SPEAKER VERIFICATION USING SIMPLIFIED AND SUPERVISED I-VECTOR MODELING

Ming Li, Andreas Tsiartas, Maarten Van Segbroeck and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA

ABSTRACT

This paper presents a simplified and supervised i-vector modeling framework that is applied in the task of robust and efficient speaker verification (SRE). First, by concatenating the mean supervector and the i-vector factor loading matrix with respectively the label vector and the linear classifier matrix, the traditional i-vectors are then extended to label-regularized supervised i-vectors. These supervised i-vectors are optimized to not only reconstruct the mean supervectors well but also minimize the mean squared error between the original and the reconstructed label vectors, such that they become more discriminative. Second, factor analysis (FA) can be performed on the pre-normalized centered GMM first order statistics supervector to ensure that the Gaussian statistics sub-vector of each Gaussian component is treated equally in the FA, which reduces the computational cost significantly. Experimental results are reported on the female part of the NIST SRE 2010 task with common condition 5. The proposed supervised i-vector approach outperforms the i-vector baseline by relatively 12% and 7% in terms of equal error rate (EER) and norm old minDCF values, respectively.

Index Terms— Speaker verification, Simplified i-vector, Supervised i-vector

1. INTRODUCTION

Joint factor analysis (JFA) [1, 2, 3] has contributed to state-of-the-art performance in text independent speaker verification (SRE). It is a powerful and widely used technique for compensating the variability caused by different channels and sessions.

Recently, total variability i-vector modeling has gained significant attention in SRE due to its excellent performance, low complexity and small model size [4]. In this approach, a single factor analysis is firstly used as a front-end to generate a low dimensional total variability space (i.e. the i-vector space) which jointly models speaker and channel variabilities [4]. Within the i-vector space, variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [5], Linear Discriminative Analysis (LDA) and Nuisance Attribute Projection (NAP) [6], are performed to reduce the variability for the subsequent probabilistic LDA (PLDA) modeling [7, 8]. Since i-vectors cover speaker and channel variabilities all together as one unsupervised method, these variability compensation methods are required in the back end. This motivates us to investigate a joint optimization to minimize the weighted sum of both the re-construction and the classification error simultaneously.

In this work, the traditional i-vectors are extended to the label regularized supervised i-vectors by concatenating the mean supervector and the i-vector factor loading matrix, respectively with the label vector and the linear classifier matrix at the end. Compared to the traditional i-vectors, this joint optimization can discriminatively select the top eigen directions related to the given labels such that the

non-relevant information is reduced in the i-vector space (e.g. noise, variabilities from undesired sources) and performance is improved.

Moreover, the i-vector training and extraction algorithms are computationally expensive which limits practical usage, especially for large sized GMM model and data set [9, 10]. Both [10] and [9] used pre-calculated Universal Background Model (UBM) weighting vector to approximate each utterance's 0th order GMM statistics vector to avoid the computationally expensive GMM componentwise matrix operations in the SRE task. This approximation resulted a processing speed up by a factor 10 to 25 at the cost of a significant performance degradation (17% EER) [9]. By enforcing the approximation at both training and extraction stage, the performance degradation can be reduced [10] in condition where no or very little mismatch between train/test data and UBM data exists. In this work, we investigated an alternative robust and efficient solution which is not based on the UBM weights vector.

We performed factor analysis (FA) on the pre-normalized GMM first order statistics supervector to ensure each Gaussian component's statistics sub-vector is treated equally in the FA which reduces the computational cost by a factor 40. This way, each utterance is represented by a single pre-normalized supervector as the feature vector plus one total frame number to control its importance against the prior. Each component's statistics sub-vector is normalized by its own occupancy probability square root, thus it avoids the mismatch between global pre-calculated average weighting vector ([10] adopted the UBM weights) and each utterance's own occupancy probability distribution vector. Furthermore, since there is only a global total frame number inside the matrix inversion, we can create a global cache table of the resulting matrices against its log value. The log domain is chosen since the smaller the total frame number, the more important it is against the prior. By looking at the table, the speed of extracting an i-vector for each sentence can be increased by another 3 times at the cost of a small quantization error. The larger the table, the smaller this quantization error.

1.1. Relation to prior work

First, the traditional i-vector approach [4] is extended to the label-regularized supervised i-vector framework to improve the performance. Second, an alternative simplified approximation method is provided for both i-vector and supervised i-vector modeling. The difference with [9, 10] is that our method does not rely on the UBM weights vector. Third, although this simplified supervised i-vector method was also introduced in another submission (language identification on highly noisy data [11]), its application on the SRE task and traditional NIST environment is new. As an extension to the work of [11], we demonstrate that the design of both the label vector and classification matrix can be flexibly changed regarding different applications (such as identification versus verification). We propose another new design with speaker specific sample mean i-vector as the label vector especially for verification purpose in this work.

2. METHODS

2.1. The i-vector (IV) baseline

In the total variability space, there is no distinction between the speaker effects and the channel effects. Rather than separately using the eigenvoice matrix $\mathbf{V}$ and the eigenchannel matrix $\mathbf{U}$ [1], the total variability space simultaneously captures the speaker and channel variabilities [4]. Given a C component GMM UBM model λ with $\lambda_c = \{p_c, \mu_\mathbf{c}, \boldsymbol{\Sigma}_\mathbf{c}\}, c = 1, \cdots, C$ and an utterance with a L frame feature sequence $\{\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_L}\}$, the 0th and centered 1st order Baum-Welch statistics on the UBM are calculated as follows:

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

$$\mathbf{F_c} = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c}) \qquad (2)$$

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

$$\tilde{\mathbf{F}}_\mathbf{c} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \qquad (3)$$

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

$$\tilde{\mathbf{F}} = \mathbf{Tx}, \qquad (4)$$

where $\mathbf{T}$ is a rectangular total variability matrix of low rank and $\mathbf{x}$ is the so-called i-vector [4]. Considering a C-component GMM and D dimensional acoustic features, the total variability matrix $\mathbf{T}$ is a $CD \times K$ matrix which can be estimated the same way as learning the eigenvoice matrix $\mathbf{V}$ in [12] except that here we consider that every utterance is produced by a new speaker [4].

Given the centered mean supervector $\tilde{\mathbf{F}}$ and total variability matrix $\mathbf{T}$, the i-vector is computed as follows [4]:

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

where $\mathbf{N}$ is a diagonal matrix of dimension $CD \times CD$ whose diagonal blocks are $N_c\mathbf{I}, c = 1, \cdots, C$ 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 $\mathbf{T}$ [4]. Covariance $\boldsymbol{\Sigma}$ is also updated iteratively.

2.2. Label-regularized supervised i-vector (SUP-IV)

The i-vector training and extraction can be re-interpreted as a classic factor analysis based generative modeling problem. For the jth utterance, the prior and the conditional distribution is defined as following multivariate Gaussian distributions:

$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), \qquad P(\tilde{\mathbf{F}}_\mathbf{j}|\mathbf{x_j}) = \mathcal{N}(\mathbf{Tx_j}, \mathbf{N_j^{-1}}\boldsymbol{\Sigma}) \qquad (6)$$

therefore, the posterior distribution of i-vector $\mathbf{x}$ given the observed $\tilde{\mathbf{F}}$ for the jth utterance is:

$$P(\mathbf{x_j}|\tilde{\mathbf{F}}_\mathbf{j}) = \mathcal{N}((\mathbf{I} + \mathbf{T^t}\boldsymbol{\Sigma}^{-1}\mathbf{N_j T})^{-1}\mathbf{T^t}\boldsymbol{\Sigma}^{-1}\mathbf{N_j}\tilde{\mathbf{F}}_\mathbf{j}, (\mathbf{I} + \mathbf{T^t}\boldsymbol{\Sigma}^{-1}\mathbf{N_j T})^{-1}). \qquad (7)$$

The mean of the posterior distribution (point estimate) is adopted as the i-vector which is the same as equation (5).

The traditional i-vectors are extended to the label-regularized supervised i-vectors by concatenating the label vector and the linear classifier matrix at the end of the mean supervector and the i-vector factor loading matrix, respectively. These supervised i-vectors are optimized not only to reconstruct the mean supervectors well but also to minimize the mean square error between the original and the reconstructed label vectors, and thus can make the supervised i-vectors become more discriminative in terms of the regularized label information.

$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), P\left(\left[\begin{array}{c} \tilde{\mathbf{F}}_\mathbf{j} \\ \mathbf{L_j} \end{array}\right] | \mathbf{x_j}\right) = \mathcal{N}\left(\left[\begin{array}{c} \mathbf{Tx_j} \\ \mathbf{Wx_j} \end{array}\right], \left[\begin{array}{c} \mathbf{N_j^{-1}}\boldsymbol{\Sigma}_\mathbf{1} \\ n_j^{-1}\boldsymbol{\Sigma}_\mathbf{2} \end{array}\right]\right) \qquad (8)$$

In (6,7,8), $\mathbf{x_j}$, $\mathbf{N_j}$, $\tilde{\mathbf{F}}_\mathbf{j}$ and $\mathbf{L_j}$ denote the jth utterance's i-vector, $\mathbf{N}$ matrix, mean supervector and label vector, respectively. $\boldsymbol{\Sigma}_\mathbf{1}$ and $\boldsymbol{\Sigma}_\mathbf{2}$ denote the variance for CD dimensional mean supervector and M dimensional label vector, respectively. $n_j = \sum_{c=1}^{C} N_{cj}$ where N_{cj} denotes the N_c for the jth utterance. The reason for using a global scalar n_j is that each label vector dimension is treated equally in terms of frame length importance, the variance $\boldsymbol{\Sigma}_\mathbf{2}$ is adopted to capture the variance of label vectors. We define two types of label vectors as follows:

Supervised type 1: $\qquad \mathbf{L}_{ij} = \left\{ \begin{array}{ll} 1 & \text{if utterance } j \text{ is from class } i \\ 0 & \text{otherwise} \end{array} \right.$

$$(9)$$

For type 1 label vectors, we want the regression matrix $\mathbf{W}$ to correctly classify the class labels. Suppose there are M speaker classes, $\mathbf{L_j}$ is a M dimensional binary vector with only one non-zero element with the value of 1 and $\mathbf{W}$ is a $M \times K$ linear classification matrix.

Supervised type 2: $\qquad \mathbf{L}_j = \bar{\mathbf{x}}_{\mathbf{s_j}}, \mathbf{W} = \mathbf{I}. \qquad (10)$

Type 2 label vectors are the sample mean vector of all the supervised i-vectors from the same speaker index in the last iteration $\bar{\mathbf{x}}_{\mathbf{s_j}}$ (similar to the one in WCCN). The reason is to reduce the within class covariance and help all the supervised i-vectors to move towards their class sample mean. Therefore, $M = K$ in this case.

The log likelihood of the total Γ utterances is:

$$\sum_{j=1}^{\Gamma} ln(P(\tilde{\mathbf{F}}_\mathbf{j}, \mathbf{L_j}, \mathbf{x_j})) = \sum_{j=1}^{\Gamma} \{ln(P\left(\left[\begin{array}{c} \tilde{\mathbf{F}}_\mathbf{j} \\ \mathbf{L_j} \end{array}\right]|\mathbf{x_j})) + ln(P(\mathbf{x_j}))\} \qquad (11)$$

Combining (10) and (11) together and removing non-relevant items, we can get the objective function J for the Maximum Likelihood (ML) EM training:

$$J = \sum_{j=1}^{\Gamma} (\tfrac{1}{2}\mathbf{x_j^t x_j} + \tfrac{1}{2}(\tilde{\mathbf{F}}_\mathbf{j} - \mathbf{Tx_j})^t\boldsymbol{\Sigma}_\mathbf{1}^{-1}\mathbf{N_j}(\tilde{\mathbf{F}}_\mathbf{j} - \mathbf{Tx_j}) \qquad (12)$$
$$-\tfrac{1}{2}ln(|\boldsymbol{\Sigma}_\mathbf{1}^{-1}|) + \tfrac{1}{2}(\mathbf{L_j} - \mathbf{Wx_j})^t\boldsymbol{\Sigma}_\mathbf{2}^{-1}n_j(\mathbf{L_j} - \mathbf{Wx_j}) - \tfrac{1}{2}ln(|\boldsymbol{\Sigma}_\mathbf{2}^{-1}|))$$

The solution is as follows:

$$\begin{aligned} E(\mathbf{x_j}) &= (\mathbf{I} + \mathbf{T^t}\boldsymbol{\Sigma}_\mathbf{1}^{-1}\mathbf{N_j T} + \mathbf{W^t}\boldsymbol{\Sigma}_\mathbf{2}^{-1}n_j\mathbf{W})^{-1} \\ &\quad (\mathbf{T^t}\boldsymbol{\Sigma}_\mathbf{1}^{-1}\mathbf{N_j}\tilde{\mathbf{F}}_\mathbf{j} + \mathbf{W^t}\boldsymbol{\Sigma}_\mathbf{2}^{-1}n_j\mathbf{L_j}), \end{aligned} \qquad (13)$$

$$E(\mathbf{x_j x_j}^t) = E(\mathbf{x_j})E(\mathbf{x_j})^t + (\mathbf{I} + \mathbf{T^t}\boldsymbol{\Sigma}_\mathbf{1}^{-1}\mathbf{N_j T} + \mathbf{W^t}\boldsymbol{\Sigma}_\mathbf{2}^{-1}n_j\mathbf{W})^{-1}. \qquad (14)$$

Type 1: $\qquad \mathbf{W}_{new} = [\sum_{j=1}^{\Gamma} n_j\mathbf{L_j}E(\mathbf{x_j}^t)][\sum_{j=1}^{\Gamma} n_jE(\mathbf{x_j x_j}^t)]^{-1} \qquad (15)$

For the $\mathbf{T}$ matrix, we employed the strategy used in [4] to update component by component since N_{cj} is also a scalar.

$$\mathbf{T_c}_{new} = [\sum_{j=1}^{\Gamma} N_{cj}\tilde{\mathbf{F}}_\mathbf{cj}E(\mathbf{x_j}^t)][\sum_{j=1}^{\Gamma} N_{cj}E(\mathbf{x_j x_j}^t)]^{-1} \qquad (16)$$

In (16), $\mathbf{T_c}$ denotes the $[(c-1)D+1:cD]$ rows sub-matrix of $\mathbf{T}$ and $\tilde{\mathbf{F}}_{\mathbf{cj}}$ is the $[(c-1)D+1:cD]$ elements sub-vector of $\tilde{\mathbf{F}}_{\mathbf{j}}$.

$$\mathbf{\Sigma_1} = \frac{diag\{\sum_{j=1}^{\Gamma}(\mathbf{N_j}(\tilde{\mathbf{F}}_{\mathbf{j}} - \mathbf{T}_{new}E(\mathbf{x_j}))\tilde{\mathbf{F}}_{\mathbf{j}}^t)\}}{\Gamma} \quad (17)$$

Type 1: $\quad \mathbf{\Sigma_2} = \dfrac{diag\{\sum_{j=1}^{\Gamma}(n_j(\mathbf{L_j} - \mathbf{W}_{new}E(\mathbf{x_j}))\mathbf{L_j}^t)\}}{\Gamma}$ (18)

Type 2: $\quad \mathbf{\Sigma_2} = \dfrac{diag\{\sum_{j=1}^{\Gamma}(n_j(\mathbf{L_j} - E(\mathbf{x_j}))^t(\mathbf{L_j} - E(\mathbf{x_j})))\}}{\Gamma}$ (19)

These 2 variance vectors describe the energy that can not be represented by factor analysis and control the importance in the joint optimization objective function (12). $\mathbf{\Sigma_2}$ for the type 2 label vectors is just the diagonal elements of the within class covariance matrix in WCCN. After several iterations of EM training, the parameters are learned. For the subsequent supervised i-vector extraction, we let $\mathbf{\Sigma_2}$ to be infinity since we do not know the ground truth label information. This will revert equation (13) back to equation (5). After the supervised i-vector extraction, the classification methods steps are the same as in the traditional i-vector modeling.

There are some obvious extensions of this supervised i-vector framework. We can make $\mathbf{L}$ as the parameter vector that we want to perform regression with (e.g. ages [13, 14], paralinguistic measures [15]) to make the proposed framework suitable for regression problems. Moreover, if the classification or regression relation is not linear, we can use non-linear mapping as a preprocessing step before generating $\mathbf{L}$.

2.3. Simplified i-vector (SIM-IV)

I-vector training and extraction is computationally expensive. Let the GMM size, feature dimension, factor loading matrix size to be C, D, and K, respectively. The complexity for generating a single i-vector is $O(K^3 + K^2C + KCD)$ [10]. In this work, we make two approximations to reduce the complexity.

The K^3 term comes from the matrix inversion while the K^2C term is from $\mathbf{T^t\Sigma^{-1}N_jT}$ in equation (5). When C is large, this K^2C term's computational cost is huge. The fundamental reason is that each Gaussian component λ_c has different N_c for each utterance j which means some sub-vectors $\mathbf{F_{cj}}$ have less variance than others in $\tilde{\mathbf{F}}_{\mathbf{j}}$ and need utterance specific intra mean supervector re-weighting in the objective function. We first decompose the $\mathbf{N_j}$ vector into $\mathbf{N_j} = n_j\mathbf{m_j}$ where $n_j = \sum_{c=1}^{C} N_{cj}$, $m_{cj} = N_{cj}/n_j$ and $\sum_{c=1}^{C} m_{cj} = 1$. $\mathbf{m_j}$ is the re-weighting vector and n_j (total frame number) controls the confidence at the global level. Our motivation is to re-weight each utterance's mean supervector with its own $(\mathbf{m_j})^{1/2}$ before the factor analysis step which makes each dimension of the new supervector $\hat{\mathbf{F}}_{\mathbf{j}}$ be treated equally in the approximated modeling (21,22).

$$\hat{\mathbf{F}}_{\mathbf{c}} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_{\mathbf{c}})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}[\frac{N_{cj}}{n_j}]^{\frac{1}{2}}, \hat{\mathbf{F}}_{\mathbf{j}} = \mathbf{m_j}^{1/2}\tilde{\mathbf{F}}_{\mathbf{j}} \quad (20)$$

So the intra supervector imbalance is compensated by this pre-weighting, and each utterance is represented by $\hat{\mathbf{F}}_{\mathbf{j}}$ as the general feature vector and n_j as the confidence value for the subsequent machine learning algorithms. We perform factor analysis in the following way by linearly projecting this new normalized supervector $\hat{\mathbf{F}}$ on a dictionary $\hat{\mathbf{T}}$:

$$\hat{\mathbf{F}} = \hat{\mathbf{T}}\hat{\mathbf{x}}, P(\hat{\mathbf{x}_{\mathbf{j}}}) = \mathcal{N}(\mathbf{0}, I) \quad (21)$$

$$P(\hat{\mathbf{F}}_{\mathbf{j}}|\hat{\mathbf{x}_{\mathbf{j}}}) = \mathcal{N}(\hat{\mathbf{T}}\hat{\mathbf{x}_{\mathbf{j}}}, \mathbf{m_j}(n_j\mathbf{m_j})^{-1}\mathbf{\Sigma}) = \mathcal{N}(\hat{\mathbf{T}}\hat{\mathbf{x}_{\mathbf{j}}}, n_j^{-1}\mathbf{\Sigma}) \quad (22)$$
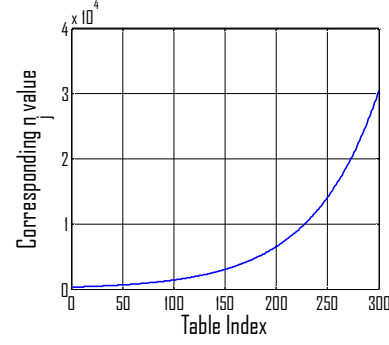

Fig. 1. The n_j quantization curve in the log domain, 300 indexes.

Table 1. Complexity of the proposed methods for a single utterance (GMM size $C = 1024$, feature dimension $D = 36$, $\mathbf{T}$ matrix rank $K = 500$, table index size is 300, label vector dimension $M = 2543, 500$ (type 1,2), time was measured on a Intel I7 CPU with a single thread and 12 GB memory)

Methods	Approximated complexity	Time
IV	$O(K^3 + K^2C + KCD)$	2.82s
Type 1 SUP-IV	$O(K^3 + K^2C + K(CD + M))$	2.85s
SIM-IV without table	$O(K^3 + KCD)$	0.062s
SIM-IV with table	$O(KCD)$	0.022s
SIM-SUP-IV without table	$O(K^3 + K(CD + M))$	0.066s
SIM-SUP-IV with table	$O(K(CD + M))$	0.023s

Therefore, the posterior distribution of the i-vector $\hat{\mathbf{x}}$ given the observed $\hat{\mathbf{F}}$ is:

$$P(\hat{\mathbf{x}_{\mathbf{j}}}|\hat{\mathbf{F}}_{\mathbf{j}}) = \mathcal{N}((\mathbf{I} + \hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{F}}_{\mathbf{j}}), (\mathbf{I} + \hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{T}})^{-1}). \quad (23)$$

From the above equation (point estimate mean vector), we can find that the complexity is reduced to $O(K^3 + KCD)$ since n_j is not dependent on any GMM component. By replacing the 1st GMM statistics supervector $\tilde{\mathbf{F}}_{\mathbf{j}}$ with the pre-normalized supervector $\hat{\mathbf{F}}_{\mathbf{j}}$ and setting $\mathbf{N_j}$ to a scalar n_j, the i-vector (equation (5)) and the supervised i-vector (equation (13-19)) training equations become the proposed simplified i-vector and simplified supervised i-vector solutions, respectively. Moreover, since the entire term $(\mathbf{I} + \hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}$ in equation (23) only depends on the scalar total frame number n_j, we can create a global table of this quantity against the log value of n_j. The reason to choose log domain is that the smaller the total frame number, the more important it is against the prior. If n_j is very large compared to the prior, then the two n_j in $(\mathbf{I} + \hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}^t\mathbf{\Sigma}^{-1}n_j\hat{\mathbf{F}}_{\mathbf{j}}$ get canceled. By enabling the cache table lookup, the complexity of each utterance's i-vector extraction is further reduced to $O(KCD)$ with a small table index quantization error. The larger the table, the smaller this error. Figure 1 shows the quantization distance curve. We can see that the quantization error is relatively small when n_j is small.

It is worth noting that for best accuracy, we only perform approximation using the global cache table for training purposes. When in testing mode, equation (23) is sill employed.

3. EXPERIMENTAL RESULTS

We performed experiments on the NIST 2010 speaker recognition evaluation (SRE) corpus [16]. Our focus is the female part of the common condition 5 (a subset of tel-tel) in the core task. We used equal error rate (EER), the normalized old minimum decision cost

Table 2. Corpora used to estimate the UBM, total variability matrix, JFA factor loading matrix, WCCN, LDA, PLDA and the normalization data for NIST 2010 task condition 5.

	Switchboard	NIST04	NIST05	NIST06	NIST08
UBM		✓	✓		
T	✓	✓	✓	✓	✓
JFA V	✓				
JFA U		✓	✓	✓	✓
JFA D		✓			
WCCN	✓	✓	✓	✓	✓
LDA		✓	✓	✓	✓
PLDA		✓	✓	✓	✓
Znorm		✓	✓		
Snorm					✓
Tnorm				✓	

Table 3. Performance of the proposed methods for the 2010 NIST SRE task female part condition 5

Method	LDA	WCCN	PLDA	S norm	EER%	norm minDCF new	norm minDCF old
IV	×	×	×	×	9.02	0.724	0.409
IV	250	×	×	✓	7.87	0.668	0.307
IV	250	✓	×	✓	3.91	0.454	0.190
IV	250	✓	✓	✓	**3.37**	**0.415**	**0.165**
T1SUP-IV	250	×	×	✓	7.64	0.640	0.278
T1SUP-IV	250	✓	×	✓	4.01	0.425	0.170
T1SUP-IV	250	✓	✓	✓	**2.95**	**0.420**	**0.154**
SIM-IV	×	×	×	×	8.94	0.758	0.374
SIM-IV	250	×	×	✓	7.96	0.696	0.311
SIM-IV	250	✓	×	✓	4.79	0.527	0.213
SIM-IV	250	✓	✓	✓	**3.45**	**0.545**	**0.192**
T1SSIV	×	×	×	×	8.65	0.746	0.341
T1SSIV	250	×	×	✓	7.06	0.654	0.289
T1SSIV	250	✓	×	✓	3.95	0.518	0.197
T1SSIV	250	✓	✓	✓	**3.13**	**0.541**	**0.176**
T2SSIV	×	×	×	×	6.45	0.645	0.285
T2SSIV	250	×	×	✓	5.35	0.575	0.228
T2SSIV	250	✓	×	✓	4.51	0.549	0.195
T2SSIV	250	✓	✓	✓	**3.06**	**0.569**	**0.179**
T2SSIV	250	×	✓	✓	3.08	0.581	0.189

T1SSIV: type 1 SIM-SUP-IV, T2SSIV: type 2 SIM-SUP-IV

Table 4. Performance of the proposed systems in fusion

ID	Systems	EER%	norm minDCF new	norm minDCF old
1	JFA linear scoring ZTnorm	3.62	0.414	0.193
2	IV LDA WCCN PLDA Snorm	3.37	0.415	0.165
3	T1SUP-IV LDA WCCN PLDA Snorm	2.95	0.420	0.154
4	T1SSIV LDA WCCN PLDA Snorm	3.13	0.541	0.176
5	Fusion ID 1 + ID 2	2.77	0.372	0.152
6	**Fusion ID 1 + ID 3**	**2.53**	**0.370**	**0.146**
7	Fusion ID 1 + ID 4	2.82	0.377	0.162

value (norm old minDCF) and norm new minDCF as the metrics for evaluation [16]. 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. We employed a Czech phoneme recognizer [17] to perform the voice activity detection (VAD) by simply dropping all frames that are decoded as silence or speaker noises. Feature warping is applied to mitigate variabilities.

The training data for NIST 2010 task included Switchboard II part1 to part3, NIST SRE 2004, 2005, 2006 and 2008 corpora on the telephone channel. The description of the dataset used in each step is provided in Table 2. The gender-dependent GMM UBMs consist of 1024 mixture components. The JFA baseline system [1, 2, 3] is trained using the BUT toolkit [18] and linear common channel point estimate scoring [19] is adopted. The speaker factor size and channel factor size is 300 and 100, respectively. ZTnorm was applied on JFA subsystem while Snorm was employed in i-vector subsystem. The PLDA implementation is based on the UCL toolkit [7] where the sizes of speaker loading matrix $\mathbf{U}$ and variability loading matrix $\mathbf{G}$ are 250 and 80, respectively. Simple weighted linear summation is adopted here as the score level fusion. Other parameter settings are reported in the captain of Table 1.

The results of the i-vector baseline and the proposed supervised, simplified as well as the simplified supervised i-vector methods are shown in Table 3. We can observe that LDA, PLDA, and Snorm contributed to increase the performance for all the systems. WCCN reduced the EER by more than 40% for all systems except the type 2 simplified supervised i-vector (T2SSIV). For T2SSIV, WCCN is not that important since the label regularized joint optimization already includes the within class covariance in the objective function (equation 10,12,13,19). This was reflected by a 30% EER reduction (I-vector 9.02%, T2SSIV 6.45%) in the cosine distance raw scoring without any back end processing. Furthermore, type 1 supervised i-vector (T1SUP-IV) and type 1 simplified supervised I-vector (T1SSIV) outperformed IV and SIM-IV by 5%-10% relatively for all the modeling configurations (3.37% and 3.45% EER vs 2.95 and 3.13% EER). Also as shown in Table 4 (ID 6 vs 5), after fusing with JFA baseline, SUP-IV still outperformed IV baseline by 9% relative EER reduction. Therefore, by adding label information in the i-vector training indeed improves the performance. The less improvement of T2SSIV compared to T1SSIV might be due to the diagonal version of $\mathbf{\Sigma_2}$ against the triangular WCCN matrix.

Moreover, simplified supervised i-vector systems (T1SSIV and T2SSIV) achieved better EER but worse norm cost compared to the i-vector baseline. However, the computationally cost is reduced by around 120 times. And after fusing with JFA system (Table 4 ID 7 vs 5), this gap is reduced to only 3% to 6% relatively. Therefore, simplified supervised i-vector has the potential to replace the computational expensive i-vector baseline when fusing with JFA system.

It is worth noting that the supervised version of all the systems only performed better on EER and norm old minDCF values. How to further reduce the norm new minDCF is our current focus. Future work also includes applying the non-simplified type 2 supervised i-vector as well as evaluating different label vector designs.

4. CONCLUSION

This paper presents a simplified and supervised i-vector modeling framework that can be applied in the SRE task. First, traditional i-vectors are extended to label-regularized supervised i-vectors by concatenating the mean supervector with the label vector, and the i-vector factor loading matrix with the linear classifier matrix. The proposed label-regularization makes the supervised i-vectors more discriminative such that their performance is improved. Second, factor analysis (FA) can be performed on the first-order statistics supervector of the pre-normalized centered GMM, to ensure that the statistics sub-vector of each Gaussian component are treated equally in the FA. This step significantly reduces the computational cost and makes the proposed method appealing for practical deployment.

5. REFERENCES

[1] 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.

[2] P. Kenny, G. Boulianne, P. Dumouchel, and P. Ouellet, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[4] N. Dehak, P.J. 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] H. Aronowitz and O. Barkan, "Efficient approximated i-vector extraction," in *Proc. ICASSP*, 2012, pp. 4789–4792.

[10] O. Glembek, L. Burget, P. Matejka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. ICASSP*, 2011, pp. 4516–4519.

[11] M. Li and S. Narayanan, "Simplified supervised i-vector modeling and sparse representation with application to robust language recognition," *Computer Speech & Language*, Oct 2012, submitted.

[12] 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.

[13] M. Li, K.J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, 2012.

[14] M. Bahari, M.H.and McLaren, H.V. Hamme, and D.V. Leeuwen, "Age estimation from telephone speech using i-vectors," in *Proc. Interspeech*, 2012.

[15] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and languagełstate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[16] National Institute of Standards and Technology, "The NIST Year 2010 Speaker Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.

[17] 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.

[18] L. Burget, M. Fapšo, and V. Hubeika, "BUT system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop*, 2008, pp. 1–4, Software available at http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo.

[19] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proc. ICASSP*, 2009, pp. 4057–4060.