*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Extended Variability Modeling and Unsupervised Adaptation for PLDA Speaker Recognition

*Alan McCree, Gregory Sell, and Daniel Garcia-Romero*

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA
alan.mccree@jhu.edu, gsell@jhu.edu, dgromero@jhu.edu

## Abstract

Probabilistic Linear Discriminant Analysis (PLDA) continues to be the most effective approach for speaker recognition in the i-vector space. This paper extends the PLDA model to include both enrollment and test cut duration as well as to distinguish between session and channel variability. In addition, we address the task of unsupervised adaptation to unknown new domains in two ways: speaker-dependent PLDA parameters and cohort score normalization using Bayes rule. Experimental results on the NIST SRE16 task show that these principled techniques provide state-of-the-art performance with negligible increase in complexity over a PLDA baseline.

**Index Terms**: speaker recognition, i-vector, probabilistic linear discriminant analysis, Bayesian speaker comparison

## 1. Introduction

State-of-the-art speaker recognition systems model i-vectors [1] with variations of Probabilistic Linear Discriminant Analysis (PLDA) [2, 3, 4]. While much recent work has focused on using Deep Neural Networks to improve these i-vector extractors over the traditional acoustic ones (e.g. [5, 6]), there remains a need to fundamentally improve this PLDA model. In this paper, we present extensions to the basic model to jointly address the issues of non-independent speaker model enrollment cuts as well as limited duration of both enrollment and test cuts. In addition, we present two novel approaches to unsupervised domain adaptation in the presence of multiple unlabeled domains with only a limited amount of adaptation data. Since the latest NIST Speaker Recognition Evaluation, SRE16, invokes all of these issues, we validate performance of these new methods on this task.

The organization of this paper is as follows. Section 2 presents necessary mathematical background based on our previously-presented alternative formulation of PLDA which we refer to as Bayesian speaker comparison (BSC). Extension to the model to include non-independent enrollment cuts, audio duration, and their combination are then presented in Section 3. Section 4 describes our two new methods for exploiting limited unlabeled data for adapting the system parameters. In Section 5, experimental results validating all of these new methods on NIST SRE16 are presented. Finally, Section 6 provides our conclusions.

## 2. Background

Before introducing the new extensions to the PLDA model, we first introduce our notation by reviewing the equations for Bayesian speaker comparison (BSC) in i-vector space [7].

### 2.1. Bayesian Speaker Comparison

The generative model for BSC is the same as for PLDA: an observed i-vector $\mathbf{z}_n$ from a given speech cut is assumed to have been generated by a speaker model $\mathbf{m}_i$ corrupted by a channel noise $\mathbf{c}_n$:

$$\mathbf{z}_n = \mathbf{m}_i + \mathbf{c}_n, \tag{1}$$

where both the speaker and channel are drawn from Gaussian distributions:

$$\mathbf{m}_i \sim \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_m) \quad \text{and} \quad \mathbf{c}_n \sim \mathcal{N}(0, \mathbf{\Sigma}_c).$$

Under this model, the likelihood ratio for speaker $S_i$ on test i-vector $\mathbf{z}_n$ is given by

$$LR(S_i, \mathbf{z}_n) = \frac{p(\mathbf{z}_n|S_i)}{p(\mathbf{z}_n)} .$$

For the numerator of this ratio, we use the predictive distribution for this speaker given the enrollment data. Given this two Gaussian model with known covariances and a set of $N$ enrollment cuts, the posterior distribution of the speaker model $\mathbf{m}_i$ is Gaussian [8] with mean:

$$\mathbf{m}_d = \mathbf{\Sigma}_m \left(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{ml}\right)^{-1} \bar{\mathbf{z}}_{ml} + \mathbf{\Sigma}_{ml} \left(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{ml}\right)^{-1} \mathbf{m}_0 \tag{2}$$

and covariance:

$$\mathbf{\Sigma}_d = \mathbf{\Sigma}_m \left(\mathbf{\Sigma}_m + \mathbf{\Sigma}_{ml}\right)^{-1} \mathbf{\Sigma}_{ml} \tag{3}$$

where $\bar{\mathbf{z}}_{ml} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n$ and $\mathbf{\Sigma}_{ml} = \frac{\mathbf{\Sigma}_c}{N}$ represent the maximum likelihood (ML) mean estimate and the covariance of this estimator. Note that as the number of enrollment cuts becomes large, this posterior distribution approaches a delta function at the sample mean.

The likelihood ratio numerator, the predictive distribution, is again Gaussian:

$$\mathbf{z}_n|S_i \sim \mathcal{N}(\mathbf{m}_d, \mathbf{\Sigma}_c + \mathbf{\Sigma}_d). \tag{4}$$

The denominator is the likelihood that the test cut represents a random speaker in a random channel, and since both are independent and Gaussian, this is also Gaussian:

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_m + \mathbf{\Sigma}_c). \tag{5}$$

These equations are the ones used for Bayesian Speaker Comparison in [7]. Although they provide the same answer as the solution for PLDA, they differ significantly in form. Traditional PLDA scoring directly computes the likelihood ratio

http://dx.doi.org/10.21437/Interspeech.2017-1586

between the same vs. different hypotheses, and never explicitly computes model parameter distributions. We have previously found in language recognition that writing the equations in this way makes the relationship between Gaussian scoring and PLDA more clear and facilitates the use of MMI discriminative training [9]. Even for speaker recognition, we believe that an explicit probabilistic enrollment process provides valuable insight as well as facilitating speaker-dependent processing as shown in Section 4.

### 2.2. Dimension Reduction and Diagonalization

To reduce computation, this work uses diagonal covariance matrices as in [9], based on the fact that two symmetric matrices can be simultaneously diagonalized with a linear transformation [10]. This process is similar to Linear Discriminant Analysis (LDA) and is given by:

1. perform eigendecomposition $\mathbf{\Sigma}_c = E_1 \Lambda_1 E_1^T$
2. transform $\mathbf{\Sigma}_m$ with $\mathbf{\Sigma}'_m = \Lambda_1^{-\frac{1}{2}} E_1^T \mathbf{\Sigma}_m E_1 \Lambda_1^{-\frac{1}{2}}$
3. perform eigendecomposition $\mathbf{\Sigma}'_m = E_2 \Lambda_2 E_2^T$
4. (optional) keep only principal components
5. final transform: $\mathbf{z}'_n = E_2^T \Lambda_1^{-\frac{1}{2}} E_1^T \mathbf{z}_n$

In this transformed space, $\mathbf{\Sigma}_c = I$ and $\mathbf{\Sigma}_m = \Lambda_2$. Since the error criterion for LDA is to maximize $tr(\mathbf{\Sigma}_c^{-1}\mathbf{\Sigma}_m)$, keeping only the eigenvectors corresponding to the largest eigenvalues in step 4 finds the same subspace as traditional LDA, although the linear transformation of this diagonalized LDA is not identical. In general the LDA criterion only specifies the optimal subspace, not the coordinates within it.

## 3. The Extended Model

Starting from this mathematical background, we now introduce two new extensions to the basic BSC model to deal with the issues of multiple enrollment cuts and finite enrollment and test cut duration.

### 3.1. Non-independent Enrollment Cuts

One outstanding issue with the use of the PLDA model for speaker recognition is how to handle enrollment with multiple cuts for a speaker. According to the the theoretical model, each cut represents a new draw from the channel distribution and as a result the model uncertainty reduces quickly as more cuts are added. In practice, better performance is often achieved by ignoring this effect, instead pretending that only one cut is present and using the average of the enrollment i-vectors to represent this cut. Clearly one reason for this is the typical NIST experiment design; for example in the NIST SRE10 8c condition all eight enrollment cuts are from the same telephone number while the test trials are from a different number. We can extend PLDA to account for this effect by removing the independence assumption from the channel noise term.

Our extension of the generative model for dependent enrollment cuts still follows the original additive channel noise model of Eq. 1, but we now model $\mathbf{c}_n$ with a Markov chain. For each new enrollment cut $n$, we either reuse the same channel noise from cut $n-1$ with probability $P_r$ or we draw a new one from the channel distribution with probability $1 - P_r$. This can be thought of as the prior probability of a same or different phone number in enrollment. In principle a similar extension could be applied to the testing formulas as well, but we prefer to assume that the test cut is always from a different number.

For this new model, the enrollment still follows Eq. 2 with the same ML mean, but the covariance of this ML mean estimator now decreases more slowly with increasing number of enrollment cuts:

$$\begin{aligned}
\mathbf{\Sigma}_{ml} &= E\left\{(\bar{\mathbf{z}}_{ml} - \mathbf{m}_i)(\bar{\mathbf{z}}_{ml} - \mathbf{m}_i)^T\right\} && (6) \\
&= \frac{1}{N^2} E\left\{\sum_{n=1}^{N} \mathbf{c}_n \sum_{m=1}^{N} \mathbf{c}_m^T\right\} && (7) \\
&= \frac{\mathbf{\Sigma}_c}{N}\left(1 + 2\sum_{j=1}^{N-1} \frac{(N-j)}{N} P_r^j\right) && (8)
\end{aligned}$$

since $E\left\{\mathbf{c}_n \mathbf{c}_{n+j}^T\right\} = P_r^j \mathbf{\Sigma}_c$. Note that this Markov chain assumption gives the same result as an autoregressive noise process to model correlation between observations (see Equation 10 of our previous duration work [11]). For single-cut enrollment, this model reduces to the basic one, and for multiple cuts the new channel probability allows continuous variation between the two extremes of “by-the-book PLDA scoring” ($P_r = 0$) and “average i-vector scoring” ($P_r = 1$).

### 3.2. Duration

A second need in speaker recognition is to model the finite duration of both enrollment and test speech cuts. To incorporate duration into the BSC/PLDA model, we extend the generative model for each observed i-vector by adding an additional observation noise term:

$$\mathbf{z}_n = \mathbf{m}_i + \mathbf{c}_n + \mathbf{o}_n$$

with nonstationary independent Gaussian observation noise having a cut-dependent covariance

$$\mathbf{o}_n \sim \mathcal{N}(0, \mathbf{\Sigma}_n).$$

In previous work, the i-vector posterior covariance has been used for $\mathbf{\Sigma}_n$ [12, 13]. While effective, this has the computational disadvantage that a full rank covariance matrix must be stored along with each i-vector, and it becomes impossible to perform joint diagonalization since there are more than two covariance matrices. We propose a simpler model: that the observation noise covariance is simply a scaled version of the channel covariance, where the scaling factor is a function of cut duration. We have previously had success with a duration-dependent score calibration using a function of the form $t_n/(t_n + T_0)$ where $t_n$ represent the duration of cut $n$ and $T_0$ is a free parameter defining the typical duration behavior of the system (and therefore the amount of observation noise) [11, 14]. Since this is equivalent to a test covariance scaled by $(t_n + T_0)/t_n = 1 + T_0/t_n$, our observation noise model is given by

$$\mathbf{\Sigma}_n = \frac{T_0}{t_n}\mathbf{\Sigma}_c$$

and the total noise for each cut is

$$\mathbf{\Sigma}_c + \mathbf{\Sigma}_n = (1 + T_0/t_n)\mathbf{\Sigma}_c$$

Note that this function has the desired uncertainty behavior: it is very large for short durations, while for long cuts the observation noise disappears and we revert to the traditional PLDA model.

For duration-modeling PLDA, the extension to the testing likelihood ratio is given by:

$$\mathbf{z}_n | S_i \sim \mathcal{N}(\mathbf{m}_d, \mathbf{\Sigma}_c + \mathbf{\Sigma}_n + \mathbf{\Sigma}_d).$$

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_m + \mathbf{\Sigma}_c + \mathbf{\Sigma}_n)$$

For our simplified model, these become

$$\mathbf{z}_n | S_i \sim \mathcal{N}(\mathbf{m}_d, (1 + T_0/t_n)\mathbf{\Sigma}_c + \mathbf{\Sigma}_d) \quad (9)$$

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_m + (1 + T_0/t_n)\mathbf{\Sigma}_c). \quad (10)$$

If only the test cut has short duration, these are the only modifications needed. However, this model can also handle enrollment cut durations by extending the speaker model posterior distribution estimation. With general observation noise, this mean and covariance are considerably more complicated than the basic BSC model [15]. However, for our simpler model the equations are straightforward. For the duration-extended BSC model, the enrollment still follows Eq. 2 but the ML mean weights each cut inversely to the total noise within it:

$$\bar{\mathbf{z}}_{ml} = W^{-1} \sum_{n=1}^{N} w_n \mathbf{z}_n \quad (11)$$

where $w_n = (1 + T_0/t_n)^{-1}$ and $W = \sum_{n=1}^{N} w_n$. For the associated covariance with independent channel and observation noises,

$$\begin{aligned} \mathbf{\Sigma}_{ml} &= \frac{1}{W^2} E \left\{ \sum_{n=1}^{N} w_n(\mathbf{c}_n + \mathbf{o}_n) \sum_{m=1}^{N} w_m(\mathbf{c}_m + \mathbf{o}_m)^T \right\} \\ &= \frac{1}{W^2} \sum_{n=1}^{N} w_n^2 (w_n^{-1}) \mathbf{\Sigma}_c \\ &= W^{-1} \mathbf{\Sigma}_c \end{aligned}$$

Notice that this model reverts to traditional PLDA as either the cut duration goes to infinity or the parameter $T_0$ goes to zero.

### 3.3. Combination

While the number of cuts and duration modeling each serve their own purpose, it is more powerful to combine the two. In this way, same phone number cuts are modeled as only differing in observation noise due to finite duration, while different number cuts use an additional channel noise. For this joint model, the test likelihood ratio is still given by the duration model Equations 9 and 10. The enrollment ML mean also uses the noise-weighted sum in Eq. 11, but the corresponding covariance needs to account for the fact that observation noise is independent across cuts while channel noise is Markov:

$$\begin{aligned} \mathbf{\Sigma}_{ml} &= \frac{1}{W^2} E \left\{ \sum_{n=1}^{N} w_n(\mathbf{c}_n + \mathbf{o}_n) \sum_{m=1}^{N} w_m(\mathbf{c}_m + \mathbf{o}_m)^T \right\} \\ &= \frac{\mathbf{\Sigma}_c}{W^2} \left( W + 2 \sum_{j=1}^{N-1} P_r^j \sum_{n=1}^{N-j} w_n w_{n+j} \right) \end{aligned}$$

since $E\left\{(\mathbf{c}_n + \mathbf{o}_n)(\mathbf{c}_{n+j} + \mathbf{o}_{n+j})^T\right\} = P_r^j \mathbf{\Sigma}_c$.

This extended model is simple to implement since it preserves the joint diagonal properties of the original PLDA and only requires two additional parameters. $T_0$ controls the amount of observation noise which increases uncertainty with short durations, and $P_r$ models the probability that enrollment cuts are from the same phone channel.

## 4. Unsupervised Adaptation

Previous work has shown great success in mitigating domain shifts in speaker recognition [16, 17]. In particular, adaptive centering and whitening followed by adaptation of PLDA parameters was very effective in the Domain Adaptation Challenge [16]. Even in the absence of speaker labels for the adaptation data, generating approximate labels by hierarchical clustering using an initial PLDA-based metric allowed the same adaptation techniques to work nearly as well [17]. However, these techniques are only designed to perform well on the new domain, regardless of degradation to the old one. In addition, they assume two things:

- each audio cut is labeled with a domain
- a rich and diverse training set is provided in each domain with multiple cuts per speaker.

In this work we explore unsupervised adaptation techniques which do not require these assumptions.

### 4.1. PLDA Adaptation with Unknown Domain

As in the Domain Adaptation Challenge, we assume a scenario where we are given a large, fully-labeled set of training data for a PLDA model, but the actual enrollment and test data is from one or more new domains for which we have a smaller amount of unlabeled training data. However, unlike in that work, these new speakers could come from a number of different domains, and we are not given labels for these domains.

We propose a simple solution to this problem: *speaker-dependent PLDA*. For each enrollment speaker $S_i$, we find the K nearest neighbors in the unlabeled adaptation data, assume these represent the appropriate speaker domain $D_i$, and adapt the PLDA parameters for this domain. Nearest neighbors are defined by Euclidean distance from the enrollment sample mean in the jointly diagonalized subspace. We use these speaker-dependent parameters for the enrollment process, and save them so they can also be used at test time. Since our BSC formulation separates the enrollment and test process, the nearest-neighbor search and parameter adaptation are performed offline during enrollment. Note that this new domain affects both numerator and denominator in the likelihood ratio:

$$LR(S_i, \mathbf{z}_n) = \frac{p(\mathbf{z}_n | S_i, D_i)}{p(\mathbf{z}_n | D_i)}.$$

The previous discussion assumes that the test impostors will have been selected to come from the same domain as the enrollment speaker, a common NIST SRE paradigm. More generally, we can assume a prior probability of out-of-domain impostors $P_{ood}$ and use the extended testing formula:

$$LR(S_i, \mathbf{z}_n) = \frac{p(\mathbf{z}_n | S_i, D_i)}{P_{ood} p(\mathbf{z}_n) + (1 - P_{ood}) p(\mathbf{z}_n | D_i)}.$$

Overall this speaker-dependent PLDA provides a powerful and flexible technique. In limited-data scenarios, simplified versions can be used. In particular, in our experiments so far with this approach we adapted only the mean parameter $\mathbf{m}_0$ because of the small unlabeled data pile available.

### 4.2. Closed Set Non-target Scoring

Another way to use adaptation data is as non-target or impostor models. In language identification, it is a standard practice to use the closed-set of languages $L_k$ to define the overall likelihood of the data [9] with

$$p(\mathbf{z}_n) = \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{z}_n|\, L_k)$$

Since we do not typically know the full set of possible models in speaker recognition (and in fact are prohibited by NIST SRE evaluation rules from using other enrollment speakers), we can instead use a large number of out-of-set models on the grounds that:

$$p(\mathbf{z}_n) = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{z}_n|\, S_k)$$

Since this approach can be viewed as a form of cohort score normalization using Bayes' rule, we call it *Bnorm.* Using this to replace Equation 5 in the testing likelihood ratio is convenient for unsupervised adaptation with limited unlabeled training data, since it allows the enrollment model to compete against comparable cohorts rather than a large Gaussian distribution trained in a mismatched domain.

In contrast to the more traditional techniques of model or score normalization (znorm, Tnorm and Snorm), Bnorm has a simple theoretical justification. In addition, we have found no need for adaptive cohort selection with this method, since distant speaker models simply contribute nothing in the linear probability sum in contrast to their significant impact on logarithmic score statistics. Finally, in practice we find some performance improvement by combining the traditional open set scoring with this cohort approach by linear fusion of the two different log likelihood ratios.

## 5. Experimental Results

For experimental analysis of these techniques, we report performance on the 2016 NIST Speaker Recognition Evaluation (SRE) [18]. Like previous SREs, this evaluation focused on telephone speech recorded from a variety of handset types. However, for the first time this evaluation used entirely non-English data recorded outside North America. This introduced a domain adaptation element to the evaluation, since the fixed training condition required systems to be built primarily from older North American English speech material. The only new material allowed was a small (200 call) labeled development set in two related languages (Cebuano and Mandarin) and an unlabeled set from about 2200 calls mostly in the two test languages of Tagalog and Cantonese. Two additional elements of this evaluation are relevant to this work. First, single cut and three cut enrollment conditions were combined, in contrast to previous SREs where single and multicut enrollment were treated as separate conditions. Second, short durations were exercised by test segments from 10 to 60 seconds and to a lesser extent by enrollment segments of approximately 60 seconds.

Our baseline system for these experiments is a PLDA acoustic i-vector system. Like many other sites, we experimented with DNN i-vector systems as well but were unable to improve over this acoustic baseline system in this SRE16 fixed data training scenario. The UBM is trained on the unlabeled data segments, the i-vector extractor (T matrix) was trained on Fisher English, and PLDA parameters were estimated using speakers from past NIST SREs. This gender-independent system uses 40 MFCC features and a 2048 mixture UBM with 600-dimensional i-vectors and full-rank PLDA. Based on our previous domain adaptation work, this system uses in-domain whitening and length normalization with the unlabeled development data, which we found to be essential to attain a good baseline for this task. We report performance using the NIST speaker detection primary metric $C_{primary}$, which sums the balanced detection errors at two different operating points of $P_{target} = 0.01$ and $0.005$.

Table 1: *Performance of baseline and enhanced PLDA systems on NIST SRE16 task.*

| System | $C_{primary}$ |
|---|---|
| Baseline | 0.76 |
| Duration model (D) | 0.73 |
| Multi-cut enrollment (MC) | 0.72 |
| Duration+multicut (D+MC) | 0.72 |
| D+MC+local mean (L) | 0.69 |
| D+MC+Bnorm (B) | 0.69 |
| D+MC+B+L | 0.67 |

Results for the NIST metric $C_{primary}$ are shown in Table 1. Since the test segments are relatively short, the new duration model with the value used in [14] of $T_0 = 3$ provides noticeable improvement. Markov modeling of the enrollment channel with $P_r = 0.5$ also gives gains. The joint modeling of both duration and enrollment maintains this improvement. Speaker-dependent PLDA using 200 nearest neighbors from the unlabeled adaptation list to generate a local mean $\mathbf{m}_0$ provides further improvement. The alternative unsupervised adaptation method of cohort normalization with Bayes' rule provides similar gain, and the combination of all these techniques provides the best performance with $C_{primary} = 0.67$.

These numbers are very competitive with top individual system performance numbers presented at the NIST SRE16 workshop. Two top techniques emerged from that workshop: adaptive score normalization and unsupervised clustering. Adaptive score normalization was also popular in NIST SRE08, but has since fallen out of favor since it is fundamentally heuristic and notoriously unreliable across different experiments. Unsupervised clustering is a well-justified technique that has often proven effective, however it was shown by NIST at the workshop that this list does not contain nearly enough cuts per speaker to be useful for PLDA estimation, and in fact it was critical to cluster to far fewer than the true number of speakers to attain good performance. The new techniques presented in this paper provide comparable performance using a consistent theoretical framework that we expect to generalize well to new tasks.

## 6. Conclusion

This paper has presented a number of extensions to the PLDA model to improve speaker recognition performance. With only two additional parameters, we can model the same/different phone number enrollment process as well as the duration of both enrollment and test cuts, while still preserving the computational advantages of joint diagonalization. For unsupervised adaptation, the combination of speaker-dependent PLDA mean parameters and Bayes' rule cohort score normalization provides a simple and effective approach for unknown new domains. Performance results on NIST SRE16 confirm that these principled techniques provide state-of-the-art performance for this task.

## 7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, May 2011.

[2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

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

[4] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. ICASSP*, 2011, pp. 4832–4835.

[5] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[6] D. Garcia-Romero and A. McCree, "Insights into deep neural networks for speaker recognition," in *Interspeech*, 2015.

[7] B. J. Borgstrom and A. McCree, "Discriminatively trained Bayesian speaker comparison of i-vectors," in *Proc. ICASSP*, 2013.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2001.

[9] A. McCree, "Multiclass discriminative training of i-vector language recognition," in *Proc. Odyssey*, 2014, pp. 166–172.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[11] A. McCree, F. Richardson, E. Singer, and D. Reynolds, "Beyond frame independence: Parametric modeling of time duration in speaker and language recognition," in *Proc. Interspeech*, 2008, pp. 767–770.

[12] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md Jahangir Alam, and Pierre Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7649–7653.

[13] Sandro Cumani, Oldřich Plchot, and Pietro Laface, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 846–857, 2014.

[14] A. McCree, G. Sell, and D. Garcia-Romero, "Augmented Data Training of Joint Acoustic/Phonotactic DNN i-vectors for NIST LRE15," in *Proc. of IEEE Odyssey*, 2016.

[15] B. J. Borgstrom and A. McCree, "Supervector Bayesian speaker comparison," in *Proc. ICASSP*, 2013.

[16] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4047–4051.

[17] 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.

[18] "The NIST year 2016 speaker recognition evaluation plan," `http://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016`.