LEAP Diarization System for the Second DIHARD Challenge

Prachi Singh[1], *Harsha Vardhan M A*[1], *Sriram Ganapathy*[1], *Ahilan Kanagasundaram*[2]

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science.
[2]University of Jaffna, Sri Lanka.

{prachisingh,sriramg}@iisc.ac.in, harshavardhan.ma@gmail.com,ahilan@eng.jfn.ac.lk

Abstract

This paper presents the LEAP System, developed for the Second DIHARD diarization Challenge. The evaluation data in the challenge is composed of multi-talker speech in restaurants, doctor-patient conversations, child language acquisition recordings in home environments and audio extracted YouTube videos. The LEAP system is developed using two types of embeddings, one based on i-vector representations and the other one based on x-vector representations. The initial diarization output obtained using agglomerative hierarchical clustering (AHC) done on the probabilistic linear discriminant analysis (PLDA) scores is refined using the Variational-Bayes hidden Markov model (VB-HMM) model. We propose a modified VB-HMM model with posterior scaling which provides significant improvements in the final diarization error rate (DER). We also use a domain compensation on the i-vector features to reduce the mis-match between training and evaluation conditions. Using the proposed approaches, we obtain relative improvements in DER of about 7.1% relative for the best individual system over the DIHARD baseline system and about 13.7% relative for the final system combination on evaluation set. An analysis performed using the proposed posterior scaling method shows that scaling results in improved discrimination among the HMM states in the VB-HMM.

Index Terms: Speaker Diarization, i-vector, x-vector, HMM-VB, PLDA.

1. Introduction

Speaker diarization, the task of identifying who spoke when in a multi-talker speech recording, is receiving increased attention in the recent years. It has several potential applications such as surveillance, forensics, information retrieval, rich transcription for automatic speech recognition (ASR) systems and in call center applications. The task of speaker diarization is challenging in noisy and channel degraded environments where speech is corrupted by background noise. The variations in domain/speaking style of the speech also affect the diarization performance [1].

In the previous decade, NIST had performed a series of evaluations in the topic of speaker diarization on meeting room conditions [2]. The diarization tasks on other domains like broadcast news were also pursued [3]. However, most of these diarization system development efforts focused on systems that were aimed to operate on isolated domains. The first among the series of recent initiatives to benchmark diarization systems, the first DIHARD challenge, proposed a challenge where the performance of a diarization system was evaluated on a range of complex realistic operational scenarios. The foundational work for this evaluation came up from an analysis of the challenging scenarios for state-of-art diarization systems [4] from doctor-patient conversations, child language acquisition data [5], meeting speech, dinner party conversations in restaurants etc. In addition, the more comfortable evaluation criterion which simply ignored overlapping speech regions and had a liberal collar were updated for the DIHARD challenge with an error evaluation on overlapping regions without any collar. This initiative has been revamped with more challenging recordings in the second DIHARD challenge [6]. This paper describes the speaker diarization system development efforts by the LEAP team.

For the first DIHARD challenge, many of the successful systems used the neural network based speech embedding (x-vector) representation [7]. The x-vector representations replaced the previously employed GMM based i-vector features and were extracted for fixed length chunks of duration 1.5sec. The pairwise scores on segment x-vectors are computed using a probabilistic linear discriminant analysis (PLDA) scoring [8]. The PLDA score matrix is clustered with an agglomerative hierarchical clustering (AHC) [9]. The AHC clusters are further refined using a Variational-Bayes hidden Markov model (VB-HMM) to improve the diarization performance [10].

In the DIHARD-I challenge, a previous effort had explored the use of i-vectors for estimating the number of speakers [11]. Another approach used a neural network based domain classifier and optimized the diarization system on each domain separately [12]. The speaker diarization approach using binary shift keying was explored by [13]. The x-vector approach with HMM-VB refinement and the fusion with i-vector representations was attempted in [14].

In this paper, our major contributions are,

- Posterior scaling for VB-HMM - In this approach, we boost the zeroth order statistics before the VB-HMM likelihood computation. This posterior scaling improves the speaker separation in the VB-HMM model which results in significant reduction in the overall DER.
- Domain compensation - The i-vector embeddings have a mismatch between training conditions and the DIHARD development dataset which can be compensated using variance normalization.

The rest of the paper is organized as follows. Section 2 introduces the dataset used in training and testing the models. Section 3 describes the x-vector baseline system. In Section 4, we describe the proposed approaches which improves over baseline including posterior scaled HMM-VB model (section 4.1) and i-vector domain mismatch compensation (section 4.2) for speaker diarization. Section 5 describes the systems implemented by LEAP team using proposed approaches. In Section 6, we report the experiments and results on the DIHARD challenge. This is followed by a summary of the work in Section 7.

2. Dataset

DIHARD II track 1 single channel development dataset, which is a superset of DIHARD I development dataset is garnered

from diverse sources such as monologues, map task dialogues, broadcast interviews, sociolinguistic interviews, meeting speech, speech in restaurants, clinical recordings, extended child language acquisition recordings from LENA vests, and YouTube videos as mentioned in [15]. The training data for embedding extraction and pairwise scoring models are VoxCeleb-1 [16] and VoxCeleb-2 [17] datasets. These datasets jointly possess around a million utterances of speaker annotated, single speaker audio files, amounting to around 2000 hours of audio in total.

3. Baseline System

The baseline system for track 1 is inspired by the JHU's Kaldi recipe [18].

Feature Extraction: It involves extraction of 24 dimensional Mel-Frequency Cepstral Coefficients(MFCCs) with delta and double delta appended making feature dimension 72 for the i-vector system [19] and 30 dimensional MFCCs alone for x-vector system [7]. A sliding mean normalization was applied over a $3s$ window.

Segment representation: The speech segments ($1.5s$ with $0.75s$ shift) are converted to 400 dimensional i-vectors or 512 dimensional x-vectors.

PLDA training and scoring: Probabilistic Linear Discriminant Analysis (PLDA) is used to model speaker and channel variability space. To adapt the PLDA matrix for DIHARD dataset. PCA transformation trained on DIHARD development dataset is applied to the training set followed by length normalization. An utterance level PCA is applied before PLDA scoring for dimensionality reduction [20].

Agglomerative Hierarchical Clustering (AHC) : The AHC hierarchically clusters the segments based on speaker similarity scores (PLDA scores) and merges the clusters that represent the same speaker identity. The AHC stopping criterion is determined using DIHARD development data.

4. Proposed Methods

We propose two new methods which provides significant improvement on baseline. First approach is a modification of VB-HMM model [10] in which we scale the $zeroth$ order statistics obtained from GMM posteriors, which enhances the emission probability and thereby helps to make the HMM state posteriors more discriminative. Another approach involves domain normalization, applied to i-vector features as introduced in 1.These approaches are described below:

4.1. Posterior Scaled VB-HMM

4.1.1. Variational Bayes speaker diarization

As proposed in [21, 10], the VB-HMM model is a Hidden Markov Model with eigenvoice priors. Each string of states of the HMM represents a speaker in an utterance and transitions between states correspond to speaker turns. To avoid frequent speaker turns, each speaker can be constrained to have minimum number of states. The HMM's speaker specific state is modeled from Gaussian Mixtures Model (GMM) distribution adapted from a Universal Background Model (UBM-GMM) with eigenvoice prior similar to i-vector model [19]. This constrains the GMM mean parameters to remain in a lower-dimensional subspace. All the speaker independent UBM-GMM model parameters with C mixtures like super-vector means $\boldsymbol{\mu}^{ubm}$, covariance $\boldsymbol{\Sigma}$, component weights w^{ubm} and total variability matrix $\boldsymbol{V}$ are pre-trained. Only $\boldsymbol{\mu}_s$, the super-vector of concatenated Gaussian component means for speaker s are speaker specific. The high dimensional super-vectors are adapted using the i-vector model approach in which the shift in the mean is captured by speaker factor $\boldsymbol{y}_s$.

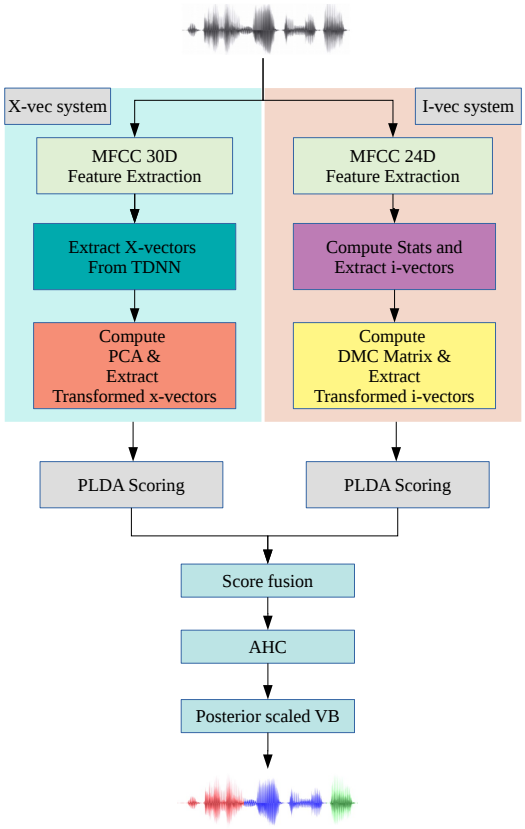

Figure 1: *Block schematic of the i-vector and x-vector systems. Pipleline explained for DIHARD development set.*

$$\boldsymbol{\mu}_s = \boldsymbol{\mu}^{ubm} + \boldsymbol{V}\boldsymbol{y}_s \tag{1}$$

For each given conversation recording, a HMM is constructed using some initial assignment of frames to the speaker states where number of speaker states will be an upper bound on the possible number of speakers in that recording. Then, each iteration of VB training will refine the HMM state specific distributions and re-segment the frames based on the posteriors obtained after the forward-backward algorithm.

4.1.2. VB-HMM with posteriors scaling

VB-HMM model gives frame level resolution but it also leads to frequent speaker turns even with the minimum duration constraint in the HMM. In order to supress this effect, one can operate on 200-500ms duration. In this work, we propose to scale the $zeroth$ order statistics in the VB-HMM modeling.

- **Segment Representation and computation of statistics with scaling:**

After feature extraction and removal of non-speech regions using oracle Speech Activity Detection (SAD), frames are uniformly segmented into non-overlapping M segments represented as $X = x_1, x_2, \ldots, x_M$ where each segment x_m will have T frames. Using the UBM-GMM we compute $zeroth$,

$first$ and $second$ order statistics using the scaled posteriors as follows:

$$\hat{\boldsymbol{N}}_{mc} = \sum_t \beta \zeta_{tc}^{(m)} \quad \boldsymbol{F}_{mc} = \sum_t \zeta_{tc}^{(m)} \left(\boldsymbol{x}_{mt} - \boldsymbol{\mu}_c^{ubm} \right) \tag{2}$$

$$\boldsymbol{S}_{mc} = diag \left(\sum_t \zeta_{tc}^{(m)} \left(\boldsymbol{x}_{mt} - \boldsymbol{\mu}_c^{ubm} \right) \left(\boldsymbol{x}_{mt} - \boldsymbol{\mu}_c^{ubm} \right)^{\intercal} \right)$$

where, $\zeta_{tc}^{(m)}$ represents posterior probability of c^{th} mixture given the t^{th} frame of segment x_m and x_{mt} is the t^{th} frame of the m^{th} segment. The scaling factor β introduced here allows to increase weightage of the posterior in the zeroth order statistics. This scaling factor is later shown to be effective in improving the discriminability of HMM state posteriors.

- **Segment enhancement:**

As proposed in [22], to further smooth the segment representation with its neighbors, we computed weighted average of the statistics around its neighbourhood as follows:

$$\begin{aligned} \bar{\boldsymbol{N}}_m &= \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) \hat{\boldsymbol{N}}_{m+\Delta m} \\ \bar{\boldsymbol{F}}_m &= \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) \boldsymbol{F}_{m+\Delta m} \\ \bar{\boldsymbol{S}}_m &= \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) \boldsymbol{S}_{m+\Delta m} \end{aligned} \tag{3}$$

where $P(\Delta m) = e^{-\lambda|\Delta m|}$. For the segment enhancement, we use $\lambda = 0.8$ and $\Delta M = 1$.

- **Update speaker factor y_s:**

The approximate posterior distribution $p(y_s|x_m)$ is Gaussian with mean $\bar{\alpha}_s$ and precision matrix $\bar{L}_s$ given as:

$$\bar{\boldsymbol{L}}_s = \boldsymbol{I} + \boldsymbol{V}^{\intercal} \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{N}}(s) \boldsymbol{V} \ , \ \bar{\boldsymbol{\alpha}}_s = \bar{\boldsymbol{L}}_s^{-1} \boldsymbol{V}^{\intercal} \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{F}}(s) \tag{4}$$

where $\boldsymbol{\Sigma}$ is a block diagonal covariance matrix. $\bar{\boldsymbol{N}}(s)$ and $\bar{\boldsymbol{F}}(s)$ are the speaker dependent Baum-Welch statistics, which are obtained by taking the segment assignment probability q_{ms} into consideration. Speaker factor y_s is the MAP estimate of Gaussian distribution which is the mean of Gaussian distribution $\bar{\alpha}_s$. Scaling $zeroth$ order statistics scales down the $\bar{\alpha}_s$.

$$\bar{\boldsymbol{N}}(s) = \sum_{m=1}^{M} q_{ms}^{\tau} \bar{\boldsymbol{N}}_m \ , \ \bar{\boldsymbol{F}}(s) = \sum_{m=1}^{M} q_{ms}^{\tau} \bar{\boldsymbol{F}}_m \tag{5}$$

where q_{ms}^{τ} is the posterior probability of speaker state s given segment x_m at τ^{th} iteration and q_{ms}^{0} denote the initial posterior probability.

- **Update emission probability $p(x_m|y_s)$:**

Emission probability is given by, $\ln p(x_m|y_s) = \bar{G}_m + \bar{H}_{ms}$ where, $\boldsymbol{N}_m = \hat{\boldsymbol{N}}_m$ with $\beta = 1$ in

$$\begin{aligned} G_m &= \sum_{c=1}^{C} N_m \ln \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} - \frac{1}{2} tr(\Sigma^{-1} S_m) \\ \bar{G}_m &= \sum_{\Delta m=-\Delta M}^{\Delta M} P(\Delta m) G_{m+\Delta m} \\ \bar{H}_{ms} &= \bar{\boldsymbol{\alpha}}_s^{\intercal} \boldsymbol{V}^{\intercal} \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{F}}_m - \frac{1}{2} tr(\boldsymbol{V}^{\intercal} \bar{\boldsymbol{N}}_m \boldsymbol{\Sigma}^{-1} V [\boldsymbol{L}_s^{-1} + \bar{\boldsymbol{\alpha}}_s \bar{\boldsymbol{\alpha}}_s^{\intercal}]) \end{aligned} \tag{6}$$

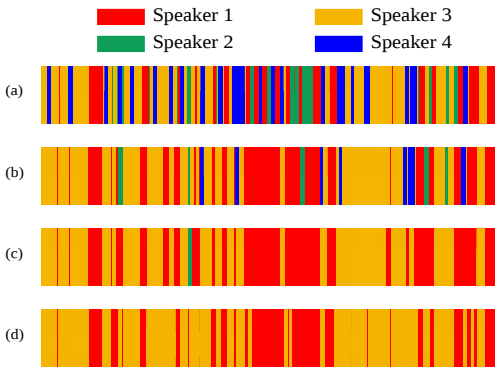

Figure 2: *VB-HMM diarization output of one DIHARD file from VAST domain and the ground truth segmentation. Plots (a),(b),(c) for $\beta = 1$, $\beta = 8$, $\beta = 24$ respectively, while (d) is the ground truth segmentation.*

Here we are using enhanced $first$ and $second$ order statistics but scaled and enhanced $zeroth$ order statistics as given in equation (3). The transition probabilities and speaker specific state probabilities q_{ms} are updated as mentioned in [10] using the forward-backward algorithm which outputs posterior probabilities $(q_{ms}^{\tau+1})$ for each state and each segment.After the convergence, speaker labels are assigned per segment by taking argmax$_s$ q_{ms} Figure 2. shows the effect of scaling factor on the output frame labels as we increase from $\beta = 1$ to $\beta = 24$. Setting the scaling factor is a tradeoff between the frequency of speaker turns and the number of speakers retained in the final diarization output. We observe that on increasing the value of β, lesser number of speakers are retained in the output while reducing the value of β results in spurious speaker turns.

4.2. I-vector system with domain mismatch compensation

In baseline system, PLDA parameters are trained using out-of-domain train data (VoxCeleb1 and VoxCeleb2) which are single speaker recordings with different recording environment then the DIHARD dataset. We propose here domain mismatch compensation(DMC) method which has been found to be useful in speaker verification task for short utterances [23]. We extract 400 dimensional i-vectors of $3s$ duration from training set and i-vectors of $1.5s$ with $0.75s$ shift from the DIHARD Development set. We compute Domain mismatch variance (Q) as follows:

$$\begin{aligned} \boldsymbol{Q} = &\frac{1}{N_t} \sum_{i=1}^{N_t} (\boldsymbol{y}_t^{(i)} - \langle \boldsymbol{y}_d \rangle)((\boldsymbol{y}_t^{(i)} - \langle \boldsymbol{y}_d \rangle)^T \\ &+ \frac{1}{N_d} \sum_{i=1}^{N_d} ((\boldsymbol{y}_d^{(i)} - \langle \boldsymbol{y}_t \rangle)(\boldsymbol{y}_d^{(i)} - \langle \boldsymbol{y}_t \rangle)^T \end{aligned} \tag{7}$$

where $\boldsymbol{y}_t^{(i)}$, $\boldsymbol{y}_d^{(i)}$ are out-domain (Voxceleb) and in-domain (DIHARD dev) i-vectors respectively. $\langle \boldsymbol{y}_d \rangle$ and $\langle \boldsymbol{y}_t \rangle$ are means of respective i-vectors. N_t and N_d are the numbers of training and development i-vectors respectively. The decorrelation transform, $\boldsymbol{D}$ is estimated using the Cholesky decomposition of $\boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{Q}^{-1}$. Then the training, development and evaluation i-vectors are transformed using $\boldsymbol{D}$ (Domain Mismatch Compensation Matrix) given by, $\hat{\boldsymbol{y}} = \boldsymbol{D}^T y$. The domain compensation did not benefit x-vector features and was employed only for i-vector features.

Table 1: *Domain-wise DER - individual system [VB-HMM (x-vec. init.)]. fused system [VB-HMM (i-vec + x-vec init.)].*

System	Dev											Eval	
	LIB.	**SEED.**	**CIR**	**ADO.**	**SCO.**	**DCI.**	**RT04**	**SLX**	**MIX6**	**VAST**	**YP**	**ALL**	**ALL**
Baseline [15]	12.22	33.74	51.41	16.05	14.64	6.92	33.39	15.84	12.82	37.19	5.80	23.70	25.99
Individual	3.08	33.10	45.65	19.87	6.10	11.04	27.92	14.37	10.18	38.71	3.24	21.08	23.57
Fused	4.48	32.86	45.53	16.88	5.26	8.45	27.71	14.28	10.26	37.03	3.04	20.56	21.90

Table 2: *DER(JER) performances for system configurations indicating the improvements from the proposed approaches.*

System config.	Dev DER(JER)
i-vectors	24.21 (52.89)
i-vectors, with DMC	**23.79** (51.03)
VB (x-vec. init)	24.72 (51.85)
posterior scaled VB(x-vec. init.)	21.15 (51.10)
posterior scaled VB (x-vec init.)+ seg. enh.	**21.08** (49.63)

Table 3: *DER(JER) of individual and fused systems. *VB is the posterior scaled VB-HMM with segment enhancement*

Individual System	Dev DER(JER)	Eval DER(JER)
Baseline[15]	23.70 (56.20)	25.99 (59.51)
i-vectors (DMC) with AHC	23.79 (51.03)	24.49 (51.32)
x-vectors with AHC	22.03 (49.59)	23.80 (52.22)
VB-HMM (x-vec init)	**21.08** (49.63)	**23.57** (53.37)
Fused system	**Dev DER(JER)**	**Eval DER(JER)**
ivec + xvec (AHC)	21.24 (46.72)	22.37 (49.32)
*VB (ivec. + xvec. init.)	**20.56** (47.43)	**21.90** (49.93)

5. System Description

This section gives the systems description using the above proposed methods for DIHARD challenge (track 1). The speaker diarization output from a i-vector/x-vector based AHC system is used for initializing the posterior-scaled VB-HMM[1] given in section (4.1). The block diagram shown in figure 1 gives a brief overview about the stages involved in the DIHARD development dataset processing to get the final speaker assignment. The implementation details are given below.

I-vector system: It involves extraction of i-vectors using MFCCs as the front end features and then applying Domain Mismatch Compensation transform (section 4.2) to the training i-vectors for PLDA training.

X-vector system: This system is similar to the baseline system except that the recording level PCA preserves 30% of dimensions (versus 10% in baseline)

System fusion: Here we take weighted average of PLDA score matrix from (0.7 times) x-vector and (0.3 times) i-vector systems and perform AHC. The output of this system is given as initialisation to the VB-HMM system.

6. Experiments & Results

Here we describe all the experiments involving system described above and overall performance using Diarization Error Rate (DER) as the primary metric along with Jaccard Error Rate (JER) as the secondary metric. The scoring script of evaluation is provided by the DIHARD challenge organizers [15].

The baseline system performs PCA on each recording before the PLDA scoring. In our experiments, we preserve about 50% of PCA dimensions in i-vector and 30% of dimensions in x-vector system which proved useful. We adapt the PLDA matrix using the domain-compensation transformation which further helped to improve the DER of i-vector system (Table 2).

Table 2 also shows the improvements on the VB-HMM using the posterior scaling approach. Without the posterior scaling, the VB-HMM model ($loop = 0.9$, $mindur = 1$, $T = 20$) was inferior to the x-vector baseline system. The posterior scaling provides about 19.4 % relative improvements over the basic implementation of the VB-HMM. The best choice of the VB-HMM hyper-parameters from development set are used for the evaluation dataset ($\beta = 24$, $loop = 0.5$, $mindur = 1$, $T = 20$). If we increase $beta$, keeping $loop = 0.9$, it results in high smoothing of speakers. To compensate for that, we used $loop = 0.5$ as the optimal loop probability. The segment enhancement using the Poisson distribution (last row of Table 2) gives minor improvements in DER, but improved the JER results considerably.

Table 3 shows the best individual system and the final fused system used in the LEAP submission to DIHARD challenge. The first two rows report the results for the i-vector (with domain compensation) and the x-vector system based on AHC based diarization. The x-vector system performs better than the i-vector system and this is used as the initialization to the VB-HMM. The best individual system is the posterior scaled VB-HMM. For fusing the individual systems, we performed a weighted average of the i-vector and x-vector PLDA score matrices. The fused PLDA scores are used in the AHC clustering. As seen in Table 3, the fusion improved the DER and JER results. The final submitted system used the AHC based segmentation from the fused scores of the i-vector and x-vector system.

The domain-wise performance of the best individual system and the fused system are compared with the x-vector baseline in Table 1 for the DIHARD development dataset. The details of the domains are also part of the second DIHARD challenge [15]. Results show that the best individual system improves the baseline on most of the domains. The fused system further improves the DER results. On the DIHARD evaluation dataset, the best individual system and the fused system improve the baseline relatively by about 7.1 % and 13.7% respectively.

7. Conclusion

In this paper, we have presented the details of the diarization system developed by the LEAP team. The novel components of the proposed system include the posterior scaling approach in the VB-HMM and the domain compensation for the i-vector features. The mathematical framework and analysis of posterior scaling for VB-HMM reveals that the scaling decreases emission probabilities and leads to increasing the impact of transition probabilities hence frames get aligned to more dominant speakers. Various experiments on the DIHARD dataset highlight the improvements obtained for the proposed approaches. The final system submission also improves the baseline model on most of the challenging domains in the DIHARD dataset.

[1] https://github.com/iiscleap/LEAP_Diarization

8. References

[1] K. Church, W. Zhu, J. Vopicka, J. Pelecanos, D. Dimitriadis, and P. Fousek, "Speaker diarization: a perspective on challenges and opportunities from theory to practice," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4950–4954.

[2] J. S. Garofolo, C. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus." in *LREC*. Citeseer, 2004.

[3] 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 *Interspeech*, 2013.

[4] N. Ryant, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta *et al.*, "Enhancement and analysis of conversational speech: JSALT 2017," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5154–5158.

[5] E. Bergelson, "Bergelson seedlings homebank corpus," *doi*, vol. 10, p. T5PK6D, 2016.

[6] N. Ryant et al. (2019), "DIHARD corpus. Linguistic Data Consortium." in *Proceedings of INTERSPEECH*, 2019.

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

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

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

[10] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on bayesian hmm with eigenvoice priors," in *Proceedings of Odyssey*, 2018, pp. 147–154.

[11] I. Vinals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the number of speakers with variational bayesian plda in the dihard diarization challenge," in *Proc. INTERSPEECH*, 2018, pp. 2803–2807.

[12] Z. Zajıc, M. Kunešová, J. Zelinka, and M. Hrúz, "Zcu-ntis speaker diarization system for the dihard 2018 challenge," in *Proc. INTERSPEECH*, 2018, pp. 2788–2792.

[13] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first dihard challenge," in *Proc. INTERSPEECH*, vol. 2018, 2018, pp. 2813–2817.

[14] 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.

[15] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines." in *Proceedings of Interspeech*, 2019.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

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

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[19] N. Dehak, P. J. 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.

[20] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5045–5049.

[21] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.

[22] X. Chen, L. He, C. Xu, Y. Liu, T. Liang, and J. Liu, "Vb-hmm speaker diarization with enhanced and refined segment representation," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 134–139.

[23] M. H. Rahman, A. Kanagasundaram, I. Himawan, D. Dean, and S. Sridharan, "Improving plda speaker verification performance using domain mismatch compensation techniques," *Computer Speech & Language*, vol. 47, pp. 240–258, 2018.