Speaker Clustering with Penalty Distance for Speaker Verification with Multi-Speaker Speech

Rohan Kumar Das, Jichen Yang and Haizhou Li
National University of Singapore, Singapore
E-mail: {rohankd, eleyji, haizhou.li}@nus.edu.sg

***Abstract*—Speaker verification in a multi-speaker environment is an emerging research topic. Speaker clustering, that separates multiple speakers, can be effective if a predetermined threshold or the number of speakers present in a multi-speaker utterance is given. However, the problem in practice does not provide the leverage for either of the factors. This work proposes to handle such a problem by introducing a penalty distance factor in the pipeline of traditional clustering techniques. The proposed framework first uses traditional clustering techniques to form speaker clusters for a given number of speakers. We then compute the penalty distance based on Bayesian information criterion that is used for merging alike clusters in a multi-speaker utterance. The studies are conducted on speakers in the wild (SITW) and recent NIST SRE 2018 databases that contain multi-speaker conversational speech in noisy environments. The results show the effectiveness of the proposed penalty distance based refinement in such a scenario.**

I. Introduction

Speaker verification (SV) refers to authenticate a speaker's identity claim using voice samples [1], [2]. Traditional SV approaches focus on tasks in which the test utterance contains speech from a single speaker. However, there can be an interest of application towards detecting a speaker in a multi-speaker environment [3]. In the recent years, SV in a multi-speaker noisy scenario has gained attention in the face of increasing demand from real-world applications [4]–[7]. This problem statement is closely associated with speaker diarization that deals with identifying who speaks when [8], [9], but with a different end goal, which is to detect the presence of the target speaker. In this work, we focus on such a scenario, where the speaker's identify has to be verified from multi-speaker speech.

It was reported that the false alarm rate is nearly double when there are two speakers in a test utterance [3]. The speakers in the wild database (SITW) was introduced to investigate such challenges in a multi-speaker environment [10]. It is collected in real-world scenarios that are mostly uncontrolled and contains background noise, which is depicted by the name of the database. The multi-speaker test scenario contains a single speaker for training and one or more speakers during testing [10], [11]. A similar task of SV with multi-speaker environment is also included in the recent NIST SRE 2018 challenge [12]. All these recent challenge trends show the significance of investigating SV in a multi-speaker environment.

Traditional speaker diarization systems follow the modules of voice activity detection (VAD), segmentation and speaker clustering in a pipeline [13]. In the recent DIHARD diarization challenge, the latest techniques are investigated with reference to these modules by different research groups [14]–[17]. For SV in a multi-speaker condition, the works of [18], [19] in SITW challenge evaluation showed that diarization can be useful for multi-speaker test condition. The result under such scenario is reported assuming that a specific number of speakers are present in the test trials. However, the number of speakers in a real-world application may vary. Therefore, there is a need to have different number of final speaker clusters from a multi-speaker test trial.

Agglomerative hierarchical clustering (AHC) is the most widely used method for speaker clustering [20]–[22]. This method considers each divided segment as a cluster and then the nearest clusters are combined together to form a new cluster. The process of merging the clusters involves a stopping criterion, which is based on the number of speakers in the segment or some threshold [23]. However, the estimation of a predetermined threshold is difficult for a SV task with multi-speaker speech. Furthermore, a deviation in it may lead to clustering error that can propagate in subsequent iterations [24]. Various works on diarization show that speaker linking is useful for improving the performance [25]–[27]. In the problem at hand, as the number of speakers is unknown and varies across the test trials, a clustering refinement can be useful. The penalty distance that is derived from Bayesian information criterion (BIC) has been found to be effective to estimate the speaker change points [28], [29]. We believe that the penalty distance can be used to refine the speaker clusters to find if the clusters belong to same or different speakers.

In this work, we propose a framework for SV in a multi-speaker noisy environment. A robust VAD is designed with a deep neural network (DNN) setup, followed by speech segmentation. The AHC is considered for speaker clustering using the segmented speech. The penalty distance based refinement is then performed for each multi-speaker test trial to correctly find the number of speakers. The rest of the SV framework follows the x-vector based system architecture to authenticate a trial [30]. We use the multi-speaker test conditions of SITW and NIST SRE 2018 corpora for the studies. The contribution of this work lies in the proposal of a framework using a penalty distance to refine the speaker clusters for SV in a multi-speaker environment.

The remainder of the paper is organized as follows. Section II presents the details of penalty distance based refinement for speaker clustering. Section III describes the proposed multi-speaker SV framework with speaker clustering

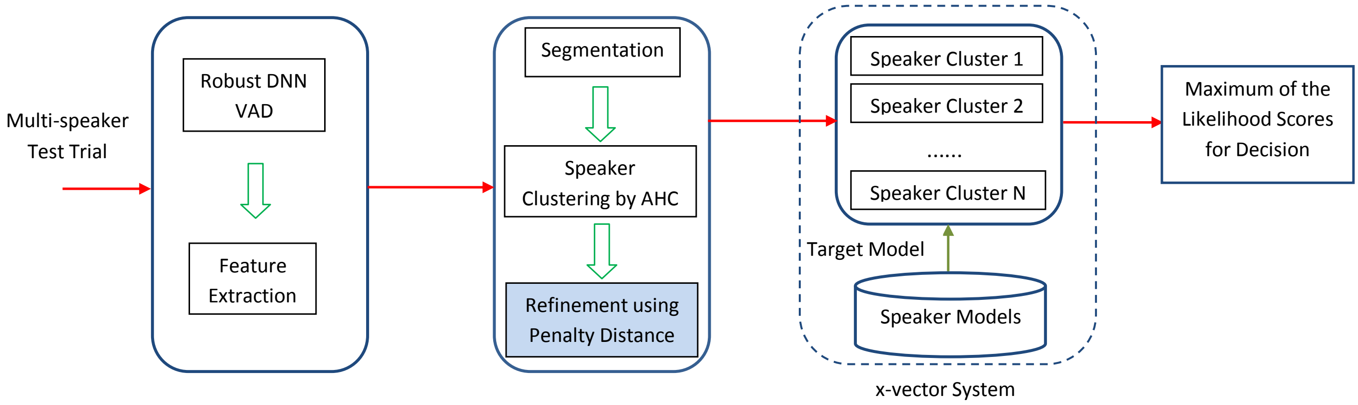

Fig. 1. Block diagram of the proposed framework for multi-speaker speaker verification with speaker clustering.

refinement. The details of the experiments is mentioned in Section IV. Section V reports the results and discussions. Finally, the paper is concluded in Section VI.

II. Speaker Clustering with Penalty Distance

As discussed in the introduction, the number of speakers in a multi-speaker test trial is unknown. The classical and popular clustering techniques like AHC requires the number of speakers in the utterance or a predetermined threshold for convergence. However, previous studies show that it is safer to use the number of speakers to minimize the clustering error during the merging process of segments as computation of threshold is difficult [24]. In the current problem at hand, considering a fixed number of speakers for all the multi-speaker test trial may lead to serious errors. Therefore, we introduce penalty distance for clustering refinement.

The BIC is proposed for the task of speaker change detection in [28]. It is represented mathematically as the following,

$$BIC = N\text{log}|\mathbf{\Sigma}| - N_1\text{log}|\mathbf{\Sigma}_1| - N_2\text{log}|\mathbf{\Sigma}_2| - \lambda P \quad (1)$$

where N_1, N_2 denote the number of frames for the two given speech segments and N stands for frame number of the merged segment. Considering all the segments follow Gaussian distribution, $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}$ represent their corresponding covariance, respectively. The λ is a penalty factor and penalty P for feature dimension d is given by,

$$P = \frac{1}{2}\big(d + \frac{1}{2}d(d+1)\big)\text{log}N \quad (2)$$

The merging of two segments depends on the BIC value. When the obtained BIC value is negative, the segments are merged, otherwise they are unaltered. For the sake of simplicity, many studies consider the penalty factor $\lambda = 1$. However, this factor can be tuned for different environments [28].

Another way of determining the penalty factor is computing it when, $BIC = 0$. We refer the penalty factor as penalty distance λ_0 for such a condition. From Equation (1) and (2) and making $BIC = 0$, we obtain the penalty distance λ_0 as,

$$\lambda_0 = \frac{4(N\log|\mathbf{\Sigma}| - N_1\text{log}|\mathbf{\Sigma}_1| - N_2\text{log}|\mathbf{\Sigma}_2|)}{(d^2+3d)\text{log}N} \quad (3)$$

This kind of penalty distance has been investigated for speaker change detection problem to obtain the speaker change points [29]. The larger the value of λ_0, the larger is the possibility of speaker change point between two speech segments. In this current work, the penalty distance is applied on the speaker clusters obtained after AHC to judge whether the clusters belong to same or different speakers. The speaker cluster refinement using this penalty distance is expected to be effective in the pipeline of verifying a multi-speaker test trial, where the number of speakers is unknown. Next, we discuss the integration of this module to the SV framework with multi-speaker speech.

III. Multi-speaker Speaker Verification with Speaker Clustering

The SV framework with multi-speaker testing in noisy environment requires three major components. The first one deals with finding the speech regions from the utterance, which has to be accurate so that it does not transfer the errors to the speaker clustering later. The importance of having a robust VAD in such scenarios is showed in [31]. In this regard, a robust DNN based VAD is implemented to handle this by using several hours of background data. The VAD is followed by segmentation of the speech. The studies of [32] showed that 0.5 to 1 second is the optimal length of a segment for speaker clustering as there is a very less chance of getting a large number of speaker change points in that small duration. The segmentation is followed by AHC to merge all similar segments for a given number of speakers in the utterance. To refine the final speaker clusters obtained through the clustering technique, the penalty distance is applied to merge alike clusters. The merging process compares penalty distance to a threshold. However, this threshold is easier to calculate using a development set, rather than the conventional AHC approach, which calculates the threshold from the test speech segment.

TABLE I
SUMMARY OF SITW AND SRE 2018 VAST CORPORA.

Database Subset	# Utterances		# Total Trials
	Enroll	Test	
SITW core-multi Dev	696	1,287	636,918
SITW core-multi Eval	1,202	2,275	2,010,683
SRE 2018 VAST Dev	10	27	270
SRE 2018 VAST Eval	101	315	31,815

Additionally, at this stage the chance of error is less as the number of final clusters is very small compared to the original 0.5 second segments of a multi-speaker test trial.

The SV systems have advanced a lot in recent decade from factor analysis approaches to the deep learning methods [30], [33]–[37]. In this work, we have used standard x-vector based architecture for speaker modeling [30]. The x-vectors corresponding to the final number of speaker clusters after penalty distance based refinement from multi-speaker speech segments are extracted. We consider that each cluster largely represents one speaker in the multi-speaker speech. Their likelihood with respect to the target speaker model is then computed and the maximum likelihood score among them is finally considered for decision. Figure 1 shows the overview of the proposed framework for SV in a multi-speaker noisy environment. Next, we describe the details of SV system with multi-speaker speech.

IV. EXPERIMENTS

This section discusses the details of the developed SV system. The database, robust VAD along with experimental setup are described in the following subsections.

A. Database

In this work, the SITW and NIST SRE 2018 corpora are considered as they have multi-speaker test condition for SV studies. The SITW corpus contains two subsets, development and evaluation set that contain 119 and 180 speakers, respectively, totaling a population of 299 speakers [10]. The speech examples of SITW database are collected in various practical noisy environments. It contains a core-multi evaluation condition that refers to single speaker for enrollment and one or more speakers in an utterance during testing. The amount of data for enrollment varies from 6 seconds to 180 seconds, whereas that for multi-speaker testing varies from 6 seconds to 10 minutes. We consider this condition for validating our ideas.

The NIST SRE 2018 database contains a subset that deals with audio from video (afV) and is referred to as video annotation for speech technology (VAST) corpus [38]. YouTube videos in various scenarios are used to extract these afV examples of the VAST corpus and they may contain multiple speakers. The speaker time marks for the enrollment utterances are provided. However, no information is provided for the test trials that makes it as the task of SV in a multi-speaker environment. The development set of this VAST corpus contains a small population of 10 speakers data, whereas the evaluation set is comparatively larger having data from 101 speakers. Table I shows the detailed composition of the SITW and VAST database used in the study.

B. Experimental setup

The databases considered for the study are collected in various noisy environments. Therefore, an effective method is required to identify the speech regions. In this regard, we implemented a robust VAD using a DNN architecture. The DNN is trained with VoxCeleb1, MUSAN and RIRS noise datasets [39]–[42]. The 39-dimensional (13-base+13-Δ+13-$\Delta\Delta$) mel frequency cepstral coefficient (MFCC) features and 3-dimensional zero crossing rate features are extracted from these databases to learn the speech and non-speech models with the DNN.

The 30-dimensional MFCC features are extracted for the speech regions obtained after robust VAD for every short-term processed frame of 25 ms. We note that the delta and double-delta coefficients are not considered in this case. For the multi-speaker test trials, we segment the speech regions into 0.5 second duration segments for speaker clustering followed by penalty distance based refinement. We obtain the threshold for speaker cluster refinement on the development set of SITW database.

The x-vector system used in this work follows the architecture in [30]. We used the wideband 16 kHz data from VoxCeleb1 and VoxCeleb2 corpora to train the x-vector extractor [39], [40]. The 512-dimensional x-vectors are then extracted with this extractor for every utterance in the enrollment set and each final speaker cluster obtained from the multi-speaker test trials. The back-end of the system considers a 150-dimensional linear discriminant analysis to reduce the dimension of the x-vectors followed by probabilistic linear discriminant analysis classifier to compute the likelihood scores against the target speakers. Finally, we consider the maximum score obtained from the speaker clusters of a multi-speaker trial as the decision to report the system performance.

The performance of the studies conducted in this work are reported in terms of equal error rate (EER), minimum cost (minC) and actual cost (actC) as per the protocols mentioned in the evaluation plan of the respective database [10], [38]. The likelihood scores are calibrated with Bosaris[1] toolkit to minimize the actual cost [43]. For the problem at hand, there is no ground truth available for the speaker change points and time marks in the test trials. Therefore, we cannot compute the performance of diarization in terms of measures like diarization error rate for the AHC output.

V. RESULTS AND DISCUSSIONS

This section describes the experimental results and analysis associated with SV in a multi-speaker environment. We first consider the core-multi condition of SITW database for the studies. A baseline system with x-vector modeling is initially developed without conducting any speaker clustering on multi-speaker test trials. Then we apply traditional AHC based

[1] https://sites.google.com/site/bosaristoolkit/

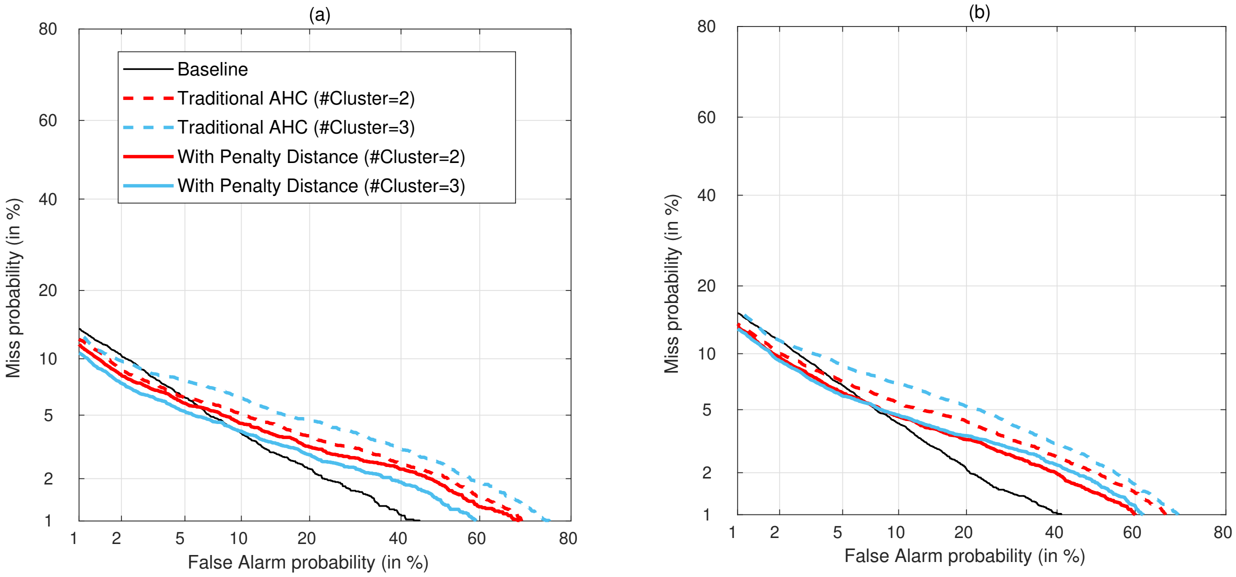

Fig. 2. DET plots for different frameworks on SITW database (a) Development Set (b) Evaluation Set.

TABLE II
PERFORMANCE FOR MULTI-SPEAKER TEST CONDITION ON SITW DATABASE.

System	Subset	EER	minC	actC
Baseline	Dev Eval	5.76 6.00	0.390 0.436	0.393 0.436
# Speaker Cluster = 2				
Traditional AHC	Dev Eval	6.02 6.53	0.405 0.433	0.434 0.455
with Penalty Distance	Dev Eval	5.62 5.86	0.376 0.401	0.386 0.401
# Speaker Cluster = 3				
Traditional AHC	Dev Eval	7.09 7.69	0.405 0.445	0.435 0.460
with Penalty Distance	Dev Eval	5.17 5.77	0.370 0.406	0.381 0.406

clustering technique to obtain speaker clusters from the test trials and follow the SV pipeline. In our experiments, we consider there are two and three speakers in the test trials for two different studies of clustering. This is based on the previous studies reported on the SITW database by different groups that showed results under assumption of different number of speakers [18], [19]. Finally, the proposed framework with speaker cluster refinement using penalty distance is implemented on those two conditions for comparison.

Table II reports the results for comparison of different SV frameworks in multi-speaker environment on SITW database. It can be observed that since the number of speakers in the test trials is unknown, assuming two or three speakers across all the trials does not help to achieve an improved result by performing speaker clustering in a traditional manner. Some of the trials can have two or more speakers and some trials may consist only a single speaker. In such cases, the amount of test data are incorrectly distributed into the clusters for verification in a multi-speaker environment. This in turn can lead to a decrease in system performance. However, on introducing penalty distance for speaker cluster refinement, gains are achieved as can be observed from Table II. The trials having fewer number of speakers than that assumed for clustering are benefited due to the refinement based on penalty distance that reflects in the result for both the conditions.

Furthermore, we note that the performance of the proposed framework is similar for assumption of speaker clusters two and three on comparing the different performance metrics. This shows a better stability and convergence with the penalty distance based refinement. Figure 2 shows the detection error tradeoff (DET) curves for different frameworks on SITW database [44]. It illustrates the gain achieved by the proposed penalty distance based cluster refinement method over the traditional way of clustering.

We then consider the VAST subset of recent NIST SRE 2018 database for SV in multi-speaker environment studies. It is to be noted that the evaluation protocols of SITW and VAST database are different to compute the associated cost [10], [38]. The prior target probability is 0.01 and 0.05 for SITW and VAST database, respectively. However, the miss and false alarm probabilities are 1.0 for both corpora. Table III shows the comparison of performance for different system frameworks for VAST database using its evaluation protocol. The VAST corpus also shows similar trend like previous study on SITW database. The performance of the proposed framework shows improved results compared to the traditional clustering that depicts the benefit gained with penalty distance based cluster refinement. We note that there are very few trials for the development set of VAST database. Therefore, the studies on that small subset may not be that conclusive. However,

TABLE III
PERFORMANCE FOR MULTI-SPEAKER TEST CONDITION ON NIST SRE 2018 VAST SUBSET.

System	Subset	EER	minC	actC
Baseline	Dev	11.11	0.383	0.527
	Eval	15.96	0.547	0.641
# Speaker Cluster = 2				
Traditional AHC	Dev	18.52	0.481	0.519
	Eval	15.87	0.550	0.647
with Penalty Distance	Dev	5.35	0.383	0.490
	Eval	14.29	0.518	0.591
# Speaker Cluster = 3				
Traditional AHC	Dev	18.52	0.481	0.519
	Eval	14.90	0.612	0.703
with Penalty Distance	Dev	7.41	0.370	0.453
	Eval	14.21	0.515	0.586

the benefit of the proposed framework can be observed by the studies on SITW and evaluation set of VAST database. The initial version of the system developed using penalty distance based speaker cluster refinement for VAST subset of NIST SRE 2018 is also submitted as a subsystem of I4U consortium [45].

VI. CONCLUSIONS

We study a framework for SV with multi-speaker speech by using penalty distance for the refinement of speaker clusters. This penalty distance is derived from BIC that is applied on top of the widely used AHC based speaker clustering method. The proposed framework with introduction of penalty distance is evaluated on SITW and recent NIST SRE 2018 VAST database. We have used x-vector system for the experimental studies. The results reveal that the proposed framework is able to work effectively for SV in a multi-speaker environment on comparing to the results obtained without penalty distance based refinement and baseline framework.

VII. ACKNOWLEDGEMENTS

This research is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and as the result of a research collaboration between the Human Language Technology Lab at NUS and Kuai Shang Tong Technology Co. Ltd. All the authors are corresponding authors of this paper.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[2] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov 2015.

[3] A. F. Martin and M. A. Przybocki, "Speaker recognition in a multi-speaker environment," in *EUROSPEECH 2001*, 2001, pp. 787–790.

[4] G. Sell and A. McCree, "Multi-speaker conversations, cross-talk, and diarization for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, March 2017, pp. 5425–5429.

[5] G. Frewat, C. Baroud, R. Sammour, A. Kassem, and M. Hamad, "Android voice recognition application with multi speaker feature," in *18th Mediterranean Electrotechnical Conference (MELECON) 2016*, 2016, pp. 1–5.

[6] Y. Wang and W. Sun, "Multi-speaker recognition in cocktail party problem," in *Communications, Signal Processing, and Systems*, Q. Liang, J. Mu, M. Jia, W. Wang, X. Feng, and B. Zhang, Eds. Singapore: Springer Singapore, 2019, pp. 2116–2123.

[7] Haris B. C., G. Pradhan, A. Misra, S. R. M. Prasanna, R. K. Das, and R. Sinha, "Multivariability speaker recognition database in indian scenario," *International Journal of Speech Technology*, vol. 15, no. 4, pp. 441–453, Dec 2012.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.

[9] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[10] M. Mclaren, L. Ferre, D. Castant, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Interspeech 2016*, 2016, pp. 818–822.

[11] ——, "The 2016 speakers in the wild speaker recognition evaluation," in *Interspeech 2016*, 2016, pp. 823–827.

[12] S. O. Sadjadi, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *Proc. Interspeech 2019*.

[13] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Fast single- and cross-show speaker diarization using binary key speaker modeling," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 23, pp. 2286–2297, 2015.

[14] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. molkov, O. Novotn, K. Vesel, O. Glembek, O. Plchot, L. Moner, and P. Matjka, "BUT system for DIHARD speech diarization challenge 2018," in *Interspeech 2018*, 2018, pp. 2798–2802.

[15] 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 2018*, 2018, pp. 2808–2812.

[16] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD challenge," in *Proc. Interspeech 2018*, 2018, pp. 2813–2817.

[17] Z. Zajc, M. Kuneov, J. Zelinka, and M. Hrz, "ZCU-NTIS speaker diarization system for the DIHARD 2018 challenge," in *Proc. Interspeech 2018*, 2018, pp. 2788–2792.

[18] O. Novotny, P. Mateka, O. Plchot, O. Glembek, L. Burget, and J. H. Cernocky, "Analysis of speaker recognition systems in realistic scenarios of the SITW 2016 challenge," in *Interspeech 2016*, 2016, pp. 828–832.

[19] Y. Liu, Y. Tian, L. He, and J. Liu, "Investigating various diarization algorithms for speaker in the wild (SITW) speaker recognition challenge," in *Interspeech 2016*, 2016, pp. 853–857.

[20] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Interspeech 2007*, 2007, pp. 1853–1856.

[21] 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 Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, Nov 2008.

[22] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, 2017, pp. 4930–4934.

[23] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[24] X. Chen, L. He, C. Xu, Y. Liu, T. Y. Liang, and J. Liu, "VB-HMM speaker diarization with enhanced and refined segment representation," in *Odyssey 2018*, 2018, pp. 134–139.

[25] M. Huijbregts and D. A. van Leeuwen, "Large-scale speaker diarization for long recordings and small collections," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 404–413, Feb 2012.

[26] P. Karanasou, M. J. F. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P. C. Woodland, and C. Zhang, "Speaker diarisation and longitudinal linking in multi-genre broadcast data," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 660–666.

[27] M. Ferrs, S. Madikeri, P. Motlicek, and H. Bourlard, "System fusion and speaker linking for longitudinal diarization of TV shows," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5495–5499.

[28] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop 1998*, 1998, pp. 127–132.

[29] J. Yang, X. Yao, and Z. Fu, "Using penalty distance for speaker change point detection," *Journal of Zhognkai university of Agriculture and Engineering*, vol. 24, pp. 32–34, 2011.

[30] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 2018, pp. 5329–5333.

[31] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Role of voice activity detection methods for the speakers in the wild challenge," in *2017 Twenty-third National Conference on Communications (NCC)*, March 2017, pp. 1–6.

[32] A. Sholokhov, T. Pekhovsky, O. Kudashev, A. Shulipa, and T. Kinnunen, "Bayesian analysis of similarity matrics for speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, 2014, pp. 106–110.

[33] 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. 788–798, 2011.

[34] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, May 2014, pp. 1695–1699.

[35] N. Kumar, R. K. Das, S. Jelil, B. K. Dhanush, H. Kashyap, K. S. R. Murty, S. Ganapathy, R. Sinha, and S. R. M. Prasanna, "IITG-Indigo system for NIST 2016 SRE challenge," in *Proc. Interspeech 2017*, 2017, pp. 2859–2863.

[36] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, 2017, pp. 999–1003.

[37] L. Xu, R. K. Das, E. Ylmaz, J. Yang, and H. Li, "Generative x-vectors for text-independent speaker verification," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018, pp. 1014–1020.

[38] "NIST 2018 Speaker Recognition Evaluation Plan," NIST, USA, 2018.

[39] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.

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

[41] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[42] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, March 2017, pp. 5220–5224.

[43] N. Brümmer and E. de Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *CoRR*, vol. abs/1304.2865, 2013.

[44] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1895–1898.

[45] K. A. Lee, V. Haütamaki, T. Kinnunen, H. Yamamoto, K. Okabe, V. Vestman, J. Huang, G. Ding, H. Sun, A. Larcher, R. K. Das, H. Li, M. Rouvier, P.-M. Bousquet, W. Rao, Q. Wang, C. Zhang, F. Bahmaninezhad, H. Delgado, and M. Todisco, "I4U submission to NIST SRE 2018: Leveraging from a decade of shared experiences," in *Proc. Interspeech 2019*, 2019.