INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Distribution Free Formulation of the Total Variability Model

Ruchir Travadi and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California
travadi@usc.edu, shri@sipi.usc.com

Abstract

The Total Variability Model (TVM) [1] has been widely used in audio signal processing as a framework for capturing differences in feature space distributions across variable length sequences by mapping them into a fixed-dimensional representation. Its formulation requires making an assumption about the source data distribution being a Gaussian Mixture Model (GMM). In this paper, we show that it is possible to arrive at the same model formulation without requiring such an assumption about distribution of the data, by showing asymptotic normality of the statistics used to estimate the model. We highlight some connections between TVM and heteroscedastic Principal Component Analysis (PCA), as well as the matrix completion problem, which lead to a computationally efficient formulation of the Maximum Likelihood estimation problem for the model.
Index Terms: Total Variability Model, ivector

1. Introduction

The Total Variability Model (TVM) is a popular framework used in audio signal processing, where the variability arising in the distribution of feature vectors across different sequences is captured in a fixed dimensional vector representation. It has been used in a wide variety of applications including speaker recognition [1], language identification [2, 3], acoustic model adaptation for speech recognition [4, 5], and also for inferring paralinguistic information such as cognitive load [6].

TVM is formulated as a generative model, where the distribution of feature vectors is assumed to be a Gaussian Mixture Model (GMM), with its mean parameters varying across different sequences (Section 2). This assumption seems natural when we view the model from a historical perspective - models that preceded TVM in the domain of speaker recognition were based upon the assumption of data distribution being a GMM. Early efforts in this area began by modeling the distribution of every speaker as a separate GMM [7]. This was followed by the method of training a combined GMM over all training data, also known as a Universal Background Model (UBM), and then adapting it to different speakers, in order to solve the problem of data sparsity [8].

Later, generative models began to be used as a front-end for discriminative methods. Initially, the statistics used for UBM adaptation were stacked into the so-called supervectors, and used as input features for a Support Vector Machine (SVM) [9]. This paved the way for Joint Factor Analysis (JFA) [10], which attempted to reduce the dimensionality of supervectors, and separate source and channel factors into smaller dimensional vectors. That eventually led to TVM formulation [1], where both source and channel variability are captured within a single small dimensional vector, on which variability compensation methods are applied subsequently.

In this paper, we show that it in order to formulate TVM, it is not necessary to make specific assumptions about the form of distribution over the feature vectors (Section 3). This involves a change in perspective - instead of formulating TVM as a generative model over *feature vectors*, we formulate it as a generative model over the *statistics* derived from those feature vectors. Asymptotically, the distribution of statistics can be shown to be Gaussian, regardless of the actual distribution of the feature vectors. Then, by assuming that the expected values of these statistics lie along a linear subspace (equivalent to making the subspace assumption on mean supervectors in conventional TVM), it is possible to arrive at a model formulation that is equivalent to conventional TVM.

This distribution free formulation leads to connections between TVM and heteroscedastic PCA as well as the matrix completion problem [11] (Section 4). One important consequence of these connections is that the Maximum Likelihood problem for obtaining the Total Variability matrix could be posed in a form that can be solved in a computationally efficient manner by means of a stochastic subgradient descent (SSGD) algorithm (Section 5).

In addition, the distribution free formulation generalizes TVM, allowing it to stay theoretically valid for statistics derived using arbitrary posterior weights. Indeed, methods have already been proposed in literature where posteriors used to derive statistics in TVM are obtained from sources other than a UBM [12, 13, 14]. The formulation provided in this paper provides a theoretical justification, validating the procedure used in these methods. We propose a few other possibilities that can be considered for future research based on this formulation in Section 7.

2. Conventional TVM Formulation

Let $\mathbf{X} = \{\mathbf{X}_u\}_{u=1}^{U}$ be the collection of acoustic feature vectors in a dataset comprising U utterances, where $\mathbf{X}_u = \{\boldsymbol{x}_{ut}\}_{t=1}^{T_u}$ denotes the feature vector sequence of length T_u from a specific utterance u. Let D be the dimensionality of each feature vector: $\boldsymbol{x}_{ut} \in \mathbb{R}^D$.

In the Total Variability Model (TVM), it is assumed that with every utterance u, there is an associated vector $\boldsymbol{w}_u \in \mathbb{R}^K$, known as the *ivector* for that utterance, such that the conditional distribution of $\boldsymbol{x}_{ut}$ given $\boldsymbol{w}_u$ is a Gaussian Mixture Model (GMM) with C components, and parameters $\{p_c, \boldsymbol{\mu}_{uc} = \boldsymbol{\mu}_c + \mathbf{T}_c\boldsymbol{w}_u, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$ where $p_c \in \mathbb{R}, \boldsymbol{\mu}_c \in \mathbb{R}^D, \mathbf{T}_c \in \mathbb{R}^{D\times K}$ and $\boldsymbol{\Sigma}_\mathbf{c} \in \mathbb{R}^{D\times D}$. The prior distribution for $\boldsymbol{w}_u$ is assumed to be standard normal:

$$\boldsymbol{w}_u \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Let $\boldsymbol{M}_0, \boldsymbol{M}_u \in \mathbb{R}^{CD}$, known as supervectors, denote vectors consisting of stacked global and utterance-specific component means $\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_{uc}$ respectively. Then, TVM can be summarized as:

$$\boldsymbol{M}_u = \boldsymbol{M}_0 + \mathbf{T}\boldsymbol{w}_u$$

where $\mathbf{T} \in \mathbb{R}^{CD\times K}$ is given as: $\mathbf{T} = \left[\begin{array}{c:c:c:c} \mathbf{T}_1^\mathsf{T} & \dots & \mathbf{T}_C^\mathsf{T} \end{array}\right]^\mathsf{T}$

http://dx.doi.org/10.21437/Interspeech.2017-219

3. Distribution Free Formulation

In this section, we show that it is possible to drop the requirement on data distribution being a GMM from the TVM formulation. Instead of formulating TVM as a model for *observed data*, we show that it can be formulated as a model for *observed statistics*, which asymptotically follow a Gaussian distribution. In particular, we show that:

1. The distribution of Baum-Welch statistics used in TVM parameter estimation is asymptotically Gaussian, regardless of the feature space distribution (Section 3.1).
2. The likelihood function for the *observed statistics* under the asymptotic Gaussian approximation is similar to the likelihood function of *observed data* under GMM assumption in conventional TVM, where the difference between the two likelihood expressions does not depend on the subspace matrix $\mathbf{T}$ (Section 3.2).
3. The expression for posterior distribution of the ivector obtained from the distribution free formulation is also the same as that in conventional TVM (Section 3.3).

3.1. Asymptotic Distribution of Baum Welch Statistics

Let $\Gamma : \mathbb{R}^D \mapsto [0,\infty)^C$ be an arbitrary function, and $\gamma_{utc} = \Gamma_c(\boldsymbol{x}_{ut})$, where Γ_c denotes the c^{th} component of Γ. For example, one possible choice for Γ could correspond to component posterior probabilities obtained from a GMM: $\gamma_{utc} = p(c|\boldsymbol{x}_{ut})$. Let $p_{uc},\ \boldsymbol{\mu}_{uc},\ \boldsymbol{\Sigma}_{uc}$ define the following expected values (with respect to the distribution of $\boldsymbol{x}_{ut}$):

$$p_{uc} = \mathbb{E}\left[\gamma_{utc}\right], \quad \boldsymbol{\mu}_{uc} = \frac{1}{p_{uc}}\mathbb{E}\left[\gamma_{utc}\,\boldsymbol{x}_{ut}\right]$$

$$\boldsymbol{\Sigma}_{uc} = \frac{1}{p_{uc}}\mathbb{E}\left[(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{\mu}_{uc})(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{\mu}_{uc})^{\mathsf{T}}\right]$$

If γ_{utc} were to correspond to GMM component posterior probabilities $p(c|\boldsymbol{x}_{ut})$, then $\boldsymbol{\mu}_{uc}$ would denote the component GMM means. Let statistics $\boldsymbol{N}_u$, $\boldsymbol{F}_u$ be defined as:

$$N_{uc} = \sum_{t=1}^{T_u}\gamma_{utc} \qquad \boldsymbol{N}_u = [\,N_{u1}\ \ldots\ N_{uC}\,]$$

$$\boldsymbol{F}_{uc} = \frac{1}{N_{uc}}\sum_{t=1}^{T_u}\gamma_{utc}\,\boldsymbol{x}_{ut} \qquad \boldsymbol{F}_u = \left[\begin{array}{c:c:c} \boldsymbol{F}_{u1}^{\mathsf{T}} & \ldots & \boldsymbol{F}_{uC}^{\mathsf{T}} \end{array}\right]^{\mathsf{T}}$$

We can rearrange the expression for $\boldsymbol{F}_{uc}$ as follows:

$$\begin{aligned}\boldsymbol{F}_{uc} &= \frac{1}{N_{uc}}\sum_{t=1}^{T_u}\gamma_{utc}\,\boldsymbol{x}_{ut} \\ &= \frac{1}{\sqrt{N_{uc}}}\sqrt{\frac{T_u}{N_{uc}}}\left[\frac{1}{\sqrt{T_u}}\sum_{t=1}^{T_u}(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{\mu}_{uc})\right] + \\ &\quad \frac{T_u p_{uc}}{N_{uc}}\boldsymbol{\mu}_{uc}\end{aligned}$$

By the multivariate Central Limit Theorem (CLT), as $T_u \to \infty$

$$\frac{1}{\sqrt{T_u}}\sum_{t=1}^{T_u}(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{\mu}_{uc}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, p_{uc}\,\boldsymbol{\Sigma}_c) \quad (1)$$

where $\xrightarrow{d}$ denotes convergence in distribution. Similarly,

$$\sqrt{\frac{T_u}{N_{uc}}} \xrightarrow{d} \frac{1}{\sqrt{p_{uc}}}, \quad \frac{T_u\,p_{uc}}{N_{uc}} \xrightarrow{d} 1 \quad (2)$$

By Slutsky's theorem, it follows that:

$$\sqrt{\frac{T_u}{N_{uc}}}\frac{1}{\sqrt{T_u}}\sum_t(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{\mu}_{uc}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_c) \quad (3)$$

Thus, the statistics $\boldsymbol{F}_{uc}$ are asymptotically distributed normally, regardless of the distribution of the feature vectors $\boldsymbol{x}_{ut}$ (as long as it satisfies the conditions for the multivariate CLT to hold), and the choice of the function Γ:

$$\boldsymbol{F}_{uc}|N_{uc} \sim \mathcal{N}(\boldsymbol{\mu}_{uc}, N_{uc}^{-1}\boldsymbol{\Sigma}_c) \quad (4)$$

3.2. Likelihood Function

Similar to conventional TVM, we assume that the covariance matrices are tied across all utterances: $\boldsymbol{\Sigma}_{uc} = \boldsymbol{\Sigma}_c$, and that the mean supervectors lie along a subspace $\mathbf{T}$, with a standard normal prior on ivectors $\boldsymbol{w}_u$:

$$\boldsymbol{\mu}_{uc} = \boldsymbol{\mu}_{0c} + \mathbf{T}_c\,\boldsymbol{w}_u, \quad \boldsymbol{w}_u \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

We also assume that the component-wise statistics $\boldsymbol{F}_{uc}$ are conditionally independent of each other given the ivector $\boldsymbol{w_u}$ and the statistics $\boldsymbol{N}_u$. Let $\Theta = \{\mathbf{T}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_C\}$ denote the parameter space of the model. Then, the total log likelihood of all observed statistics $\{\boldsymbol{F}_u\}_{u=1}^U$ given $\{\boldsymbol{N}_u\}_{u=1}^U$ and Θ, marginalized over the ivectors is given as:

$$\mathcal{L}(\Theta) = \sum_{u=1}^{U}\log\left[\,\mathbb{E}_{\boldsymbol{w}_u}\left[f(\boldsymbol{F}_u|\boldsymbol{N}_u, \boldsymbol{w}_u, \Theta)\,\right]\right] \quad (5)$$

where f denotes the conditional likelihood of $\boldsymbol{F}_u$ given $\boldsymbol{N}_u, \boldsymbol{w}_u$ and the parameters Θ. The expression in equation (5) can be simplified to:

$$\mathcal{L}(\Theta) = \mathcal{L}_1(\Theta) + \mathcal{L}_2(\Theta) + \mathcal{L}_3(\Theta) \quad (6)$$

where $\mathcal{L}_1(\Theta),\ \mathcal{L}_2(\Theta),\ \mathcal{L}_3(\Theta)$ are given as:

$$\mathcal{L}_1(\Theta) = -\frac{1}{2}\sum_{u=1}^{U}\left[\boldsymbol{F}_{u0}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\boldsymbol{F}_{u0} + \sum_{c=1}^{C}\log|\boldsymbol{N}_u^{-1}\boldsymbol{\Sigma}_c|\right]$$

$$\mathcal{L}_2(\Theta) = \frac{1}{2}\sum_{u=1}^{U}\boldsymbol{F}_{u0}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{T}\left(\mathbf{I} + \mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\mathbf{T}\right)^{-1}\mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_{u0}$$

$$\mathcal{L}_3(\Theta) = -\frac{1}{2}\sum_{u=1}^{U}\log|\mathbf{I} + \mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\mathbf{T}|$$

Here, $\boldsymbol{F}_{u0} = \boldsymbol{F}_u - \boldsymbol{M}_0$, and $\boldsymbol{\Sigma}, \mathbf{N}_u$ are $CD \times CD$ block diagonal matrices with c^{th} block given by $\boldsymbol{\Sigma}_c$ and $N_{uc}\,\mathbf{I}$. It can be seen that the terms $\mathcal{L}_2(\Theta)$, $\mathcal{L}_3(\Theta)$ are identical to the corresponding terms in the expression of data likelihood in conventional TVM ([15], equation (4)). The term $\mathcal{L}_1(\Theta)$ differs between the two expressions, but it does not depend on $\mathbf{T}$ (however, it does depend on $\boldsymbol{\Sigma}$).

3.3. Posterior distribution of the ivector

Under the model assumptions made in Section 3.2, the posterior distribution of the ivector given statistics $\boldsymbol{F}_u, \boldsymbol{N}_u$ is given as:

$$\boldsymbol{w}_u|\boldsymbol{F}_u, \boldsymbol{N}_u, \Theta \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}_u}\mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\boldsymbol{F}_{u0}, \boldsymbol{\Sigma}_{\boldsymbol{w}_u}\right) \quad (7)$$

where $$\boldsymbol{\Sigma}_{\boldsymbol{w}_u} = \left(\mathbf{I} + \mathbf{T}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\mathbf{T}\right)^{-1}$$

which is identical to the expression for MAP estimate of the ivector in the conventional TVM.

4. Connections to Heteroscedastic PCA and Matrix Completion

4.1. Heteroscedastic PCA

It is well known that PCA can be interpreted in a probabilistic manner, by viewing the observed data as being generated by addition of white Gaussian noise to a vector lying along a low dimension subspace [16]. In case of TVM, the asymptotic distribution of statistics can be interpreted similarly, by viewing the statistic $\boldsymbol{F}_u - \boldsymbol{M}_0$ as being generated by addition of Gaussian noise to its expected value $\mathbf{T}\boldsymbol{w}_u$ that lies along the subspace $\mathbf{T}$:

$$\boldsymbol{F}_{uc} - \boldsymbol{\mu}_{0c} = \mathbf{T}_c \boldsymbol{w}_u + \boldsymbol{\epsilon}_{uc}, \quad \boldsymbol{\epsilon}_{uc} \sim \mathcal{N}(\mathbf{0}, N_{uc}^{-1}\boldsymbol{\Sigma}_c) \tag{8}$$

However, the additive noise in this case has a heteroscedastic nature, where the covariance matrices vary across different component dimensions and across different utterances, being inversely proportional to the value of N_{uc}.

4.2. Matrix Completion

Matrix completion is the problem of estimating unknown entries of a matrix from a few known entries [11]. It is common in such a case to impose a low rank assumption on the matrix involved. It can therefore be viewed as a special case of heteroscedastic PCA, where the columns of the low rank matrix correspond to vectors along a subspace, and their observed values are corrupted by a heteroscedastic noise, whose variance is zero in known entries (or perhaps a small constant, as in [17]) and infinite in unknown entries. The estimation of Total Variability matrix $\mathbf{T}$ is a similar problem, where the expected value of the matrix of statistics $\mathbf{F}_0 = \left[\, \boldsymbol{F}_{10} \,\vdots\, \ldots \,\vdots\, \boldsymbol{F}_{U0} \,\right]$ is a low rank matrix $\mathbf{M} = \left[\, \boldsymbol{T}\boldsymbol{w}_1 \,\vdots\, \ldots \,\vdots\, \mathbf{T}\boldsymbol{w}_U \,\right] = \mathbf{T}\mathbf{W}$, and we are interested in recovering this low rank structure from noisy observations. However, the covariance of added noise in this case, unlike the matrix completion problem, is somewhere between zero and infinity, scaled by a variable factor N_{uc}^{-1}.

4.3. TVM Parameter Estimation

One of the problems associated with Maximum Likelihood estimation of $\mathbf{T}$ is that the log likelihood function given in (6) is non-convex in $\mathbf{T}$. However, from the perspective of connection established between TVM and matrix completion, we could potentially break down the task of estimating $\mathbf{T}$ into two parts. First, we could recover $\mathbf{M}$ by maximizing the likelihood of $\mathbf{F}_0$ given $\mathbf{M}$, while constraining $\mathbf{M}$ to be low-rank:

$$\begin{aligned} \underset{\mathbf{M}}{\text{minimize}} \quad & \sum_{u=1}^{U} (\boldsymbol{F}_{u0} - \boldsymbol{M}_u)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_u (\boldsymbol{F}_{u0} - \boldsymbol{M}_u) \\ \text{subject to} \quad & rank(\mathbf{M}) \leq k \end{aligned}$$

Then, to recover $\mathbf{T}$, we just need an appropriate factorization of $\mathbf{M} = \mathbf{T}\mathbf{W}$. Let $\mathbf{M} = \mathbf{U}_M \mathbf{D}_M \mathbf{V}_M^{\mathsf{T}}$ be the truncated Singular Value Decomposition (SVD) of $\mathbf{M}$. We can choose :

$$\mathbf{T} = \frac{1}{\sqrt{U-1}} \mathbf{U}_M \mathbf{D}_M, \quad \mathbf{W} = \sqrt{U-1}\mathbf{V}_M^{\mathsf{T}}$$

This obviously satisfies $\mathbf{M} = \mathbf{T}\mathbf{W}$. In addition, if the columns of $\mathbf{M}$ are zero mean, it can be easily verified that columns of $\mathbf{W}$ are also zero mean and have an identity covariance, satisfying the statistical prior assumptions made on ivectors.

The advantage of this procedure is that the optimization objective is convex in $\mathbf{M}$, with a gradient expression that is easy to compute, allowing us to make use of efficient gradient descent methods for the purpose of optimization. We describe an optimization algorithm for this problem in the next section.

5. Optimization Algorithm

The gradient of the objective with respect to $\boldsymbol{M}_u$ is:

$$\nabla_{\boldsymbol{M}_u}(obj) = 2\,\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_u(\boldsymbol{F}_{u0} - \boldsymbol{M}_u) \tag{9}$$

Although the objective is convex in $\mathbf{M}$, the rank constraint makes the overall optimization problem non-convex. There are two ways to deal with this issue:

1. **Projected Gradient Descent**: After each update, project the current estimate back to rank k, by truncating the SVD to k largest singular values.
2. **Nuclear Norm Regularization**: Substitute the rank constraint with a constraint on the nuclear norm, as is commonly done in matrix completion. Then, by introducing a lagrangian for the constraint, the problem can be formulated as:

$$\underset{\mathbf{M}}{\text{minimize}} \sum_{u=1}^{U} (\boldsymbol{F}_{u0} - \boldsymbol{M}_u)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_u (\boldsymbol{F}_{u0} - \boldsymbol{M}_u) + \lambda |\mathbf{M}|_\star$$

There is a common issue associated with both approaches: the operation of projecting to a low rank space, as well as that of computing the subgradient of nuclear norm, requires obtaining an SVD factorization of $\mathbf{M}$, which is computationally expensive. However, this can be handled efficiently by the method of stochastic subgradient descent suggested in [18], which efficiently maintains a low-rank SVD factorization of $\mathbf{M}$ at every step. The SVD factorization after a gradient update can be efficiently computed from the one prior to the update, by solving small SVD and QR factorization problems. While the method in [18] is proposed for solving the second type of formulation, the first formulation could be solved using the same method, by simply setting the nuclear norm regularizer λ to zero.

We conducted experiments using three choices for $\boldsymbol{\Sigma}$: $\boldsymbol{\Sigma}_{UBM}$, which is given by the UBM, $\boldsymbol{\Sigma}_1$ given below, and $\boldsymbol{\Sigma}_2$ which is a sample estimate of the expected value in Section 3.1:

$$\boldsymbol{\Sigma}_1 = \frac{1}{N_c} \sum_{u=1}^{U} \sum_{t=1}^{T_u} \gamma_{utc} \left(\boldsymbol{x}_{ut} - \boldsymbol{F}_{uc}\right) \left(\boldsymbol{x}_{ut} - \boldsymbol{F}_{uc}\right)^{\mathsf{T}}$$

$$\boldsymbol{\Sigma}_2 = \sum_{u=1}^{U} \sum_{t=1}^{T_u} \frac{(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{F}_{uc})(\gamma_{utc}\,\boldsymbol{x}_{ut} - p_{uc}\,\boldsymbol{F}_{uc})^{\mathsf{T}}}{U\,N_{uc}}$$

5.1. Modifications Introduced

The algorithm in [18] is presented with matrix completion as the target application. However, one of the differences in our case is that the gradient is proportional to entries in $\boldsymbol{N}_u$, which can have a wide dynamic range across utterances. In order to deal with this issue, we modify the objective and the sampling process to avoid instability. We use the following gradient:

$$\nabla_{\boldsymbol{M}_u}(obj) = 2\,\boldsymbol{\Sigma}^{-1}\boldsymbol{P}_u(\boldsymbol{F}_{u0} - \boldsymbol{M}_u) \tag{10}$$

where $\boldsymbol{P}_u = \frac{1}{T_u}\boldsymbol{N}_u$, and instead of sampling a batch randomly, we sample with a weight proportional to T_u. The resulting process still produces an unbiased estimate of the original gradient, but avoids the instability introduced due to dynamic range of $\boldsymbol{N}_u$. In addition, we also used gradient clipping to avoid instability when gradient magnitude is large. The remaining details of the algorithm can be obtained from the description of the procedure SSGD-Matrix-Completion in [18].

6. Experimental Results

6.1. Experimental Setup and Database

We conducted experiments for the task of Speaker Recognition (SRE) on the NIST SRE 2008 database, for the short2-short3 evaluation condition. The overall setup consisted of extracting the ivectors as a front-end, followed by a Probabilistic Linear Discriminant Analysis (PLDA) classifier.

For each frame, we obtained 20-dimensional MFCC vectors, concatenated with delta and delta-delta coefficients. A UBM of 2048 components, and TVM with ivector dimension of 400 were trained on a combination of SRE 04-06, Switchboard Cellular and Fisher data. We compared our training method with two other approaches proposed in literature: the EM algorithm, and the method of Randomized SVD (RSVD) proposed in [15] as a faster alternative to EM. For our proposed algorithm, we chose the method of projected gradient descent described in Section 5, setting the nuclear norm regularization coefficient $\lambda = 0$.

The EM and RSVD training algorithms were implemented using the Kaldi toolkit [19]. The proposed algorithm was implemented using MATLAB, except for the SVD initialization of $\mathbf{M}$, which was implemented in Kaldi. All implementations were parallelized over 24 processors.

6.2. Results

The results on male and female trials, for all eight subsets of the core test trials specified in SRE 08 are presented below in Tables 1 and 2 respectively. Here, T_θ and T_w correspond to the total amount of time taken for parameter estimation and ivector extraction respectively.

Table 1: *Equal Error Rate (%) for Male Trials*

	EM	**RSVD**	**SSGD**		
T_θ	10.5 hrs	**30 min**	2 hrs		
T_w	1.13 sec	**0.12 sec**	1.13 sec		
Subset	*Covariance Matrix*				
	$\mathbf{\Sigma}_{EM}$	$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_{UBM}$	$\mathbf{\Sigma}_2$
1	**4.98**	9.14	7.61	7.75	9.02
2	**0.81**	1.21	1.21	1.61	2.02
3	**5.07**	9.44	7.78	7.82	9.13
4	**5.01**	9.80	5.01	5.01	6.83
5	**4.69**	9.84	5.94	5.94	8.12
6	**5.15**	6.75	6.18	6.75	6.75
7	**3.64**	5.24	3.87	4.33	4.78
8	2.63	3.07	2.63	**2.19**	3.07
Avg	**4.00**	6.81	5.03	5.18	6.22

From comparing EM and RSVD, it can be seen that although RSVD estimation is much faster than EM, there is a significant increase in EER. While the EER results on Language Identification (LID) task in [15] were very similar for EM and RSVD, the same is not true for this SRE task. One of the reasons is that the estimation procedure for RSVD relies on making an assumption $\boldsymbol{N}_u \approx T_u\, p_c$. However, we found that this assumption was not met for our training data in this experiment.

The stochastic subgradient descent (SSGD) approach resulted in a mid-way performance between EM and RSVD. both in terms of time taken for estimation, as well as the Equal Error Rate (EER) obtained. This trend was observed almost uniformly across all 8 trial subsets.

The reason for a drop in performance between EM and SSGD can be attributed to the sub-optimal choices of $\mathbf{\Sigma}$: The EM algorithm optimizes the likelihood jointly with respect to $\mathbf{T}$ as well as $\mathbf{\Sigma}$, whereas SSGD uses a fixed value of $\mathbf{\Sigma}$ to estimate $\mathbf{T}$. It is possible to modify the proposed algorithm for and efficient joint optimization of $\mathbf{T}$ and $\mathbf{\Sigma}$, but our efforts in this direction are in an early stage, and we consider it as a part of future work.

Table 2: *Equal Error Rate (%) for Female Trials*

	EM	**RSVD**	**SSGD**		
T_θ	13.5 hrs	**40 min**	2.5 hrs		
T_w	1.13 sec	**0.12 sec**	1.13 sec		
Subset	*Covariance Matrix*				
	$\mathbf{\Sigma}_{EM}$	$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_{UBM}$	$\mathbf{\Sigma}_2$
1	**6.34**	10.41	10.59	10.60	10.75
2	**1.19**	1.79	1.79	1.49	1.49
3	**6.39**	10.63	10.72	10.68	10.85
4	**6.46**	12.16	9.01	9.31	8.71
5	**6.61**	13.22	7.81	7.69	9.86
6	**6.26**	8.65	7.82	7.93	7.54
7	**4.06**	5.45	5.32	5.32	5.58
8	**4.74**	6.05	6.05	5.79	5.79
Avg	**5.26**	8.55	7.39	7.35	7.57

7. Conclusion and Future Work

We have introduced a formulation of TVM that does not rely on assuming a specific form of distribution on the feature vectors, or make any assumption about the source of posteriors used for obtaining the statistics, by showing that the statistics used in TVM estimation follow a Gaussian distribution asymptotically.

The formulation also leads to connections between TVM and other problems such as heteroscedastic PCA and matrix completion. We have presented an algorithm for estimation of the total variability matrix that is based on its connection to the matrix completion problem. While the algorithm speeds up the estimation process significantly compared to EM, the sub-optimal choice of covariance matrix causes a slight drop in performance, which we plan to address in our future work.

The repercussions of the distribution free formulation are wider than just the connections and the training algorithm discussed in this paper. It provides a theoretical justification for previous work in literature where the statistics for TVM were derived from a Deep Neural Network (DNN) rather than the conventional approach of using a UBM [12, 13, 14]. However, these methods mostly rely upon DNNs trained for Automatic Speech Recognition (ASR) to get the posterior weights. Instead, weights γ_{utc} could potentially be optimized in a data driven discriminative fashion to magnify the desired source of variability, and suppress the undesired sources.

In addition, this formulation could potentially be exploited to address the problem of data sparsity for short test segments. In conventional TVM, posteriors γ_{utc} form a soft partitioning of the input space, since they sum to 1 for every frame. However, it is possible to use weights that form overlapping regions instead. For a fixed value of segment length, this would result in higher values of N_{uc}, thereby reducing the variability of $\boldsymbol{F}_{uc}$ (since its covariance is inversely proportional to N_{uc}). In turn, that would help reduce the variability in estimate of the ivector, which usually causes a significant performance drop for small segments. We intend to pursue further research in these directions in the future.

8. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *INTERSPEECH*, 2011, pp. 861–864.

[3] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, 2011, pp. 857–860.

[4] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[5] A. W. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with I-vector inputs," in *ICASSP*, 2014, pp. 225–229.

[6] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *INTERSPEECH*, 2014, pp. 751–755.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[9] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[10] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[11] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[12] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[13] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Odyssey*, 2014.

[14] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Odyssey*, 2014, pp. 293–298.

[15] R. Travadi and S. Narayanan, "Non-Iterative Parameter Estimation for Total Variability Model Using Randomized Singular Value Decomposition," in *INTERSPEECH*, 2016, pp. 3221–3225.

[16] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[17] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[18] H. Avron, S. Kale, S. Kasiviswanathan, and V. Sindhwani, "Efficient and practical stochastic subgradient descent for nuclear norm regularization," in *ICML*, 2012.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," 2011.