashion when labeled in-domain data is available.

For the DIHARD challenge, we found supervised thresholding to be a consistently effective approach. However, in the case of wideband i-vectors, a small gain of roughly 0.2% DER was attainable with unsupervised threshold estimation using the eval data. The technique did not appear to be as effective for x-vectors, but that asymmetry has been seen in past work as well [7].

2.7. Domain-specific Processing

The gains found from utilizing dev data for SAD retraining or representation whitening suggested that specialized systems could yield improvements for this challenge. Continuing that logic, we tried domain-specific processing for both our SAD system and for learning the AHC threshold.

For SAD, separate final DNN layers were learned for every domain using the known labels of the dev data. A domain estimation was also learned with a simple logistic regression classifier using the mean of the features across the file. Running the SAD for the estimated domain resulted in an improvement on SAD performance for held-out dev data (improving miss/false alarm from 7.3%/4.1% to 6.1%/4.2%), but these gains did not map to eval, resulting in an overall degradation in DER performance on the baseline i-vector system (43.9% to 45.0%).

Similarly, an attempt was made to use domain-specific thresholds to stop AHC merges. In this case, the thresholds were learned on known dev labels, and a domain estimator was again trained, this time using the segment i-vectors. In this case, the threshold was defined by a linear combination of domain-specific thresholds weighted by the posterior probability for each domain in the estimation. While performance improved in held-out dev experiments, the metrics again degraded for eval.

These two domain-specific experiments modified different modules of the pipeline, differed in their domain estimation (though both performed at over 80% accuracy on held-out dev data), and differed in making hard or soft domain assignments. In both cases, performance was shown to improve on dev, but degraded on eval. This outcome suggests that either the domain-specific processes are overfitting the dev data (either in domain estimation or domain-specific systems themselves), or the presence of several unseen domains in the eval data is causing enough gain in error to overcome the improvements seen in dev. Either way, we were unable to capitalize on domain-specific processing for the challenge.

2.8. System Fusion

Finally, gains were found by fusing the PLDA scores of multiple systems prior to AHC clustering. The scores themselves were fused via a weighted sum with coefficients learned to optimize performance on the dev set. After fusion, the scores were clustered in the same fashion as for an individual system.

Table 2 shows the eval scores for a fusion system as well as for each of the individual systems within the fusion. As can be seen, the majority of the performance is given by an x-vector system alone, but the combination does still result in an improvement.

2.9. VB Refinement

After segment clustering (which is the outcome of all modules discussed to this point), it is often helpful to refine the marks, as the initial segmentation was likely too quantized. This step would seem to be especially important in the case of the DIHARD challenge, where the practice of unscored collars around speaker transitions was abandoned, requiring more precise labeling. However, the VB refinement previously found to be highly effective for Callhome [5] was initially detrimental to system performance.

Fixing the VB refinement for this data took several steps. First, parameters were re-learned with wideband microphone data (in this case, VoxCeleb). However, in order to yield actual improvements, it was also required to de-couple the stopping criteria from the optimization criteria. In previous work, the process was permitted to iterate until convergence. However, for this data, that leniency was found to degrade performance, and, in fact, only permitting one pass was a better solution. This development not only helped VB refinement more consistently improve performance, but it also raised interesting questions about the underlying models for microphone diarization. The fact that the optimization criteria is less connected to improved diarization suggests that there is an incorrect assumption, and this is an area that warrants further analysis. It is possible that the changing dynamics of far-field microphone recordings (non-stationary noise, moving sources, etc.) violate the total variability assumption that diarization is finding a speaker in the same channel. If the channel is indeed varying throughout the recording, this would not only explain the degradation in VB refinement as compared to Callhome, but also the improvements found here by training PLDA to recognize speakers across channels (as is done for speaker recognition) instead of speakers within channels. Again, this requires more analysis, but the possibility is quite interesting and could have important implications for future diarization of microphone/far-field recordings.

Table 2 shows the performance of the final clustering systems with and without VB refinement (system details in the next section). Comparing each system with and without VB refinement shows that the addition is quite advantageous to performance in all conditions. The gains diminish with better clustering, but the process is clearly valuable for all cases.

2.10. Final Submissions

The aggregation of all these factors can be seen in Table 2, where performance results for the final JHU submissions are shown.

System	Track 1		Track2	
	Eval DER	Dev DER	Eval DER	Dev DER
All same speaker	39.01	35.97	55.93	48.69
Initial	31.56	26.58	50.78	40.89
WB i-vector, no VB	28.06	21.74	40.42	33.72
WB x-vector, no VB	25.94	20.03	39.43	31.80
Fusion, no VB	25.50	19.54	39.00	31.79
WB i-vector, with VB*	25.06	19.69	37.41	31.29
WB x-vector, with VB*	23.73	18.20	37.29	29.84
Fusion, with VB*	23.99	18.17	37.19	30.31

Table 2: *Dev and Eval performance measured in DER for both tracks for several JHU systems, including the final submitted systems (marked with *). Details of the initial system are described in Section 2.1 and details of the final submissions are described in Section 2.10*

The wideband i-vector system is trained on 16kHz VoxCeleb data for UBM, T, and PLDA, while the wideband x-vector system is trained on an aggregation of 16kHz data from VoxCeleb, Mixer 4/5, Librispeech, European Pariliament videos, and various broadcast news corpora. The x-vector embedding dimension for this system is 256. Thresholds are learned in a supervised fashion, with the dev threshold learned with cross validation, and the eval threshold learned on dev as well. VB refinement is then used as described above on all systems, and all submission systems used the wideband SAD TDNN with the final layer retrained with dev data.

The fusion is then the combination of the wideband i-vector system and two wideband x-vector systems (128 dimensions and 256 dimensions), fused as described above. Interestingly, there are gains from fusion without VB refinement, but with VB refinement included, the single x-vector system is essentially equivalent to the fusion.

With all factors combined, the improvements from the initial system to the final submission are clear in both tracks. In Track 1, our performance improved by 7.83% DER, while Track 2 performance improved by 13.59% DER, an improvement of approximately 25% relative in both cases. This process was a valuable experience in generalizing diarization to simultaneously perform in a variety of environments, and these gains show that the process was a success. At the same time, the error rates are still high (especially for Track 2) and so it is clear that there is still plenty of opportunity for the future.

3. Future Work

As the descriptions above should demonstrate, the initial experience of the JHU team for the inaugural DIHARD challenge was mostly devoted to updating systems to work in the challenging microphone conditions of the evaluation. And while this was an important and worthwhile process, there were many longer-term research directions that were largely left for future work.

The first and most obvious of these is that we were unable to devote resources to handling the overlapping speech that is frequently present in real conversational dynamics. Roughly 8% of the absolute error in our systems was from overlapping speech, which accounted for at least a fifth of our error in Track 2, and a third in Track 1. However, the challenge of overlapping speech is not trivial, as it will likely require a complete rethinking of the diarization process, since our current system simply does not allow for multiple speakers to be responsible for the same frame of speech.

Although significant gains were made during the challenge in SAD performance, this remained another source of significant error. It is also somewhat disappointing that improvements in SAD required retraining on truth marks provided with the dev data. Ideally, SAD systems should work reasonably well without requiring in-domain supervision, and this is a bar that our systems were unable to clear. A robust and widely-applicable SAD is also an area of continuing research.

It is also the case that, given the nature of this evaluation, we do not yet know the details of the successes and failures of our systems beyond the overall performance, since eval truth was not provided to teams. As a result, an important step for future work will be to better understand the sources of error in the eval set. This is especially important for understanding the effect of the unseen domains, as well as measuring the degree of overfitting to the specific nuances of the dev data. So, to some extent, understanding directions for future work is also a goal of future work, since the receipt of the truth marks for the eval data in the near future will allow for a deeper understanding of the shortcomings of the submitted systems.

4. Conclusions

The inaugural DIHARD challenge provided an opportunity to measure diarization performance in challenging conditions without the benefit of knowing the answers. This process was a valuable experience in rethinking systems to work in more general conditions, but also in confirming the effectiveness of our general pipeline. By the completion of the evaluation, we had trained more effective diarization systems with wideband data, learned how to make VB refinement more effective in microphone conditions, and built a single x-vector system that was essentially our best performing diarization system.

5. Acknowledgements

The authors would like to thank the organizers of the DIHARD challenge for an interesting and look forward to future iterations.

6. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–98, May 2011.

[2] S. Ioffe, "Probabilistic Linear Discriminant Analysis," in *ECCV*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer-Verlag, 2006, pp. 531–42.

[3] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[4] G. Sell and D. Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[5] ——, "Diarization Resegmentation in the Factor Analysis Subspace," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[6] 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," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2016.

[7] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker Diarization Using Deep Neural Network Embeddings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[8] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, 2017.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition Using Data Augmentation," submitted to ICASSP 2018.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannermann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proceedings of Interspeech*, 2017.

[12] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–39, March 1989.

[13] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[14] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.