Duration Dependent Covariance Regularization in PLDA Modeling for Speaker Verification

Weicheng Cai[2,3], *Ming Li* [1,2] , *Lin Li* [4], *Qingyang Hong* [4]

[1]SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, China
[2]SYSU-CMU Shunde International Joint Research Institute, China
[3]School of Information Science and Technology, Sun Yat-sen University, China
[4]School of Information Science and Technology, Xiamen University, China
liming46@mail.sysu.edu.cn

10.21437/Interspeech.2015-278

Abstract

In this paper, we present a covariance regularized probabilistic linear discriminant analysis (CR-PLDA) model for text independent speaker verification. In the conventional simplified PLDA modeling, the covariance matrix used to capture the residual energies is globally shared for all i-vectors. However, we believe that the point estimated i-vectors from longer speech utterances may be more accurate and their corresponding covariances in the PLDA modeling should be smaller. Similar to the inverse 0^{th} order statistics weighted covariance in the i-vector model training, we propose a duration dependent normalized exponential term containing the duration normalizing factor μ and duration extent factor ν to regularize the covariance in the PLDA modeling. Experimental results are reported on the NIST SRE 2010 common condition 5 female part task and the NIST 2014 i-vector machine learning challenge, respectively. For both tasks, the proposed covariance regularized PLDA system outperforms the baseline PLDA system by more than 13% relatively in terms of equal error rate (EER) and norm minDCF values.

Index Terms: PLDA, covariance regularization, i-vector, speaker verification, duration

1. Introduction

Total variability i-vector modeling has gained significant attention in both speaker verification (SV) and language identification (LID) domains due to its excellent performance, compact representation and small model size [1, 2, 3]. In this modeling, first, zero-order and first-order Baum-Welch statistics are calculated by projecting the MFCC features on those Gaussian Mixture Model (GMM) components using the occupancy posterior probability. Second, in order to reduce the dimensionality of the concatenated statistics vectors, a single factor analysis is adopted to generate a low dimensional total variability space which jointly models language, speaker and channel variabilities all together [1]. Third, within this i-vector space, variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [4], Linear Discriminative Analysis (LDA) and Nuisance Attribute Projection (NAP) [5], are performed to reduce the variability for the subsequent modeling methods (e.g., Support Vector Machine [6], Sparse Representation [7], Probabilistic Linear Discriminant Analysis (PLDA) [8, 9, 10], etc.).

Conventionally, in the i-vector framework, the tokens for calculating the zero-order and first-order Baum-Welch statistics are the MFCC features trained GMM components. Such choice of token units may not be the optimal solution. Recently, the generalized i-vector framework [11, 12, 13, 14, 15] has been proposed. In this framework, the tokens for calculating the zero-order statistics have been extended to tied triphone states, monophone states, tandem features trained GMM components, bottleneck features trained GMM components, etc. The features for calculating the first-order statistics have also been extended from MFCC to feature level acoustic and phonetic fused features [13]. The phonetically-aware tokens trained by supervised learning can provide better token separation and discrimination. This enables the system to compare different speakers' voices token by token with more accurate token alignment, which leads to significant performance improvement on the text independent speaker verification task [11, 12, 13, 14, 15].

After i-vectors are extracted, among the aforementioned supervised learning techniques, PLDA is widely adopted and considered as the state-of-the-art back-end modeling approach [8, 9, 10, 16, 17, 18, 19, 20]. PLDA is a generative model that incorporates both within-speaker and between-speaker variations. Generally, we model the i-vectors with a Gaussian distribution assumption(G-PLDA). After we learned the model parameters by expected maximization (EM) algorithm, the scoring is based on a hypothesis testing framework.

Recently, It is shown in[21] that the performance of PLDA on short utterance is degraded. Duration variability has also been investigated in the i-vector space using PLDA model [17][22][23][24]. This motivates us to incorporate the speech duration information directly into the PLDA model training and generate a more accurate model.

In the standard simplified PLDA modeling [10], the within-speaker variations can be considered as the residual that can't be interpreted by the speaker space. The covariance matrix used to model these residuals is globally shared by all i-vectors, no matter whether the corresponding utterances' durations are long or short. We believe that the point estimated i-vectors from longer speech utterances may be more accurate and their corresponding covariances in the PLDA modeling should be smaller. Motivated by the inverse 0^{th} order statistics weighted covariance in the i-vector model training[25][26], we propose a duration dependent normalized exponential term containing the duration

This research is supported in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), SYSU-CMU Shunde International Joint Research Institute and CMU-SYSU Collaborative Innovation Research Center.

normalizing factor μ and duration extent factor ν to regularize the covariance in the PLDA modeling. The numerical value of μ and ν are tuned to achieve good performance. Specially, when ν is set to constant 0, this model converges back to the baseline PLDA model.

2. Methods

2.1. I-vector

In the total variability space, there is no distinction between the speaker effects and the channel effects. Rather than separately using the eigenvoice matrix $\boldsymbol{V}$ and the eigenchannel matrix $\boldsymbol{U}$[27], the total variability space simultaneously captures the speaker and channel variabilities[2]. Given a C component GMM UBM model λ with $\lambda_c = \{p_c, \boldsymbol{u_c}, \boldsymbol{\Sigma_c}\}$, $c = 1, ..., C$ and an utterance with a L frame feature sequence $\{\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_L}\}$, the 0^{th} and centered 1^{st} order Baum-Welch statistics on the UBM are calculated as follows:

$$N_c = \sum_{t=1}^{L}(P(c|\boldsymbol{y_t}, \lambda) \tag{1}$$

$$\boldsymbol{F_c} = \sum_{t=1}^{L}(P(c|\boldsymbol{y_t}, \lambda)(\boldsymbol{y_t} - \boldsymbol{u_c}) \tag{2}$$

where $c = 1, ..., C$ is the GMM component index and and $P(c|\boldsymbol{y_t}, \lambda)$ is the occupancy probability for $\boldsymbol{y_t}$ on λ_c. The corresponding centered mean supervector $\tilde{\boldsymbol{F}}$ is generated by concatenating all the $\tilde{\boldsymbol{F}}_{\boldsymbol{c}}$ together:

$$\tilde{\boldsymbol{F}}_{\boldsymbol{c}} = \frac{\sum_{t=1}^{L}(P(c|\boldsymbol{y_t}, \lambda)(\boldsymbol{y_t} - \boldsymbol{u_c})}{\sum_{t=1}^{L}(P(c|\boldsymbol{y_t}, \lambda)} \tag{3}$$

The centered GMM mean supervector $\tilde{\boldsymbol{F}}$ can be projected as follows:

$$\tilde{\boldsymbol{F}} \rightarrow \boldsymbol{T}\boldsymbol{x} \tag{4}$$

where $\boldsymbol{T}$ is a rectangular total variability matrix of low rank and $\boldsymbol{x}$ is the so-called i-vector[2]. Considering a C-component GMM and D dimensional acoustic features, the total variability matrix $\boldsymbol{T}$ is $CD \times K$ matrix which can be estimated the same way as learning the eigenvoice matrix $\boldsymbol{V}$ in [28]except that here we consider that every utterance is produced by a new speaker[2].

Given the centered mean supervector $\tilde{\boldsymbol{F}}$ and total variability matrix $\boldsymbol{T}$, the i-vector is computed as follows[2]:

$$\boldsymbol{x} = (\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\tilde{\boldsymbol{F}} \tag{5}$$

where $\boldsymbol{N}$ is a diagonal matrix of dimension $CD \times CD$ whose diagonal blocks are $N_c\boldsymbol{I}$,$c = 1, ..., C$ and $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $CD \times CD$ estimated in the factor analysis training step. It models the residual variability not captured by the total variability matrix $\boldsymbol{T}$[2]. Covariance $\boldsymbol{\Sigma}$ is also updated iteratively.

2.2. PLDA Baseline

2.2.1. PLDA Training

We assume that the training data consists of j utterances from i speakers and denote the j^{th} i-vector of the i^{th} speaker by $\boldsymbol{\eta_{ij}}$. We assume that the data are generated in the following way[8]:

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\phi}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{ij} \tag{6}$$

The speaker term $\boldsymbol{\phi}\boldsymbol{\beta}_i$ is dependent on the speaker index and the noise term $\boldsymbol{\epsilon}_{ij}$ is independent for every i-vector and used to model the within-speaker variabilities.

Suppose there are M_i i-vectors from the i^{th} speaker, the sufficient statistics are denoted as follows:

$$\widetilde{\boldsymbol{x}}_i = \sum_{j=1}^{M_i}(\boldsymbol{\eta_{ij}}) \tag{7}$$

$$\boldsymbol{F}_i = \frac{\widetilde{\boldsymbol{x}}_i}{M_i} \tag{8}$$

For the i^{th} speaker the prior and conditional distribution is defined as following multivariate Gaussian distributions:

$$P(\boldsymbol{F}_i|\boldsymbol{\beta}_i) = \boldsymbol{N}(\boldsymbol{\phi}\boldsymbol{\beta}_i, \frac{\boldsymbol{\Sigma}}{M_i}) \qquad P(\boldsymbol{\beta}_i) = \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{9}$$

The Expectation Maximization(EM) algorithm is employed in the modeling training. In the E-step, the posterior distribution of the hidden variable β_i given the observed $\boldsymbol{F_i}$ is

$$P(\boldsymbol{\beta}_i|\boldsymbol{F}_i) = \boldsymbol{N}((\boldsymbol{I} + \boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{M_i}\boldsymbol{\Sigma}^{-\boldsymbol{1}}\boldsymbol{\phi})^{-\boldsymbol{1}}\boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{M_i}\boldsymbol{\Sigma}^{-\boldsymbol{1}}\boldsymbol{F_i}, \boldsymbol{I} + \boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{M_i}\boldsymbol{\Sigma}^{-\boldsymbol{1}}\boldsymbol{\phi}) \tag{10}$$

Then, in the M-step, to maximize the conditional expectation of the following log-likelihood,

$$\log \quad \{\prod_{i=1}^{N}\prod_{j=1}^{M_i}(P(\boldsymbol{\eta}_{ij}, \boldsymbol{\beta}_i)\} \tag{11}$$

the updated ϕ and Σ are calculated as follows

$$\boldsymbol{\phi} = \Big(\sum_i M_i\boldsymbol{F}_i\boldsymbol{E}(\boldsymbol{\beta}_i^T)\Big)\Big(\sum_i M_i\boldsymbol{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T)\Big)^{-1} \tag{12}$$

$$\boldsymbol{\Sigma} = \frac{\sum_i\sum_j \boldsymbol{\eta}_{ij}[\boldsymbol{\eta}_{ij}^T - \boldsymbol{E}(\boldsymbol{\beta}_i)^T\boldsymbol{\phi}^T]}{\sum_i M_i} \tag{13}$$

2.2.2. PLDA Scoring

In the speaker verification task, given a trial with two i-vectors $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$, we are interested in testing two alternative hypotheses. $\boldsymbol{H}_1$: both $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$ are from the same speaker and they share the same speaker identity latent variable $\boldsymbol{\beta_i} = \boldsymbol{\beta_j}$; $\boldsymbol{H}_0$: they come from different speakers and the underlying hidden variables $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are different[8][10]. The verification score can now be computed as the loglikelihood ratio of these two hypotheses.

$$score = \log \quad \frac{P(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j|\boldsymbol{H}_1)}{P(\boldsymbol{\eta}_i|\boldsymbol{H}_0)P(\boldsymbol{\eta}_j|\boldsymbol{H}_0)} \tag{14}$$

Since the corresponding distribution are all multivariate Gaussians, the score can be denoted in quadratic terms[10]:

$$\begin{aligned} score &= \log \boldsymbol{N}(\begin{bmatrix}\boldsymbol{\eta}_i\\ \boldsymbol{\eta}_j\end{bmatrix}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{tot} & \boldsymbol{\Sigma}_{ac}\\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_{tot}\end{bmatrix}) \\ &\quad - \log \boldsymbol{N}(\begin{bmatrix}\boldsymbol{\eta}_i\\ \boldsymbol{\eta}_j\end{bmatrix}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{tot} & 0\\ 0 & \boldsymbol{\Sigma}_{tot}\end{bmatrix}), \\ &= \boldsymbol{\eta}_i^T\boldsymbol{Q}\boldsymbol{\eta}_i + \boldsymbol{\eta}_j^T\boldsymbol{Q}\boldsymbol{\eta}_j^T + 2\boldsymbol{\eta}_i^T\boldsymbol{P}\boldsymbol{\eta}_j + const, \end{aligned} \tag{15}$$

where $\boldsymbol{\Sigma}_{tot}$, $\boldsymbol{\Sigma}_{ac}$,$\boldsymbol{P}$,$\boldsymbol{Q}$ are denoted as follows:

$$\begin{aligned}\boldsymbol{\Sigma}_{tot} &= \boldsymbol{\phi}\boldsymbol{\phi}^T + \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}_{ac} &= \boldsymbol{\phi}\boldsymbol{\phi}^T \\ \boldsymbol{Q} &= \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\ \boldsymbol{P} &= \boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}\end{aligned} \tag{16}$$

2.3. Covariance Regularized PLDA

In the PLDA model training, to consider the duration variabilities among different utterances, we add a duration dependent normalized exponential term on top of the conventional covariance $\boldsymbol{\Sigma}$ to model the variance of ϵ_{ij}, assuming that those utterances with longer duration generally have smaller covariances.

$$\boldsymbol{\Sigma} \to (\frac{d_{ij}}{\mu})^{-\nu}\boldsymbol{\Sigma} \tag{17}$$

where μ is a duration normalizing factor, and ν is a duration extent factor that describes the regularization strength. We can adapt these two coefficients to discover the influence of the duration dependent covariance regularization to the PLDA system performance. It is worth nothing that when $\mu \neq 0$ and ν is set to 0, the regularized covariance becomes the original globally shared$\boldsymbol{\Sigma}$.

Hence the conditional distribution of the i^{th} speaker's average i-vector $\boldsymbol{F_i}$ is regularized as following :

$$P(\boldsymbol{F}_i|\boldsymbol{\beta}_i) = \boldsymbol{N}(\boldsymbol{\phi}\boldsymbol{\beta}_i, \quad \frac{\sum_{j=1}^{M_i}(\frac{d_{ij}}{\mu})^{-\nu}}{M_i^2}\boldsymbol{\Sigma}) \tag{18}$$

And the posterior distribution of the hidden variable $\boldsymbol{\beta}$ given the observed $\boldsymbol{F}$ is:

$$\begin{aligned}&P(\boldsymbol{\beta}_i|\boldsymbol{F}_i) = \\ &\boldsymbol{N}\Bigg\{\left(\boldsymbol{I} + \frac{M_i^2}{\sum_{j=1}^{M_i}(\frac{d_{ij}}{\mu})^{-\nu}}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)^{-1}\frac{M_i}{\sum_{j=1}^{M_i}(\frac{d_{ij}}{\mu})^{-\nu}}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}, \\ &\left(\boldsymbol{I} + \frac{M_i^2}{\sum_{j=1}^{M_i}(\frac{d_{ij}}{\mu})^{-\nu}}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\right)^{-1}\Bigg\}\end{aligned} \tag{19}$$

The complete data log likelihood of all the training utterance is denoted as follows:

$$\boldsymbol{J} = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\{\log(P(\boldsymbol{\eta}_{ij}|\boldsymbol{\beta}_i)) + \log(P(\boldsymbol{\beta}_i))\} \tag{20}$$

In the M-step, after removing those non-relevant items, we need to maximize the following conditional expected complete data log likelihood $E(J)$:

$$\begin{aligned}\boldsymbol{E(J)} = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\{&-\frac{\log|\boldsymbol{\Sigma}|}{2} - \frac{(\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{\eta}_{ij}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_{ij}}{2} \\ &- \frac{(\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{tr}(\boldsymbol{\phi}\boldsymbol{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^{\boldsymbol{T}})\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1})}{2} + (\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{\eta}_{ij}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}\boldsymbol{E}(\boldsymbol{\beta}_i)\}\end{aligned} \tag{21}$$

By letting the derivatives of $\boldsymbol{E(J)}$ towards $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma^{-1}}$ to be 0, we can get regularized updating equation for $\boldsymbol{\phi}$ and $\boldsymbol{\Sigma}$ as following:

$$\boldsymbol{\phi} = (\sum_i\sum_j(\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{\eta}_{ij}\boldsymbol{E}(\boldsymbol{\beta_i})^T)(\sum_i\sum_j(\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{E}(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T))^{-1} \tag{22}$$

$$\boldsymbol{\Sigma} = \frac{\sum_i\sum_j(\frac{d_{ij}}{\mu})^{\nu}\boldsymbol{\eta}_{ij}(\boldsymbol{\eta}_{ij}^T - \boldsymbol{E}(\boldsymbol{\beta}_i)^T\boldsymbol{\phi}^T)}{\sum_i M_i} \tag{23}$$

when $u \neq 0$ and v is set to 0, the whole aforementioned covariance regularized training framework becomes back the original solutions as in 2.2.1.

Among μ and ν these two parameters, ν plays a more important role because μ only serves as a constant reference duration used for pre-normalization. In order to make ν more comparable on different data sets, we set μ as the averaged duration of all the PLDA training utterances denoted as $\bar{d}$ without loss of generality.

$$\boldsymbol{\Sigma} \to (\frac{d_{ij}}{\bar{d}})^{-\nu}\boldsymbol{\Sigma} \tag{24}$$

For the CR-PLDA scoring, we ignore the duration information and still adopt the baseline PLDA scoring method in 2.2.2 due to its simplification, robustness and efficiency.

3. Experimental Results

3.1. NIST SRE 2010

We first conducted experiments on the NIST 2010 speaker recognition evaluation(SRE) corpus[29]. Our focus is the female part of the common condition 5(a subset of tel-tel) in the core task with the original trials(not the extended ones). We used the equal error rate(EER), the 2008(old) and 2010(new) normalized minimum decision cost value(norm minDCF) as the metrics for evaluation[29]. We adopt the hybrid-GMM-hybrid feature level fusion strategy in[13]. For cepstral feature extraction, a 25ms Hamming window with 10ms shifts was adopted. Each utterance was converted into a sequence of 36- dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. For phonetic feature extraction, we employed an English phoneme recognizer[30] to perform the voice activity detection(VAD) and output the frame level monophone states posterior probability. After log, PCA and MVN, the resulted 52 dimensional tandem features are fused with MFCC at the feature level to get the 88 dimensional hybrid feature[13]. Feature warping is applied to mitigate variabilities.

The i-vector training data for the NIST 2010 task include Switchboard II part1 to part3, NIST SRE 2004, 2005, 2006 and 2008 corpora on the telephone channel. We trained a genderdependent GMM UBM model with 1024 mixture components. The PLDA models were trained on a subset of Switchboard II part1 to part3 and NIST 2008 corpora on the telephone channel, which amounted to 1,898 speakers and 15,480 speech files. The averaged post VAD duration for these utterances is 9.55 seconds. The dimensionality of the i-vectors and the rank of the speaker-specific subspace in the PLDA model are 500 and 150, respectively.

The performances of the proposed CR-PLDA systems on the NIST SRE 2010 common condition 5 female part task with different ν parameters are shown in Table 1. It is observed that the CR-PLDA system with ν = 1.5 outperformed the PLDA

baseline by 20% and 14% relative error reduction in terms of EER and norm new minDCF, respectively. Furturemore, Figure 1 shows the Detection Error Trade-off (DET) curves of the CR-PLDA system (ID 7) and the baseline PLDA system (ID 1). We can find out that the proposed CR-PLDA method achieves significant performance enhancement.

Table 1: Performance of the proposed CR-PLDA systems on the NIST SRE 2010 common condition 5 female part task

ID	Coefficient	EER (%)	norm minDCF	
			new[29]	old[31]
1	$\boldsymbol{\nu = 0}$	**2.82**	**0.311**	**0.126**
2	ν=0.25	2.61	0.288	0.125
3	ν=0.5	2.81	0.291	0.127
4	ν=0.75	2.84	0.294	0.125
5	ν=1	2.56	0.274	0.126
6	ν=1.25	2.32	0.271	0.126
7	ν=1.4	2.31	0.266	0.126
8	$\boldsymbol{\nu = 1.5}$	**2.26**	**0.268**	**0.125**
9	ν=1.75	2.26	0.280	0.125
10	ν=2	2.40	0.271	0.120

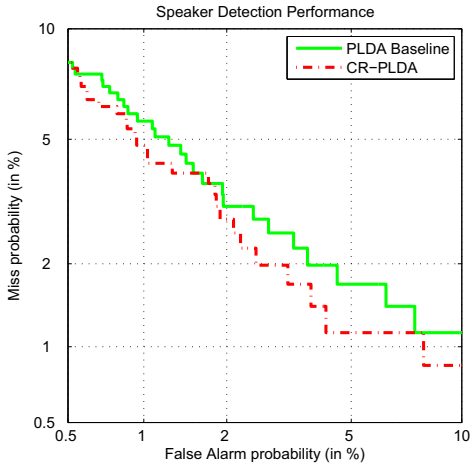

Figure 1: DET curves of the baseline PLDA system(ID 1) and the CR-PLDA system (ID 7) on the NIST SRE 2010.

3.2. NIST machine learning i-vector challenge 2014

For the NIST 2014 i-vector challenge, the 600 dimensional ivectors were trained by previous years NIST SRE data and provided by the organizers. Along with each i-vector, utterance duration information was also provided. The development data containing 36,573 speech files and 4,959 speakers were used to train the PLDA model. In the testing phase, there are 6,530 target i-vectors from 1,306 speakers and 9,634 test i-vectors. The trials were divided into two separate subsets, namely the progress subset, and the evaluation subset [32].

The performances of the proposed CR-PLDA systems on the NIST 2014 i-vector machine learning challenge are shown in Table 2. The baseline system and the best performing system on both progress and evaluation sets are highlighted. When ν is set to 1.5, the proposed CR-PLDA system achieves 13% and 9% relative EER reduction against the PLDA baseline on the the progress and evaluation subset, respectively. Figure 2 demonstrates a smooth and consistent improvement of the proposed CR-PLDA system along the DET curves for both sets. It is worth noting that the best performance was achieved when $\nu = 1.5$ rather than $\nu = 1$. In the i-vector training, the covariance is weighted by the inverse 0^{th} order statistics ($\nu = 1$). Our future works would include applying different covariance regularization parameters in the i-vector model training.

Table 2: Performance of the proposed CR-PLDA systems on the NIST 2014 i-vector machine learning challenge

ID	Coefficient	EER (%)		norm minDCF[32]	
		prog	eval	prog	eval
1	$\boldsymbol{\nu = 0}$	**3.19**	**2.85**	**0.252**	**0.233**
2	ν=0.25	3.22	2.81	0.252	0.231
3	ν=0.5	3.19	2.77	0.251	0.229
4	ν=0.75	3.13	2.71	0.248	0.222
5	ν=1	3.10	2.62	0.243	0.220
6	ν=1.25	3.10	2.62	0.255	0.220
7	ν=1.4	2.89	2.58	0.241	0.220
8	$\boldsymbol{\nu = 1.5}$	**2.76**	**2.58**	**0.240**	**0.221**
9	ν=1.75	2.86	2.62	0.257	0.237
10	ν=2	3.25	2.85	0.304	0.292

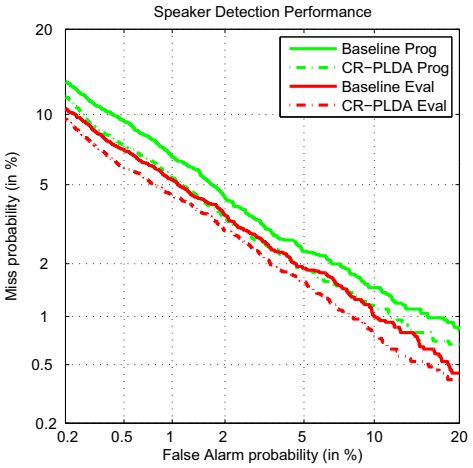

Figure 2: DET curves of the baseline PLDA system(ID 1) and the CR-PLDA system(ID 7) on the NIST i-vector machine learning challenge 2014.

4. Conclusion

This paper presents a covariance regularized PLDA training method by taking the duration information directly into the PLDA modeling. We believe that the point estimated i-vectors from longer speech utterances may be more accurate and their corresponding covariances in the PLDA modeling should be smaller. Similar to the inverse 0^{th}[28] order statistics weighted covariance in the i-vector model training, we propose a duration dependent normalized exponential term containing the duration normalizing factor and the duration extent factor to regularize the covariance in the PLDA modeling. If duration extent factor equals to 0, the corresponding CR-PLDA method is exactly the same as the conversional simplified PLDA baseline. The best performance is achieved when the duration normalizing factor and the duration extent factor are set to the averaged duration and 1.5, respectively.

5. References

[1] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. INTERSPEECH*, 2011, pp. 857–860.

[2] 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, no. 4, pp. 788–798, 2011.

[3] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *Proc. INTERSPEECH*, 2011, pp. 861–864.

[4] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. INTERSPEECH*, 2006, vol. 4, pp. 1471–1474.

[5] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[6] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, vol. 1, pp. 97–100.

[7] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proc. INTERSPEECH*, 2011, pp. 4548–4551.

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

[9] P. Matejka, O. Glembek, F. Castaldo, MJ Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Proc. ICASSP*, 2011, pp. 4828–4831.

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

[11] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014.

[12] P Kenny, V Gupta, T Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014.

[13] M. Li and Wenbo Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Proc. INTERSPEECH*, 2014.

[14] L.F. D'Haro, R. Cordoba, C. Salamea, and J.D. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 5379–5383.

[15] H.P. Wang, C.C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta mlp features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, 2013.

[16] R. Saeidi, K.A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.M. Bousquet, E. Khoury, et al., "I4u submission to nist sre 2012: A large-scale collaborative effort for noise-robust speaker verification," 2013.

[17] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, P. Dumouchel, et al., "Plda for speaker verification with utterances of arbitrary duration," in *Proc. ICASSP*. IEEE, 2013, pp. 7649–7653.

[18] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP*. IEEE, 2012, pp. 4253–4256.

[19] G. Liu, T. Hasan, H. Boril, and J. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Proc. ICASSP*. IEEE, 2013, pp. 7755–7759.

[20] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification," *Digital Signal Processing*, 2014.

[21] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "Plda based speaker recognition on short utterances," in *Proc. Odyssey*, 2012.

[22] T. Hasan, R. Saeidi, J. Hansen, and D. Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. ICASSP*, 2013.

[23] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improveing plda speaker verification with limited development data," in *Proc. ICASSP*, 2014.

[24] P. Rajan, A. Afanasyev, V. Hautamaki, and T. Kinnumen, "From Single to Multiple Enrollment i-Vectors: Practical PLDA Scoring Variants for Speaker Verification," *Digital Signal Processing*, vol. 31, pp. 93–101, 2014.

[25] M. Li and S. Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Computer speech and language*, vol. 28, pp. 940–958, 2014.

[26] M. Li, "Speaker verification with the mixture of gaussian factor analysis based representation," in *Proc. ICCSSP*, 2015.

[27] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[28] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[29] NIST, "The nist 2010 speaker recognition evaluation plan," *www.itl.nist.gov/iad/mig/tests/spk/2010/index.html*, 2010.

[30] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006, pp. 325–328, Software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[31] NIST, "The nist 2008 speaker recognition evaluation plan," *www.itl.nist.gov/iad/mig/tests/spk/2008/index.html*, 2008.

[32] NIST, "The 2013-2014 speaker recognition ivector machine learning challenge," *https://ivectorchallenge.nist.gov*, 2014.