Toeplitz Inverse Covariance based Robust Speaker Clustering for Naturalistic Audio Streams

Harishchandra Dubey, Abhijeet Sangwan, John H. L. Hansen+

Robust Speech Technologies Lab (RSTL), Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX- 75080, USA

`{Harishchandra.Dubey, Abhijeet.Sangwan, John.Hansen}@utdallas.edu`

Abstract

Speaker diarization determines *who spoke and when?* in an audio stream. In this study, we propose a model-based approach for robust speaker clustering using i-vectors. The i-vectors extracted from different segments of same speaker are correlated. We model this correlation with a Markov Random Field (MRF) network. Leveraging the advancements in MRF modeling, we used Toeplitz Inverse Covariance (TIC) matrix to represent the MRF correlation network for each speaker. This approaches captures the sequential structure of i-vectors (or equivalent speaker turns) belonging to same speaker in an audio stream. A variant of standard Expectation Maximization (EM) algorithm is adopted for deriving closed-form solution using dynamic programming (DP) and the alternating direction method of multiplier (ADMM). Our diarization system has four steps: (1) ground-truth segmentation; (2) i-vector extraction; (3) post-processing (mean subtraction, principal component analysis, and length-normalization) ; and (4) proposed speaker clustering. We employ cosine K-means and movMF speaker clustering as baseline approaches. Our evaluation data is derived from: (i) CRSS-PLTL corpus, and (ii) two meetings subset of the AMI corpus. Relative reduction in diarization error rate (DER) for CRSS-PLTL corpus is 43.22% using the proposed advancements as compared to baseline. For AMI meetings IS1000a and IS1003b, relative DER reduction is 29.37% and 9.21%, respectively.

Index Terms: Expectation maximization, i-Vector, Markov random field, Peer-led team learning, Speaker clustering, Toeplitz Inverse Covariance.

1. Introduction

Speaker diarization answers "who spoke and when?" in a multi-speaker audio stream [1]. Some of the practical applications of diarization technology include information retrieval [2], broadcast news, meeting conversations, telephone calls, VoIP, digital audio logging [3] and interaction analysis in Peer-Led Team Learning (PLTL) groups [4, 5, 6, 7]. Diarization is a challenging task for naturalistic audio streams as they contain short conversational turns, overlapped speech, noise and reverberation [8, 9]. Variable length speech segments and unsupervised clustering step are two main challenges from modeling perspective. NIST Rich Transcription (RT) evaluations involved broadcast news data and meeting recordings for diarization study [1]. Summed-channel telephone speech from NIST SRE evaluations have been used in diarization studies [4].

Most diarization systems have following five components: (i) speech activity detection (SAD) - to remove non-speech from audio [10, 11]; (ii) speaker change detection/segmentation; (iii) speaker modeling [12, 13]; (iv) speaker clustering; and (v) re-segmentation - that re-aligns each speech frame using speaker models obtained from the clustering module.

Speaker clustering is the most important component in diarization pipeline due to which, several speaker clustering approaches have been studied. These include, but not limited to, bottom-up agglomerative hierarchical clustering (AHC) [14], Cosine K-means [15], Cosine mean-shift [16], top-down approach [17], integer linear programming based method [18], Viterbi-based models [19], PLDA i-vector scoring [20], mixtures of von Mises-Fisher Distributions [21] *etc*. Recent advancements in speaker clustering still remain inadequate for naturalistic audio streams such as CRSS-PLTL corpus [6] as those do not exploit sequential correlation in speaker turns belonging to same speaker. In this study, we explore the correlation between different speech segments (i-vectors) belonging to same speaker for diarization task. We formulate speaker clustering as a structured inverse covariance estimation problem known as Toeplitz graphical lasso. Our work is motivated by recent success of such approaches in unsupervised analysis of physiological signals and real-world driving data [22]. We perform comparison of proposed Toeplitz Inverse Covariance (TIC) speaker clustering with cosine K-means and movMF baseline methods [21]. In this study, we employ i-vector features for modeling the speakers. Naturalistic audio data from CRSS-PLTL corpus is used in our experiments [23]. PLTL is a STEM education model where a peer-leader facilitate problem-solving among 6-8 students in weekly sessions [24]. We also choose two meetings subset of AMI corpus for comparative evaluations [25].

2. i-Vector Speaker Model

Diarization involve extracting i-Vectors from short speech-segments (typically 1s) unlike speaker verification where complete utterance is used. Numerous techniques exist for clustering i-Vectors using cosine distance [16]. The i-vector framework combines the speaker and channel variability sub-spaces of linear distortion model into a total-variability space represented by matrix $\mathbf{Tmat}$ [13, 26]. A speaker and channel dependent Gaussian Mixture Model (GMM) supervector, $\mathbf{S}$ is decomposed as

$$\mathbf{S} = \mathbf{S_{ubm}} + \mathbf{Tmat} \cdot \mathbf{w} \qquad (1)$$

where $\mathbf{S_{ubm}}$ is the Universal Background Model (UBM) supervector [13]. In Eqn. 1, the latent variables, $\mathbf{w}$ are low-rank factors known as i-vectors. The total-variability matrix, $\mathbf{Tmat}$ is a low-rank projection matrix that maps high-dimensional speaker supervector to a low-dimensional total-variability space [13, 26]. $\mathbf{Tmat}$ is estimated using stan-
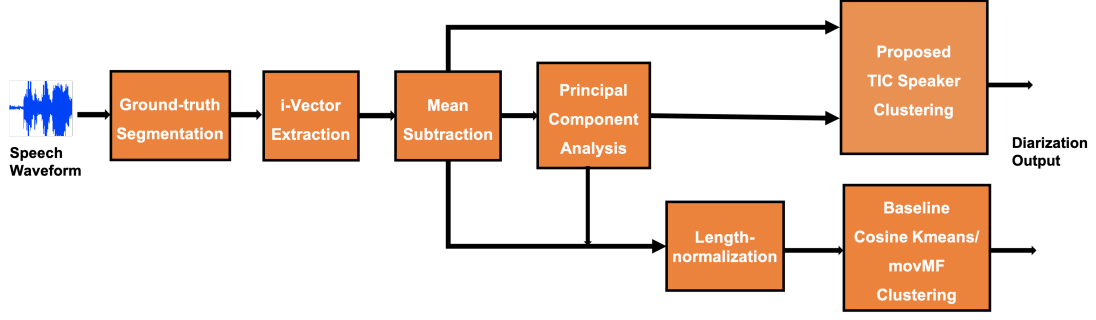
Figure 1: *Block diagram of diarization pipeline employing proposed Toeplitz Inverse Covariance (TIC)-based speaker clustering. We perform mean subtraction in all experiments. Length-normalization is required only for cosine K-means and movMF [21] baselines.*

Algorithm 1 Assign-Clusters

Input: GIVEN $\beta \geq 0$, $-LogL(i,j)$ = negative log-likelihood for i-th feature vector when it is assigned to j-th speaker cluster. K is the number of speakers. Time stamp of i-vectors (speaker features) is from 1 to T.
Output: FinalPath i.e. cluster assignment for each i-vector.
METHOD:
  1: **Initialize**
  2:     previous_cost = list of K zeros
  3:     current_cost = list of K zeros
  4:     previous_path = list of K empty lists
  5:     current_path = list of K empty lists
  6:
  7:   **for** i = 1, . . . , T **do**
  8:    **for** j = 1, . . . , K **do**
  9:     min_index = index of minimum {previous_cost}
 10:     **if** previous_cost[min_index] + β > previous_cost[j] **then**
 11:       current_cost[j] = previous_cost[j] $-LogL(i,j)$
 12:       current_path[j] = previous_path[j].append[j]
        else
 13:        current_cost[j] = previous_cost[minIndex] + β $-LogL(i,j)$
 14:       current_path[j]=previous_path[min_index].append[j]
 15:     previous_cost = current_cost
 16:     previous_path = current_path
 17:    final_min_index = index of minimum {current_cost }
 18:    FinalPath = current_path[final_min_index]
 19:    **return** FinalPath

---

dard expectation maximization (EM) algorithm on background speech data [27].

3. Proposed Speaker Clustering

Toeplitz Inverse Covariance (TIC)-based clustering was found to be suitable for segmenting the real-world time-series data such as fitness-tracker and driving data [22]. Such temporal data has complicated structure where the underlying sequences of few fixed states repeat in definitive patterns. Robust speaker clustering task possess a similar property, i.e., a small set of speakers (such as 10 for PLTL) repeat throughout the audio stream taking different conversational turns. In this section, we describe the TIC based speaker clustering followed by discussion of experimental results in next section.

In the proposed approach, each unique speaker is represented as a correlation network modeled using MRF. Such MRF network capture the interrelationship between different i-vectors (segment-level) belonging to corresponding (one)

Algorithm 2 Proposed TIC-based Speaker Clustering

Input: GIVEN: (i) Algorithm 1 for assigning i-vectors to speaker clusters; (ii) i-Vectors (features) for time 1 to T.
METHOD:
  1: **Initialize**
  2:     Speaker cluster parameters, $\boldsymbol{\Theta}$
  3:     Diarization output, spk_labels = cluster assignment set $\mathbb{C}$
  4: **Repeat**
  5:
  6:     **E-step:** Assign feature vectors to speaker clusters using Algorithm 1 i.e., map i-vectors $\Rightarrow spk\_labels$ (see section 3.1.)
  7:
  8:     **M-step:** Update speaker clusters (model) parameters by solving the Toeplitz Graphical Lasso (see section 3.2.) i.e., $spk\_labels \Rightarrow \boldsymbol{\Theta}$
  9:
 10: **Until** Convergence
 11: **return** ($\boldsymbol{\Theta}$, spk_labels)

---

speaker. Thus, each speaker is modeled by their corresponding MRF. A variant of standard expectation maximization (EM) algorithm is leveraged for estimating the MRF model for each speaker. First of all, the speaker models are initialized. Next, EM iterations run alternately between Expectation (E-step) and Maximization (M-step). During the E-step, feature vectors are assigned to speaker clusters. Next in the M-step, model parameters are updated using dynamic programming (DP) and alternating direction method of multipliers (ADMM) [22]. This process of E-step followed by M-step is repeated until convergence. Essentially, the proposed TIC approach views the temporal order of i-vectors as a sequence of states characterized by their MRF correlation network. Unlike centroid-based approaches such as cosine K-means, proposed approach models the inner structure present in the i-vector feature space. This leads to enhanced clustering performance. Each MRF models the time-invariant partial correlation structure in all feature vectors belonging to the corresponding speaker [22].

3.1. E-step: Assign feature vectors to clusters (Algorithm 1)
In this step, we fix the current value of cluster (model) parameter i.e. inverse covariance matrix, $\boldsymbol{\Theta}$ and solve the following combinatorial problem for obtaining the cluster assignment set, $\mathbb{C} = \{C_1, C_2, \cdots, C_K\}$ given by:

$$\text{minimize} \sum_{i=1}^{K} \sum_{X_t \in C_i} N(X_t, \Theta_i) + \beta 1_{X_{t-1} \notin C_i} \quad (2)$$

The Eqn. 2 assign each of the T i-vectors to one of the K speaker clusters by jointly maximizing both the log likelihood and temporal smoothness. Temporal smoothness parameter β ensures neighboring blocks are assigned to the same speaker. Thus, β is a regularization parameter (switching penalty) that controls the trade-off between two objectives. For large value of $\beta \to \infty$, all speaker features are grouped together in a single cluster.

Given the model parameters (i.e., inverse covariance matrices) Θ_i's for each of the K speaker clusters, we solve the optimization problem in Eqn. 2 that assigns the T i-vectors, X_1, X_2,....., X_T to K clusters in a way that maximizes the likelihood of data while minimizing the number of times cluster assignment changes across time. Given K potential cluster assignments for each of the T data points, this combinatorial optimization problem has KT possible mappings of feature vectors to speaker clusters. We assign each i-vector to a unique cluster using the approach presented in the Algorithm 1. This method is equivalent to finding the Viterbi path with minimum cost for the feature sequence of length T.

3.2. M-step: Toeplitz Graphical Lasso (Algorithm 2)

Each speaker cluster is modeled as a Gaussian inverse covariance matrix, $\Theta_i \in \mathbb{R}^{nb \times nb}$ where nb is the feature dimension. Since the inverse covariance matrix show the conditional independence structure between the variables, Θ_i defines a Markov Random Field (MRF) that encodes the structural representation present in all feature vectors of the i-th speaker cluster. Sparse graphical representations prevent overfitting in addition to fetching interpretable results [28]. In the M-step of EM algorithm, our objective is to estimate the K inverse covariance matrices, $\mathbf{\Theta} = \{\Theta_1, \Theta_2, \cdots, \Theta_K\}$ using the cluster assignment sets, $\mathbb{C} = \{C_1, C_2, \cdots, C_K\}$, where $C_i \subset \{1, 2, ..., T\}$ obtained from the previous E-step. We can estimate each Θ_i in parallel thus saving execution time. The negative log likelihood of feature vector, X_t to be assigned for i-th cluster with model parameter Θ_i is denoted as $N(X_t, \Theta_i)$. We can write it in terms of each Θ_i as follows [22]:

$$\sum_{X_t \in C_i} N(X_t, \Theta_i) = -|C_i|(\log \det(\Theta_i) + \operatorname{tr}(S_i \Theta_i)) + \gamma \quad (3)$$

where $|C_i|$ is the number of feature vectors assigned to the i-th cluster, S_i is the empirical covariance of these feature vectors, and γ is a constant independent of Θ_i. Now, the M-step of EM algorithm becomes

$$\text{minimize} - \log \det(\Theta_i) + \operatorname{tr}(S_i \Theta_i) + \frac{1}{|C_i|}||\lambda \circ \Theta_i||_1$$

$$\text{subject to } \Theta_i \in \mathbb{T} \quad (4)$$

where $\mathbb{T}$ is a set of blockwise Toeplitz matrix. Eqn. 4 represents a convex optimization problem known as Toeplitz graphical lasso [22]. It is a type of graphical lasso problem with additional constraint of block Toeplitz structure on the inverse covariance matrices. Here, λ is a regularization parameter matrix for handling the trade-off between two objectives: (i) maximizing the log likelihood, and (ii) ensuring Θ_i to be sparse. For invertible matrix S_i, likelihood objective lead Θ_i to be near S_i^{-1}. The inverse covariance matrix, Θ_i is constrained to be block Toeplitz and λ is a $nb \times nb$ matrix so that it can regularize each sub-block of Θ_i differently. A separate Toeplitz graphical lasso is solved for each speaker cluster at every iteration of the EM algorithm. Since we require several iterations of the EM algorithm before speaker clustering converges, a fast and efficient
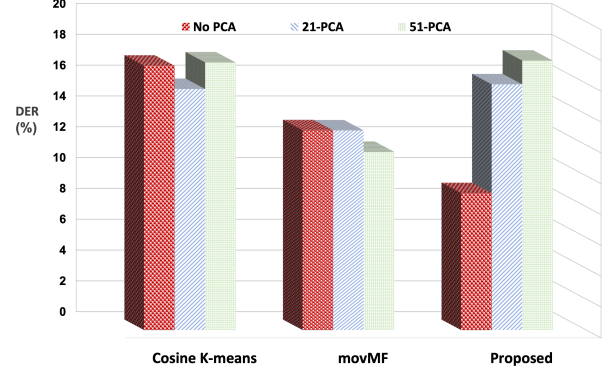

Figure 2: *PLTL results: DER (%) for proposed and baseline methods. No PCA, 21-PCA and 51-PCA represent cases where no dimensional reduction is performed, where 21 principal components and 51 principal components are chosen after PCA, respectively. Proposed approach achieves significant reduction in DER as compared to both baselines.*
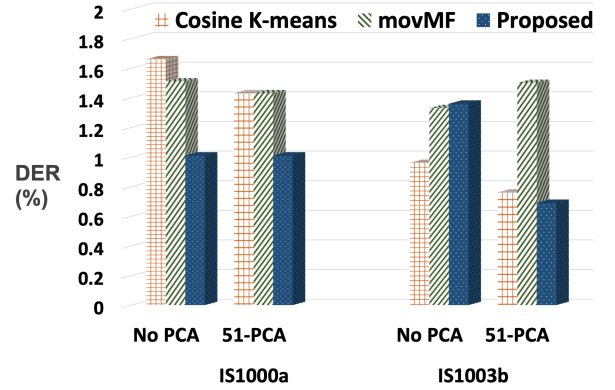

Figure 3: *AMI results: DER (%) for two meetings namely IS1000a and IS1003b. No PCA and 51-PCA represent cases where no dimension reduction is performed, and 51 principal components are chosen after PCA, respectively. Proposed speaker clustering gives less than 1% DER for both IS1000a and IS1003b (best case).*

approach based on alternating direction method of multipliers (ADMM) is employed for this purpose [22]. For more details on solving this Toeplitz graphical lasso, please refer to Section 4.2 in [22].

4. Experiments, Results & Discussions

In this study, we focus on speaker clustering and hence ground-truth segmentation information is adopted for all experiments to avoid errors from SAD and speaker segmentation steps (see Fig. 1). We conduct evaluations on: (i) CRSS-PLTL, and (ii) two meetings subset of AMI corpus, as detailed below.

4.1. CRSS-PLTL Eval Set

It contain audio recordings of 80 minute PLTL sessions that are held once per week. PLTL is a student-led STEM education paradigm popular in US universities for undergraduate courses. Each session has 6-8 students plus a peer-leader who direct the discussions for collaborative problem solving among students [29]. CRSS-PLTL corpus [6] is a collection of audio recordings from five groups studying an undergraduate chem-

istry course over 11 weeks. Each participant in PLTL sessions wear a LENA device that collects the naturalistic audio [30]. We choose the channel corresponding to peer-leader for single-stream diarization studies reported in this paper. Our evaluation data consists of 80 minute audio with 8 speakers. We remove the overlapped speech, laughter, cough *etc.* from audio signal before passing it through our diarization system. Majority of the conversational turns are short (1-1.5s) for this data. Human annotators labeled the ground-truth speaker identity with start and end time stamps.

4.2. AMI Meeting Eval Set

Augmented Multi-party Interaction (AMI) corpus contain audio and video data recorded from meetings. Our AMI evaluation set consists of *headset audio* of two meetings namely `IS1000a` (26 min.) and `IS1003b` (27 min.) from the popular 12-meetings subset of AMI corpus [25, 31]. Both meetings involve four speakers discussing the design of a new remote control device.

4.3. Evaluations & Baseline Systems

Fig. 1 illustrate our diarization pipeline. We use frame-level 20-MFCC features extracted from 40ms windows with 10ms skip-rate. A UBM with 512 components is trained for i-vector extraction [13]. Given the short conversational-turns in PLTL, we train an i-vector extractor of dimension 75. For all experiments in this paper, we perform mean subtraction on i-vectors where the mean is computed from entire PLTL session/AMI meeting (see Fig. 1). Mean subtraction is followed by principal component analysis (PCA) for de-correlation and dimension reduction. We do PCA based dimension reduction to compare the performance over 21 and 51 components.

We recently proposed Mixtures of von Mises-Fisher Distributions (movMF) based speaker clustering for naturalistic audio data [21]. In this study, we choose cosine K-means [32] and movMF speaker clustering [21] as our baseline approaches. Unlike conventional K-means, cosine K-means projects the estimated cluster centroids onto unit hypersphere at the end of each M-step in the EM algorithm. Cosine K-means is a special case of movMF speaker clustering as discussed in [21]. When all mixture weights are equal and all concentration parameters are equal, the movMF speaker clustering becomes equivalent to cosine K-means. Feature vectors undergo length-normalization [33] before passing through cosine K-means [32] and movMF [21] baselines as these methods expect data lying on the unit hypersphere. On the other hand, length-normalization is optional for proposed TIC speaker clustering. We input either the mean-subtracted i-vectors or PCA processed i-vectors to the proposed TIC speaker clustering.

4.4. Results & Discussions

Diarization error rate (DER) is used for comparing the performance of different clustering approaches in this paper [34]. Mathematically, it is defined as:

$$DER = \frac{\alpha_{fa} + \alpha_{miss} + \alpha_{err}}{\alpha_{total}} \quad (5)$$

where α_{total} is the total scored time from reference annotations, α_{fa} is scored time for which a non-speech region is incorrectly marked as containing speech, α_{miss} is scored time for which a speech region is incorrectly marked as not containing speech, and α_{err} is the scored time for which a wrong speaker identity

is assigned within a speech region. In this study, we focus on speaker clustering and hence ground-truth segmentation information was adopted in all experiments. Thus, α_{fa} and α_{miss} are zero. Here, DER is only dependent on speaker confusion error, α_{err}. Overlapped speech, laughter and other miscellaneous vocalizations such as cough is removed from the audio signal before feeding into our diarization pipeline (see Fig. 1).

The proposed speaker clustering approach requires $\mathcal{O}(KT)$ time for grouping T i-vectors into K clusters. Hence, it is a scalable approach. Fig. 2 shows the DER comparison between proposed and baseline methods with various post-processing for PLTL data. Clearly, DER is reduced to almost half for proposed approach as compared to cosine K-means. The proposed method learns a correlation network modeled as MRF and hence it expects correlated data as input. After PCA, feature vectors are de-correlated and estimating the underlying correlation model using MRF is not accurate in this case. Consequently, as expected PCA degrades the performance of proposed approach on naturalistic PLTL data. The movMF is state-of-the-art for speaker clustering on PLTL data [21]. The proposed approach lead to 23.18% relative reduction in DER as compared to movMF speaker clustering. Significant DER reduction validate the efficacy of proposed Toeplitz Inverse Covariance (TIC)-based speaker clustering for naturalistic audio such as CRSS-PLTL.

Fig. 3 illustrate the DER comparison between proposed approach and baselines for two meetings subset of AMI corpus. Clearly, the best DER for both meetings are obtained by proposed approach. For both meetings, proposed approach fetch less than 1% DER showing its efficacy for real-world meeting data. For AMI meetings IS1000a and IS1003b, relative DER reduction is 29.37% and 9.21%, respectively. CRSS-PLTL data has higher levels and more varied forms of distortions as compared to AMI meeting data. Also, there are 8 speakers in PLTL data as compared to 4 in AMI meetings. These differences cause the DER to be in different ball park for both corpora.

5. Summary & Conclusions

In this paper, we leveraged the Toeplitz Inverse Covariance (TIC) estimation in speaker clustering task for naturalistic audio such as CRSS-PLTL corpus. Such audio data are rich in noise, overlapped-speech and reverberation in addition to short conversational turns (1s). The proposed approach accurately models the inner structure in i-vector features belonging to same speaker. It leads to enhanced clustering that eventually improves diarization performance. Each speaker cluster is characterized by a Markov Random Field (MRF) that captures its correlation structure. For robust speaker clustering task, the proposed approach iterates with alternating two steps of the standard Expectation Maximization (EM) algorithm: (i) **E-step** that assigns each speaker feature to a unique cluster; and (ii) **M-step** that update the model parameters using speaker clusters obtained in the E-step. We performed experiments on (i) CRSS-PLTL; and (iii) AMI meeting data. Cosine K-means and movMF were chosen as baseline clustering methods. Proposed TIC based speaker clustering fetch 43.22% relative reduction in DER for CRSS-PLTL data. AMI evaluation set has 29.37% (IS1000a) and 9.21% (IS1003b) relative reduction in DER over baseline. Significant improvements in DER show the efficacy of proposed speaker clustering for naturalistic audio streams.

6. References

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

[2] M. Huijbregts, *Segmentation, diarization and speech transcription: surprise data unraveled.* Citeseer, 2008.

[3] J. H. L. Hansen, A. Sangwan, A. Ziaei, H. Dubey, L. Kaushik, and C. Yu, "Prof-Life-Log: Monitoring and assessment of human speech and acoustics using daily naturalistic audio streams," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3010–3010, 2016.

[4] R. Aloni-Lavi, I. Opher, and I. Lapidot, "Incremental on-line clustering of speakers' short segments," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 120–127.

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

[6] H. Dubey, A. Sangwan, and J. H. L. Hansen, "Using speech technology for quantifying behavioral characteristics in peer-led team learning sessions," *Computer Speech & Language*, vol. 46, pp. 343–366, 2017.

[7] ——, "A robust diarization system for measuring dominance in peer-led team learning groups," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 319–323.

[8] J. H. L. Hansen, J. Alberte, N. Jones, H. Dubey, and A. Sangwan, "Multi-stream audio analysis for knowledge extraction and understanding of small-group interactions in peer-led team learning," *Seventh Annual Conference Peer-Led Team Learning International Society, the University of Texas at Dallas, Richardson, TX, USA*, pp. 1–1, 2018.

[9] J. H. L. Hansen, H. Dubey, and A. Sangwan, "CRSS-LDNN Long-duration naturalistic noise corpus containing multi-layer noise recordings for robust speech processing," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1797–1797, 2018.

[10] H. Dubey, A. Sangwan, and J. H. L. Hansen, "Leveraging Frequency-Dependent Kernel and DIP-based Clustering for Robust Speech Activity Detection in Naturalistic Audio Streams," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 26, no. 11, pp. 2056–2071, 2018.

[11] ——, "Robust feature clustering for unsupervised speech activity detection," in *IEEE ICASSP*, 2018, pp. 2726–2730.

[12] ——, "Transfer learning using raw waveform sincnet for robust speaker diarization," in *IEEE ICASSP, Brighton, UK.*, 2019.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] H. Sun, B. Ma, S. Z. K. Khine, and H. Li, "Speaker diarization system for RT07 and RT09 meeting room audio," in *IEEE ICASSP*, 2010, pp. 4982–4985.

[15] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *IEEE ICASSP*, 2008, pp. 4133–4136.

[16] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 217–227, 2014.

[17] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

[18] G. Dupuy, S. Meignier, P. Deléglise, and Y. Estéve, "Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization," in *ISCA Odyssey*, 2014, pp. 187–193.

[19] I. Lapidot, A. Shoa, T. Furmanov, L. Aminov, A. Moyal, and J.-F. Bonastre, "Generalized viterbi-based models for time-series segmentation and clustering applied to speaker diarization," *Computer Speech & Language*, vol. 45, pp. 1–20, 2017.

[20] I. Salmun, I. Opher, and I. Lapidot, "On the use of PLDA i-vector scoring for clustering short segments," in *Proc. Odyssey*, 2016.

[21] H. Dubey, A. Sangwan, and J. H. L. Hansen, "Robust speaker clustering using mixtures of von mises-fisher distributions for naturalistic audio streams," in *ISCA INTERSPEECH*, 2018, pp. 3603–3607.

[22] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2017, pp. 215–223.

[23] H. Dubey, L. Kaushik, A. Sangwan, and J. H. L. Hansen, "A speaker diarization system for studying peer-led team learning groups," in *ISCA INTERSPEECH*, 2016, pp. 2180–2184.

[24] J. H. L. Hansen, H. Dubey, A. Sangwan, L. Kaushik, and V. Kothapally, "UTDallas-PLTL: Advancing multi-stream speech processing for interaction assessment in peer-led team learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1869–1869, 2018.

[25] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[26] 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, 2015.

[27] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[28] S. L. Lauritzen, *Graphical models.* Clarendon Press, 1996, vol. 17.

[29] K. Carlson, D. Turvold Celotta, E. Curran, M. Marcus, and M. Loe, "Assessing the Impact of a Multi-Disciplinary Peer-Led-Team Learning Program on Undergraduate STEM Education," *Journal of University Teaching & Learning Practice*, vol. 13, no. 1, p. 5, 2016.

[30] A. Sangwan, J. H. L. Hansen, D. W. Irvin, S. Crutchfield, and C. R. Greenwood, "Studying the relationship between physical and language environments of children: Who's speaking to whom and where?" in *IEEE Signal Proc.Education Workshop, Salt Lake City, Utah*, 2015, pp. 49–54.

[31] E. Gonina, G. Friedland, H. Cook, and K. Keutzer, "Fast speaker diarization using a high-level scripting language," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 553–558.

[32] S. Zhong, "Efficient online spherical k-means clustering," in *IEEE International Joint Conference on Neural Networks*, vol. 5, 2005, pp. 3180–3185.

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

[34] "NIST DER script for RT evaluations," (Date last accessed 9-Jan-2018). [Online]. Available: as part of the Speech Recognition Scoring Toolkit (SCTK): ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.10-20151007-1312Z.tar.bz2