DOMAIN ADAPTATION USING MAXIMUM LIKELIHOOD LINEAR TRANSFORMATION FOR PLDA-BASED SPEAKER VERIFICATION

Qiongqiong Wang, Hitoshi Yamamoto and Takafumi Koshinaka

Information and Media Processing Laboratories, NEC Corporation

ABSTRACT

While i-vector-PLDA frameworks employing huge amounts of development data have achieved significant success in speaker recognition, it is infeasible to collect a sufficiently large amount of data for every real application. This paper proposes a method to perform supervised domain adaptation of PLDA in i-vector-based speaker recognition systems with available resource-rich mismatched data and small amounts of matched data, under two assumptions: (1) between-speaker and within-speaker covariances depend on domains; (2) features in one domain can be transformed into another domain by means of an affine transformation. Maximum likelihood linear transformation (MLLT) is used to infer the relationship between the datasets of two domains in training PLDA. The proposed method improves performance over that achieved without adaptation. Using a score fusion technique, it outperforms a conventional method based on linear combination.

Index Terms— PLDA, MLLT, affine transformation

1. INTRODUCTION

Probabilistic Linear Discriminant Analysis (PLDA) [1, 2, 3] is a state-of-the-art method used in speaker recognition to separate speaker factors in i-vectors [4] from such irrelevant factors as transmission channels and emotion. In order to train parameters in PLDA models, multi-session recordings from several thousand speakers are typically used. For example, research groups involved in NIST speaker recognition evaluation (SRE) [5] typically use utterances from SRE data along with Switchboard and Fisher data. However, it would be prohibitively expensive to try to collect such a large amount of in-domain (IND) data for a new domain of interest for every application. Most available resource-rich data that already exist do not match the domain of interest, that is, most is out-of-domain (OOD) data. Domain mismatch between development and evaluation data can greatly deteriorate performance in speaker recognition systems. [6] shows that, with the use of OOD data for PLDA development, the equal error rate (EER) is 3.4 times that when IND data is used. [7, 8] show a difference of 2.84 times.

Domain adaptation techniques, for adapting a resource-rich OOD system so as to produce good results in a new domain, have recently been studied with the aim of alleviating this problem. They are either supervised adaptation [6, 7, 9, 10], for which a small amount of IND data with labels is available, or unsupervised adaptation [8, 11, 12, 13], for which a large amount of IND data without labels is available. This paper focuses on the former, supervised adaptation.

Supervised domain adaptation methods can be further categorized into the following three approaches: 1) Data augmentation. For example, [9] adds IND data to a large amount of OOD data to train PLDA. 2) PLDA parameter adaptation. [7] linearly combines parameters of PLDAs trained separately with OOD data and IND data. 3) i-vector compensation. [6] applies data shifting using the statistical information about both IND and OOD data.

While 1) and 2) implicitly assume that simple interpolation is enough, such an assumption may not be true if the characteristics of OOD and IND are largely different. 3) uses different criteria, such as maximum likelihood (ML) and minimum distance, that is different from those in PLDA training. Such inconsistency may lead to some sub-optimal local solutions. More flexible adaptation methods with a single global criterion are desirable.

In this paper, we focus on supervised domain adaptation and propose a learning algorithm that automatically and simultaneously optimizes the PLDA parameters as well as those for the transformation between two domains using only one global criterion: maximum likelihood linear feature transformation (MLLT).

The remainder of this paper is organized as follows: Section 2 describes a typical speaker verification system based on i-vectors and PLDA. Section 3 introduces a method of using MLLT in PLDA. Section 4 describes our experimental setup, results, and analyses. Finally, Section 5 summarizes our work.

2. I-VECTOR AND PLDA BASED SPEAKER VERIFICATION

In an i-vector based speaker verification system [4], it is assumed that a GMM-supervector ξ , corresponding to an utterance can be modeled as

$$\xi = \bar{\xi} + Tx,$$

where x is a random vector known as the i-vector, T is a basis for the total variability space for speaker and channel variability of ξ, and $\bar{\xi}$ is the mean of ξ. It is assumed that x follows a standard normal distribution and its dimension d is lower than that of $\bar{\xi}$.

Probabilistic linear discriminant analysis (PLDA) decomposes the total variability into within-speaker and between-speaker variability. Originally introduced in [1, 2] for face recognition, PLDA has been heavily employed for speaker recognition based on i-vectors [3, 4, 14]. Besides the original PLDA formulation [2], there are two alternative variants that assume full covariance: simplified PLDA [3] and a two-covariance model [15]. In this study, we apply MLLT to the two-covariance representation of PLDA introduced in [1, 15] and used extensively in [16]. Here, it is assumed that there is a latent variable y_i representing speaker i, so that the observed utterances $\{x_{ij}\}$ of the speaker i share the same speaker variable y_i:

$$y_i \sim \mathcal{N}(m, \Phi_b), \quad x_{ij}|y_i \sim \mathcal{N}(y_i, \Phi_w),$$

where Φ_b and Φ_w are between- and within-speaker covariance matrices, respectively, m is the global mean.

In the inference phase, given i-vectors of two utterances x_1, x_2, PLDA calculates the log-likelihood ratio between two hypotheses that they are from the same speaker (H_0) or from different speakers (H_1). If the ratio is larger than a predetermined threshold, the two utterances belong to the same speaker; otherwise, they do not [1].

3. MAXIMUM-LIKELIHOOD TRANSFORMATION

We propose a method that automatically optimizes the transformation and PLDA parameters simultaneously using only one global criterion. We also present a simplified algorithm in which the optimization is conducted in two steps, on the basis of which we carry out experiments.

3.1. PLDA-MLLT

Here, we first assume that between- and within-speaker covariances depend on domains, which means the covariances $\Phi_b^{(S)}$, $\Phi_w^{(S)}$ in the source domain (domain of OOD data) and $\Phi_b^{(T)}$, $\Phi_w^{(T)}$ in the target domain (domain of IND data) will differ. We also assume that features $x^{(T)}$ in the target domain can be transformed into the source domain by applying an affine transformation,

$$\widetilde{x}^{(T)} = Ax^{(T)} + b,$$

so that in the source domain $\widetilde{x}^{(T)}$ and $x^{(S)}$ share the same covariances $\Phi_b^{(S)}$, $\Phi_w^{(S)}$ and mean $m^{(S)}$:

$$\begin{aligned} \Phi_b^{(S)} &\approx \widetilde{\Phi}_b^{(T)} &= A\Phi_b^{(T)}A^T, \\ \Phi_w^{(S)} &\approx \widetilde{\Phi}_w^{(T)} &= A\Phi_w^{(T)}A^T, \\ m^{(S)} &\approx \widetilde{m}^{(T)} &= Am^{(T)} + b, \end{aligned}$$

where $(*)^{(T)}$ represents IND (target domain); $(*)^{(S)}$ represents OOD (source domain); $\widetilde{(*)}^{(T)}$ is transformed IND in the source domain.

We aim to find the PLDA and MLLT parameters $\theta = (A, b, \Phi_b^{(S)}, \Phi_w^{(S)})$, under which the IND and OOD data are most likely. Given $N^{(S)}$ OOD training patterns $X^{(S)} = \{X_i^{(S)}|i = 1, \cdots, K^{(S)}\}$ and $N^{(T)}$ IND training patterns $X^{(T)} = \{X_i^{(T)}|i = 1, \cdots, K^{(T)}\}$ from $K^{(S)}$ and $K^{(T)}$ speakers respectively, where $X_i = \{x_{i,1}, \cdots, x_{i,n_i}\}$ is a set of n_i patterns from speaker i, the log-likelihood is

$$l(X^{(S)}, X^{(T)}) = \sum_{i=1}^{K^{(S)}} \ln P(X_i^{(S)}) + \sum_{i=1}^{K^{(T)}} \ln P(X_i^{(T)}), \quad (1)$$

$$P(X_i) = \int N(y|m, \Phi_b) \prod_{j=1}^{n_i} N(x_{ij}|y, \Phi_w) \mathrm{d}y,$$

where $P(X_i)$ is the joint probability distribution of a set of $n_i^{(S)}$ or $n_i^{(T)}$ patterns belonging to the same speaker i in the source or target domain.

With the ML criterion, parameters θ are estimated, with the Expectation Maximization (EM) algorithm, as shown in Fig. 1. Note that Φ_b, Φ_w and m in Fig. 1 represent the parameters in source domain. Since Eq. (3) has no analytic solution, an iterative method is needed in the estimation of A.

3.2. 2-step PLDA-MLLT

2-step PLDA-MLLT is a simplified and computationally less expensive formulation of the previous single-step one. It optimizes the MLLT parameters (A, b) using IND data, given the PLDA model trained with OOD data. Then the PLDA is adapted with (A, b) to the target domain, and it assigns scores to the evaluation data. We conduct the 2-step PLDA-MLLT in the following two steps:

Step 1 Apply PLDA with OOD data to estimate $\Phi_b^{(S)}, \Phi_w^{(S)}$ by maximizing the first term in the r.h.s of Eq. (1).

Step 2 Given $\Phi_b^{(S)}, \Phi_w^{(S)}$, estimate (A, b) by maximizing the second term in Eq. (1). With EM algorithm, the estimation formulation is obtained, which is a simplified form of the single-step PLDA-MLLT (Fig. 1). It has Eq. (2) as E-step and Eq. (3) – (4) as M-step.

Experiments presented in the next section were performed on the basis of this formulation.

4. EXPERIMENTS

4.1. Experimental setup

We conducted experiments on the NIST SRE 2008 [5] core task condition-6 (SRE08). The speaker verification system in

E-step: Estimate the posterior from the latest $\theta \equiv \overline{\theta} = (\overline{A}, \overline{b}, \overline{\Phi}_b, \overline{\Phi}_w)$. $L_i^{-1}\gamma_i$ and L_i^{-1} are the posterior mean and covariance of speaker i's i-vector. m_i is the mean of the i-vectors of speaker i.

$$
\begin{aligned}
L_i^{(T)} &= \bar{A}^T \bar{\Phi}_b^{-1} \bar{A} + n_i^{(T)} \bar{A} \bar{\Phi}_w^{-1} \bar{A}^T, & \gamma_i^{(T)} &= \bar{A}^T \bar{\Phi}_b^{-1} \left(m - \bar{b} + n_i^{(T)} \bar{A}^T \bar{\Phi}_w^{-1} \bar{A} {m_i}^{(T)} \right), \\
L_i^{(S)} &= \bar{\Phi}_b^{-1} + n_i^{(S)} \bar{\Phi}_w^{-1}, & \gamma_i^{(S)} &= \bar{\Phi}_b^{-1} m + n_i^{(S)} \bar{\Phi}_w^{-1} {m_i}^{(S)}.
\end{aligned} \tag{2}
$$

M-step: Update the values of the parameter $\theta = (A, b, \Phi_b, \Phi_w)$. VP denotes a vector product: $\mathrm{VP}(z) = zz^T$. $N^{(S)}$, $N^{(T)}$, $K^{(S)}$, and $K^{(T)}$ are the numbers of speakers and the numbers of utterances in OOD and IND data respectively.

$$
\Phi_b = \left(K^{(S)} + K^{(T)} \right)^{-1} \left[\sum_{i=1}^{K^{(S)}} \left({L_i^{(S)}}^{-1} + \mathrm{VP}\left({L_i^{(S)}}^{-1} \gamma_i^{(S)} - m \right) \right) + \sum_{i=1}^{K^{(T)}} \left(\bar{A} {L_i^{(T)}}^{-1} \bar{A}^T + \mathrm{VP}\left(\bar{A} {L_i^{(T)}}^{-1} \gamma_i^{(T)} + \bar{b} - m \right) \right) \right],
$$

$$
\Phi_w = \left(N^{(S)} + N^{(T)} \right)^{-1} \left[\sum_{i=1}^{K^{(S)}} \sum_{j=1}^{n_i^{(S)}} \left({L_i^{(S)}}^{-1} + \mathrm{VP}\left({L_i^{(S)}}^{-1} \gamma_i^{(S)} - x_{ij}^{(S)} \right) \right) + \overline{A} \sum_{i=1}^{K^{(T)}} \sum_{j=1}^{n_i^{(T)}} \left({L_i^{(T)}}^{-1} + \mathrm{VP}\left({L_i^{(T)}}^{-1} \gamma_i^{(T)} - x_{ij}^{(T)} \right) \right) \overline{A}^T \right],
$$

$$
\begin{aligned}
A^T \bar{\Phi}_b^{-1} A \sum_{i=1}^{K^{(T)}} & \left[{L_i^{(T)}}^{-1} + \left({L_i^{(T)}}^{-1} \gamma_i^{(T)} - \frac{1}{K^{(T)}} \sum_{i=1}^{K^{(T)}} {L_i^{(T)}}^{-1} \gamma_i^{(T)} \right) \left({L_i^{(T)}}^{-1} \gamma_i^{(T)} \right)^T \right] \\
& + A^T \bar{\Phi}_w^{-1} A \sum_{i=1}^{K^{(T)}} \sum_{j=1}^{n_i^{(T)}} \left[{L_i^{(T)}}^{-1} + \mathrm{VP}\left({m_i}^{(T)} - {L_i^{(T)}}^{-1} \gamma_i^{(T)} \right) + \mathrm{VP}\left(x_{ij}^{(T)} \right) - \mathrm{VP}\left({m_i}^{(T)} \right) \right] = \left(N^{(T)} + K^{(T)} \right) I,
\end{aligned} \tag{3}
$$

$$
b = m - \frac{1}{K^{(T)}} \sum_{k=1}^{K^{(T)}} (A {L_i^{(T)}}^{-1} \gamma_i^{(T)}). \tag{4}
$$

Fig. 1. Parameter estimation of PLDA-MLLT using the EM algorithm

the experiments was based on the i-vector and PLDA framework described in Section 2. In the system, the input speech segment was first converted to a sequence of acoustic feature vectors, each of which consisted of 60 features (MFCC 1-20 and its Δ and $\Delta\Delta$) extracted from a frame of 25ms width at a 10ms frame shift. Then an i-vector of 400 dimensions was extracted from the acoustic feature vectors, 2048-mixture universal background model (UBM), and a total variability matrix (TVM). After that, it was evaluated with a PLDA model.

The UBM and TVM were trained with the Fisher corpus [17], which contains 12,399 speakers (OOD). For MLLT training we used, as IND, data from NIST SRE04 of different sizes representing 310 speakers, under the conditions of 1side, 3sides, 8sides, 16sides, 10sec, and 30sec. We utilized the Kaldi speech recognition toolkit [18] to run all steps other than that of MLLT. We used a stochastic hill climbing algorithm [19] to estimate A in Eq. (3).

Results for the proposed method are compared with those for the method using PLDA without adaptation (Section 2) and those for a conventional method of domain adaptation employing a linear combination of PLDA parameters [7] that uses the standard EM algorithm twice to obtain IND and OOD PLDA parameters and then linearly combines them using a weight coefficient: $\Phi_{b/w} = \alpha \Phi_{b/w}^{(T)} + (1-\alpha)\Phi_{b/w}^{(S)}$.

In the experiments, we used the 2-step PLDA-MLLT formulation described in Section 3.2. We initialized A as a unit matrix and b as a zero vector. 800 EM iterations were applied to reach its convergence.

4.2. Experimental results

4.2.1. Performance gap

Fig. 2 shows the EERs for a PLDA model trained with OOD data and for one trained with IND data. There is a considerable gap in performance between a system trained on the OOD Fisher set (red line) and a system trained on matched IND SRE04 data (310 speakers). The gap is roughly 36%. As the training data decreases, performance degrades similarly to [20]. When it decreased to that for 100 speakers, the performance suffered 54% degradation and was even 12% worse than that for the system trained with OOD data. This prompted us to explore the effect of adaptation approaches.

4.2.2. PLDA-MLLT

We compared the following 7 systems (S1-S7) in 4 cases, for which 100, 150, 200, and 250 speakers in IND SRE04 were available. S3 is the system of the proposed PLDA-MLLT. S5 and S7 are its fusions at the model level and score level, respectively. S4 is a conventional method.

S1 PLDA-OOD, PLDA trained with Fisher
S2 PLDA-IND, PLDA trained with SRE04

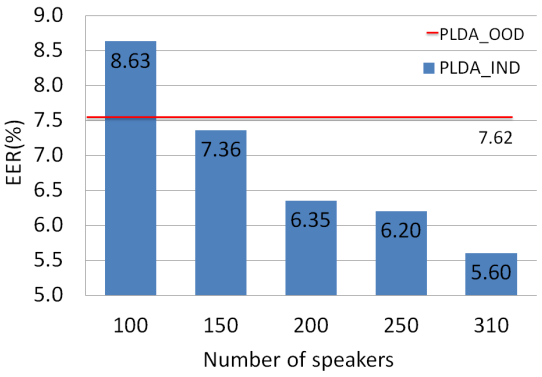

Fig. 2. Comparison of performance of PLDA model trained with OOD data and those trained with different numbers of IND data

#speaker	100	150	200	250
S1: PLDA-OOD	7.62	7.62	7.62	7.62
S2: PLDA-IND	8.63	7.36	6.35	6.20
S3: PLDA-MLLT	6.98	6.16	6.24	6.01
S4: S1+S2 (MF)	**6.57**	5.82	5.75	5.79
S5: S3+S2 (MF)	6.81	5.82	5.71	5.75
S6: S1+S2 (SF)	7.01	6.52	6.15	5.89
S7: S3+S2 (SF)	6.81	**5.81**	**5.68**	**5.66**

Table 1. Equal error rates (EERs %) for the 7 systems. Bold face denotes the best performance in each column.

S3 PLDA-MLLT, proposed single system, PLDA-OOD adapted with MLLT trained with SRE04

S4 S1+S2 (MF), conventional method [7]. Model-level fusion of S1 and S2

S5 S3+S2 (MF), model-level fusion of S2 and S3 with linear combination in the same way as S4

S6 S1+S2 (SF), score-level fusion of S1 and S2 with Bosaris toolkit [21]

S7 S3+S2 (SF), score-level fusion of S3 and S2

On the basis of the two assumptions given in Section 3.1, it may be considered that PLDA-MLLT has transformed the source domain to the target domain, so the PLDA after MLLT adaptation may be thought to represent the target domain, i.e., that it is in the same domain as is PLDA-IND. For this reason, it seemed reasonable to fuse PLDA-MLLT and PLDA-IND at both model and score levels to get a better representation of the target domain.

Table 1 shows the EERs for the 7 systems. For S4 and S5, it only shows the best performance, obtained using their optimal weight coefficients α. The relationship between their performance and α is shown in Fig. 3.

As shown in Table 1, in all 4 cases (different numbers of training speakers), the proposed PLDA-MLLT single system S3 outperforms both S1 and S2, for which PLDA models are

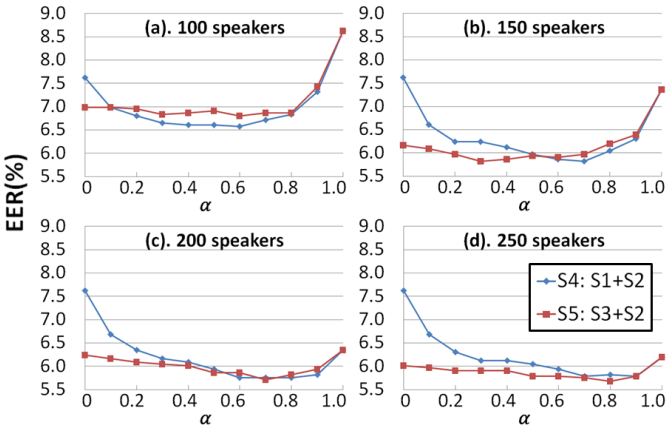

Fig. 3. Relationship between performance and weight coefficients in model-level fusion

trained with OOD Fisher data and IND SRE04 data, respectively. This indicates that MLLT indeed adapts the PLDA model to the target domain. Comparing S4 and S5 (only their best performances), which used the same model fusion technique, we found our proposed system S5 got better performance over wider ranges of α when150 or more speakers were available, i.e. S5 was less affected by α. All in all, S7 was the best over all 7 systems when 150 or more speakers were available.

5. SUMMARY

We have proposed a domain adaptation of PLDA based on maximum likelihood Linear transformation (MLLT), which assumes an affine transformation between different domains. PLDA is iteratively trained by updating the MLLT transformation automatically with a single global criterion, an approach which does not seem to have been attempted in any previous research. Experimentally, the proposed method has been shown to improve speaker recognition performance with a score fusion technique when 150 speakers or more of the target domain were available.

In the experiments, we evaluated a 2-step algorithm that had been produced as an approximation. Future issues include the implementation and evaluation of the single-step algorithm. We also intend to seek more effective numerical solutions to parameter estimation and more sophisticated nonlinear transformations to be used in adaptation.

6. ACKNOWLEDGEMENT

We thank our colleagues at work, Prof. Koichi Shinoda and Dr. Johan Rohdin in Tokyo Institute of Technology for valuable discussions and comments that greatly improved the manuscript.

7. REFERENCES

[1] S. Ioffe, "Probabilistic linear discriminant analysis," *Proceedings of the 9th European Conference on Computer Vision*, pp. 531–542, 2006.

[2] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about dentity," in *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[3] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010.

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

[5] The NIST year 2008 speaker recognition evaluation plan, "http://www.itl.nist.gov/iad/mig/tests/spk/2008/sre08 evalplan release4.pdf," 2008.

[6] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[7] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[8] S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[9] A. Misra and J. Hansen, "Spoken language mismatch in speaker verification: an investigation with NIST-SRE and CRSS Bi-ling Corpora," in *IEEE The Spoken Language Technology Workshop (SLT)*, 2014.

[10] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[11] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[12] J. Villalba and E. Lleida, "Unsupervised adaptation of PLDA by using variational bayes methods," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[13] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014.

[14] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4828–4831, 2011.

[15] N. Brummer and E. De Villiers, "The speaker partitioning problem," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[16] J. Villalba and N. Brummer, "Towards fully bayesian speaker recognition: Integrating out the between-speaker covariance," in *Interspeech*, 2011.

[17] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text," *The Fourth International Conference on Language Resources and Evaluation (LREC)*, pp. 69–71, 2004.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, N. Goel O. Glembek, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. of Workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] A. Juels and M. Wattenberg, "Hillclimbing as a baseline method for the evaluation of stochastic optimization algorithms," *MIT Press: Advances in Neural Information Processing Systems*, vol. 8, pp. 430–436, 1995.

[20] M. H. Rahman, D. Dean, A. Kanagasundaram, and S. Sridharan, "Investigating in-domain data requirements for PLDA training," in *Interspeech*, 2015.

[21] N. Brummer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011.