A PLDA Approach for Language and Text Independent Speaker Recognition

Abbas Khosravani[1], *Mohammad Mehdi Homayounpour*[1],
Dijana Petrovska-Delacréta[2], *Gérard Chollet*[3]

[1]Laboratory for Intelligent Multimedia Processing, Amirkabir University of Technology, Iran
[2]Institut Mines-Télécom, Télécom SudParis, France
[3]CNRS-LTCI, Institut Mines-Télécom, France and Intelligent Voice Ltd., England
{a.khosravani,homayoun}@aut.ac.ir

Abstract

There are many factors affecting the variability of an *i*-vector extracted from a speech segment such as the acoustic content, segment duration, handset type and background noise. The state-of-the-art Probabilistic Linear Discriminant Analysis (PLDA) aims at modelling all these sources of undesirable variability within a single covariance matrix. Although techniques such as source normalization have been proposed to reduce the effect of different sources of variability as a pre-processing for PLDA, still the performance of speaker recognition is affected under cross-source evaluation condition.

This study aims at proposing a language-independent PLDA training algorithm in order to reduce the effect of language on the performance of speaker recognition. An accurate estimation of speaker and channel subspaces from a multilingual training data set which are void of language variability can assist PLDA to work independent of the language. When evaluated on the NIST 2008 speaker recognition multilingual trials, our proposed solution demonstrates relative improvement of up to 10% in equal error rate (EER) and 6.4% in minimum DCF.

1. Introduction

Over recent years, *i*-vector representation of speech segments has been widely used by the state-of-the-art speaker recognition systems [1]. This representation provides an elegant way to map arbitrary duration speech segments into a fixed-length and low-dimensional vector that preserves the speaker information. This can be accomplished by using Factor Analysis (FA) technique to learn a low-dimensional subspace from a large collection of development data. The speaker recognition technology based on *i*-vectors currently dominates the research field due to its state-of-the-art performance, low computational cost and the suitability of *i*-vector for machine learning techniques. The recent NIST *i*-vector machine learning challenge [2] was also performed to measure state-of-the-art performance and find the most promising algorithmic approaches on the basis of *i*-vectors [3, 4, 5, 6].

Although current text-independent speaker recognition systems are considered to be independent of the language being spoken, their performance will be affected in multilingual trial conditions. This was the focus of NIST Speaker Recognition Evaluation (SRE) in 2008 [7]. Due to the availability of large corpora in English that contain a large number of speakers with multiple recordings for each speaker in different degradation conditions, we can see much better performance for English than multilingual trials [1]. The current session compensation techniques such as Within-Class Covariance Normalization (WCCN), Linear Discriminant Analysis (LDA) or Probabilistic LDA (PLDA) [8] aim to improve system robustness by alleviating unwanted variability induced by factors such as transmission channel, background noise, and speaker characteristics (health, age, language) from *i*-vectors, but they are all dependent on a development data with such characteristics. However, lack of multilingual utterances for each speaker in system development will restrict current techniques to model language source of variability and results in performance degradation.

The standard PLDA algorithm provides a powerful mechanism in extracting speaker-specific information from all other sources of undesired variability in *i*-vector space. Specifically, PLDA uses within-speaker and between-speaker variability observed in different utterances of individual speakers to find corresponding subspaces in the *i*-vector space. Therefore, PLDA requires multiple utterances for each speaker under different degradation conditions to be able to properly model all kinds of variability. However, providing such a resource might be too expensive or even unrealistic. To tackle this problem, a number of techniques have been proposed to compensate for the lack of adequate data which include source-normalization and domain-adaptation.

Source normalization technique has been proposed as to improve the estimation of within-speaker scatter matrix from a training database with insufficient variety of speaker utterances from different sources [9]. The within-speaker variability is computed as the residual total variability in the *i*-vector space that is not captured by between-speaker variability. The between speaker variability is then computed on a source conditioned basis to remove the bias toward a specific source. This technique has been incorporated into Within-Class Covariance Normalization (WCCN) [10] as well as Linear Discriminant Analysis (LDA) in order to improve the robustness of *i*-vector based speaker recognition under cross-speech-source conditions. Language-normalized WCCN (LN-WCCN) [11] was proposed as an *i*-vector pre-processing stage prior to PLDA which provides speaker recognition improvement under multilingual scenarios. Source-normalized LDA (SN-LDA) [12] technique has also been proposed to enhance standard LDA algorithm by reducing the influence of source variations on the between-speaker scatter matrix as well as incomplete representation of within-class scatter matrix due to insufficient cross-source utterances for each speaker in the training set.

Beside source-normalization, domain adaptation technique has also gained considerable attention to compensate for the cross-speech source variability of in-domain and out-of-domain data. An unsupervised adaptation of LDA matrix to unseen data domain using Within-speaker Covariance Correction (WCC)

has been proposed in [13]. In [14, 15] the authors presented a framework for supervised and unsupervised adaptation of out-of-domain PLDA parameters to produce better performance for in-domain data. After analyzing the sources of degradation, Hagai Aronowitz found that the main source of degradation is a shift in dataset [16] and based on this finding, he proposed an inter-dataset variability compensation technique to compensate for this shift. A new inter-dataset variability compensation approach has also been proposed in [17].

In this work, we extended PLDA paradigm to make it robust with respect to *i*-vectors extracted from multilingual speech utterances through direct modeling of language variability. Instead of modeling within-speaker variability in a single subspace, we proposed to separate language source of variability from all other sources of variability refered to as channel variability by using a new term for language in PLDA factorization. Therefore, each speaker utterance will be characterized by a set of speaker, language, and channel factors. A similar separation of language from channel variability for the task of language recognition based on the original Joint Factor Analysis (JFA) has been investigated in [18] where only language factors which supposed to contain all relevant language information were fed into a language classifier. By capturing and removing language source of variability, we expect the PLDA to work independent of the language being spoken in an utterance. The proposed language-independent PLDA (LI-PLDA) when provided with multilingual training data, indicates performance gain on NIST SRE'08 multilingual telephony trials compared to the standard PLDA modeling. Our experiments also showed that the use of language source normalization prior to LI-PLDA could complement the proposed method and result in even better performance.

The paper is organized as follows. Section 2 reviews the *i*-vector/PLDA speaker recognition system and presents a formal mathematical formulation, Section 3 describes the language-normalized WCCN (LN-WCCN) and our proposed language independent PLDA modeling as techniques to reduce the effect of languages. In Section 4 we explain how experiments were conducted and present the results in Section 5. Analysis and discussion is given in Section 6.

2. Speaker Recognition System

In this section we will provide a description of the main components of a speaker recognition system including *i*-vector extraction, pre-processing, modeling and scoring. Throughout the paper, vectors are represented by italic lowercase letters, matrices by upper-case bold letters and constants by italic upper-case letters.

2.1. *i*-Vector features

i-Vectors are low-dimensional representation of GMM super-vectors in a single subspace which include all characteristics of speaker and inter-session variability, named total variability matrix $\mathbf{T}$ [1]. Given an observation set $\mathcal{X}_s$, the adapted mean super-vector m_s is modeled as,

$$m_s = m_0 + \mathbf{T}w_s + \varepsilon, \tag{1}$$

where m_0 is the Universal Background Model (UBM) super-vector, essentially a speaker-independent GMM super-vector, w_s with standard normal distribution is referred to as the *i-vector*, and ε is the residual term which accounts for the variability not captured by $\mathbf{T}$. The extraction of *i*-vectors in the proposed system is based on Baum-Welch statistics calculated for a given utterance with respect to UBM components and speech frame-level Mel-Frequency Cepstral Coefficients (MFCC).

2.2. Pre-processing

In order to achieve the state-of-the-art performance, a number of techniques have been proposed as pre-possessing steps for PLDA. A common pre-processing includes within-class covariance normalization (WCCN) [10] followed by length normalization of *i*-vectors [19].

2.2.1. Within-Class Covariance Normalization (WCCN)

One of the effective pre-processing step is to normalize the within-speaker covariance matrix of *i*-vectors [10]. A within-class covariance matrix, $\mathbf{W}$, is calculated as,

$$\mathbf{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}(w_i^s - \bar{w}_s)(w_i^s - \bar{w}_s)^T, \tag{2}$$

where S is the number of speakers, each having N_s utterances and $\bar{w}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} w_i^s$ is the mean of *i*-vectors from speaker s. This technique computes a transformation matrix from the Cholesky decomposition of $\mathbf{W}^{-1} = \mathbf{B}\mathbf{B}^T$ which will normalize within-speaker scatter matrix.

2.2.2. Length-Normalization

Due to the Gaussian probability distribution assumption made by PLDA model, it has been shown that length normalization of *i*-vectors can approximately Gaussianize their distribution [19]. This has been shown to improve the performance of Gaussian PLDA to that of heavy-tailed PLDA [20].

2.3. Probabilistic Linear Discriminant Analysis (PLDA)

Probabilistic LDA (PLDA) provides a powerful mechanism to distinguish between-speaker variability which characterizes speaker information from all other sources of undesired variability that characterize distortions. To achieve this, however, it is required to provide PLDA with enough labeled data which contain multiple utterances of a speaker under different distortion.

A standard Gaussian PLDA assumes that an *i*-vector w, is modeled according to

$$w = m + \mathbf{V}y + z. \tag{3}$$

where, m is the mean of *i*-vectors, y denotes the speaker latent variable with standard normal prior and the residual z is normally distributed with zero mean and full covariance matrix $\mathbf{\Sigma}_z$. In order to estimate the parameters of the model $(\mathbf{V}, \mathbf{\Sigma}_z)$, PLDA uses the expectation-maximization (EM) algorithm [8].

After parameter estimation, for each two trial *i*-vectors w_1 and w_2, the verification score will be computed using the log likelihood ratio of the hypothesis $\mathcal{H}_s$, that both *i*-vectors are from the same speaker and the hypothesis $\mathcal{H}_d$ that they are from two different speakers,

$$score = \log\frac{p(w_1, w_2|\mathcal{H}_s)}{p(w_1, w_2|\mathcal{H}_d)}. \tag{4}$$

Considering the Gaussian assumption, the PLDA score can be computed in closed-form solution

$$\begin{aligned} score = &\log\mathcal{N}([{}^{w_1}_{w_2}]; [{}^{m}_{m}], [{}^{\mathbf{S}_T\ \mathbf{S}_B}_{\mathbf{S}_B\ \mathbf{S}_T}]) \\ &- \log\mathcal{N}([{}^{w_1}_{w_2}]; [{}^{m}_{m}], [{}^{\mathbf{S}_T\ \mathbf{0}}_{\mathbf{0}\ \ \mathbf{S}_T}]). \end{aligned} \tag{5}$$

where, $\mathbf{S}_B = \mathbf{V}\mathbf{V}^T$ and $\mathbf{S}_T = \mathbf{S}_B + \mathbf{\Sigma}_z$. For a clear exposition and a fast method to compute the score we refer you to [19].

3. Language-independent Speaker Recognition

This section provides a brief description of Language-Normalized WCCN (LN-WCCN) which has been shown to significantly improve *i*-vector/PLDA system performance in multilingual scenario [11], then it describes the proposed LI-PLDA algorithm which accounts for language variability.

3.1. Language-Normalized WCCN (LN-WCCN)

Language source normalization is an effective technique to the reduction of language dependency in the state-of-the-art *i*-vector/PLDA speaker recognition system [11]. It can be implemented by extending the Source-Normalized WCCN (SN-WCCN) [12] in order to mitigate variations that separate languages. This can be accomplished by using i-vectors language label to identify different sources during the development. Language-Normalized WCCN (LN-WCCN) utilizes source-normalized within-speaker scatter matrix $\hat{\mathbf{S}}_W$ which is estimated as the variability not captured by the between speaker scatter matrix as

$$\hat{\mathbf{S}}_W = \mathbf{S}_T - \hat{\mathbf{S}}_B. \tag{6}$$

in which $\mathbf{S}_T$ is the total scatter matrix computed as

$$\mathbf{S}_T = \sum_{n=1}^{N} w_n {w_n}^T, \tag{7}$$

where N is the total number of *i*-vectors available for development (assuming zero-mean *i*-vectors), and $\hat{\mathbf{S}}_B$ is the normalized between-speaker scatter matrix which is formulated as

$$\hat{\mathbf{S}}_B = \sum_{l=1}^{L} \sum_{s=1}^{S_l} N_s^l (m_s^l - m_l)(m_s^l - m_l)^T. \tag{8}$$

where L is the number of languages available in development data, S_l is number of speakers for language l, m_s^l is the mean of N_s^l *i*-vectors from speaker s and language l and finally m_l is the mean of all *i*-vectors of language l.

3.2. Language-independent PLDA

When *i*-vectors are extracted from multilingual utterances, the language being spoken adds additional variability to the *i*-vectors due to the differences in acoustic content. We address this problem by extending the PLDA training algorithm in order to mitigate the variability associated with languages in *i*-vector space. In this way, we proposed to add a language dependent term intended to model the language being spoken in PLDA factorization.

Given a recording of a speaker s in language l, the proposed PLDA assumes the following linear factorization for *i*-vector $w(s)$ (assume centered *i*-vectors),

$$\begin{aligned} w(s) &= \mathbf{V}y(s) + z \\ z &= \mathbf{L}x(l) + \varepsilon \end{aligned} \tag{9}$$

in which, $y(s) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the speaker-dependent component and $z \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_z)$ is a speaker-independent random vector indicating inter-session variability, $x(l) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the latent variable corresponds to language being spoken, and the residual $\varepsilon \sim \mathcal{N}(0, \mathbf{\Sigma})$ having a normal distribution with mean $\mathbf{0}$ and full scatter matrix $\mathbf{\Sigma}$ indicates channel variability. $\mathbf{V}$ and $\mathbf{L}$ are eigenvoice and eigenlanguage subspaces respectively.

The difference between this modeling and that of the standard PLDA is that the language variability associated with *i*-vector point estimation can now be expressed in the form of $\mathbf{L}x(l)$. The columns of matrix $\mathbf{L}$ contain the basis for the language subspace and the term $x(l)$ represents a point in that subspace. The idea is to estimate channel variability void of language variability.

The proposed model comprises of three statistically independent parts, the speaker specific part $\mathbf{V}y(s)$ which describes the between-speaker variability which does not depend on a particular utterance of speaker s, the language specific part $\mathbf{L}x(l)$ which only depends on the language being spoken in a particular utterance (e.g. it does not depend on a particular utterance in language l), and ε which depends on a particular utterance of speaker s and describes all other variability other than language and refers to as channel variability.

Mathematically, we can describe the model in (9) in terms of conditional probabilities:

$$p(w(s)|y(s), x(l), \lambda) = \mathcal{N}(\mathbf{V}y(s) + \mathbf{L}x(l), \mathbf{\Sigma}). \tag{10}$$

The maximum likelihood estimation of the model hyper-parameters $\lambda = \{\mathbf{V}, \mathbf{L}, \mathbf{\Sigma}\}$ are obtained from a collection of development *i*-vectors $\mathcal{W}$ with both speaker labels, $\mathcal{W}(s) = \{w_i(s)\}_{i=1}^{n_s}$, and language labels, $\mathcal{W}(l) = \{w_i(l)\}_{i=1}^{n_l}, l = 1...L$, using an EM algorithm iteratively. We should note that, training PLDA hyper-parameters λ, requires language label as well as speaker label for all *i*-vectors during development but not necessarily requires multilingual utterances for each speaker. However, evaluation will be done without such information.

We compute the posterior probability of latent variable $y(s)$ using Bayes' rule as,

$$p(y(s)|\mathcal{W}(s), \lambda) = p(\mathcal{W}(s)|y(s), \lambda)p(y(s)), \tag{11}$$

where $p(y(s)) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the conditional likelihood is given by

$$p(\mathcal{W}(s)|y(s), \lambda) = \mathcal{N}(\mathbf{V}y(s), \mathbf{\Sigma}). \tag{12}$$

We should note that in (12) the covariance matrix $\mathbf{\Sigma}$ is void of language variability. Similarly, the posterior probability of latent variable $x(l)$ is computed as,

$$p(x(l)|\mathcal{Z}(l), \lambda) = p(\mathcal{Z}(l)|x(l), \lambda)p(x(l)), \tag{13}$$

in which we defined $\mathcal{Z}(l) = \{z_i(l)\}_{i=1}^{n_l}$ which corresponds to the speaker-independent components of *i*-vectors in language l, $\mathcal{W}(l)$, $p(x(l)) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the conditional likelihood distribution is given by

$$p(\mathcal{Z}(l)|x(l), \lambda) = \mathcal{N}(\mathbf{L}x(l), \mathbf{\Sigma}). \tag{14}$$

By estimating the directions of language variability during development, although it may not include all languages, the effect of language on $\mathbf{\Sigma}$ will be reduced. In order to exclude relevant language information in computation of verification score for two *i*-vectors w_1 and w_2, we proposed to eliminate language variability $\mathbf{S}_L = \mathbf{L}\mathbf{L}^T$ in the computation of total variability $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_L + \mathbf{\Sigma}$. In this way, we expect PLDA scoring to be independent of language being spoken.

Table 1: *Performance comparison of PLDA and LI-PLDA combined by WCCN and LN-WCCN across different telephony trials of the core condition of SRE'08.*

	All Languages		English		Diff. Languages		Same Languages	
	EER	C_{det}^{min}	EER	C_{det}^{min}	EER	C_{det}^{min}	EER	C_{det}^{min}
WCCN+PLDA	4.70%	0.0236	2.34%	0.0105	5.81%	0.0263	4.48%	0.0231
LN-WCCN+PLDA	4.59%	0.0228	2.23%	0.0103	5.69%	0.0252	4.36%	0.0223
WCCN+LI-PLDA	4.42%	0.0228	**2.13%**	0.0100	5.51%	0.0252	4.25%	0.0223
LN-WCCN+LI-PLDA	**4.24%**	**0.0221**	2.20%	**0.0097**	**5.35%**	**0.0247**	**4.04%**	**0.0214**

4. Experimental Setup

4.1. Development corpora

Different corpora are used in our experiment for system development. LDC releases of Switchboard cellular: phase II, and the telephony speech data drawn from NIST 2004 and 2005 speaker recognition evaluation corpora form our development set. The aforementioned corpora contains 13338 utterances from 1108 speakers, speaking in 5 different languages including English (12047), Russian (314), Spanish (146), Arabic (488) and Mandarin (343), of whom 204 speakers have multilingual speech utterances (English and one of the other 4 languages).

4.2. Evaluation protocol

The NIST 2008 corpus is used to evaluate the proposed approach. Results are reported for telephony multilingual trials as well as English trials of SRE'08 core condition. We have reported the performance using equal error rate (EER) and minimum decision cost function (C_{dcf}^{min}) as described in NIST SRE'08 evaluation plan [7]. The evaluation protocol includes 3832 target and 33218 non-target trials in which 6377 test segments were evaluated against 3263 enrolment segments. All segments were uttered by 1336 speakers who speak mainly in 15 different language dialects which can be grouped into fewer language clusters containing the languages in development data. The majority of the utterances are in English, however, there are multilingual utterances from 468 speakers.

4.3. System configuration

For acoustic features, we used 20 MFCC features along with first and second order derivatives for a total of 60 features. These feature vectors were then passed through an energy-based speech activity detector, followed by Cepstral Mean and Variance Normalization (CMVN). We trained a full covariance, gender-independent UBM model with 2048 Gaussian on the development data. We then trained a 500-dimensional *i*-vector extractor on the same data. The open-source Kaldi software has been used for all these processing [21]. WCCN and LN-WCCN transforms as well as PLDA, were also trained on the same development data. The parameters of the PLDA model were tuned using the core condition of the SRE'05 evaluation protocol. We have set a 300-dimensional subspace for the PLDA eigenvoice and a 10-dimensional subspace for eigenlanguage latent components.

5. Results

In this section we compare the performance of PLDA with our proposed language-independent PLDA (LI-PLDA) using the SRE'08 multilingual trial set. We also report the performance of recognition on English trials as well as same-language and different-language trials to observe the effect of proposed method on the speaker recognition performance.

Speech language normalization was incorporated into our system by utilizing LN-WCCN instead of WCCN as pre-processing. To better observe the effects of normalization as pre-processing, the performance of the following four systems are evaluated:

- WCCN+PLDA: This is our baseline system which shows the state-of-the-art speaker recognition performance without considering the language of utterances.
- LN-WCCN+PLDA: This system uses source normalization in order to normalize the effect of language on *i*-vectors as a pre-processing step for the PLDA modeling as proposed in [11].
- WCCN+LI-PLDA: The proposed language independent PLDA modeling with WCCN as preprocessing.
- LN-WCCN+LI-PLDA: This system uses both the ability of source-normalization and our proposed LI-PLDA modeling to reduce the effect of language on speaker recognition.

Table 1 summarizes the results for these systems. By comparing the results, it can be observed that the performance metrics were improved through the use of language normalization which reflect previous findings about LN-WCCN [11]. In multilingual trials we can see that LN-WCCN could provide a relative improvement of 2.3% in EER and 3.4% in minimum DCF compared to the baseline system. This indicates the suppression of language variation from *i*-vectors which results in the robustness of PLDA system to multilingual speech trials. We also expect language normalization not to affect the recognition of English trials which can be seen from the results that it did not have any sensible impact on minimum DCF, yet we can see some improvement in EER. The proposed LI-PLDA when followed by LN-WCCN can also provide a better improvement than WCCN by 10% in EER and 6.4% in minimum DCF compared to the baseline system. The results indicate that the proposed solution has complemented language normalization in removing the effect of language on speaker recognition. Figure 1 compares the detection error trade off (DET) curves for each of the four systems.

6. Discussions and Conclusions

This work proposed to reduce the effect of language as a source of variability on the performance of speaker recognition by extending the PLDA training algorithm. The experiments conducted in this paper demonstrate that *i*-vector/PLDA modelling of multilingual speech data can be improved by incorporating a language-dependent term intended to model language being spoken in the PLDA training algorithm. This improvement

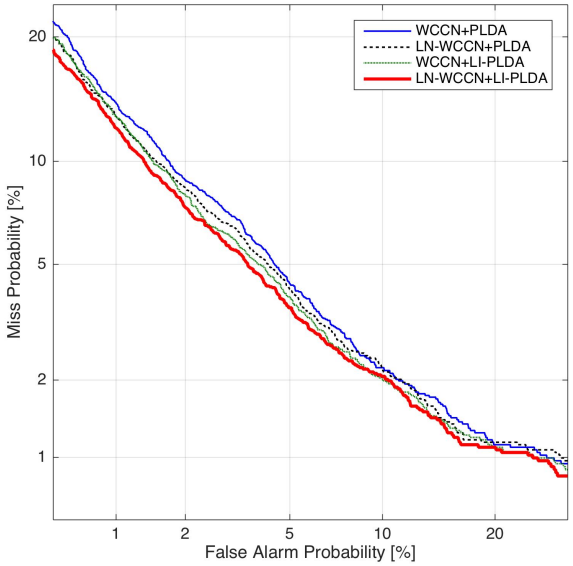

Figure 1: *Plot of DET curves for SRE'08 telephony trials of the core condition comparing the performance of the four systems.*

could be expected due to a better estimation of channel subspace which is void of language variability and as a result in a better estimation of speaker subspace. When combined with speech source normalization, LN-WCCN prior to LI-PLDA modeling was found to ameliorate the performance and offer considerable benefits. For our future work we are interested in conducting experiments using more multilingual data for development as well as experiments under additional conditions other than telephone speech.

7. Acknowledgments

This work is an original piece of research work carried out in the framework of a joint cooperation between Amirkabir University of Technology and Institut Mines-Télécom, Télécom SudParis. I express my deepest thanks to Dr. Dijana Petrovska-Delacréta and Prof. Gérard Chollet for their guidance and support. I would also want to thank Linguistic Data Consortium (LDC) for the data scholarship award which played an essential role in the fulfilment of this work.

8. References

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

[2] Craig S Greenberg, Désiré Bansé, George R Doddington, Daniel Garcia-Romero, John J Godfrey, Tomi Kinnunen, Alvin F Martin, Alan McCree, Mark Przybocki, and Douglas A Reynolds, "The nist 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[3] A Khosravani and M Homayounpour, "Linearly constrained minimum variance for robust i-vector based speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 249–253.

[4] Sergey Novoselov, Timur Pekhovsky, and Konstantin Simonchik, "Stc speaker recognition system for the nist i-vector challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 231–240.

[5] B Vesnicer, J Zganec-Gros, S Dobrisek, and V Struc, "Incorporating duration information into i-vector-based speaker-recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 241–248.

[6] Elie Khoury, Laurent El Shafey, Marc Ferras, and Sébastien Marcel, "Hierarchical speaker clustering methods for the nist i-vector challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[7] Alvin F Martin and Craig S Greenberg, "Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

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

[9] Mitchell McLaren and David Van Leeuwen, "Source-normalised-and-weighted lda for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[10] Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition.," in *Interspeech*, 2006.

[11] ML McLaren, Miranti Indar Mandasari, and David A van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Odyssey: The Speaker and Language Recognition Workshop*. 2012, pp. 55–61, Singapore : [s.n.].

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

[13] Ondrej Glembek, Jiaxin Ma, Pavel Matejka, Bing Zhang, Oldrich Plchot, Lukas Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4032–4036.

[14] 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.

[15] 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.

[16] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4002–4006.

[17] Ahilan Kanagasundaram, David Dean, and Sridha Sridharan, "Improving out-domain plda speaker verification using unsupervised inter-dataset variability compensation approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4654–4658.

[18] Brecht Desplanques, Kris Demuynck, and Jean-Pierre Martens, "Combining joint factor analysis and ivectors for robust language recognition," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 73–80.

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

[20] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010, p. 14.

[21] 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.