NEAREST NEIGHBOR DISCRIMINANT ANALYSIS FOR LANGUAGE RECOGNITION

Seyed Omid Sadjadi, Jason W. Pelecanos, Sriram Ganapathy

IBM Research, Watson Group

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

ABSTRACT

Many state-of-the-art i-vector based voice biometric systems use linear discriminant analysis (LDA) as a post-processing stage to increase the computational efficiency in the back-end via dimensionality reduction, as well as annihilate the undesired (noisy) directions in the total variability subspace. The traditional approach for computing the LDA transform uses parametric representations for both intra- and inter-class 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. In addition, the rank of the LDA projection (i.e., the maximum number of available discriminant bases) is limited to the number of classes minus 1. Accordingly, language recognition tasks on noisy data that involve only a few language classes receive limited or no benefit from the LDA post-processing. Motivated by this observation, we present an alternative non-parametric discriminant analysis (NDA) technique that measures both the within- and between-language variation on a local basis using the nearest neighbor rule. The effectiveness of the NDA method is evaluated in the context of noisy language recognition tasks using speech material from the DARPA Robust Automatic Transcription of Speech (RATS) program. Experimental results indicate that NDA is more effective than the traditional parametric LDA for language recognition under noisy and channel degraded conditions.

Index Terms— RATS, language recognition, nearest neighbor, discriminant analysis, i-vector

1. INTRODUCTION

State-of-the-art language recognition systems use i-vectors [1, 2] to represent variable-length acoustic signals in a fixed-length low-dimensional total variability subspace. I-vectors can be conveniently extracted from a variety of feature representations including cepstral (e.g., MFCCs), prosodic [3], neural network bottleneck [4, 5], and context-independent (monophones) [6] as well as context-dependent (senones) [7, 8] phone posterior feature vectors.

The i-vector approach models both signal (i.e., language) and noise (i.e., channel, session, etc) variabilities in the same total variability subspace, therefore an intersession compensation stage such as linear discriminant analysis (LDA) with the Fisher criterion [9] is typically applied on raw i-vectors to eliminate the undesired noisy directions, thereby maximizing inter-class separation [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-language) 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 that the actual distribution of i-vectors may not be necessarily Gaussian [10], and this assumption in particular is violated when short-duration speech recordings are collected in the presence of noise and channel distortions [11]. 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 [12], a nearest neighbor based 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 [13, 14] and speaker [15] recognition tasks. 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 (language) problem, the parametric LDA can provide at most $C-1$ discriminant features (i.e., the number of classes minus 1), which can render LDA less effective for language recognition tasks that involve only a few language categories. 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.

Heterocedastic LDA (HLDA) [16] is yet another alternative to LDA that removes the equal covariance constraint for class-conditional distributions. However, it still assumes Gaussian distribution for the classes. It was shown in [17] that post-processing i-vectors with HLDA can significantly improve the performance for automatic accent detection. Neighborhood component analysis (NCA) has also been studied for language recognition [2]. One major issue with NCA is that its effectiveness is highly dependent on an initial guess for the transformation matrix. In [2], the LDA solution was chosen as the initial guess, however as noted previously this can potentially limit the applicability for language recognition tasks with small number of classes.

In this study, we investigate the application of NDA for language recognition tasks under actual noisy and channel degraded conditions using speech material from the DARPA program, Robust Automatic Transcription of Speech (RATS) [18]. The RATS data consists of HF radio quality speech recordings spoken in 5 target languages. We conduct our language recognition experiments with a state-of-the-art i-vector based system [4] that uses bottleneck (BN) features extracted from a convolutional neural network (CNN) phoneme rec-

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.

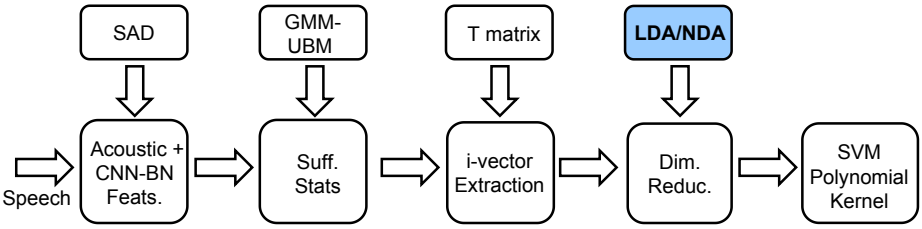

Fig. 1. Schematic block diagram of the language recognition system with CNN bottleneck features and dimensionality reduction.

ognizer (see Fig. 1). We report on average detection cost (Cavg) as defined by NIST for open-set language recognition evaluation (LRE) [19]. 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 language recognition performance.

2. 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 $\mathbf{A}$: $\mathbb{R}^d \mapsto \mathbb{R}^n$ by maximizing the ratio of the inter-class scatter to intra-class variance:

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}, \tag{1}$$

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 [20], of which the following is the most widely used,

$$\hat{\mathbf{A}} = \underset{\mathbf{A}^T\mathbf{S}_w\mathbf{A}=\mathbf{I}}{\arg\max} \left[\mathrm{tr}\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)\right], \tag{2}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between- and within- class scatter matrices, respectively. The optimization problem in (2) has an analytical solution that is a matrix whose columns are the n eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

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

$$\mathbf{S}_w = \sum_{i=1}^{C} p_i \mathrm{E}\left[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mid C_i\right] = \sum_{i=1}^{C} p_i \boldsymbol{\Sigma}_i, \tag{3}$$

where p_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are the *a priori* probability (proportional to the number of sessions per language category), 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,

$$\mathbf{S}_b = \sum_{i=1}^{C} p_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T, \tag{4}$$

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

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}] = \sum_{i=1}^{C} p_i\boldsymbol{\mu}_i. \tag{5}$$

There are three disadvantages associated with the parametric nature of the scatter matrices in (3) and (4). First, the underlying distribution of classes is assumed to be Gaussian with a common covariance matrix for all classes. Therefore, the parametric LDA does not generalize well to non-Gaussian and multi-modal (as opposed to unimodal) distributions. Second, 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 6 language categories (including 5 target and one pooled impostor classes) in the RATS program, and LDA can at most provide a 5-dimensional discriminative feature subset that may not effectively separate the language classes. Finally, because only the class centroids are taken into account for computing $\mathbf{S}_b$ in (4), the parametric LDA may not effectively capture the boundary structure between adjacent classes which is essential for classification [20].

To overcome the above noted limitations of LDA, a nearest neighbor based discriminant analysis technique was proposed in [12], 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.

3. NEAREST NEIGHBOR DISCRIMINANT ANALYSIS

To remedy the limitations identified for LDA, a nearest neighbor discriminant analysis techniques was proposed in [12]. 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[1]. More specifically, in the NDA approach, the between-class scatter matrix is defined as,

$$\tilde{\mathbf{S}}_b = \sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{l=1}^{N_i} w_l^{ij}\left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right)\left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right)^T, \tag{6}$$

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,

$$\mathcal{M}_l^{ij} = \frac{1}{K}\sum_{k=1}^{K} NN_k(\mathbf{x}_l^i, j), \tag{7}$$

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 (6) is defined as,

$$w_l^{ij} = \frac{\min\left\{d^\alpha\left(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)\right), d^\alpha\left(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j)\right)\right\}}{d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)) + d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j))}, \tag{8}$$

[1]In a recent work [21], an unsupervised channel adaptation method in i-vector space was formulated based on such local sample averages.

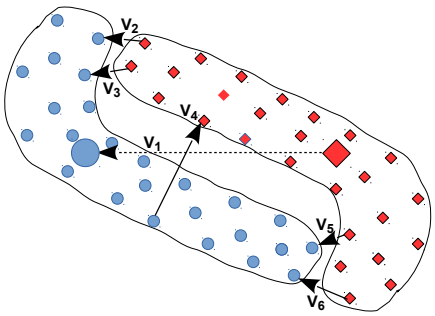

Fig. 2. Symbolic example illustrating the parametric versus nonparametric scatter between two classes. v_1 represents the global gradient of class centroids. The vectors $\{\mathrm{v}_2, \cdots, \mathrm{v}_6\}$ represent the local gradients.

where $\alpha \in \mathbb{R}$ is a constant between zero and infinity, and $d(.)$ denotes the Euclidean distance. The weighting function is introduced in (6) 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 2), while dropping off to 0 for samples that are far from the boundary (e.g., see v_4 in Figure 2). 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 (6), 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 (6). 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 2 (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.

4. EXPERIMENTS

This section provides a description of our experimental setup including speech data as well as the language recognition system used in our evaluations. We conduct our language recognition experiments using actual noisy and channel degraded speech material available from the DARPA RATS program, which is distributed by the LDC [18]. The RATS data consists of newly collected as well as existing recordings extracted from found corpora (CallFriend, Fisher, and NIST LRE) 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 three data releases are available from the LDC for system training: LDC2011E95, LDC2011E111, LDC2012E03, and a DEV-2 release which we use for evaluation (LDC2012E06). These releases contain speech spoken in five target languages: Levantine Arabic, Dari, Farsi, Pashto, and Urdu, as well as 10 impostor languages: Bengali, English, Japanese, Korean, Mandarin, Russian, Spanish, Tagalog, Thai, and Vietnamese. For system evaluation, there are 4 duration-specific conditions: 120s, 30s, 10s, and 3s. The total number of samples for duration-specific test conditions in DEV-2 set are: 1,914, 1,782, 1,715, and 1,340, respectively. Because the original data provided for training only contains 120s audio segments, we extract 30s, 10s, and 3s cuts from these to match the expected duration in evaluations. This process results in a total of 82,398 segments (duration and channel balanced) for system training.

For speech parameterization, we extract: i) 19-dimensional power normalized cepstral coefficients (PNCC) [22] from 32 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. ii) 25-dimensional BN features are extracted from a CNN based phoneme recognizer that is trained on Arabic Levantine data provided by the LDC for the keyword spotting (KWS) task in the RATS program.

These two feature vectors are then concatenated to form a 82-dimensional feature representation. For non-speech frame dropping, we employ a speech activity detector (SAD) that generates frame-level decisions using a deep neural network framework [23]. After dropping the non-speech frames, feature warping [24] is applied only to the cepstral features.

We perform our experiments in the context of a state-of-the-art i-vector based language recognition system [4]. To learn the i-vector extractor, a language and gender-independent 1024-component GMM-UBM with diagonal covariance matrices is trained using a subset of the training set (43,607 recordings). The zeroth and first order Baum-Welch statistics are then computed for each recording and used to learn a 250-dimensional total variability subspace. After extracting 250-dimensional i-vectors for the entire training set, 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, SVMs with 5^{th} order polynomial kernels are learned using the i-vectors extracted from the entire training set.

5. RESULTS

In this section we summarize our results obtained with the experimental setup presented in Section 4. Figure 3 displays t-SNE [25] scatter plots for i-vectors extracted from 120s cut in DEV-2 evaluation set. The left panel shows the language samples (i-vectors) in the

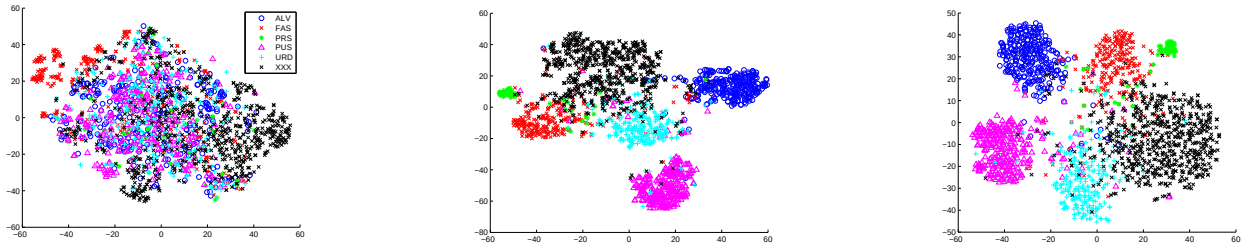

Fig. 3. Scatter plots of DEV-2 i-vectors (120s) for the original input space as well as LDA and NDA transformed spaces.

original input space (i.e., raw i-vectors). It can be seen that i) the distribution in this space is not Gaussian which is in line with the findings presented in [11], and ii) the language classes significantly overlap and can benefit from a discriminative transformation. The next two panels show scatter plots of dimensionality reduced i-vectors in the 2-dimensional plane. It can be seen that post-processing the i-vectors with either LDA or NDA increases class separation in the transformed space. Both the transforms provide a good degree of separation, however, as noted previously LDA can at most provide 5 discriminant features for language recognition on RATS. While this might be sufficient to perform language recognition on i-vectors extracted from longer duration cuts, it can result in degraded performance for shorter duration tasks. Our hypothesis is that a larger subspace is required to properly represent class separation information at the backend, hence we expect NDA to perform better than LDA, at least for shorter duration conditions (e.g., 3s).

Results of our language recognition experiments with NDA on RATS data are summarized in Tables 1, 2, for different feature dimensions as well as different number of nearest neighbors, respectively. For the sake of comparison, language recognition performance is also reported with LDA and locality preserving projection (LPP) [26] on the same tasks in Table 3 (the optimal dimensionality of each projection technique is also shown). The results are reported in terms of average detection cost (Cavg) for open-set language identification as defined in the NIST LRE. Several observations can be made from the results presented in the tables. First, irrespective of the dimensionality of the feature subspace or the number of nearest neighbors used in computing the scatter matrices, NDA based system consistently performs better than the baseline across the four duration-specific conditions. Second, it is evident from Table 2 that the number of nearest neighbors, K, can impact the performance, and should be optimized. In practice, this is typically accomplished using a development set. Third, NDA is more effective than both LDA and LPP, particularly for shorter durations. Not only does LDA provide no gain in performance for 3s test condition, it also significantly degrades the performance compared to the baseline system. As we discussed before, the superiority of NDA is due to the non-parametric representations for the scatter matrices that make 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 languages. This can specifically benefit language recognition systems that employ linear Gaussian models in the backend. Finally, as can be seen from Table 3, linear score level combination of the baseline and NDA systems with equal weights, results in further gains in performance.

Table 1. System performance, Cavg $\times 100$, with NDA ($K = 9$) at different feature dimensions.

Duration	Cavg [%]					
	140	160	180	200	220	250
120s	5.96	**5.71**	5.93	5.88	6.27	6.36
30s	6.74	**6.54**	6.75	6.89	6.95	7.04
10s	9.21	**8.77**	9.19	9.61	10.02	9.90
3s	12.72	12.75	12.88	**12.51**	12.69	12.83

Table 2. System performance, Cavg$\times$100, given the number of nearest neighbors, K, in NDA with 200-dimensional features.

Duration	Cavg [%]					
	3	5	7	9	11	13
120s	6.1	6.02	6.00	5.88	**5.84**	5.94
30s	6.97	6.90	6.95	6.89	**6.85**	6.83
10s	9.76	10.20	9.71	9.61	**9.41**	9.5
3s	12.66	12.62	12.71	**12.51**	12.8	12.57

Table 3. System performance, Cavg$\times$100, for the baseline as well as LDA and NDA based systems. Last column shows the results for the combination of the baseline and NDA based systems.

Duration	Cavg [%]				Base+
	Base-250	LDA-5	LPP-200	NDA-160	NDA
120s	6.30	5.82	6.03	**5.71**	5.82
30s	7.91	7.36	7.77	6.54	**6.04**
10s	10.56	9.85	9.89	8.77	**8.39**
3s	13.84	17.31	13.17	12.75	**12.30**

6. CONCLUSIONS

LDA has been widely applied to many state-of-the-art speaker and language recognition systems for inter-session variability compensation. However, in the recent DARPA RATS evaluations, almost none of the participants employed LDA in their language recognition systems. This is attributed to the limitations identified for the parametric scatter matrices that are formed based on class-conditional Gaussian distribution assumption. In addition, the small number of language categories in RATS limits the maximum number of discriminant bases available from LDA. We presented an alternative nearest neighbor discriminant analysis (NDA) technique that measures both the within- and between-language 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 language recognition experiments using actual noisy and channel degraded data from the RATS program. Experimental results indicated effectiveness of NDA against LDA for language recognition tasks. A clear advantage of NDA over LDA is that it is generally of full rank, making it attractive for speech applications with a limited number of classes.

7. REFERENCES

[1] 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.

[2] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. INTERSPEECH*, Florence, Italy, August 2011, pp. 857–860.

[3] A. Lawson, M. McLaren, Y. Lei, V. Mitra, N. Scheffer, L. Ferrer, and M. Graciarena, "Improving language identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 1507–1510.

[4] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. Segbroeck, and S. Narayanan, "Robust language identification using convolutional neural network features," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 1846–1850.

[5] P. Matejka, L. Zhang, T. Ng, O. Glembek, J. Ma, B. Zhang, and S. H. Mallidi, "Neural network bottleneck features for language identification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2014)*, June 2014, pp. 299–304.

[6] M. Díez, A. Varona, M. Peñagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, December 2012, pp. 274–279.

[7] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2014)*, Joensuu, Finland, June 2014, pp. 287–292.

[8] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Spoken language recognition based on senone posteriors," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 2150–2154.

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

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

[11] M. McLaren, A. Lawson, Y. Lei, and N. Scheffer, "Adaptive Gaussian backend for robust language identification," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 84–88.

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

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

[14] 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.

[15] S. O. Sadjadi, J. W. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 1860–1864.

[16] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283 – 297, 1998.

[17] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign accent detection from spoken Finnish using i-vectors," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 79–83.

[18] 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.

[19] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)," http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.

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

[21] W. Zhu, S. O. Sadjadi, and J. W. Pelecanos, "Nearest neighbor based i-vector normalization for robust speaker recognition under unseen channel conditions," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015.

[22] 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.

[23] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 3497–3501.

[24] J. W. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. The Speaker Recognition Workshop (Odyssey 2001)*, Crete, Greece, June 2001, pp. 213–218.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[26] X. He and P. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, pp. 153–160.