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

Unsupervised Discriminative Training of PLDA for Domain Adaptation in Speaker Verification

Qiongqiong Wang, Takafumi Koshinaka

Data Science Research Laboratories, NEC Corporation, Japan
q-wang@ah.jp.nec.com

Abstract

This paper presents, for the first time, unsupervised discriminative training of probabilistic linear discriminant analysis (unsupervised DT-PLDA). While discriminative training avoids the problem of generative training based on probabilistic model assumptions that often do not agree with actual data, it has been difficult to apply it to unsupervised scenarios because it can fit data with almost any labels. This paper focuses on unsupervised training of DT-PLDA in the application of domain adaptation in i-vector based speaker verification systems, using unlabeled in-domain data. The proposed method makes it possible to conduct discriminative training, i.e., estimation of model parameters and unknown labels, by employing data statistics as a regularization term in addition to the original objective function in DT-PLDA. An experiment on a NIST Speaker Recognition Evaluation task shows that the proposed method outperforms a conventional method using speaker clustering and performs almost as well as supervised DT-PLDA.

Index Terms: unsupervised, discriminative training, PLDA, domain adaptation, speaker verification

1. Introduction

Over the last decade, Probabilistic Linear Discriminant Analysis (PLDA) [1][2] has become state-of-the-art modeling used in speaker verification to separate speaker factors in i-vectors [3][4] from such irrelevant factors as transmission channels and emotion. It assumes additive speaker and channel components modeled by Gaussian distributions, and the parameters are usually optimized by generative training (GT) under the maximum likelihood (ML) criterion, using a speaker ID for each speech utterance as class information. Here, such PLDA is referred to as GT-PLDA.

Such prior Gaussian assumptions, however, have been proved inaccurate. In [5], heavy-tailed PLDA (HT-PLDA) based on the t-distribution performed much better than GT-PLDA, showing that the elements of the i-vector are, in reality, more heavy-tailed than the Gaussian distribution. Unfortunately, the extraordinarily high computational cost of HT-PLDA is a serious roadblock to application. Additionally, score calibration is often applied in GT-PLDA to adjust scores to better serve as log-likelihood ratios (LLRs) regarding whether two i-vectors are from the same speaker or not. Substantial improvements using the discriminatively trained (DT) affine transformations of the scores [6][7] have indicated that scores originally from GT-PLDA are not accurate, which is the result of inaccurate assumptions made for models, and additional mismatches between models and i-vectors have been pointed out by [8].

Whenever there is a mismatch between a model and data, DT may improve performance. Rather than application to scores alone, DT has been proposed for use on the PLDA model itself (DT-PLDA) [9][10]. The discriminative classifier used is trained to estimate the parameters of a symmetric quadratic function approximating a LLR score, from which an equivalent expression for GT-PLDA can be derived. This indicates strong connections with generative models [11]. Rather than explicitly training the PLDA model, it directly optimizes the LLR score function of the PLDA model. Thus, DT is able to avoid the Gaussian assumptions of the model [11], which allows the score function to be more general than that of a standard GT-PLDA model. Many studies have proven the effectiveness of DT-PLDA with i-vectors used for feature representation. On the other hand, due to its discriminative property, DT easily becomes over-trained [12] and thus requires training data to be matched with the target domain more so than does GT-PLDA. Considering the prohibitively high expense of collecting such a large amount of in-domain (IND) data with labels for a new domain of interest for every application, the utility of DT-PLDA can be seen to be limited despite its high capability.

Alternatively, a certain amount of matched data exists and is also easy to collect without labels. To the best of our knowledge in this regard, however, no such research studies have yet dealt with unsupervised DT-PLDA. There is, though, an approach of combining speaker clustering with supervised training, as done in GT-PLDA [13], which is easy to reach and apply to unsupervised DT-PLDA. Many clustering methods are available, including mean shift [14][15][16] and hierarchical bottom-up clustering [17]. However, with this approach, the performance of a DT-PLDA model is likely to suffer from inaccurate speaker clustering. In addition, double criteria in speaker clustering and DT-PLDA training, can achieve only sub-optimums, not global optimums.

This paper presents an unsupervised training method for DT-PLDA that estimates its parameters, as well as unknown speaker labels, on the basis of a single criterion. In order to avoid being over-trained, it uses a regularization term consisting of simple training data statistics. Experiments on NIST 2008 Speaker Recognition Evaluation (SRE08) show that the proposed method outperforms a conventional method and performs almost as well as the supervised DT-PLDA.

The remainder of this paper is organized as follows: Section 2 describes a typical speaker verification system based on i-vectors and PLDA, as well as the extension to DT-PLDA. Section 3 introduces both the proposed method of using regularization in unsupervised training of DT-PLDA and also a special case: 4-parameter DT-PLDA. Section 4 describes our experimental setup, results, and analyses of unsupervised DT-PLDA in an application of domain adaptation. Finally, Section 5 summarizes our work.

http://dx.doi.org/10.21437/Interspeech.2017-727

2. GT-PLDA and DT- PLDA

2.1. PLDA-based Speaker Verification

In an i-vector based speaker verification system [3], it is assumed that a GMM-supervector ξ corresponding to a speech utterance can be modeled as

$$\xi = \bar{\xi} + T\phi,$$

where ϕ 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 ϕ follows a standard normal distribution and that its dimension d, i.e., the rank of T, is lower than that of $\bar{\xi}$.

PLDA [1][2][5] decomposes total variability into between-class (speaker) and within-class (channel) variability. A popular configuration in speaker verification is [5][18]

$$\phi = m + Vy + Dz, \tag{1}$$

where y and z are random vectors depending, respectively, on the speaker and the channel . Speaker variability is given by V and channel variability is given by D. The elements of y and z are assumed to be independent and normally distributed. PLDA is a generative model, and its parameters are typically estimated using the ML criterion. In this paper, we call this kind of PLDA generatively trained PLDA (GT-PLDA).

For scoring two i-vectors, ϕ_i and ϕ_j, PLDA calculates a log-likelihood ratio (LLR) s_{ij} between two hypotheses: H_s – they are from the same speaker or H_d – they are from different speakers,

$$s_{ij}\left(\phi_i, \phi_j\right) = \frac{P(\phi_i, \phi_j | H_s)}{P(\phi_i, \phi_j | H_d)}. \tag{2}$$

According to Eq. (1), it is clear that pairs of i-vectors $[\phi_i \ \phi_j]$ follow a multivariate normal distribution. Calculating the mean and covariance of an i-vector pair in a target (H_s is true) and a non-target trial (H_d is true) and plugging the resulting multivariate normal distributions into Eq. (2) results in a closed-form expression of the LLR [19] given by

$$s_{ij} = \phi_i^T P \phi_j + \phi_j^T P \phi_i + \phi_i^T Q \phi_i + \phi_j^T Q \phi_j + (\phi_i + \phi_j)^T c + k, \tag{3}$$

where

$$\begin{aligned} P &= \frac{1}{2}\Sigma_{tot}^{-1}\Sigma_b(\Sigma_{tot} - \Sigma_b\Sigma_b^{-1}\Sigma_b)^{-1}, \\ Q &= \frac{1}{2}\Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_b\Sigma_b^{-1}\Sigma_b)^{-1}, \\ c &= -2(P+Q)m, \\ k &= \frac{1}{2}(log|\Sigma_{tot}| - log|\Sigma_{tot} - \Sigma_b\Sigma_{tot}^{-1}\Sigma_b|) \\ &\quad + m^T 2(P+Q)m, \end{aligned}$$

where $\Sigma_b = VV^T$, $\Sigma_w = DD^T$ are between- and within-class covariance matrices, respectively. $\Sigma_{tot} = \Sigma_b + \Sigma_w$.

2.2. Discriminative PLDA Training

Instead of using the ML criterion for training the PLDA model (m, V and D) [2][5], we can use discriminative training (DT), which directly optimizes the parameters, (P, Q, c and k) for discriminating between the same-speaker trial and a different-speaker trial. This was first proposed in [9] and [10]. Let $\theta = \text{vec}([P, Q, c, k])$, where vec($\bullet$) stacks the columns of a matrix into a column vector. In this study, we have modified the objective function into weighted loss of all the training trials:

$$E(\theta) = N(\sum_{t_{ij}=1} \frac{P_{\text{eff}}}{N_+} l_{ij} + \sum_{t_{ij}=1} \frac{1 - P_{\text{eff}}}{N_-} l_{ij}), \tag{4}$$

where $l_{ij} = l(t_{ij}, s_{ij}(\theta))$ is the loss function for a trial (ϕ_i, ϕ_j) when it is mis-recognized; N is the total number of trials in the training set, $N = N_+ + N_-$; N_+ and N_- are the numbers of target and non-target trials, respectively; P_{eff} is known as the effective prior. Such weight settings P_{eff}/N_+ and $(1 - P_{\text{eff}})/N_-$ are to follow the definition of the actual Detection Cost Function (actDCF) defined in NIST Speaker Recognition Evaluation (SRE),

$$\text{actDCF} = P_{\text{eff}} P_{\text{FR}} + (1 - P_{\text{eff}}) P_{\text{FA}}.$$

Use of the weighted loss function in Eq. (4), then, aims to improve the performance of speaker verification in terms of actDCF.

By minimizing $E(\theta)$, θ can be trained discriminatively. Using the Optimized Cutting Plane Algorithm for SVMs (OCAS) proposed in [20][21], or the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) [22], the DT-PLDA parameters are optimized by evaluating the loss function and the gradient of its error function. The gradient is given by [10]:

$$\nabla E(\theta) = \begin{bmatrix} \nabla_P E(\theta) \\ \nabla_Q E(\theta) \\ \nabla_c E(\theta) \\ \nabla_k E(\theta) \end{bmatrix} = \begin{bmatrix} 2\text{vec}(\Omega G \Omega^T) \\ 2\text{vec}([\Omega \circ (1_A G)]\Omega^T) \\ 2\text{vec}([\Omega \circ (1_A G)\Omega] 1_B) \\ 1_B^T G 1_B \end{bmatrix},$$

where 1_A is a $d \times n$ matrix of ones, and 1_B is a $n \times 1$ matrix of ones, $\Omega = [\phi_1 ... \phi_n]$, $\circ$ denotes the element-wise multiplication of two matrices, and

$$G_{ij} = \begin{cases} \frac{NP_{\text{eff}}}{N_+} \frac{\partial l(t_{ij}, s_{ij})}{\partial s_{ij}} & : t_{ij} = 1 \\ \frac{N(1-P_{\text{eff}})}{N_-} \frac{\partial l(t_{ij}, s_{ij})}{\partial s_{ij}} & : t_{ij} = -1 \end{cases}.$$

Commonly used $l(t, s)$ are the logistic loss and the hinge loss. Hinge loss optimizes the margin separation between the classes, while logistic loss minimizes the crossentropy error function. In this study we follow [9] and use the logistic loss function given by

$$l(t_{ij}, s_{ij}) = \log(1 + \exp(-t_{ij} s_{ij})).$$

3. Unsupervised Training of DT-PLDA

In practice, matched data with accurate labels required for DT-PLDA training is often difficult to collect. With estimated labels, unlabeled data can barely train a good discriminatively trained PLDA (DT-PLDA) model because it easily over-fits mislabeled samples. We propose a method of unsupervised discriminative training of PLDA that uses data statistics as a regularizer, to constrain the iterative training. It estimates the labels and PLDA parameters simultaneously. We also derive a solution to a special case of DT-PLDA: 4-parameter DT-PLDA [23], on the basis of which we carry out experiments.

3.1. Regularized Objective Function

In unsupervised training, more sophisticated modeling methods tend to converge more easily toward a local minimum of the objective function and are very sensitive to the initialization of parameters. Thus, adding a regularizer is a natural idea to constrain the unsupervised training and improve robustness. Cosine similarity, SVM scores, GT-PLDA scores, etc., which represent similarity of data, might be suitable as regularizers. In this study, we would like to explore the more fundamental data statistics, so in this paper cosine similarity is studied. We initialize PLDA parameters with generative training and then use the mean of cosine similarity of i-vectors $C(\phi_i, \phi_j)$ as a regularizer in the discriminative training. Here, we operate under the assumption that the scoring of cosine similarity and that of PLDA are weakly correlated. Thus, the addition of a regularizer such as $D(\theta)$ to the objective function Eq. (4) is able to constrain the iterative training of PLDA. The new objective function becomes

$$E'(\theta) = E(\theta) + bD(\theta), \tag{5}$$

where b is the weight for the regularization term of data statistics

$$D = -\frac{1}{N_+} \sum_{i,j;s_{ij}(\theta)>\tau} C(\phi_i, \phi_j) + \frac{1}{N_-} \sum_{i,j;s_{ij}(\theta)<\tau} C(\phi_i, \phi_j). \tag{6}$$

Note that the difference of the error function (5) from Eq. (4) is that labels are no longer available, so t_{ij} in $E(\theta)$ in Eq. (5) is replaced with the labels estimated in the previous iteration of training, using the inequalities $s_{ij}(\theta) \lessgtr \tau$. As the current PLDA parameters are already optimal for the labels, if we only have $E(\theta)$ as the objective function, the iterative training will not proceed. The regularization term helps unsupervised training avoid such settings.

In the regularization (6), although cosine similarities are constant, the addition and subtraction operations are controlled by the sgn functions w.r.t $(s_{ij}(\theta) - \tau_{ij})$. We can approximate the sgn function with the sigmoid function, in order to help the overall objective function become differentiable, and obtain

$$D \approx \left(-\frac{1}{N_-} - \frac{1}{N_+}\right) \sum_{i,j} C(\phi_i, \phi_j) \mathrm{sig}(s_{ij} - \tau) + \frac{1}{N_-} \sum_{i,j} C(\phi_i, \phi_j).$$

We can then derive the gradient of the regularization D with respect to θ, and the total gradient is

$$\nabla E'(\theta) = \nabla E(\theta) + \nabla D(\theta) = \begin{bmatrix} 2\mathrm{vec}(\Omega(G+K)\Omega^T) \\ 2\mathrm{vec}([\Omega \circ (1_A(G+K))]\Omega^T) \\ 2\mathrm{vec}([\Omega \circ (1_A(G+K))\Omega]1_B) \\ 1_B^T(G+K)1_B \end{bmatrix}, \tag{7}$$

where

$$K_{ij} = \frac{\partial D_{ij}}{\partial s_{ij}} = \left(-\frac{1}{N_-} - \frac{1}{N_+}\right) C(\phi_i, \phi_j) \frac{\exp[-(s_{ij} - \tau)]}{(1 + \exp[-(s_{ij} - \tau)])^2}.$$

In the iterative training, we assume N_+ and N_- are fixed in the calculation in each iteration, but are updated after getting new labels for the next iteration. This is done to reduce computational complexity.

3.2. Special Case: 4-Parameter DT-PLDA

We also derive a solution to unsupervised discriminative training on the basis of a special case of DT-PLDA. It has been pointed out that the large number (d^2) of DT-PLDA parameters easily caused over-fitting in training. [23] introduced several ways of constrained DT-PLDA. We chose the 4-parameter constraint to carry out our experiments.

In 4-parameter DT-PLDA, each part of the PLDA LLR score function (3) is scaled as

$$s_{ij} = a_P(\phi_i^T P \phi_j + \phi_j^T P \phi_i) + a_Q(\phi_i^T Q \phi_i + \phi_j^T Q \phi_j) + a_c(\phi_i + \phi_j)^T c + a_k k,$$

where a_P, a_Q, a_c and a_k are trained discriminatively; P, Q, c and k are obtained by generative training beforehand.

In unsupervised 4-parameter DT-PLDA, the loss function is the same as Eq. (5), but the gradient Eq. (7) changes to the following formulations, since the parameters to estimate in the discriminative training are only a_P, a_Q, a_c and a_k,

$$\nabla E'(\theta) = \begin{bmatrix} \partial_{a_P} E'(\theta) \\ \partial_{a_Q} E'(\theta) \\ \partial_{a_c} E'(\theta) \\ \partial_{a_k} E'(\theta) \end{bmatrix} = \begin{bmatrix} 1_B^T[(G+K) \circ (2\Omega^T P \Omega)]1_B \\ 1_B^T[(G+K) \circ 2\mathrm{diag}(\Omega^T Q \Omega)1_B^T]1_B \\ 1_B^T[(G+K) \circ (2\Omega)^T c]1_B \\ 1_B^T(G+K)k1_B \end{bmatrix}.$$

4. Experiments

In this section we experimentally compare our proposed unsupervised DT-PLDA to traditional supervised DT-PLDA, and a conventional method which estimates speaker labels by speaker clustering and applies supervised training.

4.1. Experimental Setup

The proposed method uses unlabeled data in training, data which is supposed to be matched with the evaluation data, and we conducted experiments on the NIST 2008 Speaker Recognition Evaluation (SRE08) Common Condition 7 (English telephone speech), with NIST SRE 2004 and 2005 (also English telephone speech) for training data. We used actual DCF (actDCF) and minimum DCF (minDCF) with the effective prior $P_{\text{eff}} = 0.0917$, given by SRE08, as well as equal error rate (EER) as evaluation metrics. See [13]-[15] for details. Thus the weights in the loss functions (4) and (5) used the same P_{eff}.

In our speaker verification system, an input speech segment was first converted to a sequence of acoustic feature vectors, each of which consisted of 60 features (20 dimensional features consisting of 0th dimension as an energy feature and 1-19th as PLP features, followed by their Δ and $\Delta\Delta$) extracted from a frame of 20 ms width for every 10 ms. An i-vector of 400 dimensions was then extracted from the acoustic feature vectors, a 2048-mixture universal background model (UBM), and a total variability matrix (TVM). We utilized the Kaldi speech recognition toolkit [16] to run these steps. Mean subtraction, whitening, and length normalization [17] were applied to the i-vector, as a pre-processing step.

For training the UBM and the T matrix, we used NIST SRE 2004 (SRE04) and 2005 (SRE05), Switchboard (SWB) II Phase

1 (SB2P1), 2 (SB2P2), and 3 (SB2P3), Switchboard Cellular Part 1 (SBCP1) and 2 (SBCP2), and Fisher. For SRE04, we used speech files included in the training lists of 1, 3, 8 and 16 single-channel conversation sides, and in the test list of 1 single-channel conversation side. For SRE05, we used speech files included in the training lists of 1, 3 and 8 two-channel conversation sides and in the test list of 1 single-channel conversation side. For the SWB datasets, we used all non-empty speech files. For training baseline GT-PLDA models, we used the same SWB data except SB2P, but excluded speech distorted by echo, crosstalk or background noise in accord with the meta-data in the databases. This gave 782 speakers with a total of 4531 utterances. For training DT-PLDA models, we used the same SRE04 and 05 data. MIXER PIN were used as unique speaker IDs for NIST SRE datasets. For the files whose MIXER PIN were missing, we used model IDs as speaker IDs. This included 371 speakers with 4563 utterances.

Given the lack of research in unsupervised discriminative training, we have chosen to use the method described below as a conventional approach with which to compare our proposed method, as it can be easily employed in the same way as GT-PLDA [13]. It contains 2 steps. Step 1: estimate speaker labels by speaker clustering; Step 2: apply supervised DT-PLDA training using the estimated labels. In our experiments, we applied mean-shift clustering, using cosine similarity, for the similarity metric in Step 1. We implemented it on the basis of the sklearn.cluster [24] module in scikit-learn [25], and determined a bandwidth (quantile $= 0.003$) which gives the closest number of speakers (376) to the true number (371).

In the experiments, we used the 4-parameter DT-PLDA formulation described in Section 3.2. In all the experiments with the supervised and unsupervised DT-PLDA, the DT-PLDA parameters were initialized by means of the GT-PLDA framework described in Section 2.1, which was trained with SWB data. The weight of the regularization was set in the range of $10^0 \sim 10^4$. In all DT-PLDA training, 5 iterations were applied, where supervised DT-PLDA achieved the best accuracy.

4.2. Results and Analyses

Table 1 shows the performance of the following 6 systems for three measures: actDCF, minDCF, and EER (%).

$S1$ supervised GT-PLDA

$S2$ supervised DT-PLDA using $S1$ as initialization

$S3$ conventional method described in 4.1

$S4$ – $S6$ are proposed unsupervised DT-PLDA with different weights of regularization b in Eq (5), using $S1$ as initialization

$S4$ $b = 1, 10, 10^2$

$S5$ $b = 10^3$

$S6$ $b = 10^4$

In a comparison of systems $S1$ and $S2$, we achieved results consistent with other studies that show DT-PLDA improves the performance of speaker verification for all of the three measures. As expected, the reduction in actDCF is much more than that in minDCF and EER, due to the fact that the objective function we used was the weighted loss function, which follows the definition of actDCF. The objective function was not meant to be responsible for improving minDCF and EER, but it did produce improvement.

In system $S3$, despite the good estimation in the number of clusters (376, close to the true number 371), it produced higher error rates for all of the three measures as compared with the GT-PLDA system, which was the initialization of the discriminative training in $S3$. This indicates that the conventional method failed in domain adaptation. In a comparison of systems $S2$ and $S3$, it suggests the fact that the conventional method is heavily affected by clustering performance. All of the three systems with the proposed method, $S4$, $S5$, and $S6$, were better for all three measures as compared with $S1$, which indicates successful adaptation of the system to the target domain. In contrast with the $S2$, the supervised DT-PLDA, which is supposed to represent the upper bound in performance (the lower bounds for the three measures), actDCF and EER in $S4, S5$, and $S6$ were higher, as expected, while minDCF even exceeded that of $S2$. We are unable to explain this, but we may note again that our objective function focuses on minimizing actDCF. In this respect, the results in Table 1 can be considered reasonable.

Systems	actDCF	minDCF	EER(%)
$S1$: supervised GT-PLDA	0.452	0.321	6.76
$S2$: supervised DT-PLDA	**0.320**	0.320	**6.71**
$S3$: conventional	0.510	0.371	7.14
$S4$: proposed (b=1, 10, 10^2)	0.339	0.309	6.76
$S5$: proposed (b=10^3)	0.336	0.309	6.77
$S6$: proposed (b=10^4)	0.351	**0.308**	6.73

Table 1: *Performance of the 6 systems. Bold face denotes the best performance in each column.*

A large difference in the weight of the regularization in $S4, S5$, and $S6$, resulting in a slight difference in speaker verification performance, indicates that cosine similarity has a certain correlation with GT-PLDA initialization. Among the three system, the lowest actDCF, was achieved in $S5$, which shows that 10^3 is the most appropriate weight for this task. The proposed method is relatively robust with respect to the weight of the regularization, and at the same time shows improvement.

5. Summary and Future Work

We have proposed unsupervised DT-PLDA that uses a regularization term derived from data statistics, to constrain the iterative training. It follows the idea of traditional DT-PLDA that uses GT-PLDA for its initialization. Working under the assumption that PLDA scoring and scoring using cosine similarity are weakly correlated, we adopted cosine similarity for the regularization in formulation and then conducted experiments. The objective function was set as the weighted loss function specifically to optimize actDCF. We have shown experimentally that the proposed method successfully adapted the system to the target domain, and performed almost as well as supervised DT-PLDA for actDCF. Given that this was the first attempt at unsupervised discriminative PLDA that we know of, we also conducted experiments for a method which is easy to employ (speaker clustering + supervised DT-PLDA). As expected, the proposed method outperformed it. Future issues include the implementation and evaluation of general unsupervised DT-PLDA. We also intend to explore the possibility of employing other data statistics such as GT-PLDA scoring and SVM scoring for regularization.

6. Acknowledgements

We especially thank my former lab-mate and good friend Dr. Johan Rohdin for his help and support. This work is greatly based on his thesis and DT-PLDA script under the supervision of Prof. Koichi Shinoda of the Tokyo Institute of Technology. We also thank our NEC colleagues for their valuable comments and advice.

7. References

[1] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 531–542.

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

[3] 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.

[4] N. Dehak, R. Dehak, P. Kenny, N.Brummer, and P. Ouellet, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," *INTERSPEECH*, pp. 1559–1562, 2009.

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

[6] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[7] N. Brummer, "Measuring, refining and calibrating speaker and language information extracted from speech," *PhD thesis, Stellenbosch: University of Stellenbosch*, 2010.

[8] P. M. Bousquet, J. F. Bonastre, and D. Matrouf, "Exploring some limits of gaussian PLDA modeling for i-vector distributions," *Odyssey*, pp. 41–47, 2014.

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

[10] S. Cumani, N. Brummer, L. Burget, and P. Laface., "Fast discriminative speaker verification in the i-vector space," *ICASSP*, pp. 4852–4855, 2011.

[11] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[12] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," *In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems (NIPS), MIT Press*, pp. 841–848, 2002.

[13] S. Dey, S. Madikeri, and P. Motlicek, "Information theoretic clustering for unsupervised domain-adaptation," *Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[14] M. Senoussaoui, P. Kenny, P. Dumouchel, and T. Stafylakis, "Efficient iterative mean shift based cosine dissimilarity for multi-recording," *Proc. of ICASSP*, 2013.

[15] I. Shapiro, N. Rabin, I. Opher, and I. Lapidot, "Clustering short push-to-talk segments," *Proc. of INTERSPEECH*, 2015.

[16] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 564–577, 2003.

[17] S. Novoselov, T. Pekhovsky, and K. Simonchik, "STC speaker recognition system for the NIST i-vector challenge," *Proc. of Odyssey*, 2014.

[18] N. Brummer and E. de Villiers, "The speaker partitioning problem," *In Odyssey*, pp. 194–201, 2010.

[19] J. Rohdin, S. Biswas, and K. Shinoda, "Constrained discriminative PLDA training for speaker verification," *In ICASSP*, 2014.

[20] V. Franc and S. Sonnenburg, "Optimized cutting plane algorithm for support vector machines," *Proc. ICML*, pp. 320–327, 2008.

[21] C. H. Teo, A. Smola, S. V. Vishwanathan, and Q. V. Le, "Bundle methods for regularized risk minimization," *J. Mach. Learn. Res*, vol. 11, pp. 311–365, 2010.

[22] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[23] J. Rohdin, "A study on discriminative training techniques for speaker verification," *PhD thesis, Tokyo Institute of Technology*, 2015.

[24] "sklearn-cluster," *URL https://sites.google.com/site/nikobrummer*.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.