Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 60–70

www.elsevier.com/locate/specom

Feature sparsity analysis for i-vector based speaker verification

Wei Li[a,*], Tianfan Fu[b], Hanxu You[a], Jie Zhu[a], Ning Chen[c]

[a] *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[b] *Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University, Shanghai 200240, China*
[c] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200240, China*

Received 30 October 2015; received in revised form 4 February 2016; accepted 4 February 2016
Available online 29 April 2016

Abstract

In recent years, the i-vector based framework has been proven to provide state-of-the-art performance in the speaker verification field. Each utterance is projected onto a total factor space and is represented by a low-dimensional i-vector. However, the degradation of performance in the i-vector space remains problematic and is commonly attributed to channel variability. Most techniques used for the channel compensation of the i-vectors, such as linear discriminant analysis (LDA) or probabilistic linear discriminant analysis (PLDA) aim to compensate for the variabilities caused by channel effects. However, in real-world applications, the duration of enrollment and test utterances by each user (speaker) are always very limited. In this paper, we demonstrate, from both analytical and experimental perspectives, that feature sparsity and imbalance widely exist in short utterances, in which case the conventional i-vector extraction algorithm, based on maximum likelihood estimation (MLE), may lead to over-fitting and decrease the performance of the speaker verification system, especially for short utterances. This prompted us to propose an improved i-vector extraction algorithm, which we term adaptive first-order Baum–Welch statistics analysis (AFSA). This new algorithm suppresses and compensates for the deviation from first-order Baum–Welch statistics caused by feature sparsity and imbalance.

We reported results on the male telephone portion of the core trial condition (short2-short3) and other short time trial conditions (short2-10sec and 10sec-10sec) on NIST 2008 Speaker Recognition Evaluations (SREs) dataset. As measured both by Equal Error Rate (EER) and the minimum values of the NIST Detection Cost Function (minDCF), 10%–15% relative improvement is obtained compared to the baseline of traditional i-vector based system.

Keywords: Speaker verification; i-vector; Total factor space; Feature variability; Adaptive first-order Baum–Welch statistics analysis (AFSA).

1. Introduction

Speaker verification technology is used to accept or reject a claimed identity by comparing two utterances. The first of these utterances is used for enrollment and is produced by the speaker with a target identity, whereas the second utterance is obtained from the speaker with a claimed identity and is used for testing purposes. In the last decade, the Gaussian mixture model based on the universal background model (GMM-UBM) framework has demonstrated strong performance and has become the most widely used method for speaker verification. In this framework, the mean vectors of each Gaussian component are commonly considered to represent most of a speaker's characteristics, whereas the other parameters, such as the weights and variances of each Gaussian component, are inherited from the UBM model (Reynolds et al., 2000). Traditional speaker models are obtained by employing a maximum a posteriori (MAP) adaptation. However, traditional MAP (or relevance MAP) treats each Gaussian component as a statistically independent distribution, which has many drawbacks in practical applications: only those components with a sufficient number of assigned speaker frames are well adapted, leaving the remaining components almost unchanged. Time-limited enrollment and test utterances or those that suffer from severe phonetic variability may lead to an obvious degradation in the performance of the verification system. Apart from

[*] Corresponding author. Tel.: +86 136 1163 6772; fax: +86 021 34205432.
E-mail address: liweisjtu@126.com (W. Li).

http://dx.doi.org/10.1016/j.specom.2016.02.008

these disadvantages, traditional MAP does not have the ability to compensate for the effects of channel distortion, especially when the enrollment and test utterances use different channels.

An extension of the GMM-UBM framework, namely the factor analysis (FA) technique (Kenny et al., 2005, 2008), attempts to jointly model the speaker components. Each speaker is represented by a *mean supervector*, which is a linear combination of the set of *eigenvoices*. Generally, only a few hundred free parameters need to be estimated, which ensures that the speaker mean supervector converges quickly by using a training utterance with a relatively short duration. Based on the FA technique, *joint factor analysis (JFA)* (Kenny, 2005; Kenny et al., 2007) decomposes the GMM supervector into a speaker component **S** and a channel component **C**, and assumes these two components to be statistically independent. Although it is known by now that channel effects are not speaker-independent (for example, experimentally, gender-dependent eigenchannel modeling has been reported to be more effective than gender-independent modeling Kenny, 2010), compared to other methods JFA has still demonstrated good performance for text-independent speaker verification tasks in past NIST speaker recognition evaluations (SREs).

Inspired by the JFA approach, the authors in Dehak et al. (2011) proposed a combined speaker and channel space by defining a novel low-dimensional space named the *total factor space*. In this space, each utterance is represented by a low-dimensional feature vector termed an i-vector. The concept of an i-vector has opened the door for new ways in which to analyze speaker and session variability. As a result, various optimization and compensation techniques and scoring methods have been proposed (Bousquet et al., 2012, 2011; Dehak et al., 2011; Kenny, 2010), all of which have improved the results obtained with the JFA approach. Of late, i-vector extraction with length normalization and probabilistic linear discriminant analysis (PLDA) has become the state-of-the-art configuration for speaker verification (Bousquet et al., 2012; Kenny, 2010).

Although the i-vector based system has dominated the speaker verification field, recent research found the adaptive relevance factor, which replaces the traditional manually tuned relevance factor, capable of boosting the MAP-based Gaussian mixture model - support vector machine (GMM-SVM) framework to obtain a performance comparable to those of the JFA and i-vector frameworks (You et al., 2012, 2013).

Despite the success of the i-vector paradigm, it still has some shortcomings, one of which is that its applicability to *text-dependent speaker verification* continues to remain difficult. In cases in which the lexical contents of enrollment and test utterances are identical, we may assert that only those components with a sufficient number of speaker frames need to be adapted. However, FA-based techniques globally adapt all the Gaussian components, including those components with sparse or no speaker frames, which results in the performance of FA-based techniques not being comparable to that of traditional MAP adaptation in text-dependent speaker verification. Related work has also shown that results obtained with the traditional MAP approach can be better than those obtained with i-vector based methods (Aronowitz, 2012; Larcher et al., 2012; Stafylakis et al., 2013).

Considering that the i-vector based system is an extension of the text-dependent speaker verification field, in the context of text-independent verification, the principal challenge of this system in terms of achieving a low error rate is that the intra-speaker variability in the estimated parameters increases considerably as a result of variability in the lexicon and the training utterance duration (Hautamäki et al., 2013) (text-dependent speaker verification can be regarded as a special case of text-independent speaker verification of short utterances).

Two main streams have emerged to address the shortcomings of i-vector applicability to short utterances: normalizing those i-vectors derived from short utterances in the low-dimensional i-vector space (Kanagasundaram et al., 2012; Kenny et al., 2013; Larcher et al., 2013) and regularizing the Baum–Welch statistics of short utterances to obtain a more robust i-vector estimation (Hautamäki et al., 2013). In the analysis presented in this paper, we attempt to show that, although the FA-based i-vector approach is capable of successfully modeling speaker variability, on some occasions the traditional i-vector extraction algorithm may lead to overfitting. This problem typically occurs when the speech frames from the target speaker are sparse. In this paper, we continue to focus on the *text-independent* speaker verification field, and propose an improved i-vector extraction algorithm we have named *adaptive first-order Baum–Welch statistics analysis* (AFSA). AFSA attempts to suppress and compensate for the biased first-order Baum–Welch statistics caused by feature sparsity. Traditionally, zero-order and first-order Baum–Welch statistics are considered as determinate values extracted from a posteriori statistics of speaker frames based on the UBM model. Our approach is to provide first-order Baum–Welch statistics with a Bayesian explanation (although we continue to refer to it as “statistics”). Our proposed method treats phonetic variability and channel variability as mutually independent distributions, and as we show in the experimental section, various channel compensation techniques continue to work efficiently. Experiments were carried out on the core condition (short2-short3) and other conditions of a short duration (short2-10sec and 10sec-10sec) of NIST 2008 SREs. The experimental results show that, by applying the AFSA algorithm to the phase of i-vector extraction, a 10%–15% relative improvement is obtained compared with a baseline system in which a traditional i-vector algorithm with the same channel compensation techniques is adopted.

2. Supervector, total factor space, and i-vector extraction

The GMM-UBM framework is commonly considered to only require adaptation of the mean vectors of each Gaussian component for a given UBM model Ω and a training utterance. A *supervector* comprises the concatenation of each of the mean vectors. In the context of the i-vector framework, speaker variability and channel variability are jointly modeled by a *total factor matrix*. This means that total variability is restricted in this linear manifold; hence, each utterance can be projected onto this total factor space and be represented by a low-dimensional feature vector.

The i-vector approach is based on the principle that each speaker- and channel-dependent GMM supervector **M** can be modeled as:

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{T}\mathbf{w}(s), \tag{1}$$

where s denotes a target speaker, **m** is a speaker- and channel-independent supervector, which is often taken from the UBM supervector, **T** is the total factor matrix with low rank, which is an expanded subspace containing speaker- and channel-dependent information (details of the process of training the total factor matrix are provided in Kenny et al., 2005), and $\mathbf{w}(s)$ is the i-vector extracted from the training utterance. For the same utterance, the supervector and i-vector allow one-to-one mapping; however, the i-vector is much shorter than the supervector, and this means that optimization and compensation techniques can simply be manipulated on the i-vector space.

3. Feature sparsity analysis

3.1. Baum–Welch statistics and adapted Gaussian mean vectors

Estimation of an i-vector requires both the first- and zero-order Baum–Welch statistics for an utterance to be extracted beforehand, and conventionally this is done based on a UBM model. Suppose we have a sequence of Łframes $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L\}$ and a UBM model Ω composed of C Gaussian components defined in some feature space of dimension F. Then, the Baum–Welch statistics for a given utterance u are obtained by

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

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

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

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

$$\mathbf{M}_c = \frac{r}{N_c + r}\mathbf{m}_c + \frac{N_c}{N_c + r}\left(\frac{\mathbf{F}_c}{N_c}\right), \tag{4}$$

where $\mathbf{M}_c$ denotes the cth component of the speaker supervector **M,** $\mathbf{m}_c$ denotes the cth component of **m,** $\mathbf{F}_c/N_c$ is the normalized first-order Baum–Welch statistics, and r is termed the *Relevance Factor*, which is an empirical value requiring manual tuning. Eq. (4) shows that the posterior mean vector of the cth Gaussian component $\mathbf{M}_c$ is an interpolation between the mean vector of the UBM Gaussian component $\mathbf{m}_c$ and the normalized first-order Baum–Welch statistics $\mathbf{F}_c/N_c$. N_c can be regarded as a conditioning factor; thus, as the number of speaker frames assigned by the cth component increases, $\mathbf{M}_c$ will approach the real statistical mean vector $\mathbf{F}_c/N_c$. In contrast, if sparse speaker frames are obtained, $\mathbf{M}_c$ will inherit more frames from $\mathbf{m}_c$.

When an i-vector is used, 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 **T** and $\mathbf{T}_c$ can be regarded as the total basis for the cth Gaussian component. Eq. (5) indicates that, in the i-vector framework, the adapted mean vector for a Gaussian component is no longer tuned by zero-order Baum–Welch statistics; in other words, all Gaussian components are adapted to the same degree, which is controlled by the i-vector.

3.2. Deviation of Baum–Welch statistics for sparse training data

Suppose the distribution of speaker frames is approximately uniform and balanced in the speaker space, the first-order Baum–Welch statistics stay within a bounded range and do not deviate much from the corresponding UBM mean vectors. Unfortunately, under real-world conditions, limiting the duration of a training utterance causes the feature sparsity and imbalance to become severe.

Verification of the extent to which the duration of an utterance can be associated with the corresponding feature distribution was conducted by extracting a designated set of Baum–Welch statistics from different male speakers. These speakers represent the conditions *10sec, short2*, and *8conv* of the NIST SRE 2008 corpus sets (a more detailed description of the diverse conditions can be obtained from Martin and Greenberg, 2009). In our designated set, 50 male speakers were included in each condition, a total of 150 speakers. Table 1 compares the average effective speech duration for each of the three conditions for training the speaker model (all silent and non-speech segments were removed.)

The feature extraction procedure and the UBM model are configured as in the experimental section. The verification process is described as follows. First, for each speaker the components of the zero-order Baum–Welch statistics vectors

Table 1
Comparison of the average effective speech duration for the three conditions.

Condition	Average speech duration
10sec	10 s
short2	1.5–2.5 min
8conv	15–20 min

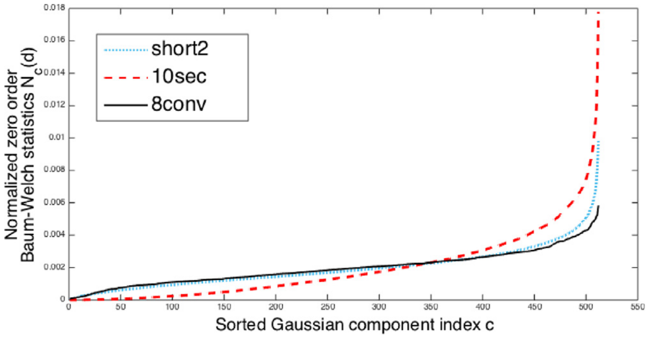

Fig. 1. Sorted distribution of zero-order Baum–Welch statistics corresponding to various conditions. The *X* and *Y* axes denote the sorted Gaussian component index and the corresponding value of the normalized zero-order Baum–Welch statistics, respectively.

are sorted in ascending order. As each component of the zero-order Baum–Welch statistics vector corresponds to a specified Gaussian component, the sorting process also leads to an alignment of all the Gaussian components. We term the order of these re-aligned Gaussian components the *zero-sorted Gaussian component order*, and it is often used in the following analysis. Next, the sorted zero-order Baum–Welch statistics vector is normalized to ensure that the summation of the components of each zero-order Baum–Welch statistics vector is equal to 1. Then, we obtain the average among the speakers in the same condition to suppress perturbation and speaker variability. Finally, normalized zero-order Baum–Welch statistics are adopted to measure their diversity among the three conditions.

The process above can be written as:

$$N_c(d) = \frac{1}{|S_d|}\sum_{j\in s_d}\frac{N_c(j)}{\sum_{c=1}^{C}N_c(j)}, \tag{6}$$

where d denotes the condition, i.e., for either 10sec, short2, or 8conv, s_d denotes the speaker set in condition d, subscript c denotes the cth index of the zero-sorted Gaussian component order for speaker j, $N_c(j)$ denotes its corresponding zero-order Baum–Welch statistics, $|S_d|$ is the total number of speakers in speaker set s_d (in our designated set, $|S_d|$ is always equal to 50 in each of the three conditions), and C is the total number of Gaussian components.

As indicated in Fig. 1, the distributions of short2 and 8conv are similarly flat; by contrast, the gradient of the 10sec condition is far steeper than those of the other two conditions, which means that shorter utterances (the 10sec condition) are more severely affected by feature sparsity and imbalance. Considering that the average duration of a 10sec portion does not last more than 10 seconds, it can be concluded that, among the utterances included in the 10sec condition, quite a number of Gaussian components do not assign a sufficient number of speaker frames to adequately support their adaptation, i.e., most speaker frames are captured by a few Gaussian components. Table 2 lists the proportion of zero-order Baum–Welch statistics assigned by the top K Gaussian components in the various conditions (in our designated experiment, the maximum value of K is 512, which is in agreement with the experimental section).

Table 2
Proportion of normalized zero-order Baum–Welch statistics assigned by the top K Gaussian components for the three conditions.

	Percentage		
Number of K	10sec	short2	8conv
80	0.469	0.333	0.293
160	0.707	0.543	0.501
240	0.857	0.706	0.672
320	**0.966**	0.803	0.781

Another problem associated with feature sparsity and imbalance is that this leads to deviation from first-order Baum–Welch statistics. We evaluated the extent of deviation by adopting the Euclidean distance between the UBM Gaussian mean vectors and the corresponding normalized first-order Baum–Welch statistics as the metric. This can be written as:

$$\begin{aligned} l_c(d) &= \frac{1}{|S_d|}\sum_{j\in s_d}\left\|\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{m}_c\right\|_2 \\ &= \frac{1}{|S_d|}\sum_{j\in s_d}\sqrt{\left(\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{m}_c\right)^T\left(\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{m}_c\right)}, \end{aligned} \tag{7}$$

where $\mathbf{m}_c$ is the UBM mean vector of the cth sorted component and $N_c(j)$ is in accordance with (6), where c denotes the cth index of zero-sorted Gaussian component order for speaker j and $\mathbf{F}_c(j)$ is the corresponding first-order Baum–Welch statistics in accordance with $N_c(j)$.

Fig. 2 shows a clear deviation of the curve of the 10sec condition from the UBM mean vectors in comparison with short2 and 8conv. This enables us to conclude that for short training utterances and sparse speaker frames (as in the 10sec condition), we are unable to obtain reliable (or unbiased) first-order Baum–Welch statistics and this involves more than half of all the Gaussian components. This is because for those Gaussian components that contain a smaller number of

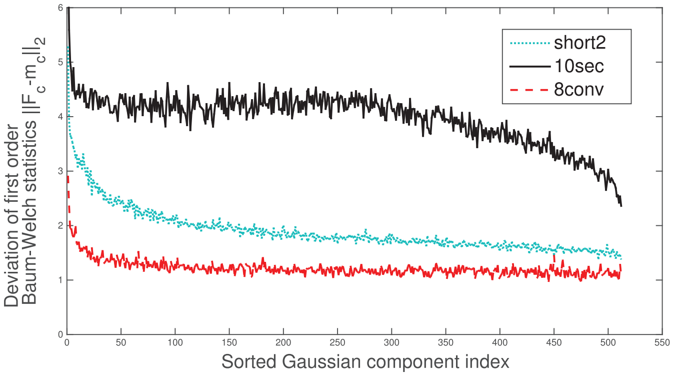

Fig. 2. Distribution of first-order Baum–Welch statistics $l_c(d)$ in accordance with $N_c(d)$ as in Fig. 1. We can see that the curve of 10sec condition deviates obviously from the UBM mean vectors in comparison with short2 and 8conv, which means that for more than 50% of all the Gaussian components, their first-order Baum–Welch statistics are questionable (or biased).

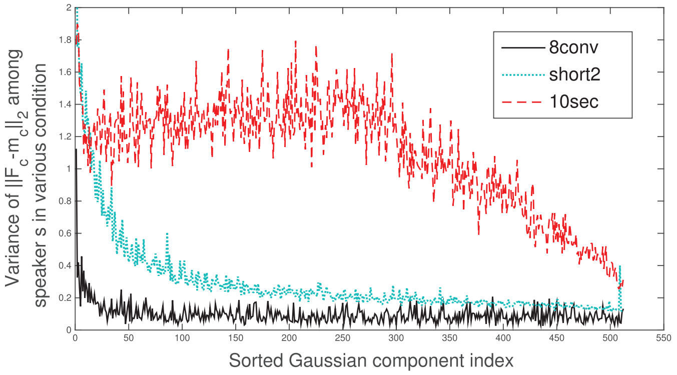

Fig. 3. Distribution of the variance of normalized first-order Baum–Welch statistics $v_c(d)$ in accordance with $N_c(d)$ as in Fig. 1.

speaker frames (corresponding to lower $N_c(d)$), their $l_c(d)$ deviate more severely than those $l_c(d)$ with a sufficient number of training frames (corresponding to higher $N_c(d)$). However, for longer training utterances (where $d \in \{short2, 8conv\}$), $l_c(d)$ show less deviation than the $l_c(d)$ in the 10sec condition.

An alternative way of measuring the deviation of normalized first-order Baum–Welch statistics is to determine the variance of $\|\mathbf{F}_c(j)/N_c(j) - \mathbf{m}_c\|_2$ among speakers for the same condition. This variance can be regarded as a metric to measure the extent of disturbance of $l_c(d)$, which can be written as:

$$v_c(d) = \frac{1}{|S_d|}\sum_{j\in s_d}\left\|\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{k}_c\right\|_2 = \frac{1}{|S_d|}\sum_{j\in s_d}\sqrt{\left(\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{k}_c\right)^T\left(\frac{\mathbf{F}_c(j)}{N_c(j)} - \mathbf{k}_c\right)}, \tag{8}$$

where

$$\mathbf{k}_c = \frac{1}{|S_d|}\sum_{j\in s_d}\frac{\mathbf{F}_c(j)}{N_c(j)}, \tag{9}$$

where $c = 1, \ldots, C$ denotes the cth index of the zero-sorted Gaussian component order for speaker j as in (6). The plots in Fig. 3 enable us to conclude that for those Gaussian components with a sufficient number of speaker frames, the variance of $l_c(d)$ among the speakers is relatively small and stable as shown in the distributions of short2 and 8conv. In contrast, the $l_c(d)$ of those Gaussian components with sparser speaker frames, i.e., lower $N_c(d)$, may experience a more severe disturbance and this leads to a higher variance of $l_c(d)$ among speakers. This is clearly indicated by the plot of the 10sec condition and the initial section of the plot of the short2 condition.

3.3. Over-fitting problem for sparse training data

As mentioned above, the i-vector framework supports one-to-one mapping of the low-dimensional i-vector to the high-dimensional speaker supervector, as shown in Eq. (1). Because extraction of the i-vector is based on the maximum likelihood criterion, (Kenny et al., 2005), a supervector that is recovered from the corresponding i-vector may be trapped in a local maximum for cases in which the training data (speaker frames) are sparse and imbalanced.

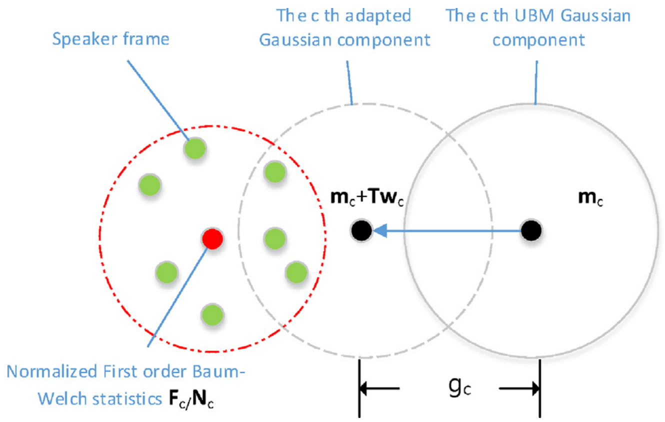
Speaker frame
The c th adapted Gaussian component
The c th UBM Gaussian component
m_c+Tw_c
m_c
Normalized First order Baum-Welch statistics F_c/N_c
g_c

Fig. 4. Schematic explanation of $G_c(d)$, which is used to measure the degree of adaptation from the UBM mean vector to the corresponding first-order Baum–Welch statistics.

The degree of adaptation of the supervector to the training data was measured by adopting the Euclidean distance between the adapted component and the corresponding UBM component, which can be written as:

$$\begin{aligned} g_c(d) &= \frac{1}{|S_d|}\sum_{j\in s_d}\|\mathbf{M}_c(j) - \mathbf{m}_c\|_2 \\ &= \frac{1}{|S_d|}\sum_{j\in s_d}\|\mathbf{m}_c + \mathbf{T}_c\mathbf{w}(j) - \mathbf{m}_c\|_2 = \frac{1}{|S_d|}\sum_{j\in s_d}\|\mathbf{T}_c\mathbf{w}(j)\|_2 \\ &= \frac{1}{|S_d|}\sum_{j\in s_d}\sqrt{(\mathbf{T}_c\mathbf{w}(j))^T(\mathbf{T}_c\mathbf{w}(j))}, \end{aligned} \tag{10}$$

where c denotes the cth zero-sorted Gaussian component order for speaker j as in (6)–(9), $\mathbf{M}_c(j)$ is the mean vector of the sorted cth adapted component for speaker j, $\mathbf{m}_c$ is the UBM mean vector of the sorted cth component, $\mathbf{T}_c$ denotes the sorted cth component of $\mathbf{T}$, and $\mathbf{w}(j)$ is the i-vector for speaker j.

An intuitive explanation of $g_c(d)$ is shown in Fig. 4, which shows how $g_c(d)$ can be used to measure "how far" the adapted components move from the UBM components. For conditions for which sufficient speaker frames are given (such as short2 and 8conv), $g_c(d)$ is always bounded in a limited range, as depicted in Fig. 5. However, when we encounter the sparse training data problem, many Gaussian components with few speaker frames may have highly biased first-order Baum–Welch statistics, in which case their $g_c(d)$ also experience severe deviation. This means that their adapted mean vectors are more prone to approaching the corresponding biased normalized first-order Baum–Welch statistics. Biased estimation of the supervector also leads to biased estimation of the i-vector.

A schematic comparison of classical MAP adaptation and i-vector adaptation on sparse training data is shown in

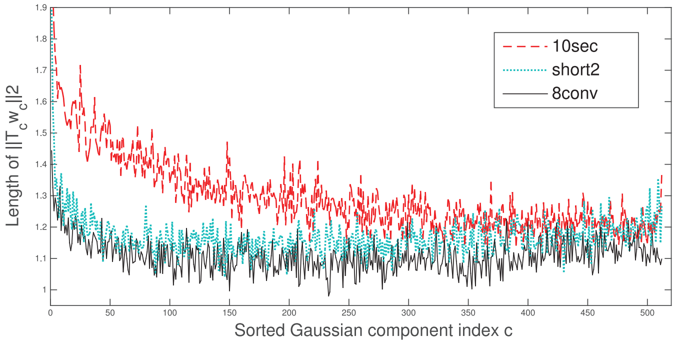

Fig. 5. Distribution of $g_c(d)$ in accordance with $N_c(d)$ as in Fig. 1. For the sparse training data condition, those components with fewer speaker frames may experience biased adaptation (over-fitting the sparse and imbalanced training data).

Figs. 6 and 7. For convenience we do not consider the overlap of Gaussian components. The left sub-figure in Fig. 6 contains the UBM model and sparse training frames, corresponding to the distribution of normalized first-order Baum–Welch statistics $\mathbf{F}_c/N_c$, whereas $c \in \{1, 2, 3\}$ is shown in the sub-figure on the right.

In relevance MAP adaptation, as shown in the left sub-figure of Fig. 7, Gaussian component 1 assigns sufficient speaker frames (corresponding to higher N_c); hence, it is adapted sufficiently. In contrast, Gaussian components 2 and 3 assign fewer speaker frames; hence, they almost remain unchanged because of the lack of sufficient and reliable statistics (corresponding to lower N_c).

On the other hand, in i-vector adaptation, as shown in the right sub-figure of Fig. 7, adaptation of a single component is no longer tuned by N_c; hence, each component is more prone to approaching its $\mathbf{F}_c/N_c$, including those components with biased statistical estimations, which is the main reason for the over-fitting problem.

4. Adaptive first-order Baum–Welch statistics analysis

As mentioned above, for a single Gaussian component, as the number of speaker frames it assigns increases, its first-order Baum–Welch statistics will be bounded and be more steady. Hence, to compensate for the deviation of first-order Baum–Welch statistics caused by feature sparsity and imbalance, the idea of adaptation is adopted, in which case the objective for compensation comprises two aspects:

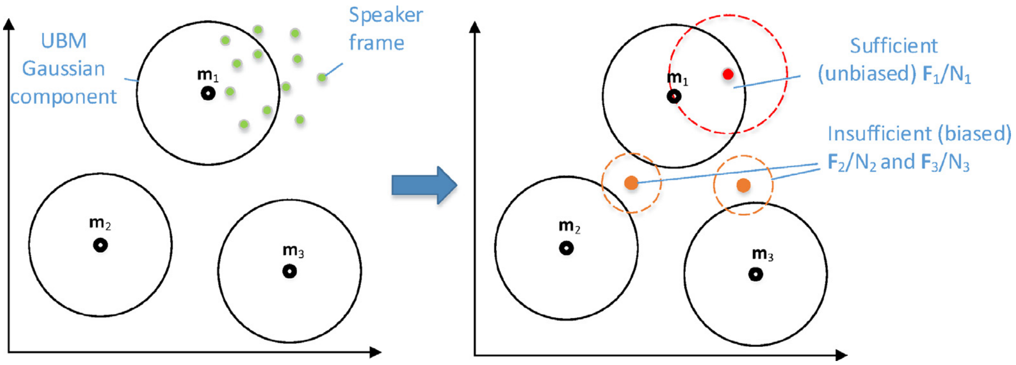

Fig. 6. Left: Sparse and imbalanced training data in GMM space. Right: Corresponding Baum–Welch statistics. Dashed contours represent the value of the zero-order Baum–Welch statistics N_c, $c \in \{1, 2, 3\}$, where larger contours represent higher N_c. Biased $\mathbf{F}_c/N_c$ corresponds to low N_c (because few speaker frames are assigned).

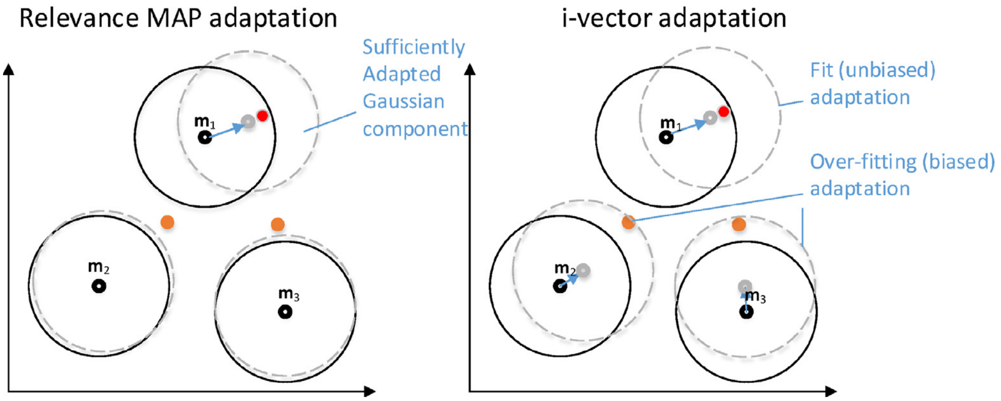

Fig. 7. Left: Traditional MAP adaptation paradigm. Because N_c is tuned, those components with biased $\mathbf{F}_c/N_c$ require either minor or no adaptation. Right: i-vector adaptation paradigm, i.e., adaptation for all components is only controlled by the low-dimensional i-vector. Hence, for excessively biased $\mathbf{F}_c/N_c$, adaptation may be trapped in a local maximum.

(1) Retain the information of Gaussian components with sufficient statistics.
(2) Suppress the possible deviation of normalized first-order Baum–Welch statistics caused by feature sparsity.

Unlike the MAP adaptation approach applied in the GMM-UBM framework, we cannot construct an "impostor-normalized first-order Baum–Welch statistic," because it is equal to the mean vectors of each UBM component, which leads to an all 0 i-vector. Besides, even if we were to adopt this kind of impostor statistics and suppress the possible biased components, it would be hard to foresee whether the suppressed normalized first-order Baum–Welch statistics would be meaningful and reliable. Recall that in Fig. 2, if sufficient training data is provided, the distribution of normalized first-order Baum–Welch statistics is bounded in a limited range. In addition, total variability can be captured by a low-dimensional space (total factor matrix **T**). Inspired by these two properties, we can construct a *sufficient normalized first-order Baum–Welch statistics space*, which we term the *S-space* for convenience. For each utterance, the corresponding normalized first-order Baum–Welch statistics is firstly projected onto this space to obtain *referential normalized first-order Baum–Welch statistics*, which takes the place of the impostor-normalized first-order Baum–Welch statistics mentioned above.

Construction of this S-space requires the construction of a large set of long recordings to enable a sufficient number of zero- and first-order Baum–Welch statistics to be extracted for each speaker. For this purpose, the corpus of the 8conv condition from NIST 2005, 2006, and 2008 SREs are appropriate development corpus sets. Each single recording of the eight recordings derived from the same speaker contains about 2 min of speech. Thus, we can concatenate the eight recordings from the same speaker to compose a single long recording for each speaker, and it is assumed that concatenation would average out the channel effect among these eight recordings. We use **H** to denote the normalized sufficient first-order Baum–Welch matrix:

$$\mathbf{H} = \left(\left(\frac{\widetilde{\mathbf{F}}(1)}{N(1)} - \mathbf{m}\right), \left(\frac{\widetilde{\mathbf{F}}(2)}{N(2)} - \mathbf{m}\right), \ldots, \left(\frac{\widetilde{\mathbf{F}}(s)}{N(s)} - \mathbf{m}\right), \right) \tag{11}$$

where $\widetilde{\mathbf{F}}(s)$ denotes the sufficient first-order statistics for speaker s. $\widetilde{\mathbf{F}}(s)/N(s)$ is the concatenation of the normalized first-order Baum–Welch statistics for each Gaussian component, respectively:

$$\frac{\widetilde{\mathbf{F}}(s)}{N(s)} = \left(\frac{\widetilde{\mathbf{F}}_1(s)}{N_1(s)}^T, \frac{\widetilde{\mathbf{F}}_2(s)}{N_2(s)}^T, \ldots, \frac{\widetilde{\mathbf{F}}_C(s)}{N_C(s)}^T, \right)^T \tag{12}$$

In (11), **m** denotes the mean supervector of the speaker-independent model, usually derived from the mean supervector of UBM. The dimensionality of **H** is $CF \times R$, where C is the number of Gaussian components, F is the dimensionality of the feature vector, and R is the rank of **H**, which is equal to or less than the number of training speakers involved for condition 8conv. Then, singular value decomposition (SVD) is applied:

$$\mathbf{H} = \mathbf{F}^{eig}\mathbf{\Sigma}\mathbf{D}, \tag{13}$$

the eigencolumn space is denoted by $\mathbf{F}^{eig}$:

$$\mathbf{F}^{eig} = \mathbf{H}(\mathbf{\Sigma}\mathbf{D})^{-1}, \tag{14}$$

where $\mathbf{\Sigma}$ is the diagonal singular value matrix and **D** is the eigenrow space of **H**. In practice, $\mathbf{F}^{eig}$ usually corresponds to a subspace of $\mathbf{H}(\mathbf{\Sigma}\mathbf{D})^{-1}$ with highest eigenvalues ($Rank(\mathbf{F}^{eig}) \leq R$), in our experiments, we did not perform dimensionality reduction, i.e. $Rank(\mathbf{F}^{eig}) = R$. For an utterance u from a target speaker (either for enrollment or test purposes), suppose its first-order Baum–Welch statistics are denoted by $\mathbf{F}(u)$. Then, according to the least-squares criterion, the objective of the optimizing function is to minimize

$$\begin{aligned}
&\min_{\boldsymbol{\phi}(u)} \sum_{c=1}^{C} N_c(u) \left\| \frac{\mathbf{F}_c(u)}{N_c(u)} - \mathbf{m}_c - \mathbf{F}^{eig}\boldsymbol{\phi}(u) \right\|_2^2 \\
&= \min_{\boldsymbol{\phi}(u)} \sum_{c=1}^{C} N_c(u) \left(\frac{\mathbf{F}_c(u)}{N_c(u)} - \mathbf{m}_c - \mathbf{F}^{eig}\boldsymbol{\phi}(u) \right)^T \\
&\times \left(\frac{\mathbf{F}_c(u)}{N_c(u)} - \mathbf{m}_c - \mathbf{F}^{eig}\boldsymbol{\phi}(u) \right),
\end{aligned} \tag{15}$$

where $\boldsymbol{\phi}(u)$ is the linear coefficients of projection, whose dimensionality is equal to the rank of $\mathbf{F}^{eig}$, and $\mathbf{m}_c$ is the mean vector of the cth UBM component. Because (15) is a convex function, $\boldsymbol{\phi}(u)$ has a unique solution (see the proof of the propositions 1 in Bishop (2006) and section 3.1.1 of Kenny et al. (2005) for an explanation as to how to obtain $\boldsymbol{\phi}(u)$). Further, $\boldsymbol{\phi}(u)$ can be written as

$$\boldsymbol{\phi}(u) = \left[\sum_{c=1}^{C} N_c(u) (\mathbf{F}^{eig})^T \mathbf{F}^{eig} \right]^{-1} \sum_{c=1}^{C} \left(\mathbf{F}_c^{eig}\right)^T (\mathbf{F}_c(u) - N_c(u)\mathbf{m}_c), \tag{16}$$

where $\mathbf{F}_c^{eig}$ is the cth component of $\mathbf{F}^{eig}$.

Finally, the referential normalized first-order Baum–Welch statistics can be written as:

$$\frac{\mathbf{F}_c^{ref}(u)}{N_c(u)} = \mathbf{F}_c^{eig}\boldsymbol{\phi}(u) + \mathbf{m}_c, \quad c \in \{1, 2, \ldots, C\}. \tag{17}$$

Following the idea of adaptation of a GMM-UBM framework, the zero-order Baum–Welch statistics $N(u)$ control the extent to which we consider the real first-order Baum–Welch statistics to be accurate; hence, the adaptive normalized first-order Baum–Welch statistics can be written as:

$$\frac{\bar{\mathbf{F}}_c(u)}{N_c(u)} = \frac{q}{N_c(u) + q} * \frac{\mathbf{F}_c^{ref}(u)}{N_c(u)} + \frac{N_c(u)}{N_c(u) + q} * \frac{\mathbf{F}_c(u)}{N_c(u)}, \tag{18}$$

where q is a tuning factor, which needs to be manually tuned. As $q \to 0$ or as we increase the amount of training data, Eq. (18) becomes asymptotically identical to the real normalized first-order Baum–Welch statistics $\mathbf{F}_c(u)/N_c(u)$.

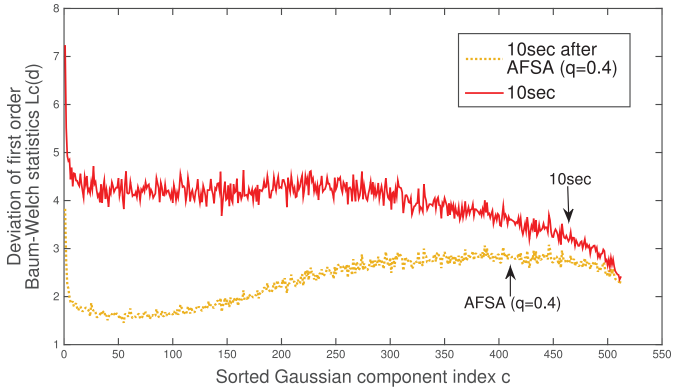

Fig. 8. Distribution of $l_c(d)$ in accordance with $N_c(d)$ as in Fig. 1. Both $l_c(d)$ were plotted using the same data of the 10sec condition as in Fig. 2. The upper plot is derived from the real first-order Baum–Welch statistics and the lower plot is derived from the AFSA adaptation.

The improved i-vector extraction formula can be written as:

$$\bar{\mathbf{w}}(u) = \left(\mathbf{I} + \mathbf{T}^t \boldsymbol{\Sigma}^{-1}\mathbf{N}(u)\mathbf{T}\right)^{-1} \sum_{c=1}^{C} \times \left[\mathbf{T}_c^t \boldsymbol{\Sigma}_c^{-1} N_c(u)\left(\frac{\bar{\mathbf{F}}_c(u)}{N_c(u)} - \mathbf{m}_c\right)\right], \qquad (19)$$

where $\mathbf{T}_c$ denotes the cth component of $\mathbf{T}$ and $\boldsymbol{\Sigma}_c$ is the cth diagonal block of $\boldsymbol{\Sigma}$ (the training of $\boldsymbol{\Sigma}$ is explained in Kenny, 2005).

Fig. 8 shows a comparison of real normalized first-order Baum–Welch statistics and adaptive normalized first-order Baum–Welch statistics. The results in the figure enabled us to conclude that AFSA efficiently compensates for the deviation caused by feature sparsity and imbalance, while retaining the statistical information for those components with relatively sufficient statistics.

5. Experiments

5.1. Database

All experiments were carried out on the core condition (short2-short3), short2-10sec, and 10sec-10sec of NIST 2008 SREs. Data from 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 (det8) for both training and testing.

5.2. Experimental setup

Our experiments operated on the Mel Frequency Cepstral Coefficients (MFCCs), with speech/silence segmentation performed according to the index of transcriptions provided by 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. The 20-dimensional feature vector was subjected to feature warping using a 3-s window (Pelecanos and Sridharan, 2001), and 20 first-order delta and 10 second-order delta were appended, equal to a total dimension of $F = 50$.

We used a male UBM containing 512 Gaussian components and the order of the total factor matrix $\mathbf{T}$ is 400. The corpora from the 2005 1conv4w and 2006 1conv4w transcription indices were used to train the UBM with a total length of 14 h speech from about 550 speakers. The corpora from the 2005 8conv4w, 2006 8conv4w, and 2008 8conv transcription indices, adding up to about 650 speakers, were used as the development datasets to train the total factor matrix, because the sessions for each speaker consist of recordings from 8 different microphones, which allows the speaker and channel variability to be modeled. For each speaker in the 2005 8conv4w, 2006 8conv4w, and 2008 8conv transcriptions, the 8 recordings were concatenated to construct a long recording containing about 15 min of speech[1] (as mentioned above, it is assumed that concatenation can average out the channel effects.) The artificially concatenated long recording set was used to construct a sufficient normalized first-order Baum–Welch statistics space $\mathbf{H}$. In our experiment, the development speaker sets that were used to train the total factor matrix and $\mathbf{H}$ space were identical. The dimensionality of $\mathbf{H}$ and $\mathbf{F}^{eig}$ were all set to 650, which were identical to the number of training speakers in the development sets. All i-vector sets (development, train and test) were extracted based on AFSA, as in (19).

Linear discriminant analysis (LDA), followed by cosine scoring (Dehak et al., 2011) and Gaussian probabilistic linear discriminant analysis (Gaussian-PLDA) (Kenny, 2010), were alternatively applied as compensation techniques. In our experiments, the optimal LDA dimension is 270, and the optimal PLDA configuration is 250 eigenvoice dimensions and 50 eigenchannel dimensions. All the decision scores were given without score normalization.

5.3. Results

Table 3 presents a comparison of the results for various trial conditions via different compensation and scoring techniques, in terms of the equal error rate (EER) and decision cost functions (DCF). The application of AFSA, starting from a relatively small tuning factor $q = 0.1$, leads to a gradual improvement in the performance of the trial conditions (with the exception of short2-short3). In the case of short2-short3, both enrollment and test utterances contain roughly 2 min speech, which is sufficient to avoid the feature sparsity problem in relation to the other two conditions.

For all trial conditions, the best performance is obtained when $q = 0.4$ (det8, Gaussian-PLDA), resulting in a 1.39% EER and 0.010 DCF for the short2-short3 condition, a 6.39%

[1] There are two optional approaches to fulfill this goal: Using `ffmpeg` toolkit to concatenate records directly, or using `IvExtractor` in ALIZE/LIA_RAL toolkit to obtain an accumulative Baum–Welch statistics from many recordings, the second approach is equivalent to record concatenation.

Table 3
Comparison of results from the trial conditions of short2-short3, short2-10sec, and 10sec-10sec from the NIST 2008 SRE data sets. LDA+cosine scoring and Gaussian-PLDA are applied as compensation and scoring techniques.

short2–short3	LDA+Cosine scoring				Gaussian-PLDA			
	All trials (det6)		English trials (det8)		All trials (det6)		English trials (det8)	
	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF
Baseline	4.71	0.025	1.80	0.014	3.27	0.018	1.41	**0.009**
AFSA, q=0.1	4.71	0.025	1.80	0.014	3.27	0.018	1.41	**0.009**
AFSA, q=0.4	**4.65**	**0.022**	**1.72**	**0.012**	**3.15**	**0.016**	**1.39**	0.010
AFSA, q=1.0	4.80	0.027	2.01	0.015	3.82	0.020	1.65	0.012
short2-10sec	LDA+Cosine scoring				Gaussian-PLDA			
	All trials (det6)		English trials (det8)		All trials (det6)		English trials (det8)	
	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF
Baseline	10.71	0.049	8.32	0.036	9.01	0.040	7.14	0.033
AFSA, q=0.1	10.44	0.048	8.09	0.034	8.82	0.039	7.02	0.032
AFSA, q=0.4	**9.96**	**0.043**	**7.28**	**0.035**	**8.30**	**0.037**	**6.39**	**0.030**
AFSA, q=1.0	10.79	0.050	8.31	0.036	9.11	0.041	7.10	0.033
10sec-10sec	LDA+Cosine scoring				Gaussian-PLDA			
	All trials (det6)		English trials (det8)		All trials (det6)		English trials (det8)	
	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF
Baseline	19.51	0.083	16.89	0.074	18.72	0.080	15.65	0.068
AFSA, q=0.1	17.52	0.076	15.67	0.069	16.84	0.071	14.12	0.061
AFSA, q=0.4	**16.39**	**0.069**	**14.28**	**0.061**	**16.02**	**0.067**	**13.37**	**0.058**
AFSA, q=1.0	17.98	0.078	16.14	0.070	17.11	0.072	15.20	0.066

EER and 0.030 DCF for the short2-10sec condition, and a 13.37% EER and 0.058 DCF for the 10sec-10sec condition.

A comparison of the best results with the corresponding baselines indicates that the best improvement is obtained for the 10sec-10sec condition, which shows more than 15% improvement compared with the baseline system. As AFSA aims to compensate for feature sparsity, it is most capable for the condition with limited training data.

Fig. 9 shows the DET curve comparison between the baseline and improved systems after applying AFSA. All results are based on the 10sec-10sec condition using the det6 male telephone dataset, LDA compensation, and cosine scoring.

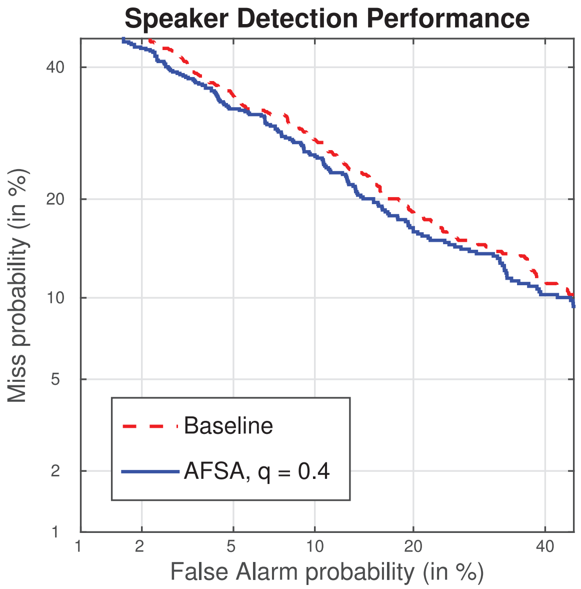

Fig. 9. Comparison of the DET curves between the baseline and improved systems after applying AFSA ($q = 0.4$) for the 10sec-10sec condition using the det6 dataset.

5.4. Conditioning of factor q to AFSA

As indicated by the results in Table 3, AFSA can improve the performance of the i-vector based system for all three trial conditions. However, if we continue to increase the value of conditioning factor q above its optimal value, the performance of all three conditions is reduced, because over-tuning of factor q may not only suppress the deviation of the normalized first-order Baum–Welch statistics, but also influence those Gaussian components with a sufficient number of speaker frames. We determined the effective tuning range of factor q by carrying out several experiments using different values of q for all three conditions. Figs. 10–12 compare the changes of EER(%) for different conditions with respect to q, with all results based on det6, LDA compensation, and cosine scoring. We notice that the effective range of factor q for the 10sec-10sec condition is wider than that of the other two conditions (short2-short3 and short2-10sec), which shows that AFSA is more effective under feature sparsity conditions. However, the verification performance is sensitive to changes in q, because factor q can suppress the deviation of

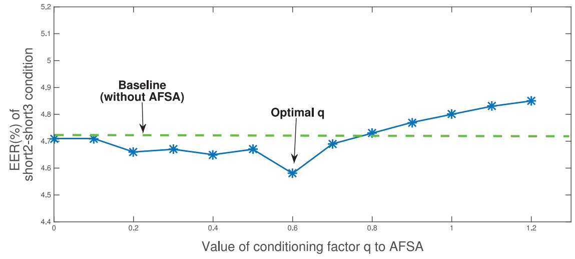

Fig. 10. EER(%) improvement of the short2-short3 (det6) condition, after conditioning by the AFSA factor q. The horizontal baseline is the EER(%) based on LDA compensation and cosine scoring, without AFSA.

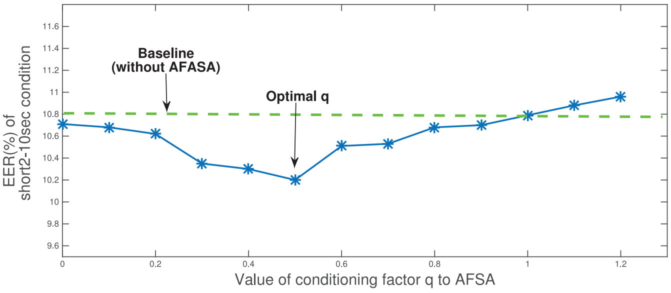

Fig. 11. EER(%) improvement of the short2-10sec (det6) condition, after conditioning by the AFSA factor q. The horizontal baseline is the EER(%) as in Fig. 10.

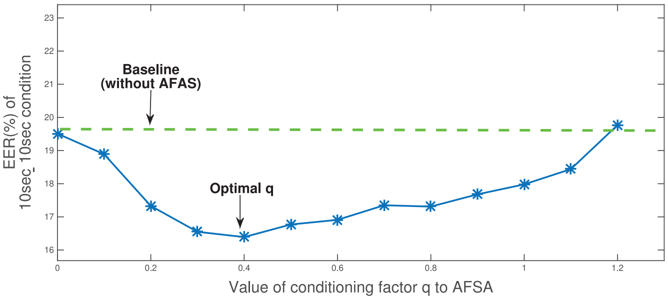

Fig. 12. EER(%) improvement of the 10sec-10sec (det6) condition, after conditioning by the AFSA factor q. The horizontal baseline is the EER(%) as in Fig. 10.

normalized first-order Baum–Welch statistics. However, overregulation of q may cause the loss of original information from speaker frames; thus, we have to reach a balance between over-fitting and under-fitting to sparse training data. Hence, real-world applications necessitate careful conditioning of q.

6. Conclusion

This paper presents an analysis of the intrinsic properties of feature sparsity and feature imbalance. Designated experiments were performed to show that these properties may lead to an over-fitting problem under the i-vector framework. We addressed this problem by proposing an adaptive first-order Baum–Welch statistics analysis (AFSA) algorithm to compensate for the deviation of normalized first-order Baum–Welch statistics caused by phonetic sparsity. Adopted normalized first-order Baum–Welch statistics may be seen as an interpolation between referential normalized first-order Baum–Welch statistics and real normalized first-order Baum–Welch statistics. This approach retains the most information of those Gaussian components with a sufficient number of speaker frames. These components also contribute the most when i-vectors are modeled.

The experimental results show that AFSA functions efficiently especially for the 10sec-10sec condition, for which an improvement of more than 15% was obtained, although this requires careful tuning of factor q and a development corpus set in which each speaker has a long recording (or at least a man-made long recording concatenated from many recordings). Nevertheless, AFSA still needs to be improved, because it shows less prominent improvements in short2-short3 and short2-10sec than in 10sec-10sec. As a comparison, related work has shown that an *adaptive relevance factor* technique of maximum a posteriori adaptation for GMM-SVM framework can produce prominent improvements in shor2-short3 and short2-10sec conditions, whose performance can be comparable or better than those obtained with i-vector based methods (You et al., 2013). Improving the robustness of factor q and methods for adapting it automatically from sparse training data as in You et al. (2012); 2013), is our next research goal.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61271349, 61371147 and 11433002.

References

Aronowitz, H., 2012. Text dependent speaker verification using a small development set. In: Proceedings of the Odyssey 2012-The Speaker and Language Recognition Workshop.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., Plchot, O., 2012. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In: Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Singapore, Singapore, pp. 157–164.

Bousquet, P.-M., Matrouf, D., Bonastre, J.-F., 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition.. In: Proceedings of the INTERSPEECH, pp. 485–488.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech Lang. Process. 19 (4), 788–798.

Hautamäki, V., Cheng, Y.-C., Rajan, P., Lee, C.-H., 2013. Minimax i-vector extractor for short duration speaker verification. In: Proceedings of the INTERSPEECH, pp. 3708–3712.

Kanagasundaram, A., Vogt, R.J., Dean, D.B., Sridharan, S., 2012. PLDA based speaker recognition on short utterances. In: Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2012). ISCA.

Kenny, P., 2005. Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. (Report) CRIM-06/08-13. CRIM, Montreal.

Kenny, P., 2010. Bayesian speaker verification with heavy-tailed priors.. In: Proceedings of the Odyssey, p. 14.

Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process. 13 (3), 345–354.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio, Speech, Lang. Process., 15 (4), 1435–1447.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of interspeaker variability in speaker verification. IEEE Trans. Audio, Speech, Lang. Process., 16 (5), 980–988.

Kenny, P., Stafylakis, T., Ouellet, P., Alam, M.J., Dumouchel, P., 2013. Plda for speaker verification with utterances of arbitrary duration. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE, pp. 7649–7653.

Larcher, A., Bousquet, P., Lee, K.A., Matrouf, D., Li, H., Bonastre, J.-F., 2012. I-vectors in the context of phonetically-constrained short utterances for speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012. IEEE, pp. 4773–4776.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2013. Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE, pp. 7673–7677.

Martin, A.F., Greenberg, C.S., 2009. NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels. In: Proceedings of the Tenth Annual Conference of the International Speech Communication Association.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. International Speech Communication Association (ISCA).

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Dig. Signal Process. 10 (1), 19–41.

Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., Dumouchel, P., 2013. Text-dependent speaker recognition using PLDA with uncertainty propagation. INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25–29, 2013, pp. 3684–3688, http://www.isca-speech.org/archive/interspeech_2013/i13_3684.html.

You, C., Li, H., Ma, B., Lee, K.-A., 2012. Effect of relevance factor of maximum a posteriori adaptation for GMM-SVM in speaker and language recognition.. In: Proceedings of the INTERSPEECH.

You, C.H., Li, H., Ma, B., Lee, K.A., 2013. A study on GMM-SVM with adaptive relevance factor and its comparison with i-vector and JFA for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE, pp. 7683–7687.