Nearest Neighbor Discriminant Analysis for Robust Speaker Recognition

Seyed Omid Sadjadi, Jason W. Pelecanos, Weizhong Zhu

IBM Research, Watson Group

{sadjadi,jwpeleca,zhuwe}@us.ibm.com

10.21437/Interspeech.2014-422

Abstract

With the advent of i-vectors, linear discriminant analysis (LDA) has become an integral part of many state-of-the-art speaker recognition systems. Here, LDA is primarily employed to annihilate the non-speaker related (e.g., channel) directions, thereby maximizing the inter-speaker separation. The traditional approach for computing the LDA transform uses parametric representations for both intra- and inter-speaker scatter matrices that are based on the Gaussian distribution assumption. However, it is known that the actual distribution of i-vectors may not necessarily be Gaussian, and in particular, in the presence of noise and channel distortions. Motivated by this observation, we present an alternative non-parametric discriminant analysis (NDA) technique that measures both the within- and between-speaker variation on a local basis using the nearest neighbor rule. The effectiveness of the NDA method is evaluated in the context of noisy speaker recognition tasks using speech material from the DARPA Robust Automatic Transcription of Speech (RATS) program. Experimental results indicate that the NDA is more effective than the traditional parametric LDA for speaker recognition under noisy and channel degraded conditions.

Index Terms: discriminant analysis, i-vector, nearest neighbor, speaker recognition

1. Introduction

Speaker recognition has evolved significantly over the past few years. The research trend in this domain has gradually migrated from joint factor analysis (JFA) based methods, which attempt to model the speaker and channel subspaces separately [1], towards the i-vector approach that models both speaker and channel variabilities in a single low-dimensional (e.g., a few hundred) space termed the total variability subspace [2]. Accordingly, there has been a growing interest to design algorithms for the compensation of the channel subspace in the i-vector paradigm. Here, irrespective of the back-end model, linear discriminant analysis (LDA) with the Fisher criterion [3] has been the most commonly adopted approach for inter-session variability compensation. Current state-of-the-art speaker recognition systems employ the Fisher LDA as a preprocessor to generate dimensionality reduced and channel-compensated features from i-vectors [4, 5, 6, 7, 8, 9, 10]. The dimensionality reduced vectors can then be conveniently modeled and scored with various classifiers such as support vector machines (SVM) [2], probabilistic LDA (PLDA) [11, 12, 13], and the simple yet effective cosine distance (CD) based method [2].

The Fisher LDA aims at finding the most discriminative feature subset through a linear transformation of the original input space. Such a transformation attempts to maximize the between-class (or inter-speaker) scatter while minimizing the within-class variation. Traditionally, parametric within- and between-class scatter matrices are formed based on the Gaussian distribution assumption for the samples in each class. However, if the class-conditional distributions are non-Gaussian, one cannot expect the use of such parametric forms to result in proper feature subsets that are capable of preserving complex structures within data needed for classification (e.g., multi-modality). It is well known in the speaker recognition community that the actual distribution of i-vectors may not necessarily be Gaussian [12]. This is in particular more problematic when speech recordings are collected in the presence of noise and channel distortions. Hence, despite its popularity, the parametric LDA may not be the best choice here.

To cope with the above noted issue associated with the parametric nature of the scatter matrices, in the seminal work of [14], a non-parametric discriminant analysis (NDA) approach was proposed for general two-class pattern recognition problems. It was later extended to multi-class problems and successfully applied in several other studies for face recognition tasks as well [15, 16]. The NDA measures both the within- and between-class scatter matrices on a local basis using the k-nearest neighbor (k-NN) rule, and unlike LDA, is generally of full rank. Note that for a C class problem, the parametric LDA can provide at most $C-1$ discriminant features (i.e., the number of classes minus 1). Nonetheless, this is less problematic in speaker recognition because typically the number of speakers in the training set exceeds the dimensionality of the total variability subspace. The non-parametric nature of the scatter matrices in the NDA inherently results in features that can preserve the local structure (e.g., the class boundaries) within data which is important for classification.

In this study, we investigate the application of NDA for robust speaker recognition. In particular, we evaluate the effectiveness of NDA against the traditional LDA in the context of speaker recognition tasks under actual noisy and channel degraded conditions using speech material from the DARPA program, Robust Automatic Transcription of Speech (RATS). The RATS data, which is distributed by the Linguistic Data Consortium (LDC) [17], consists of conversational telephone speech (CTS) recordings that have been retransmitted (through LDC's Multi Radio-Link Channel Collection System) and captured over 8 extremely degraded communication channels with distinct distortion characteristics. The distortion type seen in RATS data is nonlinear and the noise is to some extent correlated with speech. We conduct our speaker experiments with a state-of-the-art i-vector based system [10] and report on false-reject (miss) rate at 2.5% false-alarm (FA) rate as performance metric. We also explore the impact of different configuration parameters for NDA (e.g., the feature dimensionality as well as the number of neighbors in the k-NN analysis) on speaker recognition performance.

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views, opinions, findings and recommendations contained in this article are those of the authors and do not necessarily reflect the official policy or position of the DOI/NBC.

2. I-vector feature extraction

In this section we briefly describe the total variability modeling concept which was inspired by the JFA approach that attempts to estimate separate subspaces for speaker and channel variabilities. Unlike JFA, the total variability approach assumes that both speaker and channel variabilities reside in the same low-dimensional subspace. This concept is closely related to the eigenvoice adaptation technique developed for automatic speech recognition (ASR) that was first proposed in [18] based on principal component analysis (PCA), and later in [19] based on probabilistic PCA (PPCA) [20]. Given a set of observations (i.e., frames) for each speech session, the eigenvoice adaptation technique computes an offset (linear shift) in the *prior* acoustic space represented by the Gaussian mixture model (GMM) mean supervectors obtained from a universal background model (UBM). The key idea here is that variability within and across sessions can be described via a few set of parameters (a.k.a factors) in a low-dimensional subspace spanned by the columns of a low-rank rectangular matrix, $\mathbf{T}$, entitled the total variability matrix. Mathematically, the adapted mean supervector, $\mathbf{M}$, for a given set of observations can be modeled as,

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{x} + \epsilon, \tag{1}$$

where $\mathbf{m}$ is the prior mean supervector, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a multivariate random variable termed an identity[1] vector "i-vector", and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a residual noise term to account for the variability not captured via $\mathbf{T}$ ($\mathbf{\Sigma}$ is typically copied from the UBM). In other words, for the given observation set, the i-vector represents the coordinates in the total variability subspace. The procedure for training the i-vector extractor (i.e., the $\mathbf{T}$ matrix) is similar to the eigenvoice learning process described in [19] (a simplified version is presented in [21]), except each speech recording (session) is assumed to be produced by a unique speaker.

3. Linear discriminant analysis (LDA)

LDA is widely adopted in pattern recognition problems as a preprocessing stage for feature selection and dimensionality reduction. It computes an optimum linear projection $\mathbf{A}\colon \mathbb{R}^d \mapsto \mathbb{R}^n$ by maximizing the ratio of the inter-class scatter to intra-class variance:

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}, \tag{2}$$

where $\mathbf{A}$ is a rectangular matrix with n linearly independent columns. Here, the within- and between-class scatter matrices are used to formulate a class separability criterion which converts the matrices into a single statistic. This statistic takes on larger values when the between-class scatter is larger and the within-class variance is smaller. Several such class separability criteria are described in [22], of which the following is the most widely used,

$$\hat{\mathbf{A}} = \underset{\mathbf{A}^T\mathbf{S}_w\mathbf{A}=\mathbf{I}}{\arg\max} \left[\mathrm{tr}\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)\right], \tag{3}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between- and within- class scatter matrices, respectively. The optimization problem in (3) has an analytical solution that is a matrix whose columns are the n eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

[1]The term "i-vector" sometimes also refers to a vector of "intermediate" size, bigger than the underlying cepstral feature vector but much smaller than the GMM supervector.

The within-class scatter matrix measures the scatter of samples in each class around the expected value of that class as,

$$\mathbf{S}_w = \sum_{i=1}^{C} p_i \mathrm{E}\left[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mid C_i\right] = \sum_{i=1}^{C} p_i \mathbf{\Sigma}_i, \tag{4}$$

where p_i, $\boldsymbol{\mu}_i$, and $\mathbf{\Sigma}_i$ are the *a priori* probability (proportional to the number of sessions per speaker), expected value, and covariance matrix for class i. The between-class scatter matrix, on the other hand, measures the scatter of class-conditional expected values around the global mean as,

$$\mathbf{S}_b = \sum_{i=1}^{C} p_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{5}$$

where $\boldsymbol{\mu}$ is the expected value of the training samples computed as,

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}] = \sum_{i=1}^{C} p_i \boldsymbol{\mu}_i. \tag{6}$$

There are three disadvantages associated with the parametric nature of the scatter matrices in (4) and (5). First, the underlying distribution of classes is assumed to be Gaussian with a common covariance matrix for all classes. Therefore, one cannot expect the parametric LDA to generalize well to non-Gaussian and multi-modal (as opposed to unimodal) distributions. Second, notice that the rank of $\mathbf{S}_b$ is $C-1$, which means the parametric LDA can provide at most $C-1$ discriminant features. However, this may not be sufficient in applications such as language recognition where the number of language classes is much smaller than the dimensionality of the i-vectors (for instance, there are only 5 target language categories in the RATS program). Nevertheless, this may not pose a challenge for speaker recognition tasks in which the number of training speakers exceeds the dimensionality of the total variability subspace. Finally, because only the class centroids are taken into account for computing $\mathbf{S}_b$ in (5), the parametric LDA cannot effectively capture the boundary structure between adjacent classes which is essential for classification [22].

To overcome the above noted limitations of LDA, a nonparametric discriminant analysis technique was proposed in [14], that measures both the within- and between-class scatters on a local basis using a nearest neighbor rule. We provide a brief description of NDA in the next section.

4. Nonparametric discriminant analysis

To remedy the limitations identified for LDA, a nonparametric discriminant analysis techniques was proposed in [14]. In NDA, the expected values that represent the global information about each class are replaced with local sample averages computed based on the k-NN of individual samples. More specifically, in the NDA approach, the between-class scatter matrix is defined as,

$$\tilde{\mathbf{S}}_b = \sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{l=1}^{N_i} w_l^{ij}\left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right)\left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right)^T, \tag{7}$$

where $\mathbf{x}_l^i$ denotes the l^{th} sample from class i, and $\mathcal{M}_l^{ij}$ is the local mean of k-NN samples for $\mathbf{x}_l^i$ from class j which is computed as,

$$\mathcal{M}_l^{ij} = \frac{1}{K}\sum_{k=1}^{K} NN_k(\mathbf{x}_l^i, j), \tag{8}$$

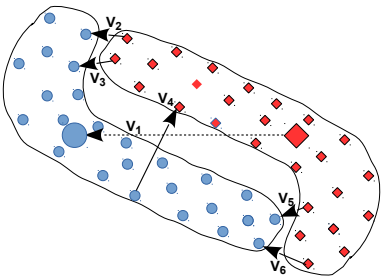

Figure 1: *Symbolic example illustrating the parametric versus nonparametric scatter between two classes.* v_1 *represents the global gradient of class centroids. The vectors* $\{\mathrm{v}_2, \cdots, \mathrm{v}_6\}$ *represent the local gradients.*

where $NN_k(\mathbf{x}_l^i, j)$ is the k^{th} nearest neighbor of $\mathbf{x}_l^i$ in class j. The weighting function w_l^{ij} in (7) is defined as,

$$w_l^{ij} = \frac{\min\left\{d^\alpha\big(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)\big), d^\alpha\big(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j)\big)\right\}}{d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)) + d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j))}, \tag{9}$$

where $\alpha \in \mathbb{R}$ is a constant between zero and infinity, and $d(.)$ denotes the Euclidean distance. The weighting function is introduced in (7) to deemphasize the local gradients that are large in magnitude to mitigate their influence on the scatter matrix. The weight parameters approach 0.5 for samples near the classification boundary (e.g., see $\{\mathrm{v}_2, \mathrm{v}_3, \mathrm{v}_5, \mathrm{v}_6\}$ shown in Figure 1), while dropping off to 0 for samples that are far from the boundary (e.g., see v_4 in Figure 1). The control parameter α determines how rapidly such decay in the weights occurs.

The nonparametric within-class scatter matrix, $\tilde{\mathbf{S}}_w$, is computed in a similar fashion as in (7), except the weighting function is set to 1 and the local gradients are computed within each class. The NDA transform is then formed by calculating the eigenvectors of $\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b$.

Three important observations can be made from a careful examination of the nonparametric between-class scatter matrix in (7). First, notice that as the number of nearest neighbors, K, approaches N_j, the total number of samples in class j, the local mean vector, $\mathcal{M}_l^{ij}$, approaches the global mean of class j (i.e., $\boldsymbol{\mu}_j$). In this scenario, if we set the weight parameters to 1, the NDA transform essentially becomes the LDA projection, which means the LDA is a special case of the more general NDA.

Second, because all the samples are taken into account for the calculation of the nonparametric between-class scatter matrix (as opposed to only the class centroids), $\tilde{\mathbf{S}}_b$ is generally of full rank. This means that unlike the LDA that provides at most $C-1$ discriminant features, the NDA generally results in d-dimensional vectors (assuming a d-dimensional input space) for the classification. As we discussed before, this is of great importance for applications such as language recognition where the number of classes is much smaller than the dimensionality of the total subspace (or the input space in general).

Finally, compared to LDA, NDA is more effective in preserving the complex structure (i.e., local and boundary structure) within and across different classes. As seen from the example shown in Figure 1 (where k is set to 1 for simplicity), LDA only uses the global gradient obtained with the centroids of the two classes (i.e., v_1) to measure the between-class scatter. On the other hand, NDA uses the local gradients (i.e., $\{\mathrm{v}_2, \cdots, \mathrm{v}_6\}$) that are emphasized along the boundary through the weighting function, w_l^{ij}. Hence, the boundary information becomes embedded into the resulting transformation.

5. Experiments

This section provides a description of our experimental setup including speech data as well as the speaker recognition system used in our evaluations. We conduct our speaker recognition experiments using actual noisy and channel degraded speech material available from the DARPA RATS program, which is distributed by the LDC [17]. The RATS data consists of CTS recordings that have been retransmitted and captured over 8 extremely degraded high-frequency (HF) radio channels, labeled A–H, with distinct noise characteristics. The type of distortion seen in RATS data is nonlinear (e.g., akin to clipping as well as amplitude compression effects) and the noise is to some extent correlated with speech. A total of five data releases are available from the LDC for the RATS speaker recognition task: LDC2012E49, LDC2012E63, LDC2012E69, LDC2012E85, LDC2012E117, which contain speech spoken in five languages: Levantine Arabic, Dari, Farsi, Pashto, and Urdu. We partitioned these data into two sets for our system training and evaluation (consisting of enrollment and test). In the RATS program a 6-sided speaker enrollment scenario is assumed, and we follow this assumption in our development setup. For system evaluation, there are 8 duration-specific tasks, of which we report on results for the following enrollment-test conditions: 120s, 30s, 10s, and 3s. It should be noted here that for the 120s condition, there are 6 enrollment sessions of 120s of speech and one test session of 120s of speech. This applies similarly to the other durations. The total number of trials for each of these conditions are: 333k, 163k, 173k, and 172k, respectively. It is worth remarking here that there are no cross-gender or cross-language trials in our evaluations.

For speech parameterization, we extract 19-dimensional power normalized cepstral coefficients (PNCC) [23] from 20 ms frames every 10 ms using a 24-channel Gammatone filterbank spanning the frequency range 125-3700 Hz. The first and second temporal cepstral derivatives are also computed over a 5-frame window and appended to the static features to capture the dynamic pattern of speech over time. This results in 57-dimensional feature vectors. For non-speech frame dropping, we employ an unsupervised speech activity detector (SAD) that generates frame-level decisions using multiple thresholds set on various basic speech features including frame log-energy, spectral divergence, and signal-to-noise ratio. After dropping the non-speech frames, global (utterance level) cepstral mean and variance normalization (CMVN) is applied to suppress the short-term linear channel effects.

We perform our experiments in the context of a state-of-the-art i-vector based speaker recognition system [10]. To learn the i-vector extractor, a gender-independent 1024-component GMM-UBM with diagonal covariance matrices is trained using a subset of the development set. The zeroth and first order Baum-Welch statistics are then computed for each recording and used to learn a 400-dimensional total variability subspace. After extracting 400-dimensional i-vectors, we either use LDA or NDA for inter-session variability compensation. The dimensionality reduced i-vectors are then centered (the mean is removed) and unit-length normalized. For scoring, a Gaussian PLDA model with full covariance residual noise term [11, 13] is learned using the i-vectors extracted from 13,158 speech segments from 743 unique speakers. The Eigenvoice subspace in the PLDA model is assumed full-rank. We include segments from 120s, 30s and 10s cuts in the PLDA training to expose our model to the duration conditions seen in the evaluation data. The number of sessions per speaker ranges from 5 to 25 segments, and we employ a one-

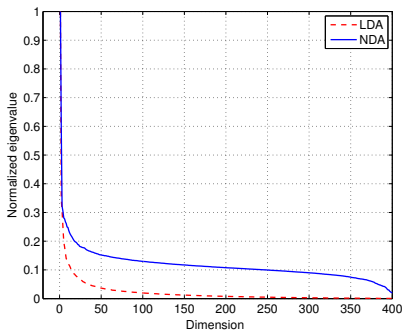

Figure 2: *Normalized and sorted eigenvalues of* $\mathbf{S}_w^{-1}\mathbf{S}_b$ *for the LDA transform (dashed), and* $\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b$ *for the NDA transform (solid) obtained with our development i-vectors.*

versus-rest strategy to compute the inter-speaker scatter matrix in (7). This provides flexibility on the number nearest neighbors used for computing the local means.

6. Results

In this section we summarize our results obtained with the experimental setup presented in Section 5. Figure 2 shows the normalized and sorted eigenvalues for $\mathbf{S}_w^{-1}\mathbf{S}_b$ in LDA (dashed), and $\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b$ in NDA (solid), which are obtained with our development i-vectors. Comparing the two curves, it is seen that the decay in the eigenvalues of the LDA transform occurs more rapidly than that of the NDA transform. In other words, the speaker-discriminative information for LDA is confined within a lower dimensional subspace compared to NDA. Note that from our discussion in Section 4, in NDA the complex local structure (as opposed to the global mean scatter) within the data is preserved and the class boundary information is embedded into the transform. Our hypothesis is that a larger subspace is required for NDA to properly represent such a complex structure. Accordingly, we expect NDA to perform best at larger feature dimensions (in the transformed space) compared to LDA.

Results of our speaker recognition experiments with LDA and NDA ($K = 9$) on RATS data are summarized in Tables 1 and 2, respectively. The results are reported in terms of false-reject (miss) rate at 2.5% false-alarm (FA) rate (miss@2.5%FA), which is a RATS program metric for the speaker recognition task. These results are generated by varying the dimensionality of the transformed subspace from 150 to 400 with an increment of 50. Two important observations can be made from the results presented in the tables. First, irrespective of the dimensionality of the feature subspace, NDA consistently performs better than LDA across the four duration-specific conditions. As we discussed before, this is due to the nonparametric representations for the scatter matrices in NDA that makes no assumption regarding the underlying class-conditional distributions. In addition, NDA is more effective in capturing the local structure and boundary information within and across different speakers. Second, LDA performs best with a 200-dimensional feature space, while the best performance for NDA is achieved with a 250-dimensional transform. As noted above, more dimensions are required for NDA to capture the boundary structure and local information.

We also investigated the impact of the number of nearest neighbors used for computing NDA on speaker recognition performance. The results are provided in Table 3 for K ranging from 3 to 13 with an increment of 2. Here, the dimensionality of the transformed subspace is fixed at 250. For longer duration tasks (i.e., 120s and 30s), the best performance is obtained with $K = 9$, while for 10s and 3s tasks the best performance is achieved with $K = 7$. It is evident that for the best performance to be achieved, the number of nearest neighbors, K, should be tuned. In practice, this is typically accomplished using a development set.

Table 1: Speaker recognition performance across the four enrollment-test conditions with LDA and different feature dimensions ranging from 150 to 400, in terms of percent miss@2.5%FA.

Enroll-Test	miss@2.5% FA [%]					
	150	200	250	300	350	400
120s-120s	3.94	**3.87**	3.91	3.97	4.14	4.17
30s-30s	9.71	**9.54**	9.81	9.83	10.16	10.28
10s-10s	23.08	**22.62**	23.15	23.29	24.09	24.58
3s-3s	**51.73**	51.89	52.63	53.17	53.98	54.83

Table 2: Speaker recognition performance across the four enrollment-test conditions with NDA ($K = 9$) and different feature dimensions ranging from 150 to 400.

Enroll-Test	miss@2.5% FA [%]					
	150	200	250	300	350	400
120s-120s	3.79	3.57	**3.49**	3.65	3.67	3.65
30s-30s	9.57	9.04	**8.89**	9.15	9.10	9.08
10s-10s	22.41	21.91	**21.85**	22.21	22.40	21.82
3s-3s	50.47	50.40	**50.40**	51.17	51.36	51.33

7. Conclusions

LDA has become an integral part of many state-of-the-art speaker recognition systems for inter-session variability compensation. Given that the actual distribution of i-vectors may not be necessarily Gaussian as well as the limitations identified for the parametric LDA, we presented an alternative nonparametric discriminant analysis (NDA) technique that measures both the within- and between-speaker variation on a local basis using the nearest neighbor rule. Unlike LDA, the NDA approach makes no specific assumption regarding the underlying class-conditional distributions. To evaluate the efficacy of NDA, we conducted speaker recognition experiments using actual noisy and channel degraded data from the RATS program. Experimental results indicated effectiveness of NDA against LDA for speaker recognition tasks. A clear advantage of NDA over LDA is that it is generally of full rank, making it attractive for speech applications (such as language recognition) with a limited number of classes. Our preliminary experiments also confirm the effectiveness of NDA for RATS language recognition tasks where the number of language classes is much smaller than the dimensionality of i-vectors.

Table 3: Speaker recognition performance for different number of nearest neighbors, K, in NDA with 250-dimensional features.

Enroll-Test	miss@2.5% FA [%]					
	3	5	7	9	11	13
120s-120s	4.03	3.81	3.58	**3.49**	3.51	3.61
30s-30s	9.72	9.54	9.18	**8.89**	8.99	9.16
10s-10s	22.80	22.41	**21.68**	21.85	21.73	21.94
3s-3s	51.04	50.65	**50.22**	50.40	50.76	50.64

8. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, 2007.

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

[3] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[4] P. Matějka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE ICASSP*, Prague, Czech, May 2011, pp. 4828–4831.

[5] A. Kanagasundaram, D. B. Dean, R. J. Vogt, M. Mclaren, S. Sridharan, and M. Mason, "Weighted LDA techniques for i-vector based speaker verification," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4781–4784.

[6] A. Kanagasundaram, D. B. Dean, S. Sridharan, and R. J. Vogt, "PLDA based speaker verification with weighted LDA techniques," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2012)*, Singapore, Singapore, June 2012.

[7] J. Yang, C. Liang, L. Yang, H. Suo, J. Wang, and Y. Yan, "Factor analysis of Laplacian approach for speaker recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4221–4224.

[8] M. McLaren and D. Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 755–766, 2012.

[9] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. IEEE ICASSP*, Vancouver, BC, May 2013, pp. 6773–6777.

[10] W. Zhu, S. Yaman, and J. Pelecanos, "The IBM RATS phase II speaker recognition system: overview and analysis," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 3137–3141.

[11] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE ICCV*, Rio De Janeiro, October 2007, pp. 1–8.

[12] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech, June 2010.

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

[14] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, 1983.

[15] M. Bressan and J. Vitria, "Nonparametric discriminant analysis and nearest neighbor classification," *Pattern Recognition Lett.*, vol. 24, no. 15, pp. 2743–2749, 2003.

[16] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, 2009.

[17] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2012)*, Singapore, Singapore, June 2012.

[18] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, 2000.

[19] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, 2005.

[20] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[21] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *Proc. INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 1242–1245.

[22] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic press, 1990.

[23] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4101–4104.