Li *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:18
DOI 10.1186/s13636-015-0061-x

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An improved i-vector extraction algorithm for speaker verification

Wei Li[1*], Tianfan Fu[2] and Jie Zhu[1]

Abstract

Over recent years, i-vector-based framework has been proven to provide state-of-the-art performance in speaker verification. Each utterance is projected onto a total factor space and is represented by a low-dimensional feature vector. Channel compensation techniques are carried out in this low-dimensional feature space. Most of the compensation techniques take the sets of extracted i-vectors as input. By constructing between-class covariance and within-class covariance, we attempt to minimize the between-class variance mainly caused by channel effect and to maximize the variance between speakers. In the real-world application, enrollment and test data from each user (or speaker) are always scarce. Although it is widely thought that session variability is mostly caused by channel effects, phonetic variability, as a factor that causes session variability, is still a matter to be considered. We propose in this paper a new i-vector extraction algorithm from the total factor matrix which we term component reduction analysis (CRA). This new algorithm contributes to better modelling of session variability in the total factor space.
We reported results on the male English trials of the core condition of the NIST 2008 Speaker Recognition Evaluation (SREs) dataset. As measured both by equal error rate and the minimum values of the NIST detection cost function, 10–15 % relative improvement is achieved compared to the baseline of traditional i-vector-based system.

Keywords: Speaker verification; i-vector; Total factor space; Phonetic variability; Component reduction analysis (CRA)

1 Introduction

In the last decade, Gaussian mixture model based on universal background model (GMM-UBM) framework has demonstrated strong performance in speaker verification. It is commonly believed that the mean vectors of GMMs represent the most characteristics of speakers [1], which are obtained by using the maximum a posteriori (MAP) adaptation. Traditional MAP (or relevance MAP) treats each Gaussian component as a statistically independent distribution, which leaves many drawbacks in the real-world application: Only those components which are observed in the speaker frames are adapted; if training and testing sessions are too short or suffer from significant phonetic variability, the performance of verification may encounter an obvious degradation; on the other hand, traditional MAP does not have solution to the effects of channel distortion, especially when the enrollment and test sessions are from different channels.

*Correspondence: liweisjtu@126.com
[1] Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Rd, 200240 Shanghai, China
Full list of author information is available at the end of the article

Extended from GMM-UBM framework, factor analysis (FA) technique [2, 3] attempts to model the speaker components jointly. Each speaker is represented by the *mean supervector* which is a linear combination of the set of *eigenvoices*. Only a few hundreds of free parameters need to be estimated, which ensures that the speaker mean supervector converges quickly by a short duration of utterance. Based on FA technique, *joint factor analysis* (JFA) [4, 5] decomposes GMM supervector into speaker component **S** and channel component **C**, JFA makes the assumption that speaker component and channel component are statistically independent, although it is known by now that in this modelling, channel effects are not speaker-independent (for example, it has been proven that gender-dependent eigenchannel modelling is more effective than gender-independent modeling [6]), JFA has demonstrated superior performance for text-independent speaker detection tasks in the past NIST Speaker Recognition Evaluation (SREs).

Inspired by JFA approach, Dehak et al. [7] propose a combination of speaker space and channel space. A new low-dimensional space named total factor space is

defined. In this new space, each utterance is represented by a low-dimensional feature vector termed i-vector. The idea of i-vector opens a new era to the analysis of speaker and session variability. Many compensation techniques and scoring methods have been proposed [6–9], which have shown better results than JFA approach. Recently, i-vector extraction with a preliminary length normalization and probabilistic linear discriminant analysis (PLDA) has become the state-of-the-art configuration for speaker verification [6, 9].

However, despite the success of the i-vector paradigm, its applicability in text-dependent speaker verification still remains questionable. In the case that the phonetic content of enrollment and test sessions is same, we may assert that only those components which are observed in the speaker frames need to be adapted, but FA-based techniques provide a global adaptation to all the Gaussian components including those unobserved Gaussian components, whose performance should be questionable compared with traditional MAP adaptation. Currently, related papers have also shown that the traditional MAP approach is superior to the i-vector-based ones [10–12].

Inspired by the drawback of i-vector in text-dependent speaker verification, we attempt to give an analysis that although FA-based i-vector approach offers better performance in modeling phonetic variability, a global adaptation to all the Gaussian components is not an optimal adaptation method. In this paper, we still focus on text-independent speaker verification, we propose a new i-vector extraction algorithm termed component reduction analysis (CRA). Compared with the traditional MAP adaptation that only adapts those observed Gaussian components, CRA approach abandons those Gaussian components which give the least contribution in modeling speaker frames, length of each basis of total factor matrix is truncated while the dimensions of the total factor space remain unchanged, extracted i-vectors also remain unchanged. No modification is needed for subsequent channel compensation and scoring methods. Experiments were carried out on the core condition of NIST 2008 SREs. Experimental results show that by applying CRA algorithm in the phase of i-vector extraction, a 10–15 % relative improvement is obtained compared with the baseline system adopting traditional i-vector algorithm and related channel compensation techniques.

This paper is organized as follow. Section 2 describes the construction of total factor matrix and the paradigm of i-vector extraction. Section 3 analyzes and verifies that phonetic variability and phonetic imbalance widely exist in speaker frames, and following this inference, we demonstrate that conventional i-vector extraction paradigm corresponding to a global adaptation of all Gaussian components is not an optimal adaptation method. Hence, an improved mechanism is introduced to compensate phonetic variability and phonetic imbalance. In Section 4, we propose our CRA algorithm applied in the i-vector extraction phase. We also propose a zero-order Baum-Welch statistics normalization approach to compensate the effects of CRA algorithm. Experiments and results are given in Section 5. Section 6 concludes the paper.

2 Total factor space and i-vector extraction

Proposed by N. Dehak, in the i-vector framework, no separate modeling of speaker and channel space is made, a single space named total factor space is constructed to model speaker and channel variability jointly. Each utterance is projected onto total factor space and is represented by a low-dimensional feature vector. The channel compensation techniques and scoring methods are carried out in this low-dimensional space, as opposed to the high-dimensional GMM supervector space for MAP adaptation and JFA.

The basic idea of i-vector approach is that each speaker- and channel-dependent GMM supervector $\mathbf{M}$ can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{Tw} \tag{1}$$

where $\mathbf{m}$ is a speaker- and channel-independent supervector, whose value is often taken from UBM supervector. $\mathbf{T}$ is the total factor matrix with low rank, which expands a subspace containing speaker- and channel-dependent information. $\mathbf{w}$ is a vector with a prior of standard Gaussian distribution. Speaker frames of an utterance are represented by posterior estimation of $\mathbf{w}$, the new feature vector $\mathbf{w}$ is named total factor, often referred to as identity vector or i-vector. The process of training the total factor matrix is detailedly explained in [2], which is similar as learning the eigenvoice matrix. In order to sufficiently estimate T-Matrix, large quantity of development corpus is necessary.

3 Phonetic variability analysis

Evolved from eigenvoice approach, i-vector approach assumes speaker- and channel-dependent GMM supervector obeys a linear combination of the basis defined by T-Matrix. The i-vector $\mathbf{w}$ can be defined by its posterior distribution conditioned to the Baum-Welch statistics for a given utterance. The posterior distribution is a Gaussian distribution, and the mean vector of this distribution is our i-vector. Following [3], the Baum-Welch statistics are extracted using the UBM model. Suppose we have a sequence of Ł frames $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L\}$ and a UBM Ω composed of C Gaussian components defined in some feature space of dimension F. The Baum-Welch statistics for a given speech utterance u are obtained by:

$$N_c = \sum_{t=1}^{L} P\left(c|\mathbf{y}_t, \Omega\right) \tag{2}$$

$$\mathbf{F}_c = \sum_{t=1}^{L} P\left(c|\mathbf{y}_t, \Omega\right) \mathbf{y}_t \tag{3}$$

where $c = 1, \ldots, C$ is the Gaussian index and $P(c|\mathbf{y}_t, \Omega)$ corresponds to the posterior probability of mixture component c generating the frame vector $\mathbf{y}_t$. (2) and (3) are named zero-order and first-order Baum-Welch statistics.

In the traditional MAP adaptation approach, the adapted mean vector of a Gaussian component can be written as:

$$\mathbf{M}_c = \frac{r}{N_c + r}\mathbf{m}_c + \frac{N_c}{N_c + r}\frac{\mathbf{F}_c}{L} \tag{4}$$

where $\mathbf{M}_c$ denotes the cth component of $\mathbf{M}$, $\mathbf{m}_c$ denotes the cth component of UBM $\mathbf{m}$, $(1/L)\mathbf{F}_c$ is the normalized first-order Baum-Welch statistics, r is termed *relevance factor* which is an empirical value and has to be manually tuned. From Eq. (4), we can see that the posterior mean vectors of the cth Gaussian component $\mathbf{M}_c$ is an interpolation between the mean of UBM Gaussian component $\mathbf{m}_c$ and the normalized first-order Baum-Welch statistics $(1/L)\mathbf{F}_c$. N_c can be regarded as a conditioning factor, as the amount of speaker frames observed by the component c increases, $\mathbf{M}_c$ will get closer to the real statistical mean vectors $(1/L)\mathbf{F}_c$.

In the i-vector approach, assumed that we have obtained the i-vector $\mathbf{w}$ for a given utterance u (the expression of $\mathbf{w}$ is in Eq. (7), the process of calculating $\mathbf{w}$ is explained in [2]), the adapted mean vector of a Gaussian component can be written as:

$$\mathbf{M}_c = \mathbf{m}_c + \mathbf{T}_c\mathbf{w} \tag{5}$$

where $\mathbf{T}_c$ denotes the cth component of $\mathbf{T}$. $\mathbf{T}_c$ can be regarded as the total basis for the cth Gaussian component. From Eq. (5), we can see that although the zero-order Baum-Welch statistics of Gaussian components vary from each other, i-vector approach adapts the mean vector of each Gaussian component in a same degree. In other words, we might say that the conditioning for each Gaussian component is identical, controlled by $\mathbf{w}$.

Although FA-based approaches have been proven to be effective in compensating the phonetic variability, as we have mentioned above, its performance is not so good as the traditional MAP adaptation in text-dependent speaker verification. An intuitive explanation could be that not all the Gaussian components should be adapted if the phonetic contents of enrollment and test are the same, only those components which are observed in speaker frames need to be adapted. Extended from text-dependent verification to text-independent verification, a rational inference could be that considering the sparsity in the real-world application and imbalanced distribution of speaker frames, it is better not to adapt those Gaussian components which contribute least in modeling speaker frames. Like we used to do in the traditional MAP adaptation, as it is uncertain to perform a full adaptation to those Gaussian components without enough observation speaker frames, those Gaussian components with minimum zero-order Baum-Welch statistics are not chosen to join in the adaptation process.

In order to verify our thought, a statistical experiment is designed based on the corpus from the core condition of NIST 2008 SREs. The core condition of the 2008 SREs is named short2-short3, each file is a 5-min telephone conversation recording containing roughly 2 min of speech.

We randomly pick out 20 male recordings which are all from different speakers, a male UBM model containing 1024 Gaussian components is used to extract the zero-order Baum-Welch statistics. The configurations of feature extraction and UBM are same as experiments section.

All the values of zero-order statistics plotted in the Fig. 1 are sorted in ascending order. For simplicity, only four lines of arbitrary 4 records out of 20 records above are plotted. We can observe from Fig. 1 that the distribution of the sorted zero-order statistics is similar to that of the exponential families. The proportion of speaker frames that each component occupies is obviously imbalanced. Table 1 gives the contribution of the sum of top N Gaussian components, N ranges from 100 to 1000. From Table 1, we can observe that in the condition of a 1024 order UBM, for a 2-min speech frames, top 300 Gaussian components are adequate to explain 61.8 % speaker frames, top 900 components can explain 99.5 % speaker frames, that is to say, 90 % top Gaussian components are almost enough to model all the speaker frames. The rest 10 % Gaussian components can be regarded as redundant ones, whose adapted mean vectors are not guaranteed to be valid.

Extended from traditional MAP adaptation approach to the i-vector adaptation approach, Gaussian components are no longer assumed to be independently distributed, total variability can be captured by a low-dimensional factor space, it means that sparse training data can give a global adaptation of total Gaussian components. A graphical explanation will be given to show that although total factor space approach is effective to compensate phonetic variability, it is still not an optimal method.

Figure 2 is a simplified description of MAP adaptation. For simplicity, we do not consider the overlap

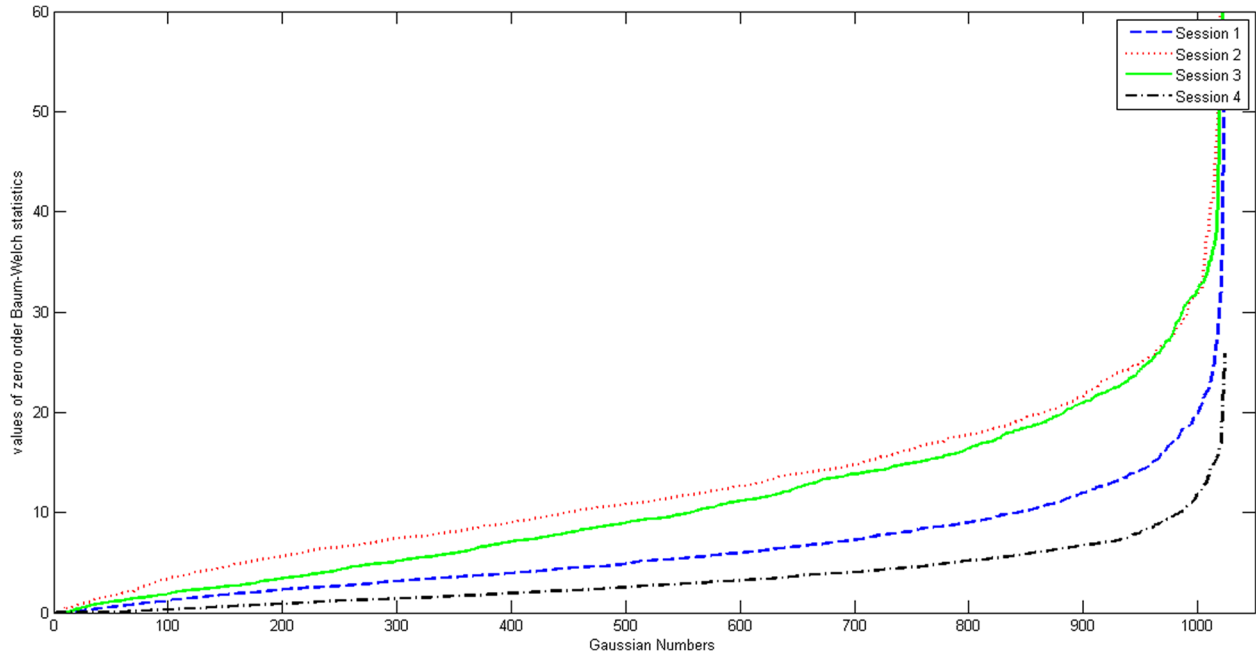

Fig. 1 The distribution of ascending zero-order Baum-Welch statistics of four records corresponding to each component

of adjacent Gaussian components and only two dimensions out of F dimensions are shown, where F is the dimensionality of mean vector of Gaussian component. According to Eq. (4), as the amount of observing speaker frames increases for a Gaussian component, the posterior distribution of mean vectors of that component will get closer to the real first-order statistics. On the other hand, those components without enough observing speaker frames only make a limited adaptation, or we can say that they are not adapted. Analogically, Fig. 3 is a description of i-vector adaptation. Evolved from eigenvoice framework, adaptation is restricted within the subspace described by the total factor matrix. Even in the case that training frames are sparse and imbalance, according to maximum likelihood estimation (MLE) of the EM auxiliary function defined in the proof of proposition 3 in [2], entire Gaussian components are adapted at same weight, adapting weight corresponding to entire Gaussian components is denoted by i-vector. Suppose that in an extreme situation of Fig. 3, most of training frames (denoted by green point) are observed by the upper Gaussian component and the rest of frames are observed by the lower left component, and following Eq. (5), we use $\mathbf{T}_c$ to denote the sub-basis of total factor matrix $\mathbf{T}$, where c denotes corresponding Gaussian component (we use $c = 1$ to denote the upper component, $c = 2$ and 3 to denote the lower left and lower right ones); according to the MLE based on i-vector framework, for components 1 and 2, we have credible first-order Baum-Welch statistics; hence, corresponding fully adapted mean vectors $\mathbf{m}_c + \mathbf{w}\mathbf{T}_c, c \in \{1, 2\}$ are also credible. However, for component 3, we have an inaccurate first-order Baum-Welch statistics (this inaccurate first-order Baum-Welch statistics also corresponds to minimal zero-order Baum-Welch statistics) because of the shortage of speaker frames; hence, corresponding fully adapted mean vector $\mathbf{m}_3 + \mathbf{w}\mathbf{T}_3$ is questionable. Moreover, from Table 1, we can conclude that "excessive" Gaussian components are used to model scarce speaker frames. Although those Gaussian components with the most observing speaker frames contribute most to the adaptation process, to those Gaussian components without enough observing speaker frames, it

Table 1 Proportion of zero-order Baum-Welch statistics of top N Gaussian components

Number of N	Percentage
100	0.2988
200	0.4791
300	0.6183
400	0.7288
500	0.8166
600	0.8848
700	0.9362
800	0.9732
900	0.9946

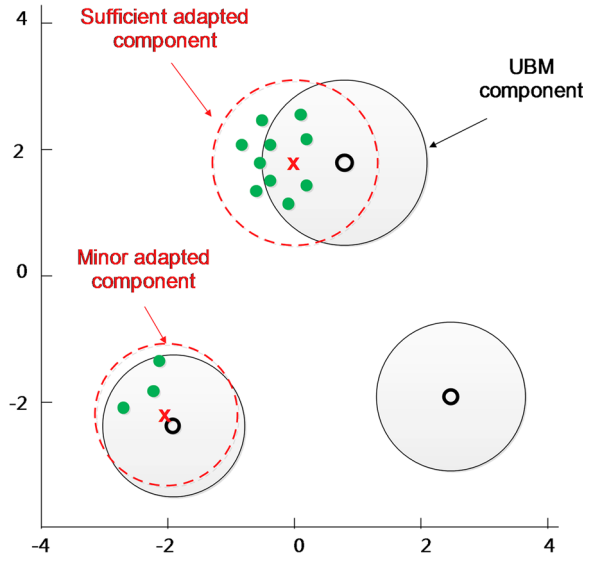

Fig. 2 A brief figure description of MAP adaptation in sparse training data. As upper Gaussian component has more speaker frames, which corresponds to higher zero-order Baum-Welch statistics N_c, the mean vector of upper component (*black dot* denotes the mean vector of Gaussian component) will be sufficiently adapted to the normalized first-order Baum-Welch statistics (*red cross* denotes the adapted mean vector and *red dashed ellipse* denotes the adapted component). In contrast, sparse speaker frames observed by lower left component correspond to smaller zero-order Baum-Welch statistics and minor adaptation. Lower right component remains still because it has no speaker frame

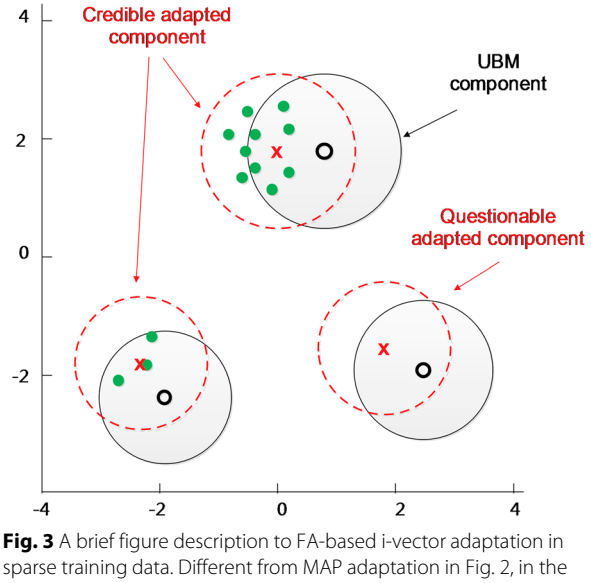

Fig. 3 A brief figure description to FA-based i-vector adaptation in sparse training data. Different from MAP adaptation in Fig. 2, in the i-vector framework, all components are adapted at the same weight. The adapted mean vectors of upper and lower left components are credible because we indeed have relative speaker frames, but the adaptation of lower right component is questionable because it suffers from a shortage of speaker frames, whose first-order Baum-Welch statistics is inaccurate

is better to abandon them rather than give a suspectable adaptation result.

4 Component reduction analysis

4.1 Implementation of component reduction analysis

Extended from our analysis above, we propose an improved i-vector extraction algorithm which we term it *component reduction analysis*. The basic idea of CRA is that those Gaussian components with minimum zero-order Baum-Welch statistics will not join in the posterior estimation of i-vector. This section describes the implementation of CRA. In order to estimate i-vector, we need to compute the centralized first-order Baum-Welch statistics based on the UBM:

$$\widetilde{\mathbf{F}}_c(u) = \mathbf{F}_c(u) - N_c(u)\mathbf{m}_c \tag{6}$$

The posterior estimation of i-vector for a given utterance u can be represented using the following equation:

$$\mathbf{w} = \left(\mathbf{I} + \mathbf{T}^t \Sigma^{-1} \mathbf{N}(u) \mathbf{T}\right)^{-1} \mathbf{T}^t \Sigma^{-1} \tilde{\mathbf{F}}_c(u) \tag{7}$$

where the process of calculating $\mathbf{w}$ can be derived from [2], $\mathbf{N}(u)$ is the diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c(u)\mathbf{I}, (c = 1, \ldots, C)$, $\mathbf{I}$ is the identity matrix of dimension $F \times F$, $\tilde{\mathbf{F}}_c(u)$ is a supervector of dimension $CF \times 1$ obtained by concatenating all the centralized first-order Baum-Welch statistics $\tilde{\mathbf{F}}_c(u)$. Here Σ is a diagonal covariance matrix of dimension $CF \times CF$ that is estimated during the training of $\mathbf{T}$. It models the residual variabilities not captured by the total variability matrix $\mathbf{T}$.

Then the zero-order Baum-Welch statistics are sorted in descending order, we use $c_1^{'}$ to denote the component with maximum value of zero-order Baum-Welch statistics, $c_2^{'}$ to denote the component with second maximum value of zero-order Baum-Welch statistics, by that analogy, $c_C^{'}$ denotes the component with minimum value of zero-order statistics. A sign function $S(c)$ is defined as:

$$S(n) = \begin{cases} 1 \text{ if } n \le R \\ 0 \text{ if } n > R \end{cases} \tag{8}$$

where R is a threshold that has to be manually tuned, which denotes the number of the components having to be discarded, and the new zero-order Baum-Welch statistics is denoted as $N_{c^{'}}(u)S(c^{'})$ where $c^{'} \in \{c_1^{'}, c_2^{'}, \ldots, c_C^{'}\}$. The update formula for the $\widetilde{w}$ is:

$$\widetilde{\mathbf{w}} = \left(\mathbf{I} + \sum_{n=1}^{C} N_{c_n^{'}}(u) S(n) \mathbf{T}_{c_n^{'}}^t \Sigma_{c_n^{'}}^{-1} \mathbf{T}_{c_n^{'}}\right)^{-1} \sum_{n=1}^{C} S(n) \mathbf{T}_{c_n^{'}}^t \Sigma_{c_n^{'}}^{-1} \widetilde{\mathbf{F}}_{c_n^{'}}(u) \tag{9}$$

$$= \left(\mathbf{I} + \sum_{n=1}^{R} N_{c_n'}(u) \mathbf{T}_{c_n'}^{t} \Sigma_{c_n'}^{-1} \mathbf{T}_{c_n'} \right)^{-1} \sum_{n=1}^{R} \mathbf{T}_{c_n'}^{t} \Sigma_{c_n'}^{-1} \widetilde{\mathbf{F}}_{c_n'}(u) \tag{10}$$

in which $\mathbf{T}_{c_n'}$ is the sub-matrix of the $c_n^{'}$th block of $\mathbf{T}$, $\mathbf{N}_{c_n'}(u)\Sigma_{c_n'}^{-1}$ is the sub-matrix of the diagonal part of the $c_n^{'}$ block of $\Sigma^{-1}\mathbf{N}(u)$, and $\tilde{\mathbf{F}}_{c_n'}(u)$ is the sub-matrix of the $c_n^{'}$th row block of $\tilde{\mathbf{F}}(u)$.

4.2 Zero-order Baum-Welch statistics normalization

Although those Gaussian components contributing least in modeling, speaker frames are removed from the adaptation formula (10), the implementation of CRA also encounters a defect. The zero-order Baum-Welch statistics corresponding to those removed Gaussian components is also removed from (10), whereas in the adaptation process of i-vector, we wish to make full use of the information of total speaker frames. For any Gaussian component c, corresponding zero-order Baum-Welch statistics $N_c(u)$ and first-order Baum-Welch statistics $F_c(u)$ are calculated from total speaker frames; hence, it seems that the best solution to cope with this defect is that we re-calculate total zero-order and first-order Baum-Welch statistics using the top R Gaussian components selected by CRA. However, it will bring almost double computational amount in the procedure of Baum-Welch statistics extraction, so we propose an approximation approach to compensate the loss of the zero-order Baum-Welch statistics which we termed zero-order Baum-Welch statistics normalization. A normalization coefficient is defined as:

$$k = N(u) / \sum_{n=1}^{C} S(n) N_{c_n'}(u) = N(u) / \sum_{n=1}^{R} N_{c_n'}(u) \tag{11}$$

After the normalization process, the update formula for the $\widetilde{w}$ is:

$$\widetilde{\mathbf{w}} = \left(\mathbf{I} + k \sum_{n=1}^{R} N_{c_n'}(u) \mathbf{T}_{c_n'}^{t} \Sigma_{c_n'}^{-1} \mathbf{T}_{c_n'} \right)^{-1} \sum_{n=1}^{R} \mathbf{T}_{c_n'}^{t} \Sigma_{c_n'}^{-1} \widetilde{\mathbf{F}}_{c_n'}(u) \tag{12}$$

This kind of normalization ensures that after applying CRA, the summation of zero-order Baum-Welch statistics of top R Gaussian components is identical to the summation of zero-order Baum-Welch statistics not applying CRA, this summation also equals to the number of total speaker frames. Although this normalization seems to be a slightly rough approximation, experimental result in the next section shows that slight improvement is obtained after applying zero-order Baum-Welch statistics normalization.

5 Experiments

5.1 Databases

All experiments were carried out on the core condition of NIST 2008 SREs. NIST 2005 and NIST 2006 were used as development datasets. Our experiments are based on male telephone data (det6) and English-only male telephone data (det7) for both training and testing. The core condition of the 2008 SREs contains 648 males.

5.2 Experimental setup

Our experiments operated on the Mel frequency cepstral coefficients (MFCCs), and speech/silence segmentation was performed according to the index of transcriptions provided by the NIST with its automatic speech recognition (ASR) tool. The MFCC frames are extracted using a 25-ms Hamming window, every 10-ms step, 19 order coefficients together with log energy, 20 first-order delta, and 10 second-order delta were appended, equal to a total dimension of $F = 50$, where we follow the configuration of [13]. All frames were subjected to cepstral mean normalization (CMN) to obey a $(0, 1)$ distribution.

We used a male UBM containing 1024 Gaussians and the order of total factor matrix T is 400, the corpus from the 2005 1conv4w transcription index and 2006 1conv4w transcription index was used to train the UBM with a total length of 17 h of speech from about 550 speakers. The corpus from the 2005 8conv4w and 2006 8conv4w transcription index was used as the development datasets to train the total factor matrix because sessions for each speaker consists of recordings from eight different microphones which is capable of modeling the speaker and channel variability. The development set of i-vectors was also extracted from the same corpus set which was used to train the T-Matrix. All the decision scores were given without normalization.

In our experiment, total sets of i-vectors were extracted with our CRA algorithm, including development set, enrollment set, and test set. We also performed experiments that only enrollment and test sets were extracted using CRA algorithm, and results show that both approaches gave better results than the baseline of traditional i-vector extraction approach, but the results of whole extraction (development set, enrollment set and test set) are best. The threshold R is tuned manually.

5.3 Results and analysis

Table 2 gives comparison results for male portion of core condition of NIST 2008 SREs. LDA and eigen factors radial (EFR) are taken as the compensation approaches, cosine scoring is taken as scoring method, and statistical results are given in terms of equal error rate (EER) and normalized minimum decision cost functions (DCF) for the two operating points as defined by NIST for the SRE 2008 evaluations.

Table 2 Comparison of cosine scoring with different compensation and normalization techniques

	Male			
	English trials		All trials	
	EER(%)	DCF	EER(%)	DCF
LDA(210)	2.50	0.0246	5.57	0.0538
LDA(210), $R = 1000$	2.48	0.0245	5.48	0.0533
LDA(210), $R = 975$	2.31	0.0226	5.37	0.0521
LDA(210), $R = 950$	2.27	0.0220	5.24	0.0514
LDA(210), $R = 925$	2.23	0.0211	5.18	0.0510
LDA(210), $R = 900$	*2.15*	*0.0207*	*5.10*	*0.0503*
LDA(210), $R = 875$	2.30	0.0224	5.12	0.054
LDA(210), $R = 850$	2.53	0.0250	5.17	0.510
EFRnorm, LDA(210)	*2.35*	*0.0233*	*5.40*	*0.0525*
EFRnorm, LDA(210), $R = 900$	2.29	0.0225	5.19	0.0512

All results are obtained with cosine scoring, LDA dimension of 210, and without EFR normalization. The baselines of our systems are 2.35 % for EER in English trials and 5.40 % for EER in all trials (italics), where EFR normalization and LDA = 210 are applied, best results after applying CRA are 2.15 % for EER in English trials and 5.10 % for EER in all trials (italics), where only LDA = 210 are applied

For our system, best LDA dimension is 210 (in the case that dimension of total factor matrix is 400), standardization before LDA compensation (EFR norm four iterations) enhances the performance and gives a baseline of 2.35 % for EER and 0.0233 for DCF in the English trials (det7), 5.40 % for EER and 0.0525 for DCF in the multi-language trials (det6, all trials). After applying CRA, as we reduce the number of components, EER and DCF decrease slightly, the best performance is obtained when the threshold R is 900, which gives a minimum EER of 2.15 and minimum DCF of 0.0207 in the English trials and a minimum EER of 5.10 and minimum DCF of 0.0503 in the multi-language trials. Here what should be highlighted is that the optimal dimension of LDA after applying CRA algorithm is still unchanged, the reason that CRA has no effect on the optimal choice of dimension of LDA is that the CRA aims at compensating the phonetic variability and phonetic imbalance, whereas LDA aims at compensating the channel variability. However, the best performance is obtained without EFR normalization. Even though CRA still works in the case of applying EFR normalization, its performance is not so good as in the case of without EFR normalization both in English trials and in multi-language trials. One possible reason for this result is that the performance of EFR normalization is highly dependent on the scale of development dataset, and in our experiment, we did not offer that much corpus to construct full-scale between- and within-speaker covariance matrices.

Table 3 gives the performance comparison of zero-order Baum-Welch statistics normalization under various threshold R. As the threshold R decreases, the effectiveness of zero-order Baum-Welch statistics normalization becomes more obvious. The reason is that the value of compensation coefficient k in (11) is getting larger as more Gaussian components are removed from (10), which provides stronger compensation effect to the loss of zero-order Baum-Welch statistics.

Figure 4 shows a comparison of DET curves (det7) of LDA + EFR baseline and our improved baseline by applying CRA algorithm without EFR

Table 3 Comparison of results with and without zero-order Baum-Welch statistics normalization

	Male			
	English trials		All trials	
	EER(%)	DCF	EER(%)	DCF
$R = 1000$, no zero-order norm	2.48	0.0245	5.48	0.0533
$R = 1000$, with zero-order norm	2.48	0.0245	5.48	0.0533
$R = 950$, no zero-order norm	2.29	0.0221	5.25	0.0516
$R = 950$, with zero-order norm	2.27	0.0220	5.24	0.0514
$R = 900$, no zero-order norm	2.19	0.0211	5.12	0.0506
$R = 900$, with zero-order norm	2.15	0.0207	5.10	0.0503

All results are obtained with cosine scoring, LDA dimension of 210, and without EFR normalization

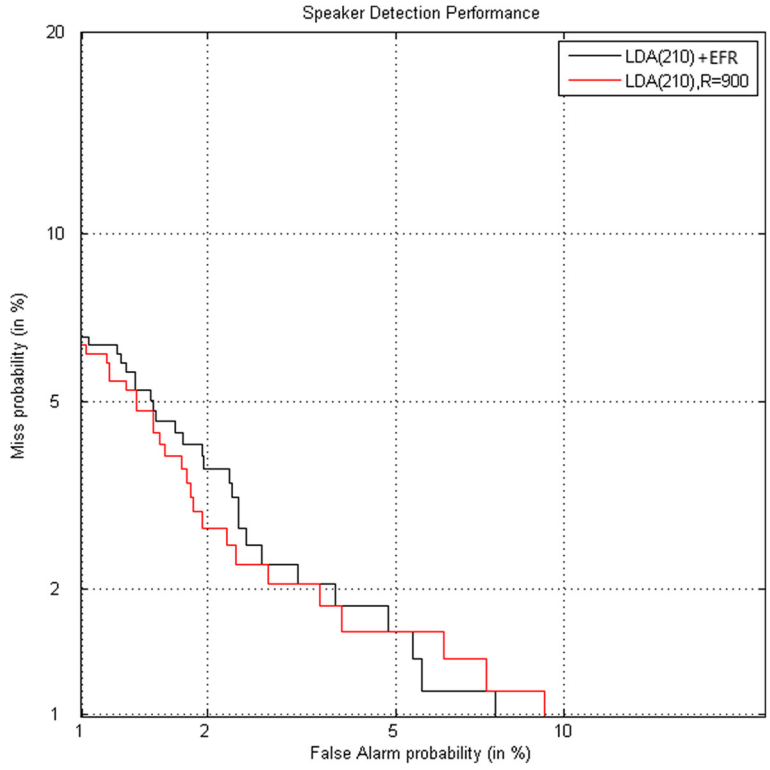

Fig. 4 DET curves comparison between LDA and cosine scoring baseline and component reduction analysis improvement

normalization. A 10–15 % relative improvement is obtained.

6 Conclusions

This paper proposes an improved i-vector extraction algorithm. We analyze the intrinsic sparsity and imbalanced distribution of speaker frames, those redundant Gaussian components, are discarded in the i-vector extraction phase so as to compensate the phonetic variability. We attempt to make a preliminary beginning to combine the advantages of traditional MAP adaptation in text-dependent speaker verification and i-vector-based framework in text-independent speaker verification. Besides speaker variability and channel variability, phonetic variability, as another impact factor, is taken into consideration. We carried out experiments on the core condition of male portion of NIST 2008 SREs. Experimental results show that a 10–15 % relative improvement is obtained.

Despite the effectiveness of our CRA algorithm, there exists many questions to be solved and be explained. First is the conflict of our CRA algorithm and the EFR normalization. More experiments have to be designed to explore the relationship and compatibility of this two techniques. Second issue is that the threshold R of the CRA is a parameter that has to be tuned manually, which is highly dependent on the length of speech frames, $R = 900$ is an optimal value in the core condition of NIST 2008 SREs, but we cannot give a similar configuration in the female portion, so do in the conditions like 10sec-10sec and short2-10sec of NIST 2008 SREs dataset, an adapted adjustment of the threshold R has to be proposed to make speaker verification more practical. On the other hand, we expect our CRA algorithm based on i-vector framework may improve the performance of i-vector-based system in text-dependent speaker verification.

Competing interests

We have stated the questions which our CRA algorithm has to face in the conclusion section. Both in theory and in practical applications, CRA still remains much work to explore. The most meaningful competing interests might be the adaptive tuning of threshold R mentioned above and how CRA can be applied in text-dependent speaker verification. We will publish them if any breakthrough is obtained.

Authors' contributions

WL and T-fan F propose the idea of CRA together. WL wrote this paper and designed all the experiments. T-fan F plotted all the experimental figures and collected statistical data. JZ, as our professor, participated in the theoretical research. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the superior open-source ALIZE/LIA_RAL platform, its powerful functionality enables us to build a state-of-the-art speaker verification system and focus on the ideas we are interested in.
This article is supported by the National Natural Science Foundation of China (61371147): A study of sparse compression and precise reconstruction of high definition audio in transform domain and its application in the mobile terminal.

This article is also supported by the National Natural Science Foundation of China (61271349): Study on music similarity model based on nonlinear dynamical characteristics analysis of acoustic signal and its application.

Author details
[1]Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Rd, 200240 Shanghai, China. [2]Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University, 800 Dong Chuan Rd, 200240 Shanghai, China.

Received: 9 July 2014 Accepted: 9 June 2015
Published online: 27 June 2015

References

1. DA Reynolds, TF Quatieri, RB Dunn, Speaker verification using adapted gaussian mixture models. Digit. Signal Process. **10**(1), 19–41 (2000)
2. P Kenny, G Boulianne, P Dumouchel, Eigenvoice modeling with sparse training data. IEEE Trans. Speech and Audio Process. **13**(3), 345–354 (2005)
3. P Kenny, P Ouellet, N Dehak, V Gupta, P Dumouchel, A study of interspeaker variability in speaker verification. IEEE Trans. Audio, Speech, and Lang. Process. **16**(5), 980–988 (2008)
4. P Kenny, Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal, (Report) CRIM-06/08-13 (2005)
5. P Kenny, G Boulianne, P Ouellet, P Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio, Speech, and Lang. Process. **15**(4), 1435–1447 (2007)
6. P Kenny, in *Odyssey*. Bayesian speaker verification with heavy-tailed priors, (2010), p. 14
7. N Dehak, P Kenny, R Dehak, P Dumouchel, P Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech, and Lang. Process. **19**(4), 788–798 (2011)
8. P-M Bousquet, D Matrouf, J-F Bonastre, in *INTERSPEECH*. Intersession compensation and scoring methods in the i-vectors space for speaker recognition, (2011), pp. 485–488
9. P-M Bousquet, A Larcher, D Matrouf, J-F Bonastre, O Plchot, in *Odyssey: The Speaker and Language Recognition Workshop*. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis (Singapore, Singapore, 2012), pp. 157–164
10. H Aronowitz, in *Odyssey 2012-The Speaker and Language Recognition Workshop*. Text dependent speaker verification using a small development set, (2012)
11. A Larcher, P Bousquet, KA Lee, D Matrouf, H Li, J-F Bonastre, in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference On*. I-vectors in the context of phonetically-constrained short utterances for speaker verification (IEEE, 2012), pp. 4773–4776
12. T Stafylakis, P Kenny, P Ouellet, J Perez, M Kockmann, P Dumouchel, Text-dependent speaker recognition using plda with uncertainty propagation. Matrix. **500**, 1 (2013)
13. J-F Bonastre, N Scheffer, D Matrouf, C Fredouille, A Larcher, A Preti, G Pouchoulin, NW Evans, BG Fauve, JS Mason, in *Odyssey*. Alize/spkdet: a state-of-the-art open source software for speaker recognition, (2008), p. 20