THE LEAP SPEAKER RECOGNITION SYSTEM FOR NIST SRE 2018 CHALLENGE

Shreyas Ramoji[1], *Anand Mohan*[1], *Bhargavram Mysore*[2], *Anmol Bhatia*[3],
Prachi Singh[1], *Harsha Vardhan*[1], *Sriram Ganapathy*[1]

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Electrical Engineering,
Indian Institute of Science, Bengaluru, India
[2]North Carolina State University, North Carolina, USA
[3]Birla Institute of Technology and Sciences (BITS) Pilani, India

ABSTRACT

The NIST Speaker Recognition Evaluation (SRE) 2018 challenge comprises an open evaluation of the text independent speaker verification task. This paper summarizes the LEAP speaker verification systems submitted to the NIST SRE 2018. For all the speaker verification approaches, the front-end feature extraction involved the use of neural embeddings from a time delay neural network (TDNN) trained on a speaker discrimination task. These features, called x-vectors, are used in multiple ways for speaker verification task. In the first approach, the x-vectors with pre-processing and dimensionality reduction, are used with probabilistic linear discriminant analysis (PLDA) scoring. The second approach applies a speaker diarization scheme on the test segments containing multiple talkers before speaker verification scoring based on PLDA. The third system uses a local pairwise LDA model for pre-processing the x-vectors which are then scored using a Gaussian back-end. With experiments on the SRE 2018 database, we show that most of the systems achieved noticeable improvements over the NIST baseline in terms of the primary cost metric. Using a system fusion of the various approaches, we obtain significant improvements over the NIST official baseline (average relative improvements of 19.7% and 20.1% for the development and evaluation set respectively).

***Index Terms*—** x-vectors, Speaker Diarization, PLDA scoring, Gaussian back-end, Dimensionality Reduction, Speaker Verification.

1. INTRODUCTION

The recent years have seen increasing demand for speaker based authentication and verification systems. In commercial and defense applications where the speaker verification system forms the first level of interface, the acceptable performance of the system relies on relatively clean recordings and with matched languages used in training and testing the systems. The performance is substantially degraded in noisy and multi-lingual environments making the downstream applications vulnerable. The NIST biannual speaker recognition evaluation (SRE) challenges provide benchmark for comparing and standardizing speaker recognition systems. The SRE 2018 challenge is the latest among the ongoing series of speaker recognition evaluations conducted. The development and evaluation data contains recordings from Call My Net 2 (CMN2) data which consists of multi-speaker data from Tagalog and Cantonese languages and the Video Annotation for Speech Technology (VAST) data containing noisy speech recordings extracted from YouTube videos. CMN2 recordings are derived from conversational telephone speech recording from various devices and Voice Over IP (VOIP) while the VAST corpora is derived from far field, multi-speaker and noisy conditions.

The traditional approach to speaker recognition used the Gaussian mixture modeling (GMM) from the training data followed by an adaptation using maximum-aposteriori (MAP) rule [1]. The adapted model is compared with the background GMM model using the log-likelihood ratio based scoring. The development of i-vectors as fixed dimensional front-end features for speaker recognition tasks was introduced in [2, 3]. The i-vectors capture long term information of the speech signal such as speaker and language. In the recent past, the i-vectors derived from deep neural network (DNN) based posterior features were attempted for SID [4]. The use of bottleneck features for front-end feature extraction derived from a speech recognition acoustic model has also shown good improvements for speaker recognition [5, 6].

Recently, neural network embeddings trained on a speaker discrimination task were also derived as features to replace the i-vectors. These features called x-vectors [7] were shown to perform better than the i-vectors for speaker recognition. Following the extraction of x-vectors/i-vectors, different speaker verification systems make use of discriminative/ generative models in the back-end for computing the scores. The most popular approaches for scoring include support vector machines (SVMs) [8], Gaussian back-end model [9, 10] and the probabilistic linear discriminant analysis (PLDA) [11]. Some efforts on pairwise generative and discriminative modeling are discussed in [12–14].

In this paper, we describe our efforts for the SRE 2018 challenge which comprised of three broad approaches. In the first approach, we use the x-vector based features with a PLDA scoring system. We process the x-vectors with various normalization and dimensionality reduction techniques such as within class covariance normalization (WCCN) [15], length normalization [16] and linear discriminant analysis (LDA) [17]. In the second approach, we leverage an i-vector based speaker diarization system by diarizing the test files (that contain multiple talkers), and scoring each of the speaker segments with the enrolment speaker utterances in a PLDA model. The final score for system submission is the maximum of all the speaker segment scores. The diarization approach is motivated by the fact that a target test recording will usually be missed when it consists of multiple speakers. Assuming the diarization to be correct, if one of the diarized segments corresponds to the target speaker, it will result in a high detection likelihood ratio and the target trial will not be missed. In the third approach, local pairwise LDA [18] attempts to

This work was funded partly by grants from the Department of Science and Technology Extra Mural Research Grant (EMR/2016/007934).

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

model pairs of x-vectors rather than single i-vectors/x-vectors done in traditional systems. In the third approach, by sampling pairs of x-vectors from the training datasets, we would have access to a very large number of trials for both the target and non-target cases. We show that using a simple two class Gaussian back-end model [14] over pairs of x-vectors can achieve significant improvements over the baseline which uses Gaussian PLDA on length normalized x-vectors.

The rest of the paper is organized as follows: In Section 2, we give an overview of the x-vector feature extraction. Section 3 details the individual systems developed for SRE 2018 highlighting the novel approaches proposed for speaker verification. In Section 4, we provide a description of the system submissions for SRE. This is followed by Section 5 where we report the SRE results. A summary of the paper is provided in Section 6.

2. X-VECTOR MODEL

The x-vector model [7, 19] is developed using the Kaldi toolkit [20]. The model consists of a time-delay neural network (TDNN) with utterance level statistics pooling followed by a fully connected neural network that maps to speaker targets. The TDNN model is trained using conversational telephone and microphone speech data extracted from the NIST 2004-2010 SRE datasets, as well as from MIXER 6, Switchboard Cellular (SWBCELL) Parts I and II, and Switchboard (SWB) Phases I, II, and III corpora. We use a 3-fold data augmentation strategy that adds two noisy versions of the original recordings to the training data [7]. The recordings are corrupted by either digitally adding noise (i.e., babble, general noise, music) or by a convolution with simulated room impulse responses [1]. The front-end features for the TDNN training consists of 23-dimensional mel frequency cepstral coefficients from 25 ms frames which are shifted every 10 ms using a 23-channel mel-scale filter bank spanning the frequency range 20 Hz - 3700 Hz. Before dropping the non-speech frames using an energy based SAD, a short-time cepstral mean subtraction is applied over a 3-second sliding window. The TDNN model has 3 layers of time delay neural network layers, two fully connected layers, a statistical pooling layer that computes the mean and standard deviation at utterance level, and 2 fully connected layers. All the layers use a rectified linear unit (ReLU) non-linearity and the model is trained to discriminate among the nearly 7000 speakers in the training set with $219,238$ speech segments. The first 5 hidden layers operate at frame-level, while the last 2 operate at segment-level. After training, the x-vector embeddings are extracted from the 512-dimensional affine component of the 6th layer (i.e., the first segment-level layer).

3. INDIVIDUAL SYSTEMS

The block schematics of the individual systems developed for SRE 2018 challenge is shown in Figure 1. Five individual systems A-E were considered for submission. System A is the x-vector PLDA baseline system with some minor changes. System B and C are x-vector PLDA systems which make use of an i-vector based diarization system. Systems D and E are Gaussian back-end systems. The detailed description of each of the systems is given below.

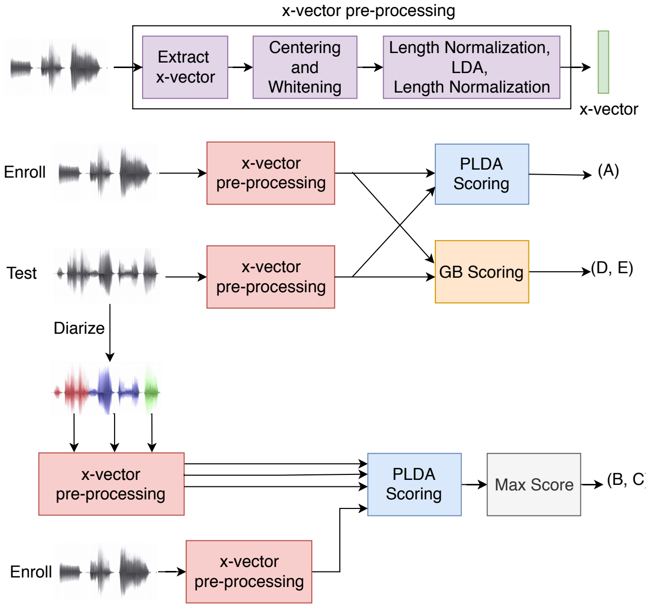

Fig. 1. Block schematic of the x-vector pre-processing pipeline and the individual system components used for SRE2018 evaluation.

3.1. System A - x-vector PLDA

The x-vectors extracted from the TDNN model are centered, unit-length normalized [16], whitened, and dimensionality reduced to 150 dimensions using LDA [17]. A Gaussian PLDA model with a full-rank eigenvoice subspace is trained using the x-vectors extracted from $127,233$ speech segments derived from the SRE and MIXER 6 datasets after discarding very short duration segments (less than 5 s). The PLDA model is adapted using the unlabelled recordings of the SRE 2018 development set. We found that the adaptation only benefited the CMN2 dataset [21] while the adaptation did not improve the VAST dataset. Hence, we use the adapted PLDA models only for the CMN2 scoring. The scores are then calibrated as described later in Section 4.

3.2. System B - Diarization based x-vector PLDA scoring

The diarization system [22] involves segmenting the test recordings into different speaker segments and then scoring the same speaker regions with the enrolment files. The following steps are involved in this system:

3.2.1. Diarization Model Training

The training datasets used are the NIST 2004-2010 SRE datasets (without augmentation), Switchboard Cellular(SWBCELL) Parts I and II, and Switchboard (SWB) Phases I, II, and III corpora. The 23 dimensional MFCC features are extracted at a window size of 25 ms and a 10 ms shift same as in System A. A Gaussian Mixture Universal Background Model (GMM-UBM) of 2048 mixtures is trained and adapted using SRE 2018 unlabeled development dataset. This is further used to estimate a 64 dimensional total variability matrix for i-vector extraction. Then, i-vectors are extracted from the training dataset to build a PLDA model. This PLDA model is also adapted using the SRE 2018 unlabeled development set.

[1] Room response functions obtained from http://www.openslr.org

3.2.2. *SRE Scoring*

For the given test recording, the 64 dimensional i-vectors are obtained at every 1.5 s with a 0.75 s shift. The i-vectors are obtained from the segments defined as speech after applying VAD. Then, the PLDA scores are computed for every pair of i-vectors from the given recording. This generates a matrix containing the similarity score of each i-vector with every other i-vector. An agglomerative hierarchical clustering (AHC) technique is used to identify speaker clusters from the PLDA scores [22]. During this clustering process, similar i-vectors are merged into a single cluster. The enrolment part of the SRE 2018 development data have been used to find the threshold used to stop the AHC procedure. The enrolment files from CMN2 source have only one speaker and files from VAST are provided with hand labeled diarization segments. For these files/segments, the AHC is done till all segments are clustered as one speaker. The lowest threshold τ_{min} needed for each enrolment file so as to cluster all frames as one speaker is computed. The threshold for diarizing the SRE18 test recordings is determined as the mode of the thresholds τ_{min} of all enrolment files.

The PLDA model trained for System A is used for this system as well. The x-vectors are extracted for enrolment recordings (as in System-A) and for each of the diarized test segments. This is illustrated in Fig 1. Each of the diarized test segment is scored with the corresponding enrolment model, and the maximum score among the diarized segments for a given trial is considered to be the final score. As the diarization errors adversely impact the SRE performance on the development data, we find that the performance of the diarization based SRE systems to be marginally worse than the baseline, but complementary nonetheless.

3.3. System C - Diarization based x-vector PLDA scoring using adapted PLDA model

This is exactly the same as System B, except that the PLDA model used in SRE scoring is adapted [21] using the unlabeled development data. We find in our experiments that the adaptation provides good improvements to the diarization based SRE system.

3.4. System D - Gaussian back-end with LDA transformed paired x-vectors

The x-vectors are extracted using the same model as System A. From a total of $127,233$ speech segments from the training datasets, $62,290$ pairs of x-vectors are randomly sampled, with the constraint that each pair is derived from the same speaker. Then, another $63,901$ pairs of segments are randomly sampled such that each pair is derived from different speakers. The x-vectors are processed with various normalization and dimensionality reduction techniques. These processed x-vectors of dimension R are concatenated into a single vector of dimension $2R$. A prior work in [14] had previously looked at pairwise generative modeling for speaker verification.

In our work, the same-speaker pairs represent target trials and different-speaker pairs represent non-target trials in the evaluation set. The paired x-vectors from same speaker pairs are modeled using a Gaussian distribution with parameters $(\boldsymbol{\mu}_t, \Sigma_t)$ and different speaker pairs are modeled by a Gaussian distribution with parameters $(\boldsymbol{\mu}_{nt}, \Sigma_{nt})$. The log-likelihood ratio (LLR) for a trial pair $\boldsymbol{x} = [\boldsymbol{x}_{enrol}^{\intercal}\ \boldsymbol{x}_{test}^{\intercal}]^{\intercal}$ is then obtained as:

$$LLR = -(\boldsymbol{x} - \boldsymbol{\mu}_t)^{\intercal} \Sigma_t^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_t) + (\boldsymbol{x} - \boldsymbol{\mu}_{nt})^{\intercal} \Sigma_{nt}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{nt})$$

For System D, the Gaussian PLDA model with a speaker factor of 100 dimensions is trained on the training dataset and the MAP estimates of the speaker factors (eigenvoice factors) are obtained for the CMN2 segments. These 100 dimensional speaker factors are used to compute the CMN2 Gaussian back-end scores. For the VAST test set, the 150 dimensional LDA x-vectors are used directly. The Gaussian means are also adapted using the labeled SRE development dataset as follows:

$$\boldsymbol{\mu}_{t(adapt)} = (1 - \lambda)\boldsymbol{\mu}_t + \lambda\boldsymbol{\mu}_{dev}$$
$$\boldsymbol{\mu}_{nt(adapt)} = (1 - \lambda)\boldsymbol{\mu}_{nt} + \lambda\boldsymbol{\mu}_{dev}$$

where $\boldsymbol{\mu}_{dev} = [\boldsymbol{\mu}_{enrol}^{\intercal}\ \boldsymbol{\mu}_{test}^{\intercal}]^{\intercal}$ is the paired mean vector of the enrolment and test x-vectors of the SRE 2018 development set. We have used an adaptation factor of $\lambda = 0.2$ for the SRE2018 evaluation.

It is worth noting that the sampling of non-target and target pairs is done in such a way that no speech segment is repeated. That is, out of ${}^{n}C_2$ pairs of segments, we choose approximately $n/2$ pairs. This procedure takes only a few seconds and also models the speakers reasonably well.

3.5. System E - Gaussian back-end with LP-LDA transformed paired x-vectors

This system is similar to System D except for the pre-processing of the paired x-vectors. In this case, the 512 dimensional x-vectors are centered, unit length normalized, whitened and reduced to 150 dimensions via Linear Pairwise Linear Discriminant Analysis (LP-LDA) [18]. A Gaussian PLDA model with a speaker factor of 100 dimensions is trained, and the MAP estimates of the speaker factors are obtained for CMN2 segments. A Gaussian PLDA model with a speaker factor of 50 dimensions is trained and the MAP estimates of the speaker factors are obtained for the VAST datasets. This is followed by the Gaussian back-end modeling and scoring as described for System D. However, the Gaussian mean adaptation is not performed for System E.

4. SYSTEMS SUBMITTED FOR SRE 2018 CHALLENGE

4.1. Score Calibration and Fusion

A linear score fusion of the different systems is done using the FoCal two class toolkit [24] where the weights and biases are obtained with a logistic regression objective. Furthermore, we calibrate the scores of the final systems using an affine transform which normalizes the within class score variance. The scores are then mean shifted such that the threshold corresponding to the minimum cost is the point where the actual cost is computed (the operating point for actual cost is given in NIST SRE 2018 evaluation plan [2]). This was performed so as to minimize the difference between C_{min} and $C_{primary}$ cost metrics used for SRE scoring. All the systems are analyzed using the equal error rate (EER), minimum detection cost (C_{min}) and the primary detection cost ($C_{primary}$) as defined in the evaluation plan.

4.2. Primary system

The primary system used a system combination of all the five individual system (System A-E). The FoCal toolkit [24] is used to get the final scores for the primary system which are then calibrated before cost metric computation.

[2] $nist.gov/sites/default/files/documents/2018/08/17/sre18\_eval\_plan\_2018-05-31\_v6.pdf$

Table 1. Results for individual systems on the SRE 2018 development and evaluation set measured using EER (%), $C_{primary}$ and C_{min}. The refernce for comparison is the official NIST baseline systems developed using i-vectors and x-vectors [23].

Systems	Dataset	Dev			Eval		
		EER (%)	$C_{primary}$	C_{min}	EER (%)	$C_{primary}$	C_{min}
NIST i-vector baseline	CMN2	12.66	0.737	0.663	13.66	0.808	0.773
	VAST	9.05	0.778	0.630	17.14	0.831	0.732
NIST x-vector baseline	CMN2	10.62	0.719	0.651	11.38	0.776	0.741
	VAST	7.41	0.704	0.572	14.22	0.837	0.721
System A	CMN2	**9.15**	**0.601**	**0.587**	10.56	**0.631**	**0.618**
	VAST	7.41	0.646	0.646	16.41	**0.701**	0.672
System B	CMN2	11.08	0.682	0.668	12.28	0.711	0.704
	VAST	7.41	0.646	0.646	15.90	0.726	0.687
System C	CMN2	9.60	0.603	0.588	11.43	0.638	0.629
	VAST	8.64	0.630	0.630	14.29	0.710	**0.603**
System D	CMN2	9.41	0.689	0.683	**10.53**	0.746	0.733
	VAST	7.41	**0.498**	**0.498**	14.60	0.742	0.688
System E	CMN2	11.25	0.771	0.749	12.13	0.838	0.829
	VAST	**3.70**	0.576	0.576	**13.02**	0.729	0.704

Table 2. Results for various fused systems submitted to the SRE 2018 on the development and evaluation set.

Systems	Dataset	Dev			Eval		
		EER (%)	$C_{primary}$	C_{min}	EER (%)	$C_{primary}$	C_{min}
Primary System	CMN2	**7.97**	**0.58**	**0.56**	**9.34**	**0.61**	**0.61**
	VAST	4.12	**0.56**	**0.56**	14.64	**0.68**	**0.64**
Contrastive System	CMN2	9.41	0.69	0.68	10.53	0.75	0.73
	VAST	**3.70**	0.58	0.58	**13.02**	0.73	0.70

4.3. Contrastive System

This system is based only on single system. The CMN2 test segments are scored using System D and the VAST segments are scored using System E. This corresponds to the best single in terms of EER on the development data.

5. RESULTS

The results for the individual systems used in SRE evaluation are provided in Table 1. The baseline system results provided by NIST [23] are also provided in the top of the table for reference. As seen here, most of the systems improve over the baseline system results. The best performance of the individual systems on the development and evaluation system for CMN2 dataset is achieved for System A (average relative improvements of 16.4% and 18.7% over the NIST baseline for the development and evaluation dataset respectively). For the VAST dataset in the SRE 2018 development, the Gaussian back-end based systems (System D and System E) improve over the PLDA based systems (Systems A,B,C). For evaluation part of SRE 2018, the diarization based System C provides the best performance in terms of C_{min}. The average relative improvement in terms of $C_{primary}$ for System-D over the NIST baseline is about 16.7% and 7.7% on the development and evaluation set respectively.

Table 2 summarizes the performances of the submitted systems that are based on system fusion. The primary system improves over the best individual systems (contrastive system) for the development and evaluation setup.The average relative improvements in terms of $C_{primary}$ for the primary system over the NIST baseline are about 20.1 % and 19.7 % respectively for the development and evaluation set. It is also worth noting the improvements observed for most of the systems in the open development set are also consistent when tested using the blind evaluation dataset.

6. SUMMARY AND CONCLUSIONS

In this paper, we have presented the detailed description of the LEAP submission to SRE 2018 evaluation. The systems developed for the challenge consisted of novel components like diarization for speaker verification, Gaussian back-end modeling and pairwise linear discriminant analysis. In particular, the Gaussian back-end outperformed the conventional PLDA modeling for the noisy VAST dataset recordings. It is important to note that the Gaussian Back-end performs better than PLDA though we do not explicitly model the speaker and channel variability like in PLDA. The individual systems developed for the NIST SRE task improved over the NIST baseline. The combined system provided considerable advancement in terms of SRE performance over the baseline system.

7. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell 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, pp. 1695–1699.

[5] Fred Richardson, Douglas Reynolds, and Najim Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[6] Seyed Omid Sadjadi, Sriram Ganapathy, and Jason W Pelecanos, "The IBM 2016 speaker recognition system," *arXiv preprint arXiv:1602.07291*, 2016.

[7] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," .

[8] William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[9] Mitchell McLaren, Aaron Lawson, Yun Lei, and Nicolas Scheffer, "Adaptive gaussian backend for robust language identification.," in *INTERSPEECH*, 2013, pp. 84–88.

[10] Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain, and Lori Lamel, "Language score calibration using adapted gaussian back-end," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[11] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, pp. 14–21.

[12] Sandro Cumani, Niko Brümmer, Lukáš Burget, Pietro Laface, Oldřich Plchot, and Vasileios Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[13] Sandro Cumani and Pietro Laface, "Large-scale training of pairwise support vector machines for speaker recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 11, pp. 1590–1600, 2014.

[14] Sandro Cumani and Pietro Laface, "Generative pairwise models for speaker recognition," in *Odyssey*, 2014, pp. 273–279.

[15] Andrew O Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Ninth international conference on spoken language processing*, 2006.

[16] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[17] Mitchell McLaren and David Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2012.

[18] Liang He, Xianhong Chen, Can Xu, Jia Liu, and Michael T Johnson, "Local pairwise linear discriminant analysis for speaker verification," *IEEE Signal Processing Letters*, 2018.

[19] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[21] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[22] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with PLDA i- vector scoring and unsupervised calibration," *Spoken Language Technology Workshop (SLT)*, pp. 413–417, December 2014.

[23] Syed Omid Sadjadi, "NIST baseline systems for the 2018 speaker recognition evaluation," 2018.

[24] Niko Brümmer, "Focal toolkit: Matlab code for evaluation, fusion and calibration of statistical pattern recognizers.," .