Explicit Modelling of Session Variability for Speaker Verification

Robbie Vogt [1], Sridha Sridharan [1]

Speech and Audio Research Laboratory
Queensland University of Technology
2 George Street, Brisbane, Australia

Abstract

This article describes a general and powerful approach to modelling mismatch in speaker recognition by including an explicit session term in the Gaussian mixture speaker modelling framework. Under this approach, the Gaussian mixture model (GMM) that best represents the observations of a particular recording is the combination of the true speaker model with an additional session-dependent offset constrained to lie in a low-dimensional subspace representing session variability.

A novel and efficient model training procedure is proposed in this work to perform the simultaneous optimisation of the speaker model and session variables required for speaker training. Using a similar iterative approach to the Gauss-Seidel method for solving linear systems, this procedure greatly reduces the memory and computational resources required by a direct solution.

Extensive experimentation demonstrates that the explicit session modelling provides up to a 68% reduction in detection cost over a standard GMM-based system and significant improvements over a system utilising feature mapping, and is shown to be effective on the corpora of recent National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations, exhibiting different session mismatch conditions.

Key words: automatic speaker verification, session variability

Email addresses: r.vogt@qut.edu.au (Robbie Vogt), s.sridharan@qut.edu.au (Sridha Sridharan).

[1] This work was supported by the Australian Research Council Grant No. DP0557387.

1 Introduction

The issue of telephone handset mismatch has been well documented in the automatic speaker verification (ASV) literature as the greatest source of verification errors. This appraisal is, however, a somewhat naïve description of the problem as mismatch is not restricted to differences in handset type as there are a myriad of possible causes of mismatch. Other examples of mismatch in a telephony environment include a number of environmental factors such as nearby sources of noise (other people, cars, TV and music) and differing room acoustics (compare a stairwell to a park) — even holding a phone with your shoulder can cause significant mismatch due to differences in microphone position relative to the mouth. This list doesn't even include many of the potential sources of mismatch introduced by the claimants themselves. All of these sources of mismatch have the potential to increase the rate of errors for a speaker verification system.

A number of techniques have been proposed to compensate for various aspects of session variability at almost every stage in the verification process with some success; a state of the art verification system will often incorporate a number of these techniques. An example system (Mason et al., 2004) from the NIST Speaker Recognition Evaluation might include feature warping (Pelecanos and Sridharan, 2001) and mapping (Reynolds, 2003) to produce more robust features as well as score compensation techniques such as H- and T-Norm (Reynolds et al., 2000; Auckenthaler et al., 2000).

These techniques fail to meet the goal stated above for different reasons, but they can be grouped into two major deficiencies.

The most common failing is only considering specific classes or sources of session mismatch; feature mapping, speaker model synthesis (SMS) (Teunen et al., 2000), and H-Norm fall into this group. These techniques all have a common theme in that they all attempt to address some form of categorical phenomena such as handset type. Assuming that the mismatch falls into categories greatly simplifies dealing with the issue but has several negative consequences.

Giving session conditions discrete labels can not generalise well. Apart from the issue that some characteristics are very difficult to describe in a discrete fashion, this can be demonstrated in general by noting that the only way to improve the representation of the mismatch encountered with these techniques is to add more variables to describe the mismatch. Modelling additional variables leads to an exponential growth in the number of categories. For example, adding a Boolean condition, such as whether the speaker is talking hands-free, will double the number of categories. Doubling the number of categories also

doubles the data required to train the method; hence the data requirements also grow exponentially.

Such categorical methods usually require ground truth information on the characteristics they model. Accurate ground truth information is often impossible to acquire after the fact and will certainly be expensive if hand transcription is necessary. Additionally automatically detecting the appropriate category for each test utterance is a necessary and error-prone process with the potential for causing verification errors by applying an inappropriate normalisation.

The second major deficiency is not actually modelling the effects of session variability but simply attempting to suppress them. Feature warping, T-Norm and Z-Norm (Auckenthaler et al., 2000) fit into this category. These methods have no knowledge of the *specific* conditions encountered in a recording but use some *a priori* knowledge of the effects the session conditions *could* have. As an example, feature warping was developed due to observing the non-linear compressing effect that additive noise has on cepstral features (Pelecanos and Sridharan, 2001). Rather than attempting to explicitly model this effect and learn how the cepstral features have been distorted for a specific session, feature warping attempts to warp every utterance back to the same (standard normal) distribution, thus losing any knowledge of the actual distortion encountered.

This article describes an approach to address the issue of mismatch in GMM-based speaker verification by explicitly modelling session variability in both the training and testing procedures and *learning* from the mismatch encountered. By directly modelling the mismatch between sessions in a constrained subspace of the GMM speaker model means, the proposed technique replaces the discrete categorisation of techniques such as feature mapping and H-Norm with a continuous vector-valued representation of the session conditions. A major strength of this approach is that the training methods used also remove the need for labelling the training data for particular conditions.

Apart from overcoming the deficiencies of previous techniques, another goal of this work is to more accurately estimate speaker parameters when multiple enrolment utterances are available: Knowing there are multiple sessions with differing conditions but the same speaker can be exploited to more accurately determine the true speaker characteristics. This is in contrast to simply agglomerating multiple sessions together for enrolment and, in effect, averaging the session conditions into the speaker model estimate.

The proposed model is described in the next section. Section 3 develops the tools and methods required for simultaneously estimating the session and speaker variables of the proposed model culminating in a novel and practi-

cal iterative approximation method based on the Gauss-Seidel method for solving linear systems. Approaches to verification scoring using the proposed model are presented in Section 4 followed by the procedure for learning the characteristics of session variability from a background population of speakers.

The proposed approach to modelling session variability is empirically evaluated and compared to the classical GMM-UBM approach and feature mapping in Section 6. Results are presented for both Switchboard-II and recent Mixer (Martin et al., 2004) conversational telephony corpora and the effects of several configuration options are explored. Finally, the results and future directions for the proposed technique are discussed.

2 Modelling Session Variability

The approach used in this work is to model the effect of session variability in the GMM speaker model space. More specifically, the particular conditions of a recording session are assumed to result in an offset to each of the GMM component mean vectors. In other words, the Gaussian mixture model that best represents the acoustic observations $\boldsymbol{X}_h$ of a particular recording session h is the combination of a session-independent speaker model with an additional session-dependent offset of the model means. This can be represented for the speaker s in terms of the $CD \times 1$ concatenated GMM component means supervectors as

$$\boldsymbol{\mu}_h(s) = \boldsymbol{\mu}(s) + \boldsymbol{U}\boldsymbol{z}_h(s), \tag{1}$$

where the GMM is of order C and dimension D.

Here, the speaker s is represented by the mean supervector $\boldsymbol{\mu}(s)$ which consists of the concatenated mixture component means, that is $\boldsymbol{\mu}(s) = \left[\boldsymbol{\mu}_1(s)^T \ \cdots \ \boldsymbol{\mu}_C(s)^T\right]^T$. To represent the conditions of the particular recording, designated with the subscript h, an additional offset of $\boldsymbol{U}\boldsymbol{z}_h(s)$ is introduced where $\boldsymbol{z}_h(s)$ is a low-dimensional representation of the conditions in the recording and $\boldsymbol{U}$ is the low-rank transformation matrix.

The presence of the term $\boldsymbol{U}\boldsymbol{z}_h(s)$ fulfils the objective of *explicitly* modelling the session conditions stated above. Also, the issues related to using a categorical approach described in the previous section are addressed by using a continuous multi-dimensional variable $\boldsymbol{z}_h(s)$ to express this model.

Further, as the observed feature vectors are assumed to be conditional on both an explicit session-dependent part and a session-independent speaker part, this model also differs from the suppressive methods such as feature warping and T-Norm.

The likelihood function in this model is ostensibly identical to the standard GMM likelihood function for the observation sequence $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, that is

$$p\Big(\boldsymbol{X}_h|\boldsymbol{\mu}_h(s)\Big) = \prod_{t=1}^{T}\sum_{c=1}^{C}\omega_c g\Big(\boldsymbol{x}_t|\boldsymbol{\mu}_{h,c}(s), \boldsymbol{\Sigma}_c\Big) \tag{2}$$

where ω_c is the mixture weight for the component c, $\boldsymbol{\mu}_{h,c}(s)$ is the portion of the supervector $\boldsymbol{\mu}_h(s)$ corresponding to component c and likewise for the component covariance matrix $\boldsymbol{\Sigma}_c$ and

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right) \tag{3}$$

is the standard Gaussian kernel.

One of the central assumptions in this formulation is that the majority of session variability can be described in a low-dimensional, linear subspace of the concatenated GMM mean vectors. In (1) this subspace is defined by the transform $\boldsymbol{U}$ from the session variables $z_h(s)$ of dimension R_z to the GMM mean supervector space of dimension CD where $R_z \ll CD$. In this work values for R_z range form 20 to 100 while $CD = 512 \times 24 = 12288$.

By knowing the effect that the session conditions can have on a speaker model, in the form of a session variability subspace, it is possible to distinguish between true characteristics of a speaker and spurious session artefacts. Assuming that the session subspace $\boldsymbol{U}$ is appropriately trained and constrained to capture only the most significant session effects then any characteristics that can be explained in the subspace will be heavily dominated by session effects and hold a minimum of reliable speaker information.

An important aspect to the session variability modelling approach is that the subspace defined by $\boldsymbol{U}$ is determined in an entirely data-driven manner using a corpus of background speakers without any requirements for labelling the session conditions. By observing the actual differences in component means for multiple recordings of the same speaker under a variety of session conditions, $\boldsymbol{U}$ can be estimated without any knowledge of the specific session characteristics actually captured. While the corpus should reflect the anticipated deployment conditions and is preferably quite large, the composition of the corpus does not need to be as carefully balanced as is required for H-Norm or feature mapping; this makes much better use of the available training corpus as much less is potentially wasted due to balancing issues.

Returning to the model described in (1), it simply states that the true speaker characteristics are described by the concatenated mean supervector $\boldsymbol{\mu}(s)$. There are a number of possibilities for how this supervector is estimated but there is one important restriction: Adaptation must be used for a subspace to describe the relationships of the component means, that is the speaker mean should comprise a shared speaker-independent mean plus a speaker-dependent

offset. That is

$$\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{d}(s)$$

where $\boldsymbol{m}$ is the speaker-independent UBM mean and $\boldsymbol{d}(s)$ is the speaker offset supervector. This requirement is necessary to ensure that the component means' relationships modelled in $\boldsymbol{U}$ hold between distinct speaker models. This restriction will not be met, for instance, with standard maximum likelihood (ML) training of GMMs using the expectation-maximisation (E-M) algorithm.

Classical relevance *maximum a posteriori* (MAP) adaptation is an example that fulfils this requirement, and this is the configuration used in this work. Another possibility is to also introduce a speaker variability subspace defined by the low rank transform matrix $\boldsymbol{V}$ and adapt within that subspace, giving $\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{x}(s)$, such as described for speaker verification by Lucey and Chen (2003). As this model contains far fewer variables than relevance MAP it potentially requires far less data to train but has the disadvantage of not asymptotically converging with an ML estimate. Kenny and Dumouchel (2004) use a combination of classical relevance MAP and subspace adaptation in a bid to get the best of both approaches, giving $\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{d}(s) + \boldsymbol{V}\boldsymbol{x}(s)$.

Ideally, the enrolment and verification algorithms will be able to accurately discern the session-independent speaker model $\boldsymbol{\mu}(s)$ in the presence of session variability. These topics will be discussed in Sections 3 and 4, respectively. This will be followed by a description of the algorthim for training the session variability transform $\boldsymbol{U}$ in Section 5.

3 Estimating the Speaker Model Parameters

3.1 Speaker Model Enrolment

The goal of the enrolment process is to get the best possible representation of a speaker. According to the model described in (1) this information is contained in the concatenated GMM mean supervector $\boldsymbol{\mu}(s)$ but this task is complicated by the prevalent conditions in the recording or recordings used for enrolment, represented by $\boldsymbol{z}_h(s)$. Therefore the purpose of enrolment is to find the set of parameters $\lambda_s = \{\boldsymbol{\mu}(s), \boldsymbol{z}_1(s), \ldots, \boldsymbol{z}_H(s)\}$ that maximise the posterior likelihood

$$\begin{aligned} p\Big(\lambda_s | \boldsymbol{X}_1(s), \ldots, \boldsymbol{X}_H(s)\Big) &= p\Big(\boldsymbol{X}_1(s), \ldots, \boldsymbol{X}_H(s) | \lambda_s\Big)\, p(\lambda_s) \\ &= p\Big(\boldsymbol{\mu}(s)\Big) \prod_{h=1}^{H} p\Big(\boldsymbol{z}_h(s)\Big)\, p\Big(\boldsymbol{X}_h(s) | \boldsymbol{z}_h(s), \boldsymbol{\mu}(s)\Big) \end{aligned} \quad (4)$$

over the H sessions available for training. This is a simultaneous optimisation problem over all variables in λ_s however it is only necessary to retain the true speaker mean $\boldsymbol{\mu}(s)$.

The likelihood function of the observation data, $p\Big(\boldsymbol{X}_h(s)|\boldsymbol{z}_h(s), \boldsymbol{\mu}(s)\Big)$ is the standard GMM likelihood of (2) with the component means given by (1) and component covariance matrices $\boldsymbol{\Sigma}_c$. It can be seen that the speaker mean supervector $\boldsymbol{\mu}(s)$ is optimised according to the MAP criterion often used in speaker verification systems (Reynolds, 1997). The prior distribution $p\Big(\boldsymbol{\mu}(s)\Big)$ in this case is derived from a UBM, as previously described by Reynolds.

The MAP criterion is also employed for optimising each of the session variability vectors $\boldsymbol{z}_h(s)$. As described by Kenny and Dumouchel (2004) the prior distribution in this case is set to be a standard normal distribution with zero mean and unit covariance in the subspace defined by the transformation matrix $\boldsymbol{U}$. The optimisation of such a criterion has previously been described for speaker recognition problems (Kenny and Dumouchel, 2004; Lucey and Chen, 2003).

Using the model described by (1) there are an infinite number of possible representations of any given value of $\boldsymbol{\mu}_h(s)$ as the range of $\boldsymbol{U}\boldsymbol{z}_h(s)$ is a subset of the range of $\boldsymbol{\mu}(s)$. This is not an issue, however as the MAP criteria ensure that there is not a "race condition" between the simultaneous optimisation criteria as the constraint imposed by the prior information ensures there is a single solution for all parameters that maximises the combined posterior probability.

An E-M algorithm is used to optimise this model as there is no sufficient statistics for mixtures of Gaussians due to the missing information of mixture component occupancy of each observation. The remainder of Section 3 only discuss the maximisation of the model parameters given an estimate of the mixture component occupancy statistics thus this is only the M step of the full E-M algorithm. A full estimation procedure using these results will be an iterative approach also including the estimation part, which is identical to the E step described by Gauvain and Lee (1996).

The following sections develop the tools necessary for speaker enrolment under the session variability modelling framework, concluding with a practical approximation method in Section 3.4. The optimisation of transform $\boldsymbol{U}$ is a separate, off-line process which is addressed in Section 5.

Suppose we wish to estimate a GMM speaker model where the concatenated mean vectors are constrained to lie in a low-dimensional subspace. The model in this situation is

$$\boldsymbol{\mu} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{z},$$

where $\boldsymbol{\mu}$ is the $CD \times 1$ concatenated supervector of the GMM component means, $\boldsymbol{m}$ is the prior mean, $\boldsymbol{z}$ is the low-dimensional, $R_z \times 1$ vector variable to optimise and $\boldsymbol{U}$ is a $CD \times R_z$ transformation matrix. For MAP estimation of this model the task is to estimate the variable $\boldsymbol{z}$ which is assumed to have a standard normal distribution with zero mean and identity covariance, that is $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Given this model and the prior distribution hyperparameters $\{\boldsymbol{m}, \boldsymbol{U}\}$ the MAP estimate maximises

$$p(\boldsymbol{X}|\boldsymbol{\mu})\, p(\boldsymbol{\mu}|\boldsymbol{m}, \boldsymbol{U}) = p(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{U})\, g(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}) \tag{5}$$

where $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ is the set of observation vectors and $g(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$ refers to evaluating the standard Gaussian kernel at $\boldsymbol{z}$.

As with relevance adaptation, there is the missing information of which mixture component produced which observation. For this reason an iterative E-M approximation is used to optimise this model. The statistics required from the expectation step using this approach are the component occupancy count n_c and sample sum vector $\boldsymbol{S}_{X;c}$ for each mixture component c, defined as

$$n_c = \sum_{t=1}^{T} P(c|\boldsymbol{x}_t) \qquad\qquad \boldsymbol{S}_{X;c} = \sum_{t=1}^{T} P(c|\boldsymbol{x}_t)\, \boldsymbol{x}_t.$$

Further, define $\boldsymbol{S}_X$ as the $CD \times 1$ concatenation of all $\boldsymbol{S}_{X;c}$ and $\boldsymbol{N}$ as the $CD \times CD$ diagonal matrix consisting of C blocks along the diagonal of $\boldsymbol{N}_c = n_c\boldsymbol{I}$ where $\boldsymbol{I}$ is the $D \times D$ identity matrix. Similarly, $\boldsymbol{\Sigma}$ is defined as the $CD \times CD$ matrix consisting of the component covariance matrices $\boldsymbol{\Sigma_c}$ along the diagonal.

With these quantities it can be shown that maximising the MAP criterion is equivalent to solving

$$(\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{U})\boldsymbol{z} = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X|m} \tag{6}$$

for $\boldsymbol{z}$ where $\boldsymbol{S}_{X|m} = \boldsymbol{S}_X - \boldsymbol{N}\boldsymbol{m}$ is the first order statistic centralised on $\boldsymbol{m}$. This can be expressed in the conventional linear algebra form of $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}$ where $\boldsymbol{A} = \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{U}$ is an $R_z \times R_z$ matrix and $\boldsymbol{b} = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X|m}$ is an $R_z \times 1$ column vector. As $\boldsymbol{A}$ is a positive definite matrix this can be straightforwardly solved for $\boldsymbol{z}$ using the Cholesky decomposition method.

3.3 Simultaneous Relevance MAP and Subspace MAP Estimation

Before presenting the solution to simultaneous relevance and subspace MAP estimation, it is helpful to present relevance adaptation in a similar form to subspace estimation using a standard normal prior. This result will be combined with the result of the previous section to simultaneously optimise in both a subspace and the full CD-sized speaker model space. Finally the solution of optimising with multiple sessions will be examined. [2]

3.3.1 Relevance MAP revisited

The relevance MAP described in by Reynolds (1997) can be expressed in the form

$$\boldsymbol{\mu} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{y} \tag{7}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{m}$ have the same meaning as in the previous section and we are optimising the $CD \times 1$ vector $\boldsymbol{y}$ to maximise the same MAP criterion as the previous section also with a standard normal prior distribution. That is, we are adapting a UBM mean vector to represent the training data available for a speaker based on a MAP criterion. For equivalence with Reynold's development of relevance adaptation, the $CD \times CD$ matrix $\boldsymbol{D}$ is set to be the diagonal matrix satisfying

$$\boldsymbol{I} = \tau \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D} \tag{8}$$

where τ is the relevance factor.

According to the solution above, this can also be formed into a standard linear system of equations, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$, with

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{I} + \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{D} \\ &= \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} (\tau \boldsymbol{I} + \boldsymbol{N}) \boldsymbol{D} && (9) \\ \boldsymbol{b} &= \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S}_{X|m}. && (10) \end{aligned}$$

Substituting back in and removing $\boldsymbol{D}^T \boldsymbol{\Sigma}^{-1}$ from both sides,

$$\begin{aligned} (\tau \boldsymbol{I} + \boldsymbol{N}) \boldsymbol{D}\boldsymbol{y} &= \boldsymbol{S}_{X|m}, && (11) \\ \boldsymbol{y}' &= (\tau \boldsymbol{I} + \boldsymbol{N})^{-1} \boldsymbol{S}_{X|m}, && (12) \end{aligned}$$

where $\boldsymbol{y}' = \boldsymbol{D}\boldsymbol{y}$ is the offset in the concatenated GMM mean space. It can be readily seen that (11) has a trivial solution as $(\tau \boldsymbol{I} + \boldsymbol{N})$ is a diagonal matrix and that it is equivalent to Reynolds' relevance MAP adaptation solution.

[2] This section follows the development presented in the unpublished work "Joint factor analysis of speaker and session variability: Theory and algorithms," by Patrick Kenny, available at `http://www.crim.ca/perso/patrick.kenny/`

3.3.2 Simultaneous optimisation of $\boldsymbol{y}$ and $\boldsymbol{z}_h$

Having shown the equivalence of relevance MAP and subspace MAP estimation techniques given the appropriate transformation matrix $\boldsymbol{D}$, we can extend the result to simultaneously optimise $\boldsymbol{y}$ and the set $\boldsymbol{z}_h$ over all observed sessions $\boldsymbol{X}_h; h = 1, \ldots, H$. (The speaker label s has been dropped in this section for clarity as we are only dealing with a single speaker at this point.)

The set of variables to optimise can be expressed in the form

$$\underline{\boldsymbol{z}} = \left[\boldsymbol{z}_1^T \cdots \boldsymbol{z}_H^T \ \boldsymbol{y}^T\right]^T, \tag{13}$$

which is a $(HR_z + CD) \times 1$ column vector. Given this definition, the standard $\boldsymbol{A}\underline{\boldsymbol{z}} = \boldsymbol{b}$ formulation of the optimisation problem can be expressed, in an analogous fashion to the previous sections, as

$$\boldsymbol{A} = \boldsymbol{I} + \underline{\boldsymbol{U}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{N}\boldsymbol{U}} \tag{14}$$

$$\boldsymbol{b} = \underline{\boldsymbol{U}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{S}}_{X|m}. \tag{15}$$

In this formulation, a combined $HCD \times (HR_z + CD)$ transformation matrix for these variables can be defined as

$$\underline{\boldsymbol{U}} = \begin{bmatrix} \boldsymbol{U} & & \boldsymbol{D} \\ & \ddots & \vdots \\ & \boldsymbol{U} & \boldsymbol{D} \end{bmatrix}, \tag{16}$$

and the statistic $\underline{\boldsymbol{S}}_X$ is defined as

$$\underline{\boldsymbol{S}}_X = \left[\boldsymbol{S}_{X,1}^T \cdots \boldsymbol{S}_{X,H}^T\right]^T \tag{17}$$

which allows the statistics of each session to be available independently. Similar definitions of the component occupancy statistics matrix $\underline{\boldsymbol{N}}$ and $\underline{\boldsymbol{\Sigma}}$ are also required, producing a $HCD \times HCD$ diagonal matrices. $\underline{\boldsymbol{N}}$ is simply the concatenation of all available $\boldsymbol{N}_h$ while $\underline{\boldsymbol{\Sigma}}$ consists of H repeats of $\boldsymbol{\Sigma}$ along the diagonal. It will also be convenient to define $\boldsymbol{S}_X = \sum_{h=1}^{H} \boldsymbol{S}_{X,h}$ and $\boldsymbol{N} = \sum_{h=1}^{H} \boldsymbol{N}_h$.

Unfortunately evaluating the solution to this equation directly is less than practical; it involves the decomposition of $\boldsymbol{A}$ which in this case is a $(HR_z + CD) \times (HR_z + CD)$ matrix. With the typical values of these dimensions this is a large task, especially as this matrix is not diagonal. It is, however, still positive definite.

Noting that the $CD \times CD$ block in the lower right region of $\boldsymbol{A}$ is diagonal and given by $\boldsymbol{I} + \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{D}$, the Strassen block matrix inversion algo-

rithm (Press, 1992) can be utilised to make the solution to this system more practical. This identity is given by

$$\begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & \boldsymbol{\gamma} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\zeta}^{-1} & -\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \\ -\boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T\boldsymbol{\zeta}^{-1} & \boldsymbol{\gamma}^{-1} + \boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \end{bmatrix} \tag{18}$$

where

$$\boldsymbol{\zeta} = \boldsymbol{\alpha} - \boldsymbol{\beta}\boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T. \tag{19}$$

Using this identity and expressing $\boldsymbol{A}$ in the form $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & \boldsymbol{\gamma} \end{bmatrix}$, the solution to the maximisation of the model is given by,

$$\begin{aligned} \underline{\boldsymbol{z}} &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ &= \begin{bmatrix} \boldsymbol{\zeta}^{-1} & -\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \\ -\boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T\boldsymbol{\zeta}^{-1} & \boldsymbol{\gamma}^{-1} + \boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T\boldsymbol{\zeta}^{-1}\boldsymbol{\beta}\boldsymbol{\gamma}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X,1|m} \\ \vdots \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X,H|m} \\ \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X|m} \end{bmatrix}. \end{aligned}$$

with

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_1\boldsymbol{U} & & \\ & \ddots & \\ & & \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_H\boldsymbol{U} \end{bmatrix} \tag{20}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_1\boldsymbol{D} \\ \vdots \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_H\boldsymbol{D} \end{bmatrix} \tag{21}$$

$$\boldsymbol{\gamma} = \boldsymbol{I} + \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{D}. \tag{22}$$

This gives

$$\boldsymbol{z}_{1,\ldots,H} = \begin{bmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_H \end{bmatrix} = \boldsymbol{\zeta}^{-1} \begin{bmatrix} \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X,1|m} - \boldsymbol{N}_1\boldsymbol{\delta}\right) \\ \vdots \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X,H|m} - \boldsymbol{N}_H\boldsymbol{\delta}\right) \end{bmatrix}, \tag{23}$$

where

$$\boldsymbol{\delta} = \boldsymbol{D}\boldsymbol{\gamma}^{-1}\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X|m} \tag{24}$$

and

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\gamma}^{-1}\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{X|m} - \boldsymbol{\gamma}^{-1}\boldsymbol{\beta}^T\boldsymbol{z}_{1,\ldots,H} \\ &= \boldsymbol{\gamma}^{-1}\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{S}_{X|m} - \sum_{h=1}^{H}\boldsymbol{N}_h\boldsymbol{U}\boldsymbol{z}_h\Big). \end{aligned} \tag{25}$$

In the case of classical relevance adaptation with $\boldsymbol{D}$ satisfying (8), these solutions can be simplified to

$$\boldsymbol{\delta} = (\tau \boldsymbol{I} + \boldsymbol{N})^{-1} \boldsymbol{S}_{X|m},$$

and

$$\boldsymbol{y} = \boldsymbol{D}^{-1}(\tau \boldsymbol{I} + \boldsymbol{N})^{-1} \Big(\boldsymbol{S}_{X|m} - \sum_{h=1}^{H} \boldsymbol{N}_h \boldsymbol{U} \boldsymbol{z}_h \Big).$$

Using this method, the inverse of $\boldsymbol{A}$ can be determined by inverting the $HR_z \times HR_z$ matrix $\boldsymbol{\zeta}$, which is much smaller than $\boldsymbol{A}$, and inverting $\boldsymbol{\gamma}$, which is large but diagonal. While inverting $\boldsymbol{\zeta}$ will be much faster than inverting $\boldsymbol{A}$ directly, the cost of this operation is $O(H^3 R_z^3)$. This cost is therefore very sensitive to both the number of sessions and size of the session subspace; both of which can potentially limit the feasibility of this model.

3.4 Gauss-Seidel Approximation Method

While a practical solution to the simultaneous MAP estimation of multiple session variables and the speaker mean offset was presented in the previous section, the solution is still expensive in terms of processing requirements. In fact, it is impractical if a reasonable number of speakers, each with a reasonable number of sessions, are to be estimated — such as is the case for a NIST evaluation that typically involves training thousands of models. Also, it is worth noting that the solutions above are just for the maximisation step of an E-M algorithm and multiple E-M iterations are usually required.

A method with more modest processing requirements is desirable. This section presents an efficient optimisation method inspired by the Gauss-Seidel method (Barrett et al., 1994).

The Gauss-Seidel method is a stationary, iterative method for solving linear systems of equations in the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ by iteratively improving an estimate of $\boldsymbol{x}$. On each iteration k, the individual elements x_i are re-estimated with the update equation

$$x_i^{(k)} = a_{ii}^{-1} \Big(b_i - \sum_{j<i} a_{ij} x_j^{(k)} - \sum_{j>i} a_{ij} x_j^{(k-1)} \Big),$$

where the superscripts (k) and $(k-1)$ refer to values on the current and previous iterations respectively. This scheme improves the convergence rate of the Jacobi method by analysing the elements in the order $x_1, x_2, \ldots$ and using the most up-to-date estimate available for the other elements.

The Gauss-Seidel approach could be used directly for solving the system $\boldsymbol{A}\underline{\boldsymbol{z}} = \boldsymbol{b}$ in Section 3.3 with $\underline{\boldsymbol{z}}$, $\boldsymbol{A}$ and $\boldsymbol{b}$ described in (13), (14) and (15), however, the approach used in this work takes advantage of the direct solutions for the constituent parts $\boldsymbol{y}, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ of $\underline{\boldsymbol{z}}$. Instead of updating each element individually each constituent vector is updated as a block; the speaker mean offset and each of the session condition variables are solved successively assuming the estimate of all other variables is fixed. In this way, the speaker mean offset $\boldsymbol{y}$ can be estimated with the usual relevance MAP adaptation equations assuming that the session conditions $\boldsymbol{z}_h$ are all known. Similarly, the session variables $\boldsymbol{z}_h$ for $h = 1, \ldots, H$ can each be estimated assuming that $\boldsymbol{y}$ is known.

As with the direct solution presented in the previous section, this is only the solution to maximising the MAP criterion and forms only the M step of an E-M algorithm. Due to the missing information of the mixture component allocations of the training data, an iterative algorithm is also required on this level to converge on the optimal result. The complete algorithm for estimating the speaker model and the session condition variables is presented in Algorithm 1.

Algorithm 1 Speaker Model Estimation

$\boldsymbol{y} \leftarrow \boldsymbol{0}; \boldsymbol{z}_h \leftarrow \boldsymbol{0}; h = 1, \ldots, H$
for $i = 1$ to Number of E-M iterations **do**
  E Step:
  for $h = 1$ to H **do**
    Calculate $\boldsymbol{N}_h$ and $\boldsymbol{S}_{X,h}$ for session $\boldsymbol{X}_h$ where $\boldsymbol{\mu}_h = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{y} + \boldsymbol{U}\boldsymbol{z}_h$
  end for
  $\boldsymbol{N} \leftarrow \sum_{h=1}^{H} \boldsymbol{N}_h$
  $\boldsymbol{S}_X \leftarrow \sum_{h=1}^{H} \boldsymbol{S}_{X,h}$
  M Step:
  for $j = 1$ to Number of Gauss-Seidel iterations **do**
    for $h = 1$ to H **do**
      $\boldsymbol{z}_h \leftarrow \boldsymbol{A}_h^{-1}\boldsymbol{b}_h$
        where $\boldsymbol{A}_h = \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_h\boldsymbol{U}$ and $\boldsymbol{b}_h = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X,h|m} - \boldsymbol{N}_h\boldsymbol{D}\boldsymbol{y}\right)$
    end for
    $\boldsymbol{y} \leftarrow \boldsymbol{A}_y^{-1}\boldsymbol{b}_y$
      where $\boldsymbol{A}_y = \boldsymbol{I} + \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{D}$ and $\boldsymbol{b}_y = \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X|m} - \sum_{h=1}^{H} \boldsymbol{N}_h\boldsymbol{U}\boldsymbol{z}_h\right)$
  end for
end for
return $\boldsymbol{y}$

In this algorithm, the expectation or E step is essentially the same as for standard E-M algorithm for GMM training with the caveat that the session statistics are gathered separately and the Gaussian means also include a session-dependent offset.

The maximisation or M step uses an iterative solution. The resulting solutions are given by

$$\boldsymbol{z}_h = \left(\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_h\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X,h|m} - \boldsymbol{N}_h\boldsymbol{D}\boldsymbol{y}\right), \tag{26}$$

$$\boldsymbol{y} = \boldsymbol{\gamma}^{-1}\boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X|m} - \sum_{h=1}^{H}\boldsymbol{N}_h\boldsymbol{U}\boldsymbol{z}_h\right). \tag{27}$$

These are, respectively, the subspace MAP and relevance MAP solutions with compensated $\boldsymbol{b}$ vectors, as emphasised on Lines 12 and 14.

Comparing these solutions with the direct solutions for multiple sessions, the solution for the speaker mean offset $\boldsymbol{y}$ takes an identical form ((25) and (27)) that is dependent on the solution to $\boldsymbol{z}_h$. This is somewhat misleading as the actual resulting values are potentially quite different due to the differing solutions for the session variables. As can be seen in (23) and (26) the direct solution for the session variables is significantly more involved; it requires the inversion of a larger matrix and also couples the results from all of the session variables. On the other hand, the iterative approximation is independent of the other sessions and relies solely on the most recent approximation of $\boldsymbol{y}$.

The initial guesses of all variables in this algorithm is chosen to be $\mathbf{0}$. This is a reasonable assumption given that the aim is to optimise a MAP criterion for each variable with the standard normal distribution as the prior. After the first iteration of the E-M algorithm, the initial guess for the Gauss-Seidel maximisation part of the algorithm will be initialised with the results of the previous iteration providing a much better guess than $\mathbf{0}$, leading to better convergence rates in subsequent iterations. This behaviour of *refining* previous estimates is a strength of an iterative approximation method.

The convergence of the standard Gauss-Seidel approach is conditional on the diagonal dominance of $\boldsymbol{A}$.[3] While $\boldsymbol{A}$ in this case is always positive definite, the values of $\boldsymbol{U}$ and $\boldsymbol{D}$ may effect it's diagonal dominance. In practice, the algorithm presented above converged for the experiments presented in this work.

The processing requirements for this algorithm grow linearly with the number of sessions used for training, which was the goal of this method, and only H matrix decompositions are necessary of size $R_z \times R_z$. A large value for R_z would be required before these decompositions start to dominate the processing time; for the values used in this study the algorithm is dominated by the E step of calculating the statistics $\boldsymbol{N}_h$ and $\boldsymbol{S}_{X,h}$ for each session (Line 5 of Algorithm 1).

[3] It is expected that the approach used in Algorithm 1 relaxes the convergence constraints to be a form of *block* diagonal dominance. This has not been investigated to date.

3.4.1 Behaviour of the Gauss-Seidel approximation

There are several interesting aspects to this algorithm that deserve some exploration.

Given that the E-M algorithm for Gaussian mixture models generally converges to a *local* optimum, it is possible for different solutions to occur for the same data with different initialisation for each iteration. The implication for the approximation method described in Algorithm 1 is the potential to converge to a different local optimum to the direct solution method of Section 3.3. While this will not happen with a fully converged G-S solution, as it will match the direct solution, it can occur if full convergence is not achieved.

So the relevant question to arise from this observation is, how many iterations of the Gauss-Seidel method are necessary for convergence? Or, more practically, how many iterations are necessary for optimal verification performance?

These questions are complicated by the fact that changing the order of evaluating the estimates in the Gauss-Seidel method will effect the intermediate approximations of the variables. The algorithm described above estimates the session variables first but can be formulated to estimate the speaker first. This should not effect the final converged result to the system of linear equations but does impact on the rate of convergence and the intermediate estimates (Barrett et al., 1994).

Figures 1 and 2 demonstrate the effect of using only one iteration of the Gauss-Seidel approximation with estimating the session variables first (as described in Algorithm 1) compared to estimating the speaker offset first. Both variants are compared to a fully converged G-S estimate and estimating the session and speaker variables independently of each other for each E-M iteration. (While the values graphed in these plots cannot be directly used to assess convergence, they are useful from the perspective of understanding and comparing methods.)

The most significant point of these figures for the current discussion is the similarity between the single iteration, session first method and the fully converged result. These results are so similar that they are almost indistinguishable in all figures. For the speaker first method this is also true of the speaker vector, $\boldsymbol{y}(s)$ from around 14 iterations of the E-M algorithm but the session vectors do not share this similarity. It would seem, however, that in the case of this example all of these methods will eventually converge to the same result.

Interestingly, the independent estimation method seems to have little in common with any of the Gauss-Seidel variants and seems unlikely to converge to the same result; the session variables seem to stabilise after only a few iterations to a very different result to the other methods while the estimate of the

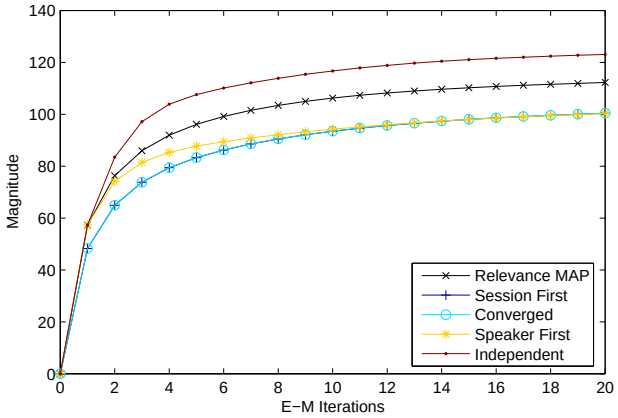

Fig. 1. Plot of the speaker mean offset supervector magnitude, $|\boldsymbol{y}(s)|$, for differing optimisation techniques as it evolves over iterations of the E-M algorithm.

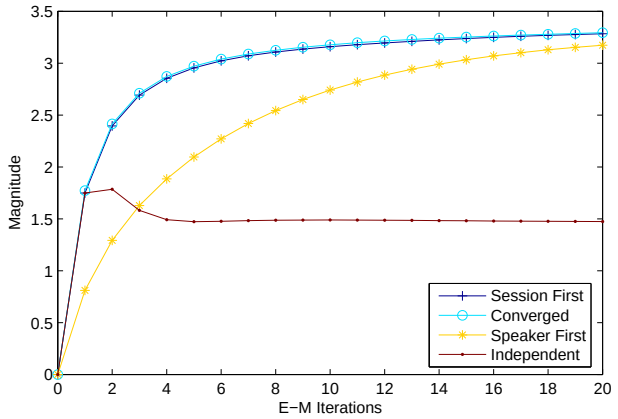

Fig. 2. Plot of the session variability vector magnitude, $|\boldsymbol{z}_h(s)|$, for differing optimisation techniques as it evolves over iterations of the E-M algorithm.

speaker variables is larger in magnitude than the standard MAP adaptation. These factors indicate that this method will indeed converge to a different local minimum to the fully converged G-S approximation.

It is not possible to draw conclusive statements based on the single example depicted above although the single iteration, session first estimate appears to be a close approximation to the fully converged estimate. This may allow for more efficient speaker enrolment procedures for equivalent verification performance. This possibility will be investigated further in Section 6.3, as will the effect on performance of the other variants described in this comparison.

4 Verification

The previous section developed the procedure for enrolling a speaker with a model incorporating session variability using simultaneous optimisation of speaker and session variables. This section extends this treatment to the verification stage of the system. To this end, the session variation introduced in the verification utterance must also be considered.

An expected log likelihood ratio (ELLR) score takes the general form

$$\Lambda_s(\boldsymbol{X}_v) = \frac{1}{T}\log\frac{\ell_s(\boldsymbol{X}_v)}{\ell_0(\boldsymbol{X}_v)} \tag{28}$$

where the $\boldsymbol{X}_v$ is the set of verification trial observations, T is the number of observation vectors, $\ell_s(\cdot)$ is the likelihood score for the speaker s and $\ell_0(\cdot)$ is the background likelihood based on the UBM. (To aid clarity, the parameterisation by $\boldsymbol{X}_v$ will be omitted for the rest of this section where it is obvious due to context.)

The simplest approach to verification under the session variability framework is to continue to use ELLR scoring as is traditionally used with GMM-UBM verification systems. By taking this approach the conditions of the verification session are completely ignored but performance gains are still possible over standard GMM-UBM systems assuming that the training procedure produced a speaker model that more accurately represents the speaker. This is a reasonable assumption given that the point of the training procedure was to separate the speaker and session contributions as separate variables rather than modelling a combination of speaker and session conditions; particularly with multiple training sessions to distinguish between speaker and session effects this should be the case. According to the model proposed in (1) this is equivalent to calculating the ratio of likelihoods

$$\ell_s = p\Big(\boldsymbol{X}_v|\boldsymbol{\mu}_v(s) = \boldsymbol{\mu}(s)\Big) \tag{29}$$

where the session variable has been set to $\boldsymbol{z} = \boldsymbol{0}$.

Using standard scoring methods with the improved training can only ever hope to address half of the mismatch issue; it may be possible to determine the speaker characteristics sans session effects but comparing this to a verification trial *with* session effects still entails mismatched conditions.

One possible way of dealing with the mismatch introduced by the verification utterance is to estimate the session variable $\boldsymbol{z}_v(s)$ of the utterance for each speaker prior to performing standard top-N ELLR scoring. Under this

approach the likelihood score for a speaker is given by

$$\ell_s = \max_{\boldsymbol{z}} p\Big(\boldsymbol{X}_v | \boldsymbol{\mu}_v(s) = \boldsymbol{\mu}(s) + \boldsymbol{U}\boldsymbol{z}\Big)\, g(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}). \tag{30}$$

This likelihood is essentially the MAP criterion used in Section 3.2 however in this case the evaluation of the likelihood is the desired result rather than determining the argument $\boldsymbol{z}$ that maximises it, although $\boldsymbol{z}$ is a necessary by-product. This verification method will be used throughout this study.

The estimation procedure for $\boldsymbol{z}$ is similar to that described in Section 3.2 with a few differences due to the context in which this estimation occurs. Often, (30) must be evaluated for several models for the same verification trial — at least the target and background model but many more if T-Norm score normalisation is to be used — so efficiency is very important.

To substantially reduce the processing required, a simplification is made in that the mixture component occupancy statistics for the observations are calculated based on the UBM rather than independently for each model to be scored. This allows for a solution that calls for only one additional pass of the verification utterance than standard top-N ELLR scoring and implies that only one $R_z \times R_z$ matrix decomposition is necessary, regardless of the number of speakers being tested. Also, only a single adaptation step is used as, without re-aligning the observation vectors, more iterations would not produce a different result.

It is interesting to note the role of the prior distribution of $\boldsymbol{z}$ in (30). While its presence is necessary to mirror the MAP criterion used for estimating the session variables in the training algorithm, the effect it has is to penalise models that require a large session compensation offset compared to those that are "closer" to matching the recording. In practice, empirical evidence suggests that the presence of the prior is insignificant in terms of verification performance as its contribution to the overall score is dwarfed by that of the observation vectors.

More sophisticated verification techniques are possible with the session variability modelling approach. Future research will investigate the effectiveness of Bayes factor techniques in conjunction with modelling session variability in a similar approach to (Vogt and Sridharan, 2004). Under this approach the speaker model parameters are not assumed to be known at testing time, but rather to have posterior distributions that have been refined by the training procedure.

5 Training the Session Variability Subspace

For the session variation modelling described in this article to be effective, the constrained session variability subspace described by the transformation matrix $\boldsymbol{U}$ must represent the types of intra-speaker variations expected between sessions. To this end, the subspace is trained on a database containing a large number of speakers each with several independently recorded sessions. Preferably this training database will include a variety of channels, handset types and environmental conditions that closely resembles the conditions on which the eventual system is to be used.

This section describes the procedure for optimising the session transform matrix $\boldsymbol{U}$ for a population of speakers by building on the results of Section 3. Firstly, a straightforward method of estimating the transform using a principal components approach is described. An E-M algorithm is then presented that fully optimises $\boldsymbol{U}$ for all of the available data.

5.1 Principal Components of Session Variability

The simplest method of estimating the session variability transform is to observe the differences of models trained for the same speaker from different recordings for a group of speakers and determine the principal components of this variation.

Given a set of recordings $\boldsymbol{X}_h(s); h = 1, \ldots, H(s)$ for a group of speakers $s = 1, \ldots, S$, a model is first estimated for each recording using classical relevance MAP adaptation. This gives a set of adapted GMM mean supervectors $\boldsymbol{\mu}_h(s)$. This set of mean supervectors then form the samples of a standard principal components analysis (PCA).

The within-class scatter matrix for this analysis is given by

$$\boldsymbol{S}_W = \frac{1}{R} \sum_{s=1}^{S} \sum_{h=1}^{H(s)} \Big(\boldsymbol{\mu}_h(s) - \bar{\boldsymbol{\mu}}(s)\Big)^T \Big(\boldsymbol{\mu}_h(s) - \bar{\boldsymbol{\mu}}(s)\Big)$$

where $\bar{\boldsymbol{\mu}}(s) = \frac{1}{H(s)} \sum_{h=1}^{H(s)} \boldsymbol{\mu}_h(s)$ is the mean of the mean supervectors for speaker s and $R = \sum_{s=1}^{S} H(s)$.

As $\boldsymbol{S}_W$ is a large $CD \times CD$ matrix, it is typically too large to directly perform eigenvalue analysis but it usually has significantly lower rank, with a maximum possible rank of R. Thus an equivalent eigenvalue problem can be constructed with an $R \times R$ matrix as described by Fukunaga (1990) (pp. 35–37).

Taking the eigenvalue decomposition of the scatter matrix gives the form $\boldsymbol{S}_W = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^T$ where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues and $\boldsymbol{X}$ is the matrix with the corresponding eigenvectors as its columns. The desired behaviour for the transform $\boldsymbol{U}$ is to whiten this scatter matrix in order to use the standard normal distribution with covariance $\boldsymbol{I}$ as the prior distribution of the session variable $\boldsymbol{z}$, therefore the desired decomposition is $\boldsymbol{S}_W = \boldsymbol{U}\boldsymbol{I}\boldsymbol{U}^T$, giving

$$\boldsymbol{U} = \boldsymbol{X}\boldsymbol{\Lambda}^{-\frac{1}{2}}. \tag{31}$$

The number of (non-zero) columns of $\boldsymbol{U}$ is at most R and is determined by the rank of the scatter matrix but in practice only the columns corresponding to the largest eigenvalues are retained.

5.2 Iterative Optimisation

Estimating the principal components of the variation observed in speaker model training provides a starting point for estimating the session subspace but, as it does not use the same simultaneous estimation training method as described in Section 3, it will not provide optimal results.

To accurately model the speaker and the session variability the session subspace must be found that maximises the total *a posteriori* likelihood of all segments in the training database by training a model for each speaker represented using the procedure in section 3. That is, $\boldsymbol{U}$ must satisfy

$$\boldsymbol{U} = \arg\max_{\boldsymbol{U}} \prod_{s=1}^{S} p\Big(\lambda_s | \boldsymbol{X}_1(s), \ldots, \boldsymbol{X}_{H(s)}(s)\Big). \tag{32}$$

As the speaker and corresponding session variables are hidden in this optimisation procedure, another E-M algorithm is used. This procedure is described in detail in (Kenny et al., 2005a), with the caveat that a modified speaker model training procedure was used.

Briefly, the iterative optimisation of the subspace proceeds as follows: Firstly, an initial estimate of $\boldsymbol{U}$ is used to bootstrap the optimisation. The PCA estimate described above is appropriate for this task as the better the initial estimate the more quickly the iterative method will converge. Then for the following iterations there are successive estimation and maximisation steps.

The E-step in this algorithm involves estimating the parameter set $\lambda_s = \{\boldsymbol{y}(s), \boldsymbol{z}_1(s), \ldots, \boldsymbol{z}_{H(s)}(s)\}$ for each speaker s in the training database using the current estimate of the session subspace transform $\boldsymbol{U}$. This estimation follows the speaker enrolment procedure described in Section 3 above.

The M-step then involves maximising (32) given the expected values for λ_s. Using the notation of Section 3.4, this maximisation is equivalent to solving the system of equations

$$\sum_{s=1}^{S}\sum_{h=1}^{H(s)} \boldsymbol{N}_h(s)\boldsymbol{U}\Big(\boldsymbol{z}_h(s)\boldsymbol{z}_h(s)^T + \boldsymbol{A}_h^{-1}(s)\Big) = \sum_{s=1}^{S}\sum_{h=1}^{H(s)} \Big(\boldsymbol{S}_{X,h|m} - \boldsymbol{N}_h(s)\boldsymbol{D}\boldsymbol{y}(s)\Big)\boldsymbol{z}_h(s)^T \quad (33)$$

for $\boldsymbol{U}$. Using the notation $\boldsymbol{U}_c$ to represent the rows of $\boldsymbol{U}$ corresponding to the cth mixture component — that is rows $cD+1$ to $(c+1)D$ — and similarly for the other variables, this can be rewritten as

$$\boldsymbol{U}_c\boldsymbol{A}_c = \boldsymbol{B}_c \quad (34)$$

where

$$\boldsymbol{A}_c = \sum_{s=1}^{S}\sum_{h=1}^{H(s)} n_{c,h}(s)\Big(\boldsymbol{z}_h(s)\boldsymbol{z}_h(s)^T + \boldsymbol{A}_h^{-1}(s)\Big) \quad (35)$$

$$\boldsymbol{B}_c = \sum_{s=1}^{S}\sum_{h=1}^{H(s)} \Big(\boldsymbol{S}_{X,c,h|m} - n_{c,h}(s)\boldsymbol{D}_c\boldsymbol{y}_c(s)\Big)\boldsymbol{z}_h(s)^T. \quad (36)$$

This system of equations can be solved in the usual way for $\boldsymbol{U}_c$.

As stated in (Kenny and Dumouchel, 2004) this optimisation converges quite slowly and requires significant processing resources, therefore determining the minimum number of iterations for adequate performance is of interest. The sensitivity of this approach to the quality of session transformation will be further investigated empirically in terms of verification performance in Section 6.4.

6 Speaker Verification Experiments

The proposed session variability modelling technique was initially evaluated on data from the Switchboard-II conversational telephony corpus. By design, this corpus exhibits a wide variety of session conditions including a variety of landline handset types used over PSTN channels in a number of locations. As participants in the collection were encouraged to use different telephones on different numbers throughout the collection, this corpus is well suited for evaluating the suitability of the session modelling methods and also training the required session subspace. A expanded version of the NIST EDT 2003

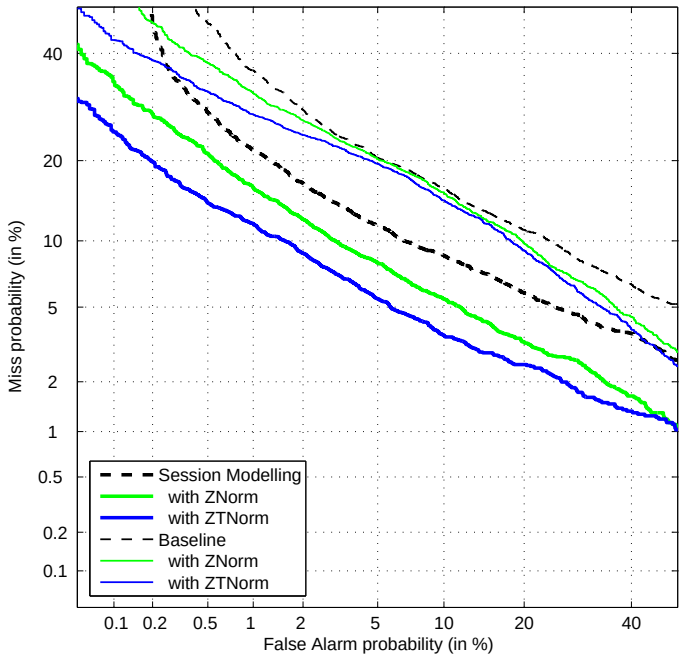

Fig. 3. DET plot of the 1-side training condition for the baseline system and session variability modelling on Switchboard-II data.

Table 1
Minimum DCF and EER of the baseline system and session variability modelling on Switchboard-II data.

System	**Raw Scores**		**Z-Norm**		**ZT-Norm**	
	Min DCF	EER	Min DCF	EER	Min DCF	EER
1-Side						
Baseline	.0458	13.6	.0415	13.0	.0367	12.7
Session Modelling	.0311	9.0	.0251	6.8	.0191	5.3
3-Side						
Baseline	.0243	5.9	.0252	5.6	.0213	5.7
Session Modelling	.0110	2.8	.0089	2.0	.0069	1.9

protocol with additional non-target trials was used for these experiments. [4]

Figure 3 shows detection error trade-off (DET) plots comparing systems with and without session variability modelling for the 1-side training condition. Table 1 presents the minimum detection cost function (DCF) and equal error rate (EER) performance corresponding to these DET plots including also the 3-side training condition.

With no score normalisation applied, the session modelling technique provided a 32% reduction in DCF for the 1-side condition and a 54% reduction in the 3-side condition with similar trends in EER. While the improvement in the 3-side

[4] The QUT EDT 2003 protocol is available from the authors on request.

training condition is very substantial, the 1-side result is at least as interesting and, in many ways, more surprising and encouraging: In the 1-side condition, there was not multiple sessions from which to gain a good estimate of the true speaker characteristics by factoring out the session variations, however, the technique successfully factored out the variations between the training and testing sessions.

Also presented are results with normalisation applied to all systems. The normalisations applied were Z-Norm to characterise the response of each speaker model to a variety of (impostor) test segments followed by T-Norm to compensate for the variations of the testing segments, such as duration and linguistic content. Again the proposed technique outperforms the baseline system, but also in fact gains more from this normalisation process than the baseline system with the improvements in DCF growing to 48% and 68% respectively for the 1- and 3-side conditions.

The benefits gained with Z-Norm score normalisation, particularly in the 1-side case, seem to imply that a model produced with the proposed technique exhibits a more consistent response to a variety of test segments from differing session conditions. In contrast, the baseline system improved little with Z-Norm while it is well known that H-Norm — utilising extra handset type labels — is more effective. [5] This difference indicates that the session modelling technique is successfully compensating for session differences such as handset type.

At the same time, the Z-Norm result indicates that there is a significant discrepancy between score distributions from different models that the normalisation is correcting.

Figure 4 compares the performance of the presented technique to a feature mapping system trained with data-driven clustering as described by Mason et al. (2005) on equivalent development data. Again, it can be seen that the session variation modelling technique has a clear advantage with a 19% improvement at the minimum DCF operating point, and similarly for the EER.

With score normalisation applied, the advantage of the session modelling method increases as Z-Norm is largely ineffective for feature mapping. Following the logic above, this indicates that feature mapping is less effective in compensating for the encountered session effects.

[5] As H-Norm is known to be more effective than Z-Norm for the baseline system it is relevant to question why H-Norm was *not* used for this comparison. One of the focuses of this work is alleviating the need for labelled corpora for training the normalisation techniques and for this purpose Z-Norm is more suitable since H-Norm requires its normalisation data to be accurately labelled for handset types.

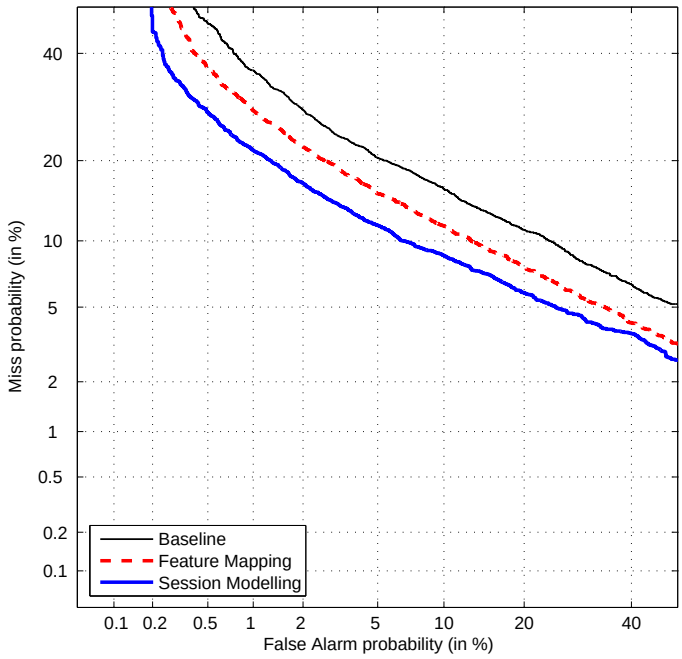

Fig. 4. Comparison of session variability modelling and blind feature mapping for the 1-side training condition.

6.1 Mixer Results

The results presented so far indicate that session modelling can produce significant gains in speaker verification performance for the conversational telephony data of Switchboard-II. This section presents results of the same system for the Mixer corpus (Martin et al., 2004) to demonstrate that this method is not exploiting hidden characteristics of Switchboard. Furthermore, the increased variety of channel conditions present — including a variety of mobile transmission types, hands-free and cordless handsets as well as cross-lingual trials — represents a significantly more challenging situation for the proposed session modelling approach to tackle. Table 2 presents results for Mixer data using a modified NIST 2004 protocol analogous to the results presented above. [6]

Due to the limited number of speakers in this database the background data was supplemented with Switchboard-II data. The UBM and session transform were trained on a combination of Switchboard-II and Mixer data with approximately equal proportions. In contrast, the background data used for Z-Norm and T-Norm statistics were restricted to Mixer. Results for all three splits are combined in these results.

[6] The QUT 2004 protocol is available from the authors on request. It provides an expanded set of trials via a more efficient use of the data provide for the NIST 2004 SRE.

Table 2
Minimum DCF and EER of the baseline system and session variability modelling on Mixer data.

System	Raw Scores		Z-Norm		ZT-Norm	
	Min DCF	EER	Min DCF	EER	Min DCF	EER
1-Side						
Baseline	.0389	10.6	.0339	9.2	.0300	9.0
Session Modelling	.0358	8.7	.0242	6.0	.0211	5.4
3-Side						
Baseline	.0183	4.2	.0183	3.8	.0146	3.5
Session Modelling	.0119	2.8	.0108	2.2	.0093	2.1

Overall the advantage gained through session modelling for this data is less than for the Switchboard-II case. Relative improvements over the reference GMM-UBM system are approximately 30% and 36% at the minimum DCF operiating point for the 1-side and 3-side conditions, respectively, and 40% reduction in EER for both conditions when full score normalisation is applied. This performance is still a significant step forward and confirms the usefulness of explicitly modelling session variability.

Interestingly, the session modelling results are actually quite consistent across the different databases, with the absolute error rates and detection costs being very similar across the corpora both with and without score normalisation. It would seem that the reduced relative improvement gained with the session modelling is actually a result of better baseline performance. This is somewhat surprising due to the stated intention of the Mixer project to produce a more challenging dataset with a wider variety of mismatch (Martin et al., 2004).

The relatively modest improvements experienced in the 3-side training condition for Mixer data (36% minimum DCF improvement compared to 68% for Switchboard-II) combined with the known increase in the variety of channel conditions suggests that the session subspace may be saturated by the observed session variabilty for this data. Increasing the variation captured in the subspace may lead to further performance gains.

6.2 Session Subspace Size

All results so far have assumed a session variability subspace of dimension $R_z = 20$. Presented in Table 3 are results obtained by varying the dimension of the session variability subspace for the 1- and 3-side training conditions of the QUT 2004 protocol.

In (Vogt et al., 2005) the importance of severely constraining the dimension of

Table 3
Minimum DCF and EER results when varying the number of session subspace dimensions, R_z.

System	**1-Side**				**3-Side**			
	Raw Scores		**ZT-Norm**		**Raw Scores**		**ZT-Norm**	
	Min DCF	EER	Min DCF	EER	Min DCF	EER	Min DCF	EER
$R_z = 10$	.0355	8.8	.0230	6.2	.0128	3.1	.0107	2.3
$R_z = 20$	**.0358**	**8.7**	.0211	5.4	.0128	3.1	.0107	2.3
$R_z = 50$	.0391	9.4	**.0174**	**4.8**	**.0104**	**2.5**	**.0073**	**1.7**

the session variability subspace was noted, citing degrading performance comparing results from the $R_z = 50$ and $R_z = 20$ cases in the 1-side condition with no score normalisation. Further experiments revealed this to not necessarily be the case. As Table 3 shows, increasing R_z from 20 to 50 results in worse performance based on the raw output scores but after normalisation is applied the situation has reversed, with $R_z = 50$ giving both superior minimum DCF and EER.

For the 3-side condition the advantage of increasing the subspace size is clear as improved performance is gained for both measures with or without score normalisation.

The implications of this result are that increasing the power of the system's ability to model session variability can provide improved performance but score normalisation may be required to realise these benefits. This leads to the conclusion that the session variability modelling method produces inherently less calibrated raw scores than the reference GMM-UBM system with standard top-N ELLR scoring, particularly as R_z is increased.

It is also apparent that it is not always possible to make accurate conclusions about the comparative performance of different configurations after normalisation based on raw system scores alone.

6.3 Comparison of Training Methods

As noted in Section 3.4 there are several possibilities for the algorithm used to simultaneously optimise the set of variables $\{\boldsymbol{y}(s); \boldsymbol{z}_h(s), h = 1, \ldots, H(s)\}$ during speaker enrolment. Results comparing several configurations for the female portion of the QUT 2004 protocol are presented in Table 4.

The configurable parameters of interest in this experiment are the number of E-M iterations with either a single iteration or 5, and the Gauss-Seidel optimisation part of the algorithm. As described in Section 3.4, the G-S configurations presented are *Converged* with multiple G-S iterations, *Speaker First*

Table 4
Minimum DCF and EER for different configurations of the speaker and session variable estimation methods for the female subset of QUT 2004 protocol.

System	**1-Side**				**3-Side**			
	Raw Scores		**ZT-Norm**		**Raw Scores**		**ZT-Norm**	
	Min DCF	EER	Min DCF	EER	Min DCF	EER	Min DCF	EER
Multiple E-M Iterations								
Baseline	.0380	9.9	.0266	7.9	.0165	4.0	.0114	3.2
Independent	.0404	9.6	.0150	4.4	.0112	3.2	.0050	1.4
Session First	.0389	9.3	.0147	4.4	.0091	2.5	.0049	1.4
Converged	.0389	9.3	.0147	4.4	.0092	2.5	.0049	1.4
Speaker First	.0392	9.4	.0148	4.5	.0093	2.6	.0049	1.4
Single E-M Iteration								
Baseline	.0319	8.5	.0281	7.9	.0169	3.6	.0134	3.3
Independent	.0221	5.4	.0141	4.2	.0058	**1.6**	.0042	**1.2**
Session First	**.0219**	**5.1**	**.0138**	**4.0**	**.0054**	**1.6**	**.0040**	**1.2**
Converged	**.0219**	**5.1**	**.0138**	**4.0**	**.0054**	**1.6**	.0041	**1.2**

and *Session First* both with a single iteration of G-S and estimating either the speaker offset or the session variable first. Also included for contrast are *Baseline* with no session variability modelling and *Independent* with the speaker and session variables independently estimated on each E-M iteration. It is advantageous from a computing resources perspective to keep both of these parameters to a minimum.

Configurations with a single E-M iteration and multiple iterations are grouped together in Table 4 with the different Gauss-Seidel configurations described in Section 3.4 explored within these groups. As noted in Section 3.4, estimating the speaker vector does not converge quickly and is seemingly far from converging even after 20 iterations in the sense of finding a final optimal speaker offset, as shown in Figures 1 and 2. For comparison purposes it was therefore impractical to wait for full convergence and a maximum of five iterations was selected based on empirical knowledge from standard GMM-UBM systems (designated *Baseline* in the table above).

Interestingly, dropping back to only one iteration of the E-M procedure gives much better performance than using more iterations across the board for all session modelling variants; more than 40% reductions in both minimum DCF and EER were observed comparing the best five-iteration system to best one-iteration system based on unnormalised scores for the 1-side training condition. Similar results were observed for the 3-side condition. While single iteration training remained ahead after score normalisation was applied, the margin was significantly reduced.

The one-iteration result is quite interesting for two reasons. Firstly this result reverses the usual trend of improved and more consistent performance from multiple-iteration MAP adaptation seen in standard GMM-UBM systems (Pelecanos et al., 2002; Vogt et al., 2003). This result indicates that the overall performance of a GMM-UBM verification system is not necessarily improved by improving the accuracy by which the target models estimate the probability distribution of the targets' speech.

Secondly, in the case of only a single Gauss-Seidel iteration (labelled *Session First*), the speaker mean supervector is effectively trained on the *residual* variability that can not be explained in the session subspace as the session variables are estimated before the speaker.

The impact of the order in which the speaker and session variables are estimated seems to make minimal difference to the overall system performance as shown by comparing the results labelled *Session First* and *Speaker First* in Table 4, which both use only a single iteration of Gauss-Seidel optimisation. [7] Ensuring this optimisation has properly converged (*Converged* in Table 4) also seems irrelevant; there is virtually nothing to separate the fully converged estimate and a single iteration of the session first estimate.

Finally, enrolment using independent optimisation of the speaker and session variables results in only a small degradation in performance compared to the Gauss-Seidel methods, as can be seen by observing the results for the systems labelled *Independent* in Table 4.

Using the results of this section, the performance of an optimised system using the session modelling techniques of this chapter was compared to the baseline system for the QUT 2004 protocol for both the 1- and 3-side training conditions. With a minimum DCF of .0158 and EER of 4.2% for the 1-side condition, this translates to relative reductions of 47% and 53% compared to the baseline system. The performance improvements in the 3-side condition are more impressive with 56% and 58% reductions in detection cost and EER respectively with absolute values of .0064 and 1.5%.

Table 5 demonstrate the performance of this system for the common evaluation condition of the NIST SRE 2005 protocol. Relative improvements in minimum DCF were achieved for this protocol that are very similar to the QUT 2004 results in both the 1- and 3-side conditions. The reductions in EER were also large although slightly less than for QUT 2004. This system is believed to be the best performing individual system submitted to NIST for evaluation in

[7] It should be noted that the results for *Speaker First* with one iteration are intentionally absent as this configuration will produce identical results to the single-iteration *Independent* system as the estimates of the session variables do not have an opportunity to feed back into the speaker variable estimate.

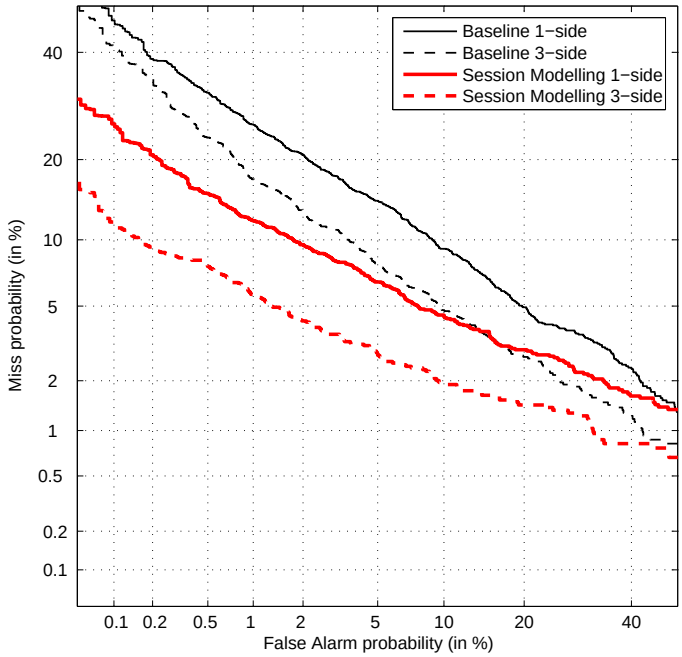

Fig. 5. DET of the English-only 1- and 3-side training conditions of the NIST SRE 2005 protocol comparing an optimised session modelling system with a baseline GMM-UBM system with score normalisation applied to both.

Table 5
Comparison of minimum DCF and EER of session modelling and baseline systems with ZT-Norm for common evaluation condition of the NIST 2005 protocol.

System	**1-Side**		**3-Side**	
	Min DCF	EER	Min DCF	EER
Baseline	.0352	9.5	.0267	6.6
Session Modelling	.0197	6.1	.0110	3.4

the 2005 SRE. [8]

6.4 Sensitivity to the Session Variability Subspace

Two aspects of performance sensitivity to the training of the session variability subspace transform $\boldsymbol{U}$ are of practical interest. Firstly, the impact of the number of E-M iterations will be investigated as the E-M training algorithm is very computationally expensive and also appears to converge quite slowly. Also, the issue of database mismatch is an important consideration as the training database for an application does not typically match the situation it is applied to. The results of these experiments are summarised in Table 6.

[8] This claim cannot be substantiated as not all sites reported the results for the individual systems that were combined for final submission, however, few sites produced *fused* results with comparable or better performance.

Table 6
Minimum DCF and EER results with varying degrees of convergence in the session variability subspace training.

System	**1-Side**				**3-Side**			
	Raw Scores		**ZT-Norm**		**Raw Scores**		**ZT-Norm**	
	Min DCF	EER	Min DCF	EER	Min DCF	EER	Min DCF	EER
Switchboard-II	.0257	6.7	.0200	5.5	.0089	2.3	.0071	1.8
1 iteration	.0247	6.1	.0178	4.9	.0076	2.0	0059	1.6
2 iterations	.0238	5.7	.0168	4.6	.0071	1.9	.0055	1.5
5 iterations	.0226	5.3	.0149	4.3	.0059	1.6	.0044	1.3
10 iterations	.0219	**5.1**	.0138	**4.0**	**.0054**	**1.6**	**.0040**	**1.2**
20 iterations	**.0213**	**5.1**	**.0134**	**4.0**	**.0054**	**1.6**	.0041	**1.2**

Contrary to the conclusions drawn by Kenny et al. (2005b), the proposed method gains significantly from allowing the E-M algorithm for training the subspace to converge, especially in the 1-side training condition. Furthermore, there does appear to be considerable sensitivity to the nature of the data used to train the subspace transform as the results using the transform trained solely on Switchboard-II data demonstrated degraded performance compared to using Mixer data (comparing the system labelled *Switchboard-II* in Table 6 to the other systems). Using Switchboard data still performs favourably compared to the reference system with no session variability modelling, again demonstrating the utility of the method. The results in Table 6 also demonstrate diminishing returns with more than 10 iterations of the E-M algorithm.

6.5 Reduced Test Utterance Length

An important part of the session modelling method is estimating the session vector $\boldsymbol{z}$ for the test utterance. While this is a low-dimensional variable, estimating it accurately will require a sufficient quantity of speech. This experiment aims to determine the minimum requirements for extracting improved results from session modelling.

Figure 6 shows the impact of reducing the test utterance length for both the session variability modelling method and standard GMM-UBM modelling with test utterance lengths of 5, 10 and 20 seconds of active speech.

These results indicate that approximately 10 seconds of speech are required to estimate the session factors sufficiently accurately to produce improved results over standard modelling and scoring practice, while 20-second trials produce advances in performance approaching those experienced with full-length testing utterances, with relative improvements of over 20% in both minimum DCF and EER. In fact the 20-second session modelling results out

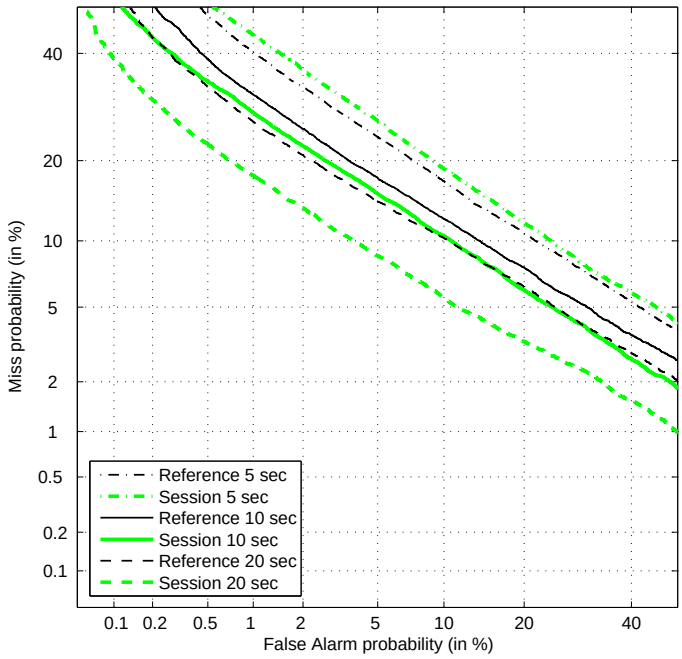

Fig. 6. DET plot for the 1-side training condition comparing baseline and session modelling results for short test utterance lengths on the QUT 2004 protocol.

perform the baseline system using full verification utterances with an average of more than 100 seconds of active speech.

7 Summary

In this article a technique was proposed to compensate for mismatch experienced in text-independent speaker verification due to inter-session variability. Explicit modelling of the prevalent conditions in training and verification sessions was introduced by adding a session-dependent variable to the speaker modelling process that was constrained to lie in a session variation subspace. Techniques were developed to incorporate this augmented model into both the speaker enrolment and verification phases of a GMM-UBM verification system.

The enrolment process involved the simultaneous optimisation of the speaker mean vector and additional session vectors for each session available for enrolment according to a maximum *a posteriori* criterion on each variable. Due to the model complexity, a direct solution to the simultaneous optimisation was shown to be very computationally expensive to the point of being impractical for large verification trials, such as a NIST evaluation. To avoid this issue, a novel iterative approximation method was proposed based on the Gauss-Seidel method for solving linear systems.

Methods for training the session variability subspace based on a database of background speakers were also described. The sensitivity of verification performance to insufficient convergence of this training was empirically investigated, as was the issue of mismatched conditions between training and testing databases.

Experiments on conversational telephony data demonstrated the effectiveness of the technique for both single and multiple training session conditions with up to 68% reduction in detection cost over a standard GMM-UBM system and significant improvements over a system utilising feature mapping. It was also observed that the session variability modelling responds particularly well to score normalisation with the Z-Norm and T-Norm approaches.

References

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Processing 10 (1/2/3), 42–54.

Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., der Vorst, H. V., 1994. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition. SIAM, Philadelphia, PA.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, California, USA.

Gauvain, J.-L., Lee, C.-H., 1996. Bayesian adaptive learning and MAP estimation of HMM. In: Lee, C.-H., Soong, F., Paliwal, K. (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic, Boston, Mass, pp. 83–107.

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

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2005b. Factor analysis simplified. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Vol. 1. pp. 637–640.

Kenny, P., Dumouchel, P., 2004. Experiments in speaker verification using factor analysis likelihood ratios. In: Odyssey: The Speaker and Language Recognition Workshop. pp. 219–226.

Lucey, S., Chen, T., 2003. Improved speaker verification through probabilistic subspace adaptation. In: Eurospeech. pp. 2021–2024.

Martin, A., Miller, D., Przybocki, M., Campbell, J., Nakasone, H., 2004. Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004. In: International Conference on Language Resources and Evaluation. pp. 587–590.

Mason, M., Vogt, R., Baker, B., Sridharan, S., 2004. The QUT NIST 2004 speaker verification system: A fused acoustic and high-level approach. In: Australian International Conference on Speech Science and Technology. pp. 398–403.

Mason, M., Vogt, R., Baker, B., Sridharan, S., 2005. Data-driven clustering for blind feature mapping in speaker verification. In: Eurospeech. pp. 3109–3112.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. In: A Speaker Odyssey, The Speaker Recognition Workshop. pp. 213–218.

Pelecanos, J., Vogt, R., Sridharan, S., 2002. A study on standard and iterative MAP adaptation for speaker recognition. In: International Conference on Speech Science and Technology. pp. 190–195.

Press, W. H., 1992. Numerical recipes in C: the art of scientific computing, 2nd Edition. Cambridge University Press, Cambridge.

Reynolds, D., 1997. Comparison of background normalizaftion methods for text-independent speaker verification. In: Eurospeech. Vol. 2. pp. 963–966.

Reynolds, D., 2003. Channel robust speaker verification via feature mapping. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 2. pp. 53–56.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10 (1/2/3), 19–41.

Teunen, R., Shahshahani, B., Heck, L., 2000. A model-based transformational approach to robust speaker recognition. In: International Conference on Spoken Language Processing. Vol. 2. pp. 495–498.

Vogt, R., Baker, B., Sridharan, S., 2005. Modelling session variability in text-independent speaker verification. In: Interspeech. pp. 3117–3120.

Vogt, R., Pelecanos, J., Sridharan, S., 2003. Dependence of GMM adaptation on feature post-processing for speaker recognition. In: Eurospeech. pp. 3013–3016.

Vogt, R., Sridharan, S., 2004. Bayes factor scoring of GMMs for speaker verification. In: Odyssey: The Speaker and Language Recognition Workshop. pp. 173–178.