The Third DIHARD Diarization Challenge

Neville Ryant[1], *Prachi Singh*[4], *Venkat Krishnamohan*[4], *Rajat Varma*[4], *Kenneth Church*[2], *Christopher Cieri*[1], *Jun Du*[3], *Sriram Ganapathy*[4], *Mark Liberman*[1]

[1]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
[2]Baidu Research, Sunnyvale, CA, USA
[3]University of Science and Technology of China, Hefei, China
[4]Electrical Engineering Department, Indian Institute of Science, Bangalore, India

nryant@ldc.upenn.edu

Abstract

This paper introduces the third DIHARD challenge, the third in a series of speaker diarization challenges intended to improve the robustness of diarization systems to variation in recording equipment, noise conditions, and conversational domain. Speaker diarization is evaluated under two segmentation conditions (diarization from a reference speech segmentation vs. diarization from scratch) and 11 diverse domains. The domains span a range of recording conditions and interaction types, including read audiobooks, meeting speech, clinical interviews, web videos, and, for the first time, conversational telephone speech. We describe the task and metrics, challenge design, datasets, and baseline systems for speech speech activity detection and diarization.

Index Terms: speaker diarization, speaker recognition, robust ASR, noise, conversational speech, DIHARD challenge

1. Introduction

Speaker diarization, often referred to as “who spoke when”, is the task of determining how many speakers are present in a conversation and correctly identifying all segments for each speaker. In addition to being an interesting technical challenge, it forms an important part of the pre-processing pipeline for speech-to-text [1] and is essential for making objective measurements of turn-taking behavior. Early work in this area was driven by the NIST Rich Transcription (RT) evaluations [2], which ran from 2002 to 2009. In addition to driving substantial performance improvements, especially for meeting speech, the RT evaluations introduced diarization error rate (DER), which remains the principal evaluation metric in this area.

After the RT evaluation series ended in 2009, diarization continued to improve (e.g., i-vectors, x-vectors, PLDA scoring), though until quite recently there was no common task for diarization, resulting in a fragmented research landscape where individual groups focused on different datasets or domains (e.g., conversational telephone speech [3, 4, 5, 6, 7], broadcast [8, 9], or meeting [10, 11]), often with slightly differing evaluation methodologies. At best, this has made comparing performance difficult, while at worst it may have engendered overfitting to individual domains/datasets, resulting in systems that do not generalize. Moreover, the majority of this work has evaluated systems using a modified version of DER in which speech within 250 ms of reference boundaries and overlapped speech are excluded from scoring. As short segments such as backchannels and overlapping speech are both common in conversation, this may have resulted in an over-optimistic assessment of performance even within these domains[1] [12].

Recently, there has been renewed interest in a diarization common task to facilitate systematic benchmarking. Whereas from 2009-2017 there were no major evaluations with a diarization component, there now is an annual diarization specific evaluation – DIHARD – as well as numerous other challenges that include a diarization component; among others, the Fearless Steps series [13, 14], the Iberspeech-RTVE challenge [15], CHiME-6 [1], and VoxSRC-20[2].

The first DIHARD challenge (DIHARD I) [16] ran in the spring of 2018 and evaluated diarization of single channel wideband recordings drawn from a diverse range of domains. As expected, state-of-the-art systems performed poorly, with final DER on the evaluation set for the top systems ranging from 23.73% [17] when provided with reference speech activity detection (SAD) marks to 35.51% [18] when forced to perform diarization from scratch – error rates rates more than double the state-of-the-art for CALLHOME [19] at the time [5, 6]. This was followed by DIHARD II [20, 21] in 2019, which was even more successful, attracting 50 teams from 17 countries and 4 continents. While DIHARD II continued the single channel diarization tracks from DIHARD I, it also collaborated with the CHiME challenge series with the addition of two new tracks focusing on conversational speech from multiple farfield microphone arrays during a dinner party scenario. All tracks continued to be challenging for participants, with the tracks that required systems to produce their own speech segmentation and dinner party data particularly challenging. In the case of the latter, the CHiME-6 data, DER of the best performing system was over 45% when provided with an oracle speech segmentation and over 58% when required to produce its own segmentation.

The third DIHARD challenge (DIHARD III), which builds upon DIHARD I and II, addresses the problem of robust diarization; that is, diarization that is resilient to variation in, among others, conversational domain, recording equipment, recording environment, reverberation, ambient noise, number of speakers, and speaker demographics. Like its predecessors, diarization system performance is evaluated under two SAD conditions: diarization from a supplied reference SAD and diarization from scratch. There are no constraints on training data, with participants allowed to use any combination of public/proprietary data for system development. Recordings are sampled from 11 de-

[1] See, for instance, the release of IBM’s diarization API in 2017. The feature worked well for simple cases, but when run by users on real inputs, the performance was found to be lacking, especially for overlaps, back-channels, and short turns.

[2] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html

manding domains ranging from clean, nearfield recordings of read audiobooks to extremely noisy, highly interactive, farfield recordings of speech in restaurants to clinical interviews with children. Unlike DIHARD II, diarization from multi-channel audio is not evaluated; parties interested in this condition should instead consult the results from track 2 of CHiME-6 [1], which is essentially a rerun of the DIHARD II multichannel condition.

In the remainder of this paper, we introduce the task (Section 2), metrics (Section 3), and data (Section 4), as well as the baseline SAD and diarization systems (Section 5). Results of the baseline systems for both tracks are reported in Section 5.2. More details may be found in the evaluation plan [22] and on the challenge website:

`https://dihardchallenge.github.io/dihard3/`

2. Task

The goal of the challenge is to automatically detect and label all speaker segments for each recording; that is: i) determine how many speakers are present; ii) for each speaker identify all corresponding speech segments. Because system performance is strongly influenced by the quality of the speech segmentation used, two different tracks are covered:

- **Track 1** – Diarization from reference SAD. Systems are provided with a reference speech segmentation that is generated by merging speaker turns in the reference diarization.
- **Track 2** – Diarization from scratch. Systems are provided with just the raw audio input for each recording session and are responsible for producing their own speech segmentation.

3. Performance Metrics

As in DIHARD I and II, the primary metric is DER [2], which is the sum of missed speech, false alarm speech, and speaker misclassification error rates. Because systems are provided with the reference speech segmentation for track 1, for this track DER exclusively measures speaker misclassification error. This is the metric used to rank systems on the leaderboards. For each system we also compute a secondary metric, Jaccard error rate (JER), originally introduced for DIHARD II. JER is based on the Jaccard similarity index [23, 24], a metric commonly used to evaluate the output of image segmentation systems, which is defined as the ratio between the sizes of the intersections and unions of two sets of segments. An optimal mapping between speakers in the reference diarization and speakers in the system diarization is determined and for each pair the Jaccard index of their segmentations is computed. JER is defined as 1 minus the average of these scores, expressed as a percentage.

All metrics are computed using version 1.0.1 of the *dscore* tool[3] without the use of forgiveness collars and with scoring of overlapped speech. For further details, please consult Section 4 of the DIHARD III evaluation plan [22] and the *dscore* repo.

4. Datasets

4.1. Overview

The development and evaluation sets consist of selections of 5-10 minute duration samples drawn from 11 domains exhibiting wide variation in recording equipment, recording environment, ambient noise, number of speakers, and speaker demographics. These domains range in difficulty from the trivial, read audiobooks recorded under clean conditions by a single speaker, to the extremely challenging, conversations between up to 6 diners recorded by a binaural microphone in restaurants with varying room acoustics and noise levels. Both adult and child speech (e.g., ADOS interviews) are represented as is speech from multiple languages (English and Chinese). For the first time, narrowband recordings are included as well as wideband recordings; in the narrowband case, all recordings are drawn from the unreleased Phase II calls from the Fisher English collection conducted as part of the DARPA EARS project. All audio is distributed via LDC as 16 kHz, monochannel FLAC files.

Table 1: *Overview of DIHARD III datasets. The* **Part.** *column indicates the partition (core or full), while the* **% speech** *and* **% overlap** *columns indicate, respectively, the percentage of speech/overlapped speech in the partition.*

Set	Part.	# rec	# hours	% speech	% overlap
dev	core	181	23.94	78.43	10.04
	full	254	34.15	79.81	10.70
eval	core	184	22.73	77.35	-
	full	259	33.01	79.11	-

The datasets are summarized in Table 1. For additional details about the domains and source drawn on for each domain, consult the DIHARD III evaluation plan [22].

4.2. Scoring partitions

For DIHARD III, we define two partitions of the evaluation data:

- **core evaluation set** – a “balanced” evaluation set in which the total duration of each domain is approximately equal
- **full evaluation set** – a larger evaluation set that uses all available selections for each domain; it is a proper superset of the core evaluation set

The core evaluation set strives for balance across domains so that the evaluation metrics are not dominated by any single domain. It mimics the evaluation set composition from DIHARD I and II. The full evaluation set includes additional material from two domains (clinical interview and CTS), potentially resulting in more stable metrics at the expense of being unbalanced. All system submissions to all tracks are scored against both sets and the results reported on the leaderboards.

4.3. Annotation

Reference diarization was produced by segmenting the recordings into labeled speaker turns according to the following guidelines:

- split on pauses $>$ 200 ms, where a pause by speaker “S” is defined as any segment of time during which “S” is not producing a vocalization of any kind[4]
- attempt to place boundaries within 10 ms of the true boundary, taking care not to truncate edges of words (e.g., utterance-final fricatives or utterance initial stops)

[3] `https://github.com/nryant/dscore`

[4] Vocalization is defined as any noise produced by the speaker by means of the vocal apparatus; e.g., speech (including yelled and whispered speech), backchannels, filled pauses, singing, speech errors and disfluencies, laughter, coughs, breaths, lipsmacks, and humming.

- where close-talking microphones exist for each speaker, perform the segmentation separately for each speaker using their individual microphone

Reference SAD was then derived from these segmentations by merging overlapping speech segments and removing speaker identification.

During DIHARD II, it was found that manual annotation to this spec required use of highly skilled and experienced annotators using multiple spectrogram displays, making the annotation extremely slow and costly. Many annotators were incapable of performing the task even after extensive training and the remainder found it extremely laborious with real time rates typically greater than 15X and sometimes exceeding 30X. Consequently, for DIHARD III we abandoned a commitment to entirely manual segmentation. Where a manual segmentation to these specs already existed (i.e., files annotated for DIHARD II), we used it. For all other data we instead produced a careful turn-level transcription, then established boundaries using a Kaldi-based forced aligner.

5. Baseline system

5.1. Speech activity detection

The baseline for track 2 uses a TDNN SAD model based on the Kaldi Aspire recipe (“egs/aspire/s5”). 40-D mel frequency cepstral coefficients (MFCCs) extracted every 30 ms using a 25 ms window are fed into a neural network consisting of 5 TDNN layers [25] followed by 2 statistics pooling layers [26]. The network context is set to approximately 1 second (left context: 0.8 sec; right context: 0.2 sec). The DNN was trained with two classes – speech and non-speech – with labels at training time derived from the reference speech segmentation for the DIHARD III DEV set. Training utilized the entire DIHARD III DEV set and was continued for 40 epochs. During inference, the posteriors of the model were converted to pseudo-likelihoods using the empirical speech/non-speech priors for the DEV set and Viterbi decoding was performed using an HMM with the following constraints: minimum speech duration: 240 ms, minimum non-speech duration: 30 ms.

5.2. Diarization

The diarization baseline is based on LEAP Lab’s submission to DIHARD II [27]. The system performs diarization by dividing each recording into short overlapping segments, extracting x-vectors [28, 29], scoring with probabilistic linear discriminant analysis (PLDA) [30], and clustering using agglomerative hierarchical clustering (AHC) [31]. The AHC ouput is then refined using Variational Bayes Hidden Markov Model (VB-HMM) [32, 33] with posterior scaling [27]. The trained models and recipes for both tracks are distributed through GitHub[5].

The x-vector extractor configuration is identical to that of [17, 29] with two exceptions: i) 30-D MFCCs are used instead of a mel filterbank; ii) the embedding layer uses 512 dimensions. MFCCs are extracted every 10 ms using a 25 ms window and mean-normalized using a 3 second sliding window. For training we use a combination of VoxCeleb 1 and VoxCeleb 2 [34, 35] augmented with additive noise and reverberation according to the recipe from [28]. Segments under 4 seconds duration are discarded, resulting in a training set with 7,323 speakers. Reverberation is added by convolution with room responses from the RIR dataset [36], while additive noises are drawn from the MUSAN dataset [37]. At test time, x-vectors are extracted from 1.5 sec segments with a 0.25 sec shift. x-vectors are centered and whitened using statistics estimated from the DIHARD III development set, followed by length normalization [38] .

Table 2: *Baseline SAD results for the core/full DEV and EVAL sets. The Part. columns indicates whether scoring was performed using the full or core DEV/EVAL set.*

Set	Part.	Miss (%)	FA (%)	Overall error (%)
dev	core	1.84	3.98	2.30
	full	1.88	4.55	2.42
eval	core	4.97	15.07	7.26
	full	4.35	14.65	6.51

The x-vectors are then clustered using AHC and a similarity matrix produced by scoring with a Gaussian PLDA model [30]. The PLDA model was trained using centered, whitened, and length normalized x-vectors extracted from VoxCeleb segments with duration $\geq$3 sec. Prior to PLDA scoring, dimensionality reduction was performed using conversation-dependent PCA [4] preserving 30% of the total variability. For each track, the stopping criteria for AHC was tuned to minimize DER on the DEV set.

We then refine the AHC output using frame-level Variational Bayes Hidden Markov Model (VB-HMM) resegmentation as described by [32, 33]. 24-D MFCCs are extracted every 10 ms using a 15 ms window; neither mean nor variance normalization are applied, nor do we use delta coefficients. We use a Universal Background Model (UBM-GMM) with 1,024 diagonal-covariance components and a total variability (V) matrix containing 400 eigenvoices. Both the UBM-GMM and V were trained using the same data as for the x-vector extractor. Following [27], posterior scaling was applied to discourage frequent speaker transitions by the VB-HMM. This scaling was accomplished by boosting the zeroth order, but not first or second order, statistics prior to VB-HMM likelihood computation. The VB-HMM is initialized separately for each recording from the result of AHC and run for one iteration. Parameters were set to the following values by tuning on the DIHARD III DEV set: scaling factor $\beta = 10$, loop probability $Ploop = 0.45$, downsampling factor $downSamp = 25$.

Table 3: *Track 1 diarization results for the core/full DEV and EVAL sets with and without VB-HMM resegmentation.*

Part.	VB-HMM reseg.	DER (%)		JER (%)	
		Dev	Eval	Dev	Eval
core	no	21.05	21.66	46.34	48.10
core	yes	20.25	20.65	46.02	47.74
full	no	20.71	20.75	42.44	43.31
full	yes	19.41	19.25	41.66	42.45

6. Baseline results

6.1. Track 1

Table 3 reports DER and JER for Track 1. DER for the full system ranges from 19.25% to 20.65% and JER from 41.66% to 47.4%. Mirroring the findings of [18, 39], VB-HMM resegmentation reliably improves DER and JER, though here the gains are relatively modest: about 1% absolute for DER and less than

[5] https://github.com/dihardchallenge/dihard3_baseline/

Table 4: *Track 2 diarization results for the core/full DEV and EVAL sets with and without VB-HMM resegmentation.*

Part.	VB-HMM reseg.	DER (%)		JER (%)	
		Dev	Eval	Dev	Eval
core	no	24.06	29.51	49.17	53.82
core	yes	22.28	27.34	47.75	51.91
full	no	24.08	28.00	45.61	49.35
full	yes	21.71	25.36	43.66	46.95

1% for JER. Possibly, the effects of VB-HMM resegmentation could be enhanced by using a UBM-GMM and variablity matrix trained on or adapted to in domain materials, though we did not explore this possibility. As expected, the unbalanced core DEV/EVAL sets are harder than their unbalanced full counterparts, though degree by which performance degrades for the core sets is less than we had expected (mean decrease of 0.87% absolute for DER and 4.6% absolute for JER) and the EVAL set is marginally harder than the DEV set.

6.2. Track 2

As seen in Table 4, for Track 2 VB-HMM is universally helpful with a much more pronounced effect than for Track 1: mean improvement of 2.24% absolute for DER and 1.92% for JER. As expected, metrics are across the board higher for Track 2, reflecting the shift from manual speech segmentation to an automatic segmentation. To better understand the interaction between the SAD and diarization components, we computed miss rate, false alarm rate, and overall error (i.e., the actual framewise error rate) for the SAD system for both sets. These results are reported in Table 2. Miss rates are quite low (below 5% across the board), but false alarm rates are much higher, especially for the EVAL set where they exceed 15%. Overall error ranges from sub 2.5% (for the DEV sets) to greater than 7% (for the core EVAL set), indicating that for Track 2 substantial gains in improvement could be attained simply by further tuning of the SAD model.

6.3. VB-HMM diarization

Recently, VB-HMM has been proposed as a method for clustering the x-vectors themselves [40] and has been used to great effect by BUT in their winning submissions to DIHARD II [39, 41]. Consequently, we also experimented with VB-HMM clustering of x-vectors, which was inserted as a step between AHC and VB-HMM resegmentation. However, we found tuning VB-HMM for x-vector clustering to be difficult and were unable to find a set of parameters for which it reliably improved on the results of AHC. In the end, we omitted this step from the challenge baseline, though we intend to continue exploring its use for future works.

6.4. Bandwidth-aware pipelines

The x-vector extractor, PLDA, UBM-GMM, and variability matrix were all trained using wideband speech data. We also experimented with separate systems for wideband and narrowband recordings with logistic classifier operating on x-vectors used to select the appropriate system for each recording. However, we observed no improvements over the baseline presented in Section 5, so opted for the system with the simpler architecture.

7. Summary

This paper provides an overview of the DIHARD III challenge as baselines for diarization and speech activity detection and results for those baselines on the challenge data.

8. References

[1] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[2] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[3] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.

[4] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proc. ICASSP*, 2016.

[5] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. ICASSP*, 2017, pp. 4930–4934.

[6] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, 2018, pp. 5239–5243.

[7] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," *Proc. ICASSP*, 2019.

[8] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Interspeech*, 2013, pp. 1477–1481.

[9] I. Viñals, A. Ortega, J. A. V. López, A. Miguel, and E. Lleida, "Domain adaptation of PLDA models in broadcast diarization by means of unsupervised speaker clustering." in *Proc. Interspeech*, 2017, pp. 2829–2833.

[10] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. ICASSP*, 2013, pp. 7746–7750.

[11] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial neural network features for speaker diarization," in *Proc. IEEE Spoken Language Technology Workshop*, 2014, pp. 402–406.

[12] R. Milner and T. Hain, "Segment-oriented evaluation of speaker diarisation performance," in *Proc. ICASSP*, 2016, pp. 5460–5464.

[13] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio." in *INTERSPEECH*, 2019, pp. 1851–1855.

[14] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "Fearless steps challenge (fs-2): Supervised learning with massive naturalistic apollo data," *arXiv preprint arXiv:2008.06764*, 2020.

[15] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[16] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," Tech. Rep., 2018. [Online]. Available: https://zenodo.org/record/1199638

[17] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD Challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[18] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolıková, O. Novotnỳ, K. Veselỳ, O. Glembek, O. Plchot *et al.*, "BUT system for DIHARD Speech Diarization Challenge 2018," in *Proc. Interspeech*, 2018, pp. 2798–2802.

[19] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone collection protocols, their uses and yields," in *Proc. EUROSPEECH*, 2003.

[20] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "Second DIHARD challenge evaluation plan," Tech. Rep., 2019. [Online]. Available: https://coml.lscp.ens.fr/dihard/2019/second_dihard_eval_plan_v1.1.pdf

[21] ——, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[22] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[23] L. Hamers *et al.*, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula." *Information Processing and Management*, vol. 25, no. 3, pp. 315–18, 1989.

[24] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity," *Systematic Biology*, vol. 45, no. 3, pp. 380–385, 1996.

[25] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[26] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs." in *Interspeech*, 2016, pp. 3434–3438.

[27] P. Singh, M. Harsha Vardhan, S. Ganapathy, and A. Kanagasundaram, "LEAP diarization system for the second DIHARD challenge," 2019, pp. 983–987.

[28] 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 *2016 IEEE Spoken Language Technology Workshop*, 2016, pp. 165–170.

[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[30] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[31] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. Audio, Speech, Language Process*, vol. 16, no. 8, pp. 1590–1601, 2008.

[32] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Proc. Odyssey*, 2018, pp. 147–154.

[33] M. Diez, L. Burget, F. Landini, and J. Černockỳ, "Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2019.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[36] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[37] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[38] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[39] 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 DIHAARD Speech Diarization Challenge," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6529–6533.

[40] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Cernockỳ, "Bayesian HMM based x-vector clustering for speaker diarization." in *INTERSPEECH*, 2019, pp. 346–350.

[41] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černockỳ, "Optimizing Bayesian MM based x-vector clustering for the Second DIHARD Speech Diarization Challenge," in *Proc. ICASSP*, 2020, pp. 6519–6523.