I-Vector/PLDA Variants for Text-Dependent Speaker Recognition

T. Stafylakis[1,2], P. Kenny[1], P. Ouellet[1], J. Perez[3], M. Kockmann[3] and P. Dumouchel[1,2]

[1] *Centre de Recherche Informatique de Montreal (CRIM), Canada,*
[2] *Ecole de Technologie Superieure (ETS), Canada,*
[3] *VoiceTrust, Germany*
themos.stafylakis@crim.ca

Abstract

The *i*-vector/PLDA approach currently dominates the field of text-independent speaker recognition and the question of how to translate this methodology to the text-dependent domain has recently become an active area of research. The essential difference between the two fields is that it is possible to do speaker recognition with enrollment and test utterances of very short duration in the text-dependent case but not in the text-independent case. The *i*-vector representation of short utterances turns out to be very sensitive to their phonetic content and this introduces a major source of nuisance variability when *i*-vectors are used in text-dependent speaker recognition. We show how, despite this complication, *i*-vector extractors can be successfully trained on short utterances (rather than on whole conversation sides as is usually done) and how this source of nuisance variability can be dealt with successfully in a PLDA classifier by making the PLDA model parameters phrase-dependent. Our results show that this phrase dependent version of PLDA is capable of outperforming the speaker-phrase version of PLDA presented in [8] on the RSR2015 dataset. We also give a detailed account of uncertainty propagation in PLDA and we show that it combines very successfully with phrase-dependent PLDA.

Keywords: Text-dependent speaker recognition, *i*-vectors, PLDA

1. Introduction

Recent research in text-dependent speaker recognition has focused primarily on the use of large scale statistical methods which were originally developed to model speaker and channel variability in text-independent speaker recognition. In [5; 6] and the papers cited there, the authors report results on the proprietary Wells Fargo dataset obtained with Nuisance Attribute Projection, Joint Factor Analysis, and *i*-vector/cosine distance methods. In [6; 7; 8], the authors developed a "speaker-phrase" version of PLDA for text-dependent speaker recognition which is designed to recognize a speaker-phrase combination rather than a speaker, using the publicly available RSR2015 dataset as a testbed.

A surprising aspect of the work reported in [6; 7; 8] is that a slight modification of the classical GMM/UBM approach (namely the HiLam model in [7]) gives results on the RSR2015 data which are hard to beat, despite the fact that no measures other than a robust front end were taken to protect against channel variability. This indicates that the session variability in RSR2015 is very benign compared with, say, the NIST datasets. On the other hand *i*-vector/PLDA methods are at a disadvantage here. Not only were *i*-vector/PLDA methods developed principally to counter serious channel effects, but *i*-vectors extracted from short utterances are very noisy in the sense that they are highly sensitive to the phonetic content of the utterances. This introduces a major source of nuisance variability which is not encountered in text-independent speaker recognition based on long utterances; moreover this type of variability does not appear to be amenable to subspace modeling (it is not low dimensional). Furthermore, the results in [6; 8] show that there are serious mismatches between *i*-vector extractors trained on NIST data and the RSR data. (A similar mismatch was found with the Wells Fargo data.) Thus getting *i*-vector/PLDA methods to work well on text dependent tasks such as the RSR data is a challenging problem.

In this article, we will show how several refinements of the *i*-vector/PLDA approach enable us to improve on the speaker-phrase PLDA model which was introduced [8] to adapt PLDA modeling from the text-independent domain to the text-dependent domain. In that paper, the performance of the speaker-phrase model was hampered by the use of an *i*-vector extractor trained on the NIST data. I-vector extractors are traditionally trained on whole sessions or conversation sides (typically of duration 2–3 minutes in the NIST data). The background and development portions of the RSR dataset do not contain nearly enough sessions to allow an *i*-vector extractor to be trained in this way. However we will see that an *i*-vector extractor can be successfully trained using short utterances (typically of duration 2–3 seconds) rather than recording sessions as units of speech.

Training an *i*-vector extractor on short utterances rather than whole conversation sides exacerbates the problem of phonetic variability in the resulting *i*-vectors. Given that a limited inventory of phrases was used in the RSR collection, we propose to deal with this problem in the *i*-vector/PLDA framework by introducing a *phrase-dependent* version of PLDA, in which the PLDA model parameters vary from one phrase to another but the speaker factors are tied across speakers. (A formally similar model was introduced in [9] for a different purpose.)

Furthermore, we explore the use of uncertainty propagation with phrase-dependent PLDA. This method was introduced in [10] in order to provide a way of dealing with utterances of widely disparate durations in text-independent speaker recognition. In that paper, space limitations prevented us from given a complete presentation of the method; we will remedy this here. Our experimental results show that generally speaking, if durations vary within a narrow range then uncertainty propagation brings little or no benefit to conventional PLDA in text-constrained speaker recognition as in text-independent speaker recognition. On the other hand, we will report some substantial improvements from uncertainty propagation in phrase-dependent PLDA.

We have also experimented with a standard PLDA model (i.e. phrase independent, without uncertainty propagation), trained with classes defined as combinations of speakers and phrases, rather than speakers alone (proposed in [8]). By training the PLDA model this way, the variations due to the phonetic content are absorbed into the between-class variability, letting the residual variable to capture only the channel variability. By noting that during run time, enrollment and test utterances are of the same phrase in each trial - hence target trials are subjected to channel variability only - the above way of labeling the training data is in line with the recognition protocol. We show that by training the phonetically constrained PLDA model in a gender-dependent way, the results are very competitive.

Although these methods have enabled us to achieve substantial improvements in performance on the RSR data, questions remain as to whether something more akin to the classical GMM/UBM approach (especially with score normalization) might not be more suited to text-dependent speaker recognition tasks than the *i*-vector/PLDA paradigm which currently holds sway in text-independent speaker recognition. We discuss this issue in the concluding section of the paper.

2. A PLDA model for text-dependent speaker recognition

In this section, we describe the phrase dependent version of PLDA which we used in experimenting with the RSR data. We assume that the phrase labels are given for all utterances in PLDA training, speaker enrollment and testing.

2.1. The i-vector space

As usual, our starting point is a universal background model (UBM), that is, a Gaussian Mixture Model (GMM) trained on a large background set in which speaker and channel variability are adequately represented. We denote the number of Gaussians in the UBM by C and the acoustic feature dimension by F so that GMM supervectors are of dimension $CF \times 1$. The *i*-vector representation originates from Joint-Factor Analysis (JFA, [1]); its role is to represent an utterance of arbitrary duration by a vector of fixed dimension which we denote by d. (Typically d is in the range 400–600.) An *i*-vector extractor is specified by a matrix $\boldsymbol{T}$ of dimension $CF \times d$.

The key idea is to assume that the (hidden) supervector $\boldsymbol{m}$ associated with an utterance can be described by the following generative model

$$\boldsymbol{m} = \boldsymbol{m}_{\text{UBM}} + \boldsymbol{T}\boldsymbol{i} \tag{1}$$

where $\boldsymbol{m}_{\text{UBM}}$ is the UBM supervector and $\boldsymbol{i}$ the i-vector whose prior distribution is assumed to be standard normal:

$$\boldsymbol{i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2}$$

This is in contrast to the independent normal priors over UBM components which is used in GMM/UBM systems based on relevance MAP. Given the model parameters $\boldsymbol{T}$ and $\boldsymbol{\Sigma}_{\text{UBM}}$ (i.e. a block diagonal covariance matrix, with blocks given by the covariance matrices of the UBM), and the zero and first order Baum-Welch statistics (denoted by $\boldsymbol{N}$ and $\boldsymbol{F}$) of an utterance r, the i-vector $\boldsymbol{i}_r$ is defined as follows

$$\boldsymbol{i} = \boldsymbol{B}^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}_{\text{UBM}}^{-1}\boldsymbol{F} \tag{3}$$

where

$$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}_{\text{UBM}}^{-1}\boldsymbol{N}\boldsymbol{T} \tag{4}$$

is the precision matrix of the i-vector. Thus, the i-vector is the posterior expectation of $\boldsymbol{i}$, given the sufficient statistics $\boldsymbol{N}$ and $\boldsymbol{F}$, [2].

2.2. The baseline PLDA model

Suppose that we have several recordings of a speaker with corresponding i-vectors $\{\boldsymbol{i}_r\}_{r=1}^R$. PLDA is a generative model that is described by the following equations

$$\boldsymbol{i}_r = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{\epsilon}_r \tag{5}$$

where

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{6}$$

is a vector of speaker factors, and the residual is distributed according to

$$\boldsymbol{\epsilon}_r \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}^{-1}) \tag{7}$$

where $\boldsymbol{D}$ is a precision matrix (i.e. an inverse covariance matrix).

This model was the predominant approach in the NIST 2012 text-independent speaker recognition evaluation campaign.

2.3. Phrase-dependent PLDA

We assume now that speakers are constrained to utter short phrases in a predefined set which we index by $l = 1, \ldots, L$. (This is the case in the RSR dataset where utterance durations are in the range 1–3 seconds [7].) Due to their short duration, the phonetic variability in such utterances is comparable or greater than the variability attributable to speaker and channel effects. This is in contrast to the NIST data, where the utterances are typically of 2–3 minutes duration so that phonetic variation is relatively minor. Thus it is natural to ask how to modify the baseline PLDA model so that it takes account of phrase labels.

The proposed phrase-dependent PLDA model is described by the following equation

$$\boldsymbol{i}_r = \boldsymbol{\mu}_l + \boldsymbol{V}_l \boldsymbol{y} + \boldsymbol{\epsilon}_r \tag{8}$$

where

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{9}$$

is a vector of speaker factors and the residual is distributed according to

$$\boldsymbol{\epsilon}_r | l \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}_l^{-1}) \tag{10}$$

where $\boldsymbol{D}_l$ is a precision matrix.

Thus, the phrase-dependent PLDA is defined by a set of L phrase dependent parameters $\{\mathcal{P}_l\}_{l=1}^L = \{\boldsymbol{\mu}_l, \boldsymbol{V}_l, \boldsymbol{D}_l\}_{l=1}^L$ which we denote by $\mathcal{P}$. Given a speaker factor vector $\boldsymbol{y}$ and phrase label l, $\boldsymbol{y}$ is mapped to a random vector in the i-vector space, through (i) a phrase-dependent affine transformation $(\boldsymbol{\mu}_l, \boldsymbol{V}_l)$ that defines its expected value, and (ii) a phrase-dependent full covariance matrix $\boldsymbol{D}_l^{-1}$. Thus, we are modeling the distribution of a single speaker as an L-component mixture PLDA model (having one component per phrase). Note that both covariance matrices (i.e. $\boldsymbol{V}_l\boldsymbol{V}_l^t$ and $\boldsymbol{D}_l^{-1}$) are phrase-dependent in our model. We have experimented on tying either $\boldsymbol{V}$ or $\boldsymbol{D}$, yet, the results showed degradation, especially when tying $\boldsymbol{V}$ (a model proposed in [9] for a different purpose).

2.4. Phrase-dependent i-vectors

An unusual aspect of the phrase-dependent PLDA model is that it supports the possibility of varying the i-vector feature space from one phrase to another. There is no reason in principle why we should not use different i-vector extractors for different phrases if that led to a more faithful representation of the data. Of course training different i-vector extractors for different phrases is not practical but adapting a common i-vector extractor to each phrase is feasible. Suppose that we have a trained an i-vector extractor in the usual way using utterances of all of the phrases. The assumption that i-vectors share a common standard normal prior as in Eq. (2), independent of the phrase identity, is obviously questionable. In a situation where phrase identities are known at training and run time, it is straightforward to estimate a non-standard normal prior for each phrase and to use this in place of the standard normal prior to extract i-vectors in a phrase dependent way. This type of i-vector extraction is only reasonable in the case of phrase-dependent PLDA. We have found that it results in modest improvements in performance although these tend to be masked when uncertainty propagation is applied to phrase-dependent PLDA models.

2.5. Uncertainty propagation

I-vectors extracted from utterances of short duration are less reliable than those extracted from utterances of longer duration and PLDA needs to take account of this if it is to handle utterances of disparate durations at run time. In [10], we outlined how to quantify the statistical noise in the i-vector extraction process that is attributable to duration variability and propagate this uncertainty into a PLDA classifier (similar work has been reported in [14; 15]). Here we will give a detailed treatment of this topic in the context of phrase dependent PLDA.

Recall that for a given utterance, extracting the corresponding i-vector consists in calculating the posterior expectation $\boldsymbol{i}$ and posterior precision matrix $\boldsymbol{B}$ of the hidden variables in the eigenvoice probability model described in [1]. The formulas for $\boldsymbol{i}$ and $\boldsymbol{B}$ are given by Eq. (3) and Eq. (4) and the posterior covariance matrix $\boldsymbol{B}^{-1}$ can be thought of as measuring the uncertainty associated with the point estimate $\boldsymbol{i}$. It is apparent from Eq. (4) that $\boldsymbol{B}$ depends on the utterance only through the zero order Baum-Welch statistics (that is, $\boldsymbol{N}$). In the case of a long utterance where all of the UBM components are visited frequently, the posterior covariance matrix will be close to zero; conversely, in the case of a short utterance, the posterior covariance will be almost as large as the prior covariance matrix (that is, the identity matrix).

Returning to our formulation of phrase dependent PLDA in Eq. (8), the simplest way of accommodating the posterior covariance matrix $\boldsymbol{B}_r^{-1}$ associated with the point estimate of the i-vector $\boldsymbol{i}_r$ is to write

$$\boldsymbol{i}_r = \boldsymbol{\mu}_l + \boldsymbol{V}_l \boldsymbol{y} + \boldsymbol{\eta}_r \tag{11}$$

where

$$\boldsymbol{\eta}_r | l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_l^{-1} + \boldsymbol{B}_r^{-1}). \tag{12}$$

The idea here is simply that the channel noise $\boldsymbol{\epsilon}_r$ and the statistical noise associated with the i-vector extraction process are statistically independent so that their covariance matrices can be added. This modification leaves the PLDA model formally unchanged, so that the calculations needed to perform speaker verification can be carried over directly. Note however that the computations with this version of PLDA will be slower than with the standard version because the fact that the covariance matrix of $\boldsymbol{\eta}_r$ varies from one recording to another means that the required matrix products cannot be precomputed.

This way of formulating uncertainty propagation is adequate for all purposes but one, namely estimating the covariance matrices $\boldsymbol{D}_l^{-1}$ during PLDA model training. It may happen in some application contexts that the utterances used for PLDA training (as distinct from the utterances encountered at run time) are all sufficiently long that the posterior covariance matrices $\boldsymbol{B}_r$ can ignored for the purposes of training the PLDA model. However, this will not be the case in general and estimating the covariance matrices $\boldsymbol{D}_l^{-1}$ in PLDA training turns out to be a tricky problem: it is clear the two types of noise (namely the statistical noise associated with i-vector extraction and the channel noise $\boldsymbol{\epsilon}$) *cannot* be conflated as in Eq. (12) if the covariance matrix of the channel noise (namely $\boldsymbol{D}_l^{-1}$) is to be estimated properly. For this we need a different formulation involving another hidden variable $\boldsymbol{x}_r$, namely

$$\boldsymbol{i}_r = \boldsymbol{\mu}_l + \boldsymbol{V}_l \boldsymbol{y} + \boldsymbol{U}_r \boldsymbol{x}_r + \boldsymbol{\epsilon}_r \tag{13}$$

where

$$\boldsymbol{x}_r \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{14}$$

and $\boldsymbol{U}_r$ is the lower-triangular Cholesky decomposition of $\boldsymbol{B}_r^{-1}$, so that $\boldsymbol{B}_r^{-1} = \boldsymbol{U}_r \boldsymbol{U}_r^t$). (That this is an equivalent formulation can be seen by taking $\boldsymbol{\eta}_r = \boldsymbol{U}_r \boldsymbol{x}_r + \boldsymbol{\epsilon}_r$.) So the idea is to use the channel factors in the original formulation of PLDA (see [21; 3]) to model the statistical noise in the i-vector extraction process. (PLDA with channel factors is no longer in general use in speaker recognition, because its performance is equivalent to the PLDA formulation with full $\boldsymbol{D}^{-1}$.)

3. Training and evaluating the model

In this section, we show how to train and evaluate the proposed model. The training algorithm is an extension of the standard EM algorithm, so that it takes into account of (i) the phrase labels of each utterance, and (ii) the uncertainty in the i-vector estimates. Note that the complexity of the algorithm is higher that in the case of standard PLDA model, because the uncertainty propagation does not allow standard acceleration tricks (matrix pre-computations) to be applied, [10].

Several approaches to uncertainty propagation have recently been proposed. In [14] and [15], uncertainty propagation is applied only to test utterances at verification time. (In the 2012 NIST speaker evaluation, enrollment and test utterances were sufficiently long that the uncertainty in estimating the corresponding i-vectors could be ignored.) This has the advantage that no change to the EM training algorithm is required but it is not adequate in situations where training utterances are of short duration. (In general, if the uncertainties associated with the point estimates of the i-vectors in the training corpus are ignored the effect will be to overestimate the residual covariance matrices $\boldsymbol{D}_l^{-1}$.)

Assume we have a training set of $\mathcal{D} = \{\mathcal{D}_s\}_{s=1}^{S}$, where $\mathcal{D}_s = \{\boldsymbol{i}_{s,r}, \boldsymbol{U}_{s,r}, l_{s,r}\}$, where $s = 1, \ldots, S$ and $r = 1, \ldots, R_s$ denote the speaker and i-vector index, respectively. Let $N_{s,l}$ denote the number of training i-vectors of the lth phrase that belongs to the sth speaker. Also, let $N_{\cdot,l} = \sum_s N_{s,l}$ and $N_{s,\cdot} = \sum_l N_{s,l}$. The mean parameters of the model $\{\boldsymbol{\mu}_l\}_{l=1}^{L}$ are directly calculated and subtracted from the corresponding i-vectors. The mean-normalized i-vectors are denoted by $\boldsymbol{f}_{s,r}$, i.e. $\boldsymbol{f}_{s,r} = \boldsymbol{i}_{s,r} - \boldsymbol{\mu}_{l_{s,r}}$.

3.1. Training: E-step

In the E-step, we condition on the current estimate of the model parameters and estimate the posterior of the hidden variables, i.e. of the speaker factors $\{\boldsymbol{y}_s\}$ and $\{\boldsymbol{x}_{s,r}\}$. The posterior is decomposed as follows

$$P(\boldsymbol{y}_s, \{\boldsymbol{x}_{s,r}\}_{r=1}^{R_s}|\mathcal{D}_s) = P(\{\boldsymbol{x}_{s,r}\}_{r=1}^{R_s}|\boldsymbol{y}_s, \mathcal{D}_s)P(\boldsymbol{y}_s|\mathcal{D}_s). \tag{15}$$

(In the terminology of [22], the first factor is the 'outer posterior' and the second factor is the 'inner posterior'.) We define and precalculate the following matrices

$$\boldsymbol{G}_{s,r}^{-1} = \boldsymbol{D}_{l_{s,r}}^{-1} + \boldsymbol{B}_{s,r}^{-1} \tag{16}$$

and

$$\boldsymbol{Q}_{s,r} = \boldsymbol{I} - \boldsymbol{U}_{s,r}^{t}\boldsymbol{G}_{s,r}\boldsymbol{U}_{s,r} \tag{17}$$

The matrix $\boldsymbol{G}_{s,r}$ is the summation of the uncertainty of the $\{s, r\}$ utterance and its corresponding covariance matrix $\boldsymbol{D}_{l_{s,r}}^{-1}$. Moreover, $\boldsymbol{Q}_{s,r}$ measures the *quality* of the utterance, in the sense that if the precision $\boldsymbol{B}_{s,r}$ is small, then $\boldsymbol{Q}_{s,r}$ approaches zero, while if $\boldsymbol{B}_{s,r}$ is large, $\boldsymbol{Q}_{s,r}$ approaches the identity matrix.

3.1.1. Speaker factors

Based on the decomposition defined in Eq. (15), the posterior of $\boldsymbol{y}_s$ is defined as follows

$$P(\boldsymbol{y}_s|\mathcal{D}_s) \propto P(\mathcal{D}_s|\boldsymbol{y}_s)P(\boldsymbol{y}_s) \tag{18}$$

where

$$P(\mathcal{D}_s|\boldsymbol{y}_s) = \prod_{r=1}^{R_s} \mathcal{N}(\boldsymbol{f}_{s,r}|\boldsymbol{V}_l\boldsymbol{y}_s, \boldsymbol{G}_{s,r}^{-1}) \tag{19}$$

Thus, $P(\boldsymbol{y}_s|\mathcal{D}_s)$ is again Gaussian distribution and its moments, namely its posterior expectation $\langle\boldsymbol{y}_s\rangle$ and precision $\boldsymbol{P}_s$ can be calculated using the following formulae

$$\boldsymbol{P}_s = \boldsymbol{I} + \sum_{r=1}^{R_s} \boldsymbol{V}_{l_{s,r}}^{t}\boldsymbol{G}_{s,r}\boldsymbol{V}_{l_{s,r}} \tag{20}$$

and

$$\langle \boldsymbol{y}_s \rangle = \boldsymbol{P}_s^{-1} \sum_{r=1}^{R_s} \boldsymbol{V}_{l_{s,r}}^t \boldsymbol{G}_{s,r} \boldsymbol{f}_{s,r} \tag{21}$$

Similarly, the second order expectation of the speaker factor vector is given by

$$\left\langle \boldsymbol{y}_s \boldsymbol{y}_s^t \right\rangle = \langle \boldsymbol{y}_s \rangle \left\langle \boldsymbol{y}_s^t \right\rangle + \boldsymbol{P}_s^{-1} \tag{22}$$

These equations are derived from the fact that the precisions are additive, and the posterior expectation $\langle \boldsymbol{y}_s \rangle$ is a matrix-weighted average, with weights being given by the precision matrices. Note also that the prior expectation of $\boldsymbol{y}_s$ is zero, which is why it has been omitted from Eq. (21). (The formulae of the PLDA algorithm without uncertainty propagation is a special case obtained by setting $\boldsymbol{U}_r$ to zero.)

3.1.2. Channel factors

We now show how to calculate the first and second order posterior moments for the “channel” factors. Note that the term “channel” factors is used only in order to highlight their similarity with the channel factors of original PLDA formulation in speaker recognition, [3]. We emphasize that the channel factors are only needed for estimating the precision matrices $\{\boldsymbol{D}_l\}_{l=1}^{L}$ during the M-step. In all other steps, including when enrolling and testing, the addition of the uncertainty to the residual covariance that is defined in Eq. (16) is sufficient.

From the decomposition of the overall posterior in Eq. (15) and the fact that $\boldsymbol{y}_s$ and $\boldsymbol{x}_{s,r}$ are a priori independent, we obtain

$$P(\boldsymbol{x}_{s,r} | \boldsymbol{y}_s, \boldsymbol{f}_{s,r}) \propto P(\boldsymbol{f}_{s,r} | \boldsymbol{x}_{s,r}, \boldsymbol{y}_s) P(\boldsymbol{x}_{s,r}) \tag{23}$$

where

$$P(\boldsymbol{f}_{s,r}|\boldsymbol{x}_{s,r},\boldsymbol{y}_s) = \mathcal{N}(\boldsymbol{f}_{s,r}|\boldsymbol{V}_{l_{s,r}}\boldsymbol{y}_s + \boldsymbol{U}_{s,r}\boldsymbol{x}_{s,r}, \boldsymbol{D}^{-1}_{l_{s,r}}) \tag{24}$$

Thus, the posterior of $\boldsymbol{x}_{s,r}$ is also Gaussian with moments defined as follows

$$\langle \boldsymbol{x}_{s,r} \rangle = \boldsymbol{K}^{-1}_{s,r}\boldsymbol{D}_{l_{s,r}}\boldsymbol{U}^t_{s,r}\left(\boldsymbol{f}_{s,r} - \boldsymbol{V}_{l_{s,r}}\langle \boldsymbol{y}_s \rangle\right) \tag{25}$$

where $\boldsymbol{K}_{s,r}$ is the posterior precision matrix given by

$$\boldsymbol{K}_{s,r} = \boldsymbol{I} + \boldsymbol{U}^t_{s,r}\boldsymbol{D}_{l_{s,r}}\boldsymbol{U}_{s,r}. \tag{26}$$

Finally, the second-order expectations are given by

$$\left\langle \boldsymbol{x}_{s,r}\boldsymbol{x}^t_{s,r} \right\rangle = \boldsymbol{K}^{-1}_r\boldsymbol{U}^t_{s,r}\boldsymbol{D}_{l_{s,r}}\boldsymbol{T}_{s,r}\boldsymbol{D}_{l_{s,r}}\boldsymbol{U}_{s,r}\boldsymbol{K}^{-1}_r + \boldsymbol{K}^{-1}_r \tag{27}$$

where

$$\boldsymbol{T}_{s,r} = \left\langle (\boldsymbol{f}_{s,r} - \boldsymbol{V}_{l_{s,r}}\boldsymbol{y}_s)(\boldsymbol{f}_{s,r} - \boldsymbol{V}_{l_{s,r}}\boldsymbol{y}_s)^t \right\rangle \tag{28}$$

which can be expressed as follows

$$\boldsymbol{T}_{s,r} = \boldsymbol{f}_{s,r}\boldsymbol{f}^t_{s,r} - \boldsymbol{V}_{l_{s,r}}\langle \boldsymbol{y}_s \rangle \boldsymbol{f}^t_{s,r} - \boldsymbol{f}_{s,r}\left\langle \boldsymbol{y}^t_s \right\rangle \boldsymbol{V}^t_{l_{s,r}} + \boldsymbol{V}_{l_{s,r}}\left\langle \boldsymbol{y}_s\boldsymbol{y}^t_s \right\rangle \boldsymbol{V}^t_{l_{s,r}} \tag{29}$$

3.2. Training: the M-step

In the M-step, the posterior expectations and precision matrices are used to update the point-estimates of the model parameters $\mathcal{P}$. The EM objective auxiliary function is given by

$$\boldsymbol{\Phi}_{\boldsymbol{D}_l,\boldsymbol{V}_l} = \sum_s \sum_r \left\langle \boldsymbol{\epsilon}^t_{s,r}\boldsymbol{D}_{l_{s,r}}\boldsymbol{\epsilon}_{s,r} \right\rangle \tag{30}$$

where $\boldsymbol{\epsilon}_{s,r} = \boldsymbol{f}_{s,r} - \boldsymbol{V}_{l_{s,r}}\boldsymbol{y}_s - \boldsymbol{U}_{s,r}\boldsymbol{x}_{s,r}$. By differentiating $\boldsymbol{\Phi}_{\boldsymbol{D}_l,\boldsymbol{V}_l}$ w.r.t. $\boldsymbol{V}_l$ and setting it equal to zero, we obtain the orthogonality relation

$$\sum_s \sum_r \left\langle \left(\boldsymbol{f}_{s,r} - \boldsymbol{V}_{l_{s,r}}\boldsymbol{y}_s - \boldsymbol{U}_{s,r}\boldsymbol{x}_{s,r}\right)\boldsymbol{y}^t_s \right\rangle = \boldsymbol{0} \tag{31}$$

which yields

$$\boldsymbol{V}_l = \boldsymbol{R}_{l,fy}\boldsymbol{R}_{l,yy}^{-1} \tag{32}$$

where

$$\boldsymbol{R}_{l,yy} = \sum_{s=1}^{S} N_{s,l}\left\langle \boldsymbol{y}_s\boldsymbol{y}_s^t\right\rangle \tag{33}$$

and

$$\boldsymbol{R}_{l,fy} = \sum_{s=1}^{S}\sum_{r:l_r=l} \boldsymbol{Q}_{s,r}\boldsymbol{f}_{s,r}\left\langle \boldsymbol{y}_s^t\right\rangle + (\boldsymbol{I}-\boldsymbol{Q}_{s,r})\boldsymbol{V}_{l_{s,r}}\left\langle \boldsymbol{y}_s\boldsymbol{y}_s^t\right\rangle . \tag{34}$$

Differentiating $\boldsymbol{\Phi}_{\boldsymbol{D}_l,\boldsymbol{V}_l}$ w.r.t. $\boldsymbol{D}_l^{-1}$ and using Eq. (32), we obtain

$$\boldsymbol{D}_l^{-1} = \frac{1}{N_{\cdot,l}}\sum_s\sum_r \left\langle (\boldsymbol{\phi}_{s,r} - \boldsymbol{V}_l\boldsymbol{y}_s)\boldsymbol{\phi}_{s,r}^t\right\rangle \tag{35}$$

where

$$\boldsymbol{\phi}_{s,r} = \boldsymbol{f}_{s,r} - \boldsymbol{U}_{s,r}\boldsymbol{x}_{s,r} \tag{36}$$

which can be expressed as follows

$$\boldsymbol{D}_l^{-1} = \frac{1}{N_{\cdot,l}}\sum_{s=1}^{S}\sum_{r:l_r=l} \left(\left\langle \boldsymbol{\phi}_{s,r}\boldsymbol{\phi}_{s,r}^t\right\rangle\right) - \boldsymbol{V}_l\boldsymbol{R}_{l,fy}^t \tag{37}$$

where

$$\left\langle \boldsymbol{\phi}_{s,r}\boldsymbol{\phi}_{s,r}^t\right\rangle = \boldsymbol{f}_{s,r}\boldsymbol{f}_{s,r}^t - \boldsymbol{f}_{s,r}\left\langle \boldsymbol{x}_{s,r}^t\right\rangle\boldsymbol{U}_{s,r}^t - \boldsymbol{U}_{s,r}\left\langle \boldsymbol{x}_{s,r}\right\rangle\boldsymbol{f}_{s,r}^t + \boldsymbol{U}_{s,r}\left\langle \boldsymbol{x}_{s,r}\boldsymbol{x}_{s,r}^t\right\rangle\boldsymbol{U}_{s,r}^t \tag{38}$$

3.3. Training: Minimum divergence step

As in the standard PLDA training algorithm, we can apply a Minimum divergence (MD) step, which enforces the condition that the speaker factors are distributed according to a standard normal prior.

This can be attained by transforming the current estimates of $\{\boldsymbol{V}_l\}_{l=1}^L$ in a way so that the empirical correlation matrix of the speaker factors

$$\boldsymbol{C}_{yy} = \frac{1}{S} \sum_{s=1}^{S} \langle \boldsymbol{y}_s \boldsymbol{y}_s^t \rangle \tag{39}$$

becomes equal to the identity matrix. To do so, the Cholesky decomposition of $\boldsymbol{C}_{yy}$ is calculated, so that we can write $\boldsymbol{C}_{yy} = \boldsymbol{H}^t \boldsymbol{H}$ and the model transformation is given by $\boldsymbol{V}_l \leftarrow \boldsymbol{V}_l \boldsymbol{H}^t$.

3.4. Training: Evaluating the evidence

The log-evidence of the model is guaranteed to increase from one training iteration to the next, so that it is very useful for monitoring convergence and debugging code. To derive an expression for the evidence, one may integrate-out the speaker and channel factors $\boldsymbol{\Theta} = (\boldsymbol{y}_s, \{\boldsymbol{x}_{s,r}\}_{r=1}^{R_s})_{s=1}^{s}$ or apply the Bayes formula

$$P(\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})}{P(\boldsymbol{\Theta}|\mathcal{D})} \tag{40}$$

for any convenient choice of $\boldsymbol{\Theta}$; we take $(\boldsymbol{y}_s, \boldsymbol{x}_{s,r}) = (\boldsymbol{0}, \boldsymbol{0})$. This yields

$$\mathcal{L}(\mathcal{D}) = \frac{1}{2} \sum_{l=1}^{L} \left(N_{\cdot,l} \log \left| \frac{1}{2\pi} \boldsymbol{D}_l \right| - \operatorname{tr}(\boldsymbol{D}_l \boldsymbol{G}_l) \right) - \alpha. \tag{41}$$

where

$$\boldsymbol{G}_l = \sum_{s=1}^{S} \sum_{r: l_{s,r}=l} \boldsymbol{f}_{s,r} \left(\boldsymbol{f}_{s,r} - \boldsymbol{U}_{s,r} \langle \boldsymbol{x}_{s,r} \rangle - \boldsymbol{V}_l \langle \boldsymbol{y}_s \rangle \right)^t \tag{42}$$

and

$$\alpha = \frac{1}{2} \sum_{s=1}^{S} \left(\log |\boldsymbol{P}_s| + \sum_{r=1}^{R_s} \log |\boldsymbol{K}_{s,r}| \right) \tag{43}$$

3.5. Calculating the Likelihood Ratio for Speaker Verification

To evaluate the model, we assume a set of enrollment utterances from the target speaker s, $\mathcal{D}_s = \{\boldsymbol{i}_{s,r}, \boldsymbol{U}_{s,r}, l_{s,r}\}_{r=1}^{R_s}$ and a test utterance $\{\boldsymbol{i}_t, \boldsymbol{U}_t, l_t\}$. (Note that we assume that l_t is given; if not it would be necessary to sum over the phrase label.) The log-likelihood ratio is formed as follows

$$\text{LLR}_{s,t} = \log \frac{P(\boldsymbol{i}_t, \{\boldsymbol{i}_{s,r}\}_{r=1}^{R_s})}{P(\boldsymbol{i}_t)P(\{\boldsymbol{i}_{s,r}\}_{r=1}^{R_s})} = \log \frac{P(\boldsymbol{i}_t | \{\boldsymbol{i}_{s,r}\}_{r=1}^{R_s})}{P(\boldsymbol{i}_t)} \tag{44}$$

where the probability density functions (pdf) are assumed to be conditioned on the model parameters $\mathcal{P}$ and on $\{\boldsymbol{U}_{s,r}, l_{s,r}\}_{r=1}^{R_s}$ and $\{\boldsymbol{U}_t, l_t\}$. To calculate the numerator (known as the predictive distribution) we first need to enroll the target speaker; this consists of estimating the first and second order statistics of $\boldsymbol{y}_s$ i.e. $(\langle \boldsymbol{y}_s \rangle, \boldsymbol{P}_s)$ given by Eq. (21) and Eq. (20). The numerator is a normal pdf evaluated at $\boldsymbol{i}_t$ with mean and covariance matrix equal to

$$\left(\hat{\boldsymbol{\mu}}_{l_t}, \hat{\boldsymbol{\Sigma}}_{l_t}\right) = \left(\boldsymbol{\mu}_{l_t} + \boldsymbol{V}_{l_t} \langle \boldsymbol{y}_s \rangle, \boldsymbol{V}_{l_t} \boldsymbol{P}_s^{-1} \boldsymbol{V}_{l_t}^t + \boldsymbol{D}_{l_t}^{-1} + \boldsymbol{U}_t \boldsymbol{U}_t^t\right). \tag{45}$$

As for the denominator, it is a normal pdf as well, evaluated at $\boldsymbol{i}_t$ with parameters

$$\left(\boldsymbol{\mu}_{l_t}^0, \boldsymbol{\Sigma}_{l_t}^0\right) = \left(\boldsymbol{\mu}_{l_t}, \boldsymbol{V}_{l_t} \boldsymbol{V}_{l_t}^t + \boldsymbol{D}_{l_t}^{-1} + \boldsymbol{U}_t \boldsymbol{U}_t^t\right). \tag{46}$$

The expression for the standard PLDA likelihood ratio can be recovered by setting the entries of all $\boldsymbol{U}$ matrices equal to zero.

Although in the experiments we assume enrollment and test utterances of the same phrase, note that the LLR expression can be used for arbitrary phrase labels for enrollment and test.

4. Experiments

4.1. The RSR2015 corpus

We evaluate the performance of the proposed methods using the RSR2015 speaker recognition dataset, [7]. This data was collected from 300 speakers, with each speaker participating in 9 recording sessions and a total of 6 types of smartphone were used to introduce channel diversity. (In the case of the NIST data, there are typically 10 sessions per speaker and channel diversity arises in a more natural way; thousands of speakers are available to model population variability.)

The RSR collection has been carefully designed to support many different types of experimentation in text-constrained speaker recognition although the channel variability in the RSR data is quite benign, so that a simple GMM/UBM benchmark implemented without score normalization or any channel modeling other than a robust front end is hard to beat.

We found that some early attempts to use UBMs and *i*-vector extractors trained on NIST data as a basis for PLDA modeling on the RSR data indicated a major dataset mismatch; our experience is consistent with the results reported in [16] and it led us to abandon this approach. In the experiments reported here, we used the RSR background and development sets (*bkg* and *dev*) for training and the evaluation (*eval*) set for testing.

As well as being divided by speaker population, the dataset is divided by lexical content into three parts. As in [7] we restrict ourselves principally to Part I which consists of repetitions of 30 short phrases taken from TIMIT. In each session, a speaker uttered each of the 30 phrases (as well as material from Parts II and III). This pattern was repeated for all of the background, development and evaluation speakers.

For our experiments on Part I, we used the evaluation test set proposed in [7] although we excluded the so-called “client different” trials since our concern is with recognizing speaker’s identities and not the lexical content of utterances. (The term “client-different” refers to a trial where a client speaker utters a phrase different from that which a speaker recognition system expects. Some means of detecting this type of event is needed to protect a text-dependent system from spoofing attacks using pre-recorded speech.)

In the evaluation test set for Part I, three repetitions of a given phrase are supplied at enrollment time and one repetition of the same phrase is supplied at test time. Different trials involve different phrases so the problem is harder than text speaker recognition with a universal, fixed pass phrase (such as the digit string $0 \ldots 9$ which was studied in [5]). On the other hand, the three utterances used to enroll a speaker came from three different sessions which makes the speaker recognition problem easier than if all of the enrollment utterances were recorded in a single session as some channel robustness is built into the enrollment procedure. (Naturally, the test and enrollment utterances in a trial were recorded in different sessions.) When “client-different” trials are excluded, the number of trials in the Part I evaluation test set is 968791 (18253 target and 950538 non-target trials, respectively). As usual, cross-gender trials are excluded.

4.2. UBM, i-vector and PLDA configurations

For the front end, we low-pass filtered speech to 0–4 KHz, we used an energy-based voice-activity detector and 60-dimensional features (mel frequency cepstral coefficients and their first and second derivatives) with short-term mean and variance normalization.

For all of our *i*-vector/PLDA experiments, we used a gender-independent universal background model with 1024 diagonal covariance Gaussians using all of the *bkg* data (Parts I, II and III) and a 400 dimensional *i*-vector extractor using all of the *bkg* and *dev* data. We scaled the zero and first order Baum-Welch statistic by a factor of 1/3 to compensate for inter-frame correlations, [13].

Although it is traditional to use the recording session (that is, a whole conversation side) as the speech unit in training *i*-vector extractors for work on the NIST datasets, the number of sessions available to us in the *bkg* and *dev* subsets of the RSR data is less than 2K which is not sufficient to train an *i*-vector extractor. Accordingly, we used the phrase as the speech unit rather than the session. This has the advantage of providing lots of training data (there are more than 130 K utterances in the *bkg* and *dev* sets) but it has the undesirable effect that *i*-vectors extracted from utterances are highly sensitive to their lexical content. As reported in [7], [8] *i*-vectors extracted from short utterances manifest this type of behavior even if whole sessions are used to train an *i*-vector extractor but the problem is exacerbated if the *i*-vector extractor is trained on short utterances. Finding a way to deal with this type of nuisance variability was one of our primary motivations for introducing phrase-dependent PLDA.

Another problem in training an *i*-vector extractor on the RSR *bkg* and *dev* sets is that the speaker population is not sufficiently diverse (there are fewer than 200 speakers in the *bkg* and *dev* sets). As a result, *i*-vector features do not represent speaker variability adequately. This type of problem has been encountered by the NIST speaker recognition community even though the datasets used for research in text-dependent speaker recognition contain several thousand speakers. (The Fisher corpora have generally not proved to be useful in text-independent speaker recognition research but it is well known that including Fisher data in training i-vector extractors is helpful because it exposes the training algorithm to more speaker variability.) Insufficient training data may prove to be a serious impediment to the development of i-vector methods on text constrained tasks such as RSR. We will return to this question in the discussion in Section 4.

As for the PLDA models, we took them to be of full rank (so that the hidden variables $\boldsymbol{x}_{s,r}$ and $\boldsymbol{y}_s$ have the same dimension as the i-vectors, namely 400), since the experiments we did with $\boldsymbol{y}_s$ dimension equal to 300, 200 and 120, respectively, showed increasing degradation in performance.

4.3. Length normalization in uncertainty propagation

It is well known that length normalization of i-vectors reinforces the Gaussian assumptions in PLDA modeling. However it is not obvious how to apply "length normalization" to posterior covariance matrices before propagating them into PLDA. We experimented with this in [10] but failed to resolve the problem satisfactorily in that paper.

In [10], we took the posterior covariance matrices in Eq. (4) to be full, and reduced them to manageable size by an LDA projection. This alleviates the problem of working with 130 K posterior covariance matrices in PLDA training but it is not a satisfactory solution, particularly as LDA projections are no longer in general use in standard i-vector/PLDA modeling with length normalization.

For our work in this paper, we suppressed the off-diagonal terms in the formula for the posterior precision matrix Eq. (4) so that the matrix $\boldsymbol{U}_r$ could be treated as diagonal (and so easily stored on disk). We implemented the length normalization transformation as $\boldsymbol{i} \leftarrow \sqrt{d}\frac{\boldsymbol{i}}{\|\boldsymbol{i}\|}$, where d is the i-vector dimensionality so as to preserve the distribution of the i-vector population (as measured by first and second order statistics), [12]. Note that with this implementation, the matrices $\boldsymbol{U}_r$ can be imported into PLDA without any other modification. This is due to the fact that by multiplying by $\sqrt{d}$, the scaling of the original i-vector space is preserved. (Equivalently, one could leave the i-vectors unchanged and divide all the matrices U_r by $\sqrt{d}$.)

Although we do not report results here, we also experimented with transforming the posterior covariance matrices using the Jacobian of the length normalization transformation, or more simply its scalar part, as proposed in [14]. This is perhaps the most natural solution to the problem of "length normalizing" covariance matrices, but our experience was that it degraded the performance by 20% approximately, in terms of Equal Error Rate (EER).

4.4. Experimental results

4.4.1. GMM/UBM benchmarks

We implemented a standard GMM/UBM approach using two 512 component UBMs trained on two different datasets: the RSR *bkg* set and the CSLU multilingual telephone corpus.[1]. Table 1 reports results on the Part I evaluation test set for various error metrics, namely the Equal Error Rate (EER) and "old" and "new" minimum normalized Detection Cost Functions,

[1] LDC distribution 2006S35

DCF$_o$ and DCF$_n$, as defined in the NIST SRE evaluation plans for 2008 and 2010. No score normalization was applied.

Table 1: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, using the UBM/GMM approach (gender independent UBM, trained on bkg+dev sets). Without score Normalization (line 1) with t-Norm (line 2).*

	female			male		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
1	1.53	0.88	3.32	2.17	1.25	4.66
2	**0.79**	**0.43**	**1.71**	**1.07**	**0.54**	**2.06**

The error rates attained by the GMM/UBM approach with t-Norm are very low and constitute a baseline performance that is hard to improve on. Similar results are reported in [7] using a modification of the GMM/UBM approach which enables left-to-right alignment constraints to be imposed so that the lexical content of utterances can be verified at the same time as speakers' identities. As in [7], we made no attempt to deal with channel variability other than using a robust front end (short term mean and variance normalization in our case).

4.4.2. Speaker-phrase PLDA

Later in the paper we will present instances where uncertainty propagation leads to substantial performance gains with phrase dependent PLDA. So it is important to be clear that our experience with uncertainty propagation in text-constrained speaker recognition has not been an unqualified success.

Table 4.4.2 shows results obtained with the variant of PLDA introduced in [8] where the speaker factor vector $\boldsymbol{y}_s$ is used to characterize a speaker-phrase combination (rather than a speaker). This model is useful in cases where enrollment and test utterance are of the same phrase (as in the Part I test set). These results appear to be better than those reported in [8] which confirms that our unorthodox approach to *i*-vector extractor training is reasonable. (In that paper, the *i*-vector extractor was trained on NIST data in the usual way rather than on RSR data. Note however that in [8] the phrases used in testing are disjoint from those used in training the PLDA model so this is not a fair comparison.)

Table 2: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, using speaker-phrase tying of speaker factors. Line 1: speaker-phrase PLDA. Line 2: speaker-phrase PLDA, i-vector extractor trained on bkg, dev & eval (cheating experiment).*

	female			male		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
1	2.45	1.11	3.44	1.28	0.65	2.51
2	1.94	0.88	2.60	1.06	0.55	2.12

For the speaker-phrase PLDA model, we should note that we have tried uncertainty propagation, however the results showed some degradation in performance (2.53% and 1.48% EER for female and male, respectively, with *i*-vector extractor trained on bkg & dev). Broadly speaking, our experience with uncertainty propagation and conventional PLDA models (as distinct from the phrase-dependent PLDA which is the principal focus of this article) is that it is beneficial in cases where enrollment and test utterances are of widely disparate durations. This was the case in [19] where we devised a speaker verification test set using utterances whose durations varied ran-

domly in the range 3–60 sec and found that uncertainty propagation gave substantial improvements in performance. We have observed similar behavior in applying PLDA to speaker diarization, where the duration of speaker turns is naturally quite variable, [20]. On the other hand, if utterance durations are controlled (as is the case for NIST SRE test sets prior to 2012 and the RSR test sets), then uncertainty propagation is unlikely to be helpful except insofar as it tends to produce reasonably well calibrated scores.
The number of speakers in the combined *bkg* and *dev* sets is only 194, which is not sufficient to train an *i*-vector extractor that adequately models inter-speaker variability. We were interested to know what would be the best performance we could get in a situation where we did not have to contend with this limitation. The results in Table , line 2 are substantially improved, showing the potential of *i*-vector based modeling when enough data to train robustly the *i*-vector extractor is available.

4.4.3. Phrase-dependent PLDA

We now turn to phrase dependent PLDA. Table 3 reports results obtained on the Part I evaluation test set using a standard, gender-independent phrase-independent PLDA model (line 1), a phrase-dependent PLDA model (line 2) and a phrase-dependent PLDA model with uncertainty propagation (line 3). Comparing with Table 3 shows that the standard PLDA gives inferior results to the GMM/UBM approach. (This is consistent with the results presented in [7], [8].) On the other hand, phrase-dependent PLDA without uncertainty propagation behaves comparably and phrase-dependent PLDA with uncertainty propagation gives much better results.

Since our *i*-vector extractor was trained with short utterances rather than

Table 3: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, using several i-vector/PLDA approaches. Line 1: phrase-independent model, no uncertainty propagation (UP). Line 2: phrase-dependent model, no UP. Line 3: phrase-dependent model, with UP. Line 4: phrase-dependent model, with UP, i-vector extractor trained on bkg+dev+eval (cheating experiment).*

	female			male		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
1	2.83	1.58	5.22	2.21	1.21	5.17
2	2.78	1.23	4.01	1.78	0.94	3.36
3	**1.64**	**0.78**	**2.81**	**1.06**	**0.61**	**2.77**
4	1.23	0.58	2.12	0.83	0.49	2.38

whole sessions, it produces *i*-vectors which are highly sensitive to the phonetic content of utterances. It is not surprising to see that phrase-dependent PLDA outperforms phrase-independent PLDA under such conditions. However, the dramatic improvement which appears to be attributable solely to uncertainty propagation is surprising. A plausible explanation for this behavior is that when phrase dependent PLDA parameters are introduced, the residual covariance matrices $\boldsymbol{D}_l^{-1}$ are inadequately estimated owing to insufficient training data but this is compensated for adding the covariance matrices $\boldsymbol{B}_r^{-1}$ supplied by the *i*-vector extractor.

Table 3, line 4 reports the results of a cheating experiment where the *i*-vector extractor is trained on RSR *bkg*, *dev* and *eval* sets. The results of the phrase-dependent PLDA model with uncertainty propagation show a substantial gain in performance even though the lack of "serious" channel variability in the RSR data does not favor the i-vector/PLDA approach. (Repeating this cheating experiment for the benchmark GMM/UBM system — that is, training the UBM on all of the RSR data rather than just *bkg* — leads to a very minor gains, as one would expect.)

These results are consistent with those reported in [16], where it was shown that i-vector methods perform very well on text constrained speaker recognition tasks provided that large amounts of training data (500+ training speakers) are available but i-vector based methods do not outperform simpler approaches if the amounts of training data available are limited.

DET curves for several of the approached discussed are depicted in Fig. 1 and Fig. 2 for female and male speakers, respectively, and show the same tendency described above. The GMM/UBM with t-norm is better by a large margin on the female speakers, while the proposed model performs very well on the male speakers. The significant improvement in performance when the *eval* set is used in training the i-vector extractor is clearly depicted for both PLDA models.

4.4.4. Other experiments

We mentioned in Section 2 that phrase dependent PLDA could support phrase-dependent i-vector extraction. A comparison between standard i-vectors and phrase-dependent i-vectors with the phrase-dependent PLDA model described in the previous section showed a minor gain in performance (3% relative improvement in terms of EER). However, phrase-dependent i-vector gave a more significant improvement in the following experiment, where we attempted to see if it would be possible to devise a text-constrained speaker recognition system in which disjoint phrase sets are used for enroll-

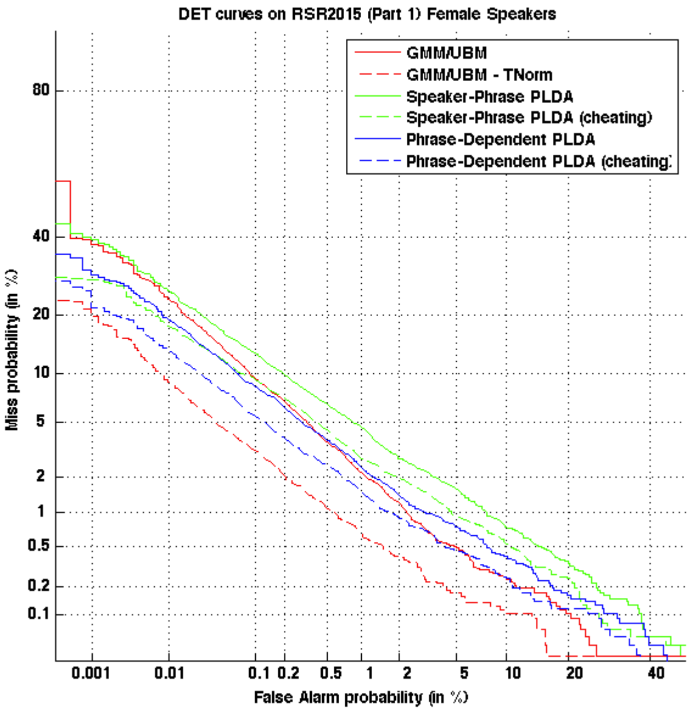

Figure 1: *DET curves on female speakers.*

ment and testing. The idea would be to randomize the selection of enrollment and test phrases as much as a limited phrase inventory (such as the 30 TIMIT phrase in RSR Part I) would allow, thereby simulating a text-independent speaker recognition system based on arbitrary, random prompts. If this idea could be made to work it could serve as a basis for designing a system which would be less vulnerable to spoofing attacks (e.g. replay attacks or voice conversion) than a text-dependent system.

For this experiment, we designated twenty of the Part I phrases as potential enrollment phrases and ten as potential test phrases. For each trial, seven enrollment phrases and one test phrase were chosen at random. Contrary to the standard Part I test set, all enrollment phrases were recorded in

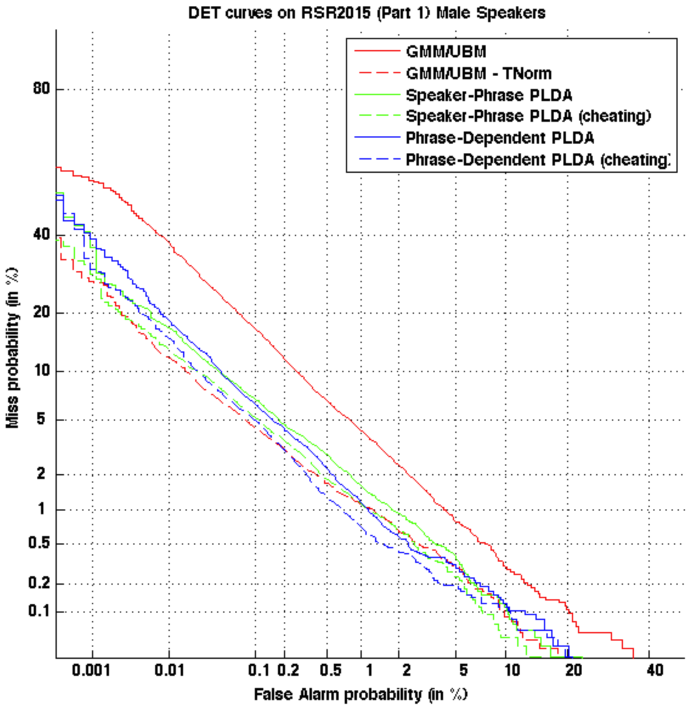

Figure 2: *DET curves on male speakers.*

the same session. The number of target and non-target trials are 44970 and 2203530 (female speakers only).

Results are reported in Table 4. We tried using phrase-independent PLDA (line 1), phrase-dependent PLDA (line 2) and phrase-dependent PLDA with phrase-dependent i-vectors (line 3), with and without uncertainty propagation in each case. The results show that this is a very difficult task and that the simplest strategy (phrase-independent PLDA without uncertainty propagation) works best. In this case, phrase-dependent i-vectors leads to uniformly better results than phrase-independent i-vectors (line 3 vs. line 2) although the differences are not great. Once again we observe that uncertainty propagation leads to dramatic improvements in the performance of phrase-dependent PLDA (EERs are halved) but even so phrase-independent PLDA performs less well than phrase-independent PLDA.

Table 4: *EER (%), normalized minDCF$_{old}$ and normalized minDCF$_{new}$ on RSR part I, for the case of disjoint enrollment and test utterance. Line 1: phrase-independent model. Line 2: phrase-dependent model. Line 3: phrase-dependent model with phrase-dependent i-vectors.*

	no UP			with UP		
	EER	DCF$_o$	DCF$_n$	EER	DCF$_o$	DCF$_n$
1	**7.75**	**4.15**	**9.24**	**7.66**	**4.12**	**9.25**
2	18.19	8.00	9.79	9.86	5.18	9.61
3	16.76	7.85	9.79	8.47	4.83	9.61

5. Conclusion

The performance of i-vector/PLDA methods on text-independent speaker recognition tasks is currently unrivaled, but our experience and that of other authors suggests that simpler methods which are much closer to a classical GMM/UBM approach may be a better fit for text-dependent speaker recognition tasks.

Although, a back-to-back comparison of the results in [8] (speaker-phrase PLDA with a UBM and i-vector extractor trained on NIST data) with those in [7] (the Hierarchical multi-Layer acoustic model, HiLam) is not possible, the HiLam system seems to perform better. HiLam is a modification of the traditional GMM/UBM approach in which each target speaker is represented by a hidden Markov model (HMM) rather than a GMM. The idea is to first adapt a conventional UBM to the speaker at enrollment time in the usual way using relevance MAP; the Gaussians in the adapted UBM are then strung together to form a 5-state HMM. The advantage of using a HMM rather than a GMM seems to be that it offers a natural way of protecting against spoofing attacks. If a fraudster has a recording of a target speaker's voice uttering a phrase other than the expected text, then a HMM-based system may be able to detect this where a GMM-based system would be more vulnerable. (A speech recognition system or some other means of protecting against this eventuality would need to be provided.) There does not seem to be any evidence that the HMM is better equipped to handle non-target trials as this term is usually understood (where a speaker other than the target speaker utters the phrase expected by the system) so that the good performance of the method appears to be attributable to the effectiveness of relevance MAP in text-dependent speaker recognition.

The speaker-phrase PLDA system presented in [8] is hampered by the mismatch between the RSR and NIST data. In this paper we have shown how to train a good *i*-vector extractor using only the background and development portions of the RSR datasets. Comparing Tables 1 and 4.4.2 shows that, with this *i*-vector extractor the results for the speaker-phrase PLDA system (without uncertainty propagation) and the GMM/UBM system (with the UBM trained on the RSR background data) are similar (although there is an inconsistency across genders). We found that substantially improved results could be obtained by using phrase-dependent PLDA with uncertainty propagation (Table 3), although uncertainty propagation does not yield consistent gains either with conventional PLDA or speaker-phrase PLDA. Thus there is a question as to whether the real benefit from uncertainty propagation in phrase-dependent PLDA is that it is masking an undertraining problem (as discussed in Section 4) and we cannot claim that our results convincingly dislodge the GMM/UBM paradigm.

Of course, if the results of the cheating experiments could be believed we would have strong evidence in favor of such a claim. Recall that we found that, if the evaluation data were included in the dataset used to train the i-vector extractor (but not the PLDA model) then performance improves dramatically. It is clear that for i-vector based methods to work, speaker variability needs to be adequately represented in the i-vector training set. This is confirmed by the results in [16] on the Wells Fargo digit data: it is only by partitioning the data set in such a way that the majority of the speakers (550 out of 750) are used for i-vector training that i-vector methods can beat simpler methods such as GMM-NAP (which require modest amounts of training data) [17], [18]. The problem seems to be a fundamental one which raises doubts about the suitability of i-vector modeling for text-dependent tasks: it seems to be practically impossible to collect sufficiently large volumes of data for training an i-vector extractor if speakers are restricted to a small vocabulary such as digit strings.

It is perhaps not surprising that GMM/UBM-based methods (with or without channel compensation techniques such as NAP) work well on text-dependent speaker recognition tasks. Contrary to the subspace priors used in JFA and i-vector modeling, relevance MAP is based on a factorial prior which treats the UBM components as being statistically independent. Thus, in the case of speaker recognition trials in which enrollment and test utterances are perfectly matched with respect to phonetic content, the mixture components which are adapted in enrollment are the same as those which are activated in testing. So relevance MAP is sensitive to the phonetic content of an enrollment utterance as well as speaker characteristics. This seems to be desirable, considering that the task in text-dependent speaker recognition is to recognize a speaker-phrase combination rather than a speaker as such.

The *i*-vector/PLDA approach arose by turning JFA into a feature extractor. This seem to be very well suited to text-independent speaker recognition tasks in which sufficiently large volumes of data are available to model speaker variation adequately and phonetic detail is largely irrelevant but there is a mounting body of evidence that the *i*-vector/PLDA approach may not be ideally suited to text-dependent speaker recognition. This raises the question of whether there might be other ways of using JFA as a feature extractor which are better suited to text-dependent speaker recognition. One approach would be to start from a JFA model that contains diagonal factors and channel factors but no speaker factors. Point estimates of the diagonal factors could play an analogous role to *i*-vectors (but rather than being low dimensional, these feature vectors would be of the same dimension as supervectors). For this type of JFA model, a training set would merely need to be rich enough to reflect channel variability adequately (but not necessarily speaker variability) and a diagonal version of PLDA could serve as a classifier for making speaker verification decisions. We intend to explore this idea in future work.

References

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification", in *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", in *IEEE Transactions on Audio, Speech & Language Processing*, 2011.

[3] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", in *Proc. Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. "Speaker verification using adapted Gaussian mixture models", in *Digital signal processing*, 2000.

[5] H. Aronowitz, "Text-Dependent Speaker Verification Using a Small Development Set", in *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Singapore, June 2012.

[6] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Y. Li, J.-F. Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification", in *Proceedings of ICASSP*, Kyoto, Japan, March 2012.

[7] A. Larcher, K.-A. Lee, B. Ma and H. Li, "The RSR2015: database for text-dependent speaker verification using multiple pass-phrases", in *Proceedings of Interspeech*, Portland (Oregon), USA, Sept. 2012.

[8] A. Larcher, K.-A. Lee, B. Ma and H. Li, ”Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[9] J. Villalba, E. Lleida, A. Ortega and A. Miguel, ”I3A SRE12 System Description”, on *SRE12 Speaker Recognition Workshop*, Oct. 2012.

[10] P. Kenny, T. Stafylakis, P. Ouellet, M.J. Alam and P. Dumouchel, ”PLDA for speaker verification with utterances of arbitrary duration”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[11] B. J. Borgstrom and A. McCree, ”Supervector Bayesian speaker comparison”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[12] D. Garcia-Romero and C. Y. Espy-Wilso, ”Analysis of *i*-vector length normalization in speaker recognition systems”, in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[13] T. Stafylakis, P. Kenny, V. Gupta and P. Dumouchel, ”Compensation for inter-frame correlations in speaker diarization and recognition”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[14] S. Cumani, O. Plchot and P. Laface, ”Probabilistic linear discriminant analysis of *i*-vector posterior distributions”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[15] B. J. Borgstrom and A. McCree, ”Supervector Bayesian Speaker Comparison”, in *Proceedings of ICASSP*, Vancouver, Canada, May 2013.

[16] H. Aronowitz and O. Barkan, ”On Leveraging Conversational Data for Building a Text Dependent Speaker Verification System”, in *Proceedings of Interspeech*, Lyon, France, August 2013.

[17] W. Campbell, Z. Karam, ”Simple and Efficient Speaker Comparison using Approximate KL Divergence”, in Proc. Interspeech, 2010.

[18] W. M. Campbell, et al., ”SVM based Speaker Verification using GMM Supervector Kernel and NAP Variability Compensation”, in Proc. ICASSP, 2006.

[19] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann and P. Dumouchel, ”Text-dependent speaker recognition using PLDA with uncertainty propagation”, in *Proc. of Interspeech*, Lyon, France, August 2013.

[20] T. Stafylakis, G. Dupuy, P. Kenny and P. Dumouchel, ”Speaker diarization using PLDA with uncertainty propagation”, (in preparation).

[21] S. J. D. Prince and J. H. Elder, ”Probabilistic linear discriminant analysis for inferences about identity,” in Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, Oct. 2007.

[22] Niko Brummer, ”EM for Probabilistic LDA”, Technical Report, Feb. 2010.