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 input 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 (6th order GMM statistics vector to avoid the computationally expensive GMM component-wise 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.

This work was supported in part by NSF and DARPA.
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 $V$ and the eigenchannel matrix $U$ [1], the total variability space simultaneously captures the speaker and channel variabilities [4]. Given a $C$ component GMM UBM model $\lambda_c = \{\mu_c, \Sigma_c, N_c\}, c = 1, \cdots, C$ and an utterance with a $L$ frame feature sequence $\{y_1, y_2, \cdots, y_L\}$, the $0^{th}$ and centered $1^{st}$ order Baum-Welch statistics on the UBM are calculated as follows:

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

$$F_c = \frac{\sum_{t=1}^{L} P(c|y_t, \lambda)(y_t - \mu_c)}{\sum_{t=1}^{L} P(c|y_t, \lambda)}.$$  (2)

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

$$\bar{F} = [\bar{x}_1^\top, \cdots, \bar{x}_C^\top]$$

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

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

$$x = (I + T^\top \Sigma^{-1} N T)^{-1} T^\top \Sigma^{-1} N \bar{F}$$

where $N$ is a diagonal matrix of dimension $CD \times CD$ whose diagonal blocks are $N_c I, c = 1, \cdots, C$ and $\Sigma$ 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 $T$ [4]. Covariance $\Sigma$ 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 $j^{th}$ utterance, the prior and the conditional distribution is defined as following multivariate Gaussian distributions:

$$P(x_j) = \mathcal{N}(0, I), \quad P(\bar{F}_j|x_j) = \mathcal{N}(T x_j, N_j^{-1} \Sigma)$$

therefore, the posterior distribution of i-vector $x$ given the observed $\bar{F}$ for the $j^{th}$ utterance is:

$$P(x_j|\bar{F}_j) = \mathcal{N}((I + T^\top \Sigma^{-1} N_j T)^{-1} T^\top \Sigma^{-1} N_j \bar{F}_j, (I + T^\top \Sigma^{-1} N_j T)^{-1})$$

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.

$$P(x_j) = \mathcal{N}(0, I), \quad P(\bar{F}_j|L, x_j) = \mathcal{N}(T x_j W x_j, N_j^{-1} \Sigma 1_n 1_n^\top)$$

In (6,7,8), $x_j, N_j, F_j$ and $L_j$ denote the $j^{th}$ utterance's i-vector, N matrix, mean supervector and label vector, respectively. $F_1$ and $F_2$ 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 $j^{th}$ 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 $\Sigma_2$ is adopted to capture the variance of label vectors. We define two types of label vectors as follows:

Supervised type 1: $L_{ij} = \begin{cases} 1 & \text{if utterance } j \text{ is from class } i \\ 0 & \text{otherwise} \end{cases}$

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

Supervised type 2: $L_j = \tilde{x}_{aj}, W = I.$

Type 2 label vectors are the sample mean vector of all the supervised i-vectors from the same speaker index in the last iteration $\tilde{x}_{aj}$ (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 $\Gamma$ utterances is:

$$\sum_{j=1}^{\Gamma} \ln(P(\bar{F}_j, L_j, x_j)) = \sum_{j=1}^{\Gamma} \{\ln(P(\bar{F}_j|L_j)) + \ln(P(x_j))\}$$

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

$$J = \sum_{j=1}^{\Gamma} \left\{ \frac{1}{2} (x_j - T x_j)^\top \Sigma^{-1} (x_j - T x_j) + \frac{1}{2}(\bar{F}_j - T x_j)^\top \Sigma^{-1} (\bar{F}_j - T x_j) \right\} - \frac{1}{2} \ln(|\Sigma_1^{-1}|) + \frac{1}{2} \ln(|\Sigma_2 - W x_j|) - \frac{1}{2} \ln(|\Sigma_2^{-1}|).$$

The solution is as follows:

$$E(x_j) = (I + T^\top \Sigma^{-1} N_j T + W^\top \Sigma_2^{-1} n_j W) -1 (T^\top \Sigma^{-1} N_j \bar{F}_j + W^\top \Sigma_2^{-1} n_j L_j),$$

$$E(x_j | x_j') = E(x_j) E(x_j') + (I + T^\top \Sigma^{-1} N_j T + W^\top \Sigma_2^{-1} n_j W)^{-1}.$$  (14)

Type 1: $W_{new} = \sum_{j=1}^{\Gamma} n_j E(x_j'|x_j))|\sum_{j=1}^{\Gamma} n_j E(x_j|x_j')^{-1}|^{-1}$

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

$$T_{new} = \sum_{j=1}^{\Gamma} n_j E(x_j)|\sum_{j=1}^{\Gamma} n_j E(x_j|x_j')^{-1}|^{-1}$$
In (16), $T_c$ denotes the $[(c-1)D+1:cD]$ rows sub-matrix of $T$ and $F_{c,j}$ is the $[(c-1)D+1:cD]$ elements sub-vector of $F_j$.

$$\Sigma_1 = \frac{\text{diag}\{(\sum_{j=1}^t (N_j (F_j - T_{new} E(x_j)) (F_j)')\}}{\Gamma}$$  \hspace{1cm} (17)

Type 1: $$\Sigma_2 = \frac{\text{diag}\{(\sum_{j=1}^t (n_j (L_j - W_{new} E(x_j)) (L_j)'))\}}{\Gamma}$$  \hspace{1cm} (18)

Type 2: $$\Sigma_2 = \frac{\text{diag}\{(\sum_{j=1}^t (n_j (L_j - E(x_j))' (L_j - E(x_j)))\}}{\Gamma}$$  \hspace{1cm} (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). $\Sigma_2$ 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 $\Sigma_2$ 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 $L$ 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 $L$.

### 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 $T^T \Sigma^{-1} N_j T$ in equation (5). When $C$ is large, this $K^2C$ term’s computational cost is huge. The fundamental reason is that each Gaussian component $\lambda_c$ has different $N_c$ for each utterance $j$ which means some sub-vectors $F_{c,j}$ have less variance than others in $F_j$ and need utterance specific intra mean supervisor weighting in the objective function. We first decompose the $N_j$ vector into $N_j = n_j m_j$ where $n_j = \sum_{c=1}^C N_{c,j}$, $m_{c,j} = N_{c,j}/n_j$ and $\sum_{c=1}^C m_{c,j} = 1$. $m_j$ 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 supervisor with its own $(m_j)^{1/2}$ before the factor analysis step which makes each dimension of the new supervector $\hat{F}_j$ be treated equally in the approximated modeling (21,22).

$$F_c = \frac{\sum_{l=1}^L P(c|y, \lambda)(y_l - \mu_c) N_{c,j}}{\sum_{l=1}^L P(c|y, \lambda)} m_j^{1/2} F_j$$  \hspace{1cm} (20)

So the intra supervector imbalance is compensated by this pre-weighting, and each utterance is represented by $F_j$ 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 $\hat{F}$ on a dictionary $T$:

$$\hat{F} = T x, P(\hat{x}_j) = N(0, I)$$  \hspace{1cm} (21)

$$P(\hat{F}_j|\hat{x}_j) = N(T \hat{x}_j, m_j (n_j m_j)^{-1} \Sigma) = N(T \hat{x}_j, n_j^{-1} \Sigma)$$  \hspace{1cm} (22)

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

$$P(\hat{x}_j|\hat{F}_j) = N((I + T^T \Sigma^{-1} n_j T)^{-1} T^T \Sigma^{-1} n_j \hat{F}_j, (I + T^T \Sigma^{-1} n_j T)^{-1})$$  \hspace{1cm} (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 $F_j$ with the pre-normalized supervector $\hat{F}_j$ and setting $N_j$ 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 $(I + T^T \Sigma^{-1} n_j T)^{-1} T^T \Sigma^{-1} n_j \hat{F}_j$ 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 $(I + T^T \Sigma^{-1} n_j T)^{-1} T^T \Sigma^{-1} n_j \hat{F}_j$ 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 still 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 1. Complexity of the proposed methods for a single utterance (GMM size $C = 1024$, feature dimension $D = 36$, $T$ matrix rank $K = 500$, table index size $= 300$, label vector dimension $M = 2543$, 500 (type 1,2), time was measured on a Intel 17 CPU with a single thread and 12 GB memory )](image-url)
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 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 subsystem while Snorm was employed in i-vector subsystem. The PLDA 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 subsystem while Snorm was employed in i-vector subsystem. The PLDA 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 subsystem while Snorm was employed in i-vector subsystem.

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 $\Sigma_2$ against the triangular WCCN matrix.

Moreover, simplified supervised i-vector systems (T1SSIV and T2SSIV) achieved better EER but worse norm cost compared to the

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Switchboard</th>
<th>NIST04</th>
<th>NIST05</th>
<th>NIST06</th>
<th>NIST08</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM</td>
<td>√</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JFA V</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JFA U</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JFA D</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WCCN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LDA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLDA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Znorm</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tnorm</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>LDA</th>
<th>WCCN</th>
<th>PLDA</th>
<th>Snorm</th>
<th>EER%</th>
<th>norm minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>9.02</td>
<td>0.724 0.409</td>
</tr>
<tr>
<td>IV</td>
<td>250</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>7.87</td>
<td>0.668 0.307</td>
</tr>
<tr>
<td>IV</td>
<td>250</td>
<td>✓</td>
<td>√</td>
<td>✓</td>
<td>3.37</td>
<td>0.415 0.165</td>
</tr>
<tr>
<td>T1SUP-IV</td>
<td>250</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>7.64</td>
<td>0.640 0.278</td>
</tr>
<tr>
<td>T1SUP-IV</td>
<td>250</td>
<td>√</td>
<td>×</td>
<td>✓</td>
<td>4.01</td>
<td>0.725 0.279</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>250</td>
<td>√</td>
<td>√</td>
<td>✓</td>
<td>2.95</td>
<td>0.420 0.154</td>
</tr>
<tr>
<td>SIM-IV</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>8.94</td>
<td>0.758 0.374</td>
</tr>
<tr>
<td>SIM-IV</td>
<td>250</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>7.96</td>
<td>0.696 0.311</td>
</tr>
<tr>
<td>SIM-IV</td>
<td>250</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4.79</td>
<td>0.527 0.213</td>
</tr>
<tr>
<td>SIM-IV</td>
<td>250</td>
<td>√</td>
<td>✓</td>
<td>✓</td>
<td>3.45</td>
<td>0.545 0.192</td>
</tr>
<tr>
<td>T1SSIV</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>7.65</td>
<td>0.746 0.341</td>
</tr>
<tr>
<td>T1SSIV</td>
<td>250</td>
<td>×</td>
<td>√</td>
<td>✓</td>
<td>7.06</td>
<td>0.654 0.289</td>
</tr>
<tr>
<td>T1SSIV</td>
<td>250</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>3.95</td>
<td>0.518 0.197</td>
</tr>
<tr>
<td>T1SSIV</td>
<td>250</td>
<td>√</td>
<td>√</td>
<td>✓</td>
<td>3.13</td>
<td>0.541 0.176</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>6.45</td>
<td>0.645 0.285</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>250</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>5.35</td>
<td>0.575 0.228</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>250</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>4.51</td>
<td>0.549 0.195</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>250</td>
<td>√</td>
<td>√</td>
<td>✓</td>
<td>3.06</td>
<td>0.509 0.179</td>
</tr>
<tr>
<td>T2SSIV</td>
<td>250</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>3.08</td>
<td>0.581 0.189</td>
</tr>
</tbody>
</table>

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

### Table 4. Performance of the proposed systems in fusion

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

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


