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.
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 \( \xi \) corresponding to a speech utterance can be modeled as

\[ \xi = \bar{\xi} + T \phi, \]

where \( \phi \) is a random vector known as the i-vector, \( T \) is a basis for the total variability space for speaker and channel variability of \( \xi \), and \( \bar{\xi} \) is the mean of \( \xi \). It is assumed that \( \phi \) 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]

\[ \phi = m + V y + D z, \] (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, \( \phi_i \) and \( \phi_j \), PLDA calculates a log-likelihood ratio (LLR) \( s_{ij} \) between two hypotheses: \( H_s \) – they are from the same speaker or \( H_w \) – they are from different speakers,

\[ s_{ij} = \phi_i^T P \phi_j + \phi_j^T P \phi_i + \phi_i^T Q \phi_i + \phi_j^T Q \phi_j + (\phi_i + \phi_j)^T c + k, \] (2)

where

\[ P = \frac{1}{2} \Sigma_{sot}^{-1} \Sigma_{sit}^{-1} \Sigma_{rot}^{-1} \Sigma_{rto}^{-1}, \]
\[ Q = \frac{1}{2} \Sigma_{sot}^{-1} (\Sigma_{sit} - \Sigma_{sot} \Sigma_{sit}^{-1} \Sigma_{rot})^{-1}, \]
\[ c = -2(P + Q)m, \]
\[ k = \frac{1}{2} (log \Sigma_{sit} - log \Sigma_{rot} - \Sigma_{sot} \Sigma_{sit}^{-1} \Sigma_{rot}), \]
\[ +m^T (2(P + Q)m), \]

where \( \Sigma_s = V V^T, \Sigma_w = D D^T \) are between- and within-class covariance matrices, respectively. \( \Sigma_{sot} = \Sigma_s + \Sigma_w \).

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 \( \theta = vec([P, Q, c, k]) \), where vec(•) 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:

\[ E(\theta) = N \sum_{i,j=1}^{l_{ij}} \frac{P_{\text{eff}} l_{ij}}{N_+} + \sum_{i,j=1}^{l_{ij}} \frac{1 - P_{\text{eff}} l_{ij}}{N_-}, \] (4)

where \( l_{ij} = l(t_{ij}, s_{ij}()) \) is the loss function for a trial \( (\phi_i, \phi_j) \) 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; \( P_{\text{eff}} \) is known as the effective prior. Such weight settings \( P_{\text{eff}}/N_+ \) 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),

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

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

By minimizing \( E(\theta) \), \( \theta \) 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]:

\[ \nabla E(\theta) = \begin{bmatrix} \nabla_P E(\theta) \\ \nabla_Q E(\theta) \\ \nabla_c E(\theta) \\ \nabla_k E(\theta) \end{bmatrix} = \begin{bmatrix} 2vec((\Omega G \Gamma T) \phi) \\ 2vec((\Omega \circ (1 \sigma (G)) \Gamma T) \phi) \\ 2vec((\Omega \circ (1 \sigma (G)) \Gamma) B) \\ G \end{bmatrix}, \]

where \( 1_A \) is a \( d \times n \) matrix of ones, and \( B \) is a \( n \times 1 \) matrix of ones, \( \Omega = [\phi_1 \ldots \phi_n] \), \( \sigma \) denotes the element-wise multiplication of two matrices, and

\[ G = \begin{bmatrix} N P_{\text{eff}} \alpha(l(t_{ij}, s_{ij})) \phi_i \phi_j \theta_{ij} \\ N (1 - P_{\text{eff}}) \alpha(l(t_{ij}, s_{ij})) \theta_{ij} \end{bmatrix} : t_{ij} = 1 \]
\[ : t_{ij} = -1. \]

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 cross-entropy 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

$$E'(\theta) = E(\theta) + bD(\theta),$$

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

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

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 $s_{ij}(\theta) \geq \tau$. 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

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

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

$$\nabla E'(\theta) = \nabla E(\theta) + \nabla D(\theta)$$

$$= \begin{bmatrix} 2\text{vec}(\Omega(G + K))\Omega^T) \\ 2\text{vec}(\Omega \circ (1_A(G + K))\Omega^T) \\ 2\text{vec}(\Omega \circ (1_A(G + K))\Omega)1_B \\ 1_B^T(G + K)1_B \\ 1_B^T(G + K)\circ (2\Omega^T P\Omega)1_B \end{bmatrix},$$

where

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

In the iterative training, we assume $N_n$ and $N_p$ 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 term of the PLDA LLR score function (3) is scaled as

$$s_{ij} = a_P(\phi_i^T P \phi_j + a_Q(\phi_i^T Q \phi_j + \phi_j^T Q \phi_j)$$

$$+ a_c(\phi_i + \phi_j)^T c + a_k k,$$

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 $a_k$.

$$\nabla E'(\theta) = \begin{bmatrix} \partial a_P E'(\theta) \\ \partial a_Q E'(\theta) \\ \partial a_c E'(\theta) \\ \partial a_k E'(\theta) \end{bmatrix} = \begin{bmatrix} 1_B^T(G + K) \circ (2\Omega^T P\Omega)1_B \\ 1_B^T(G + K) \circ (2\Omega^T Q\Omega)1_B \\ 1_B^T(G + K) \circ (2\Omega^T c)1_B \\ 1_B^T(G + K)k1_B \end{bmatrix}.$$
Table 1: Performance of the 6 systems. Bold face denotes the best performance in each column.

<table>
<thead>
<tr>
<th>Systems</th>
<th>actDCF</th>
<th>minDCF</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: supervised GT-PLDA</td>
<td>0.352</td>
<td>0.321</td>
<td>6.76</td>
</tr>
<tr>
<td>S2: supervised DT-PLDA</td>
<td>0.330</td>
<td>0.320</td>
<td>6.71</td>
</tr>
<tr>
<td>S3: conventional</td>
<td>0.351</td>
<td>0.371</td>
<td>7.14</td>
</tr>
<tr>
<td>S4: proposed (b=1, 10, 10^4)</td>
<td>0.339</td>
<td>0.309</td>
<td>6.76</td>
</tr>
<tr>
<td>S5: proposed (b=10^5)</td>
<td>0.336</td>
<td>0.309</td>
<td>6.77</td>
</tr>
<tr>
<td>S6: proposed (b=10^5)</td>
<td>0.351</td>
<td>0.308</td>
<td>6.73</td>
</tr>
</tbody>
</table>

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


