Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance

Jesús Villalba[1]*, Niko Brümmer*[2]

[1] Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
[2]AGNITIO, South Africa
villalba@unizar.es, nbrummer@agnitio.es

Abstract

We propose a variational Bayes solution to integrate out the model parameters in a generative i-vector speaker recognizer. The existing state-of-the-art in generative i-vector modelling plugs in fixed maximum-likelihood point-estimates of model parameters. This recipe may suffer from over-fitting of especially the between-speaker covariance. We show how to integrate out the between-speaker covariance and demonstrate dramatic improvements on NIST SRE 2010.

Index Terms: speaker recognition, i-vectors, variational Bayes

1. Introduction

In this paper we are interested in the *generative* approach to text-independent speaker recognition. We start by highlighting some landmark publications of the last decade: In [1], the emphasis was on modelling individual speakers, by making *point estimates* of GMM speaker models. In [2], the emphasis was on modelling within-speaker variability and point estimates were made of both speaker and channel[1] models. In [3], within and between speaker variabilities were jointly modelled. Point estimates were still made of speaker models, but channel models were integrated out. With the advent of *i-vectors* [4, 5], it became possible to integrate out both speaker and channel models, as a principled way of computing speaker detection likelihood ratio scores [6, 7] of the form:

$$R_p(\boldsymbol{\phi}_1,\boldsymbol{\phi}_2|\mathcal{M}) = \frac{P(\boldsymbol{\phi}_1,\boldsymbol{\phi}_2|\mathcal{M})}{P(\boldsymbol{\phi}_1|\mathcal{M})P(\boldsymbol{\phi}_2|\mathcal{M})} \tag{1}$$

where $\boldsymbol{\phi}_1,\boldsymbol{\phi}_2$ are feature vectors representing the two input speech segments in a speaker detection trial; and $\mathcal{M}$ represents the parameters of a generative model for *all speakers and channels*. The numerator is the likelihood that both speech segments come from the same speaker and the denominator is the likelihood that they come from different speakers. The subscript p is a mnemonic for *plug-in*, because $\mathcal{M}$ is plugged into this formula.

In [6, 7], $\mathcal{M}$ was a point estimate made by maximizing the likelihood over a large supervised development database, $\mathcal{D}$, containing several recordings of each of several hundreds of speakers. The problem is that when the point-estimate is plugged into (1), all uncertainty about the values of the model parameters is ignored. The fully *Bayesian* solution to this problem [8, 9] would be to also integrate out $\mathcal{M}$, to form the likelihood-ratio:

$$R_B(\boldsymbol{\phi}_1,\boldsymbol{\phi}_2|\mathcal{D}) = \frac{\int P(\boldsymbol{\phi}_1,\boldsymbol{\phi}_2|\mathcal{M})P(\mathcal{M}|\mathcal{D})\,\mathrm{d}\mathcal{M}}{\int P(\boldsymbol{\phi}_1|\mathcal{M})P(\boldsymbol{\phi}_2|\mathcal{M})P(\mathcal{M}|\mathcal{D})\,\mathrm{d}\mathcal{M}} \tag{2}$$

where $P(\mathcal{M}|\mathcal{D})$ is the posterior for the model parameters, given the development data. If there is little uncertainty about $\mathcal{M}$, i.e. when $P(\mathcal{M}|\mathcal{D})$ is sharply peaked, then the much simpler plug-in recipe R_p would suffice for most purposes. But if there is too much model uncertainty, then R_B can provide better accuracy. It is the aim of this paper to investigate whether there is indeed enough model uncertainty to warrant further investigation of the fully Bayesian solution.

In the rest of this paper, we introduce our model $\mathcal{M}$, show how to approximate the intractable integrals in (2) via *variational Bayes* [9] and demonstrate dramatic improvements on the telephone portion of NIST SRE 2010.

2. The two-covariance model

We adopt the approach taken in [4, 5] where each speech segment is represented by a single feature vector known as *i-vector*. We model the i-vectors with the *two-covariance model* introduced in [7], which supposes that an (observed) i-vector $\boldsymbol{\phi}$ of speaker s can be written as the sum of two hidden variables: $\boldsymbol{\phi} = \mathbf{y}_s + \mathbf{z}$, where $\mathbf{y}_s$ is denoted as the *speaker identity variable* and $\mathbf{z}$ as the *channel offset*. The identity variable remains constant, but the channel offset changes between different observations of the speaker. The model $\mathcal{M}$ is defined by the following two probability distributions:

$$P\left(\mathbf{y}|\mathcal{M}\right) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu},\mathbf{B}^{-1}\right) \tag{3}$$

$$P\left(\boldsymbol{\phi}|\mathbf{y},\mathcal{M}\right) = \mathcal{N}\left(\boldsymbol{\phi}|\mathbf{y},\mathbf{W}^{-1}\right) \tag{4}$$

where $\mathcal{N}$ denotes a Gaussian distribution; $\boldsymbol{\mu}$ is the speakers mean; $\mathbf{B}^{-1}$ is the between speaker covariance matrix and $\mathbf{W}^{-1}$ is the within speaker covariance matrix. See [7] for a closed-form expression for the plug-in likelihood-ratio (1). It is important to note that when $\mathcal{M}$ is given, all variables (hidden and observed) of different speakers are *independent*.

3. The Bayesian solution

Here we motivate and explain solutions for approximating R_B. We start by introducing some notation:

[1] It is understood that *channel* is a synecdoche, representing *everything* that causes recordings of a speaker to vary from one occasion to the next—rather than just physical transmission and recording channels.

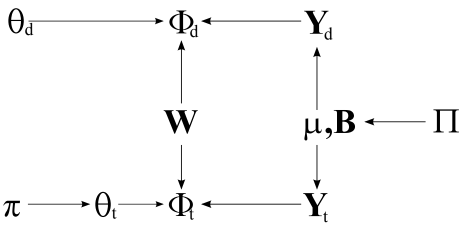

Figure 1: Graphical model showing relationships between parameters and variables.

3.1. Notation

The whole database of development i-vectors is denoted by $\mathbf{\Phi}_{\mathrm{d}}$, while the pair of test i-vectors in a speaker detection trial is denoted by $\mathbf{\Phi}_{\mathrm{t}} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$. The pooled development and test data is denoted by $\overline{\mathbf{\Phi}} = \mathbf{\Phi}_{\mathrm{d}} \cup \mathbf{\Phi}_{\mathrm{t}}$. We shall also use $\mathbf{\Phi}$ to refer in general to any of these three datasets. We assume that the speakers in the test data are not among the speakers in the development data.

Let θ_{d} be the labelling of the development dataset. It partitions the N_d i-vectors into M_d speakers. $\theta_{\mathrm{t}} \in \{\mathcal{T}, \mathcal{N}\}$ is the labelling of the test set where $\mathcal{T}$ is the hypothesis that the test i-vectors belong to the same speaker and $\mathcal{N}$ to different speakers. Let $\overline{\theta}$ be the labelling of $\overline{\mathbf{\Phi}}$ and θ any of the previous labellings. Let $P(\theta_t|\pi)$ denote the hypothesis prior.[2]

Let $\mathbf{Y}_{\mathrm{d}}$ and $\mathbf{Y}_{\mathrm{t}}$ respectively denote the hidden speaker identity variables of the development and test sets. $\mathbf{Y}$ can be used to refer to any of them.

Finally, we define $\mathcal{M} = (\boldsymbol{\mu}, \mathbf{B}, \mathbf{W})$ and $\mathcal{M}_y = (\boldsymbol{\mu}, \mathbf{B})$. We shall also need a *model prior,* denoted $P(\mathcal{M}_y|\Pi)$. The conditional independence relationships between all the variables introduced here are summarized in graphical model notation in figure 1.

3.2. The role of the model posterior

The fully Bayesian treatment requires integrating over probability distributions for all the model parameters. Our first simplifying approximation is to fix the value of $\mathbf{W}$, the within-class precision, at the maximum-likelihood point-estimate [7]. We subject only $\mathcal{M}_y = (\boldsymbol{\mu}, \mathbf{B})$ to Bayesian treatment. The motivation is that the number of development speakers is not so large compared to the i-vector dimension, so that the posterior for $\mathcal{M}_y$ may be relatively flat. In contrast, the total number of speech segments (and therefore channels) in the development data is an order of magnitude larger, which should give a more peaked posterior for $\mathbf{W}$.

We denote the *prior* for the model parameters as $P\left(\mathcal{M}_y|\Pi\right)$. We elaborate on the prior in the next two subsections. Given the prior, the fixed $\mathbf{W}$, and a supervised data set $(\mathbf{\Phi}, \theta)$, we can now denote the model *posterior* as[3] $P\left(\mathcal{M}_y|\mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right)$.

Our approach to approximating R_B revolves around the model posterior. Specifically, by using Bayes' rule and the conditional independence assumptions encoded in the graphical model of figure 1, we can express[4] the integrals of (2) in terms of the model posterior. For a detailed derivation see [10]. We can now rewrite (2) as:

$$R_B\left(\mathbf{\Phi}_{\mathrm{t}}|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = R_p\left(\mathbf{\Phi}_{\mathrm{t}}|\mathcal{M}_y, \mathbf{W}\right) \frac{P\left(\mathcal{M}_y|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{N}, \mathbf{W}, \Pi\right)}{P\left(\mathcal{M}_y|\overline{\mathbf{\Phi}}, \theta_{\mathrm{d}}, \mathcal{T}, \mathbf{W}, \Pi\right)} \tag{5}$$

where the model posteriors are conditioned on the *pooled* development and test data. Note also that the LHS does not depend on $\mathcal{M}_y$, so we can plug in any convenient value of the model in the RHS, as long the denominator is not zero.

Equation (5) gives an insightful interpretation of the Bayesian likelihood ratio as the plug-in ratio multiplied by a correction factor. We can plug in any model estimate $\mathcal{M}_y$ and the correction factor will compensate. But the correction will be noticeable only if the posterior model densities at $\hat{\mathcal{M}}_y$ are considerably different for the two alternate conditionings.

We have now transformed the problem of calculating model parameter integrals to one of calculating model parameter posteriors. Unfortunately, even for the simple two-covariance model, these posteriors cannot be expressed in closed form. We propose to use a variational Bayes (VB) approach to calculate approximate posteriors. In the next sections, we present the VB solutions for the two-covariance model assuming two different types of model priors: non-informative and conjugate.

3.3. VB with non-informative priors

3.3.1. Non-informative prior

We can assume a non-informative prior (Jeffreys prior) for the parameters $\boldsymbol{\mu}$ and $\mathbf{B}$ of the speaker Gaussian distribution [11]. A non-informative prior encodes the absence of information about $\boldsymbol{\mu}$ and $\mathbf{B}$ other than the training data. With this prior no Gaussian should be preferred over others and it should be invariant to any translation or scaling of the measurement space. These conditions are satisfied by this distribution:

$$P\left(\boldsymbol{\mu}, \mathbf{B}|\Pi\right) = \alpha \left|\frac{\mathbf{B}}{2\pi}\right|^{1/2} |\mathbf{B}|^{-(d+1)/2} \tag{6}$$

where d denotes the dimensionality of $\boldsymbol{\mu}$. Since this density does not integrate to 1, it is improper and the symbol α is used to denote a normalizing constant which approaches zero. Note that using an improper prior does not mean that the posterior will be improper.

3.3.2. VB likelihood ratio

Our VB solution approximates the joint posterior distribution for the hidden variables and model parameters by a factorized distribution of the form:

$$P\left(\mathcal{M}_y, \mathbf{Y}|\mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) \approx q\left(\mathcal{M}_y, \mathbf{Y}\right) = q\left(\mathcal{M}_y\right) q\left(\mathbf{Y}\right) \tag{7}$$

which ignores any posterior dependencies between the speaker variables $\mathbf{Y}$ and the model $\mathcal{M}_y$. Note that we are not making further factorizing assumptions or restricting the functional form of the individual factors.

We complete our recipe by using the approximation $P\left(\mathcal{M}_y|\mathbf{\Phi}, \theta, \mathbf{W}, \Pi\right) \approx q\left(\mathcal{M}_y\right)$ in (5), so that the *VB* approximation of the likelihood ratio is

$$R_{VB}\left(\mathbf{\Phi}_{\mathrm{t}}|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi\right) = R_p\left(\mathbf{\Phi}_{\mathrm{t}}|\mathcal{M}_y, \mathbf{W}\right) \frac{q_{\mathcal{N}}\left(\mathcal{M}_y\right)}{q_{\mathcal{T}}\left(\mathcal{M}_y\right)} \tag{8}$$

[2] In speaker detection, $P(\mathcal{T}|\pi)$ is usually called the *target* prior.

[3] In (2), we used a simplified notation for the model posterior. With the more detailed notation introduced here, we have $P(\mathcal{M}_y|\mathcal{D}) = P(\mathcal{M}_y|\mathbf{\Phi}_{\mathrm{d}}, \theta_{\mathrm{d}}, \mathbf{W}, \Pi)$.

[4] We use *express* rather than *solve*, because calculating these posteriors involves solving similar intractable integrals.

where $q_{\mathcal{T}}(\mathcal{M}_y)$ and $q_{\mathcal{N}}(\mathcal{M}_y)$ are the variational posteriors conditioned respectively on $\theta_\mathrm{t} = \mathcal{T}$ and $\theta_\mathrm{t} = \mathcal{N}$.

In this approximation, the property that the likelihood ratio does not depend on the plug-in model $\hat{\mathcal{M}}_y$ is no longer true. In our experiments we used the maximum likelihood point-estimate.

3.3.3. VB distributions

According to variational Bayes theory [9], given a set of visible variables $\mathbf{X}$ and hidden variables $\mathbf{Z}$, the optimum value of the factoring distribution $q_j^*(\mathbf{Z}_j)$ is given by

$$\ln q_j^*(\mathbf{Z}_j) = \mathrm{E}_{i \neq j}\left[\ln P(\mathbf{X}, \mathbf{Z})\right] + \mathrm{const} \tag{9}$$

This equation means that the log of the optimum solution for factor q_j is estimated by taking the expectation of the log joint distribution over all hidden and visible variables with respect to all other factors $q_{i \neq j}$. The additive constant is needed to normalize the distribution to integrate to one.

VB is an iterative procedure. We first initialize the factors and then cycle $q_j(\mathbf{Z}_j)$ through the factors re-estimating each one using (9) until convergence.

In our model the joint distribution of all variables is given by

$$\begin{aligned} &P(\mathbf{\Phi}, \mathcal{M}_y, \mathbf{Y}|\theta, \mathbf{W}, \Pi) = \\ &P(\mathbf{\Phi}|\mathbf{Y}, \theta, \mathbf{W})\, P(\mathbf{Y}|\mathcal{M}_y)\, P(\mathcal{M}_y|\Pi) \end{aligned} \tag{10}$$

Now, applying (9), it is straightforward to obtain our variational distributions.

The optimum for the factor $q(\mathbf{Y})$ is given by a product of Gaussian distributions:

$$q^*(\mathbf{Y}) = \prod_{i=1}^{M} q^*(\mathbf{y}_i) = \prod_{i=1}^{M} \mathcal{N}\left(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1}\right) \tag{11}$$

where the speaker identity variables $\mathbf{y}_i$ are independent. Note that we have not forced that in any way but it originates naturally from the original factorization that we have chosen. Due to the limited space available, the expressions for the parameters of the factoring distributions have been omitted from this paper. A detailed derivation of these parameters can be found in our technical report [12].

The optimum for the factor $q(\mathcal{M}_y)$ is a Gaussian-Wishart distribution.

$$q^*(\mathcal{M}_y) = \mathcal{N}\left(\boldsymbol{\mu}|\overline{\mathbf{y}}, (M\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y^{-1}, M\right) \tag{12}$$

We have to remark that for this distribution to be proper we need the number of speakers M to be larger than the i-vectors dimensionality. The parameters $\overline{\mathbf{y}}$ and $\mathbf{S}_y$ are the mean and covariance of the speaker identity variables.

3.4. VB with conjugate priors

3.4.1. Conjugate prior

The approach taken in section 3.3 has a large computational cost: for each trial and for each VB iteration, we need to re-estimate the $q(\mathbf{y}_i)$ distributions for all the speakers of the development database. In this section, we make a further approximation, by effectively fixing these $q(\mathbf{y}_i)$. This is achieved by first computing the VB posterior for $\mathcal{M}_y$, conditioned only on the development data $\mathbf{\Phi}_\mathrm{d}$, θ_d. In other words, the test data is not involved. Then we use this posterior to act as the prior, denoted $P(\mathcal{M}_y|\Pi_\mathrm{d})$, for further processing of the test data. Now for every trial, only the test i-vectors $\mathbf{\Phi}_\mathrm{t}$ are involved in the calculation of the likelihood ratio. This idea relies on the assumption that adding the trial data should not modify the posteriors of the development speaker identity variables significantly. Most development speakers have a large number of segments, so that the speaker identity posteriors are less affected by changes in the value of $\mathcal{M}_y$. In contrast, test speakers have at most two segments. In summary, our model prior is now:

$$P(\mathcal{M}_y|\Pi_\mathrm{d}) = q_\mathrm{d}(\mathcal{M}_y) \approx P(\mathcal{M}_y|\mathbf{\Phi}_\mathrm{d}, \theta_\mathrm{d}, \mathbf{W}, \Pi) \tag{13}$$

which is conditioned on the development data. The VB factor $q_\mathrm{d}(\mathcal{M}_y)$ is Gaussian-Wishart distributed, which is a conjugate prior for the Gaussian distribution. As shown in (12), it is given by

$$q_\mathrm{d}(\mathcal{M}_y) = \mathcal{N}\left(\boldsymbol{\mu}|\overline{\mathbf{y}}_\mathrm{d}, (\beta_\mathrm{d}\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_{\mathrm{d}y}^{-1}, \nu_\mathrm{d}\right) \tag{14}$$

where $\beta_\mathrm{d} = \nu_\mathrm{d} = M_\mathrm{d} > d$.

3.4.2. VB likelihood ratio

In a similar way to section 3.3.2 we define the variational likelihood ratio as

$$R_{VB}(\mathbf{\Phi}_\mathrm{t}|\mathbf{W}, \Pi_\mathrm{d}) = R_p(\mathbf{\Phi}_\mathrm{t}|\mathcal{M}_y, \mathbf{W}) \frac{q_{\mathcal{N}}(\mathcal{M}_y)}{q_{\mathcal{T}}(\mathcal{M}_y)} \tag{15}$$

The difference with (8) is that the ratio does not depend explicitly on the development data. The dependence is now implicit through the prior Π_d.

3.4.3. Variational distributions

The variational distributions are calculated in a similar way to section 3.3.3. In fact, the optimum for the factor $q(\mathbf{Y})$ is the same as in the non-informative case. The optimum for the factor $q(\mathcal{M}_y)$ is again Gaussian-Wishart

$$q^*(\mathcal{M}_y) = \mathcal{N}\left(\boldsymbol{\mu}|\overline{\mathbf{y}}', (\beta'\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}|\mathbf{S}_y'^{-1}, \nu'\right) \tag{16}$$

where $\overline{\mathbf{y}}'$ and $\mathbf{S}_y'$ are the mean and covariance of the speaker identity variables MAP adapted to the trial data.

4. Experiments

We performed experiments on the core-core det5 condition of the NIST 2010 speaker recognition evaluation (telephone-telephone, English, normal vocal effort).

As features for our system, we have used the i-vectors provided by Brno University of Technology (BUT) [13]. They are extracted using 20 short time Gaussianized MFCC plus deltas and double deltas and a 2048 component full covariance UBM. The UBM was trained using telephone data from SRE04 and SRE05 and the i-vector extractor with data from SRE04, SRE05, SRE06, Switchboard and Fisher.

We have conducted experiments with the Bayesian and the *plug-in* likelihood ratio. We have used telephone speech from SRE04, SRE05, SRE06 and Switchboard as development data for the two-covariance model. The models are gender dependent. For score normalization we used s-norm [14] with utterances from SRE04 to SRE06 (1599 male, 2530 female). We present results with the plain i-vectors and preprocessed with LDA to reduce dimensionality to 90. The LDA transform was trained with the same development data as the two-covariance model.

Table 1: *EER(%)/minDCF without score normalization (top) and with score normalization (bottom)*

	male		female	
	EER	DCF	EER	DCF
2cov	3.56	0.52	3.60	0.40
Bay2cov Jprior	**1.14**	**0.31**	**1.31**	0.37
Bay2cov Cprior	**1.15**	0.37	1.58	**0.29**
LDA90 + 2cov	2.59	0.53	3.19	0.38
LDA90 + Bay2cov Jprior	1.80	0.41	1.91	0.32
LDA90 + Bay2cov Cprior	1.84	0.42	2.00	0.34

	male		female	
	EER	DCF	EER	DCF
2cov	2.02	**0.35**	1.99	**0.48**
Bay2cov Cprior	**1.47**	0.43	**1.72**	**0.49**
LDA90 + 2cov	1.83	0.59	2.09	0.58
LDA90 + Bay2cov Cprior	1.67	0.57	1.84	0.59

The results are summarized in Table 1. The minimum DCF is calculated for the new NIST SRE10 operating point ($C_{Miss} = 1, C_{FA} = 1, P_{\mathcal{T}} = 0.001$). The first thing that we note is that the Bayesian approach produces better EER than the non-Bayesian one in all cases. The minDCF is better in the Bayesian case than in the non-Bayesian when there is no score normalization. However, it is similar or slightly worse when the scores are normalized. Another thing that we can see is that score normalization is needed for the non-Bayesian case but harmful for the Bayesian one and the best results are achieved for the Bayesian case without normalization.

We present results with non-informative (Jprior) and informative (Cprior) priors. Results with non-informative priors are shown only without s-norm due to its high computational cost. Results of both types of priors are quite similar so we can assume that the approximation explained in section 3.4.1 is good.

Finally, we analyse the results of the low dimensional i-vectors. Reducing dimensionality improves the EER of the non-Bayesian system but not the minDCF. For the Bayesian system, it is harmful in any case. We think that the Bayesian approach reduces the risk of over-training of the model and allows it to extract the information of the 400 dimensional vector in a better way. Therefore, the reduction of dimensionality is not necessary.

5. Discussion

We have presented a method to calculate the likelihood ratio of a speaker verification system in a fully Bayesian way integrating-out the parameters of the model. We have shown how this ratio can be estimated approximately for the particular case of the two-covariance model. To do that, we have adopted a VB procedure assuming non-informative and informative priors for the speaker model parameters.

We have shown results on the telephone condition of NIST SRE10. The Bayesian approach produces EER around a 35% better than the best non-Bayesian system evaluated. Another interesting point is that the Bayesian system does not need score normalization for this dataset. However, it does not give naturally well-calibrated likelihood ratios.

One way to interpret our results is that the plug-in method suffers from over-fitting of the model. This over-fitting can be compensated for by the ad-hoc dimensionality reduction given by the LDA. In contrast, the Bayesian method is more robust against over-fitting and does not need the LDA.

On the other hand, one should keep in mind that the simple Gaussian two-covariance model may not be optimal for modelling i-vector data. Indeed, it has been shown that a heavy-tailed model [6], or i-vector magnitude normalization [15] can give similar improvements to our Bayesian method.

6. Acknowledgements

We want to acknowledge BUT for providing the i-vectors used in this work. This work has been supported by the Spanish Government through national project TIN2008-06856-C0504.

7. References

[1] 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, Jan. 2000. [Online]. Available: http://dx.doi.org/10.1006/dspr.1999.0361

[2] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Černocký, “Analysis of feature extraction and channel compensation in GMM speaker recognition system,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[3] 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, May 2007.

[4] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification,” in *Interspeech 2009*, Brighton, UK, 2009.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis For Speaker Verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, 2010.

[6] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[7] N. Brümmer and E. De Villiers, “The Speaker Partitioning Problem,” in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[8] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, 1st ed. Cambridge University Press, 2003.

[9] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[10] N. Brümmer, “Bayesian PLDA,” Agnitio Labs Technical Report. Online: https://sites.google.com/site/nikobrummer/bplda.pdf, Aug. 2010.

[11] T. Minka, “Inferring a Gaussian distribution,” 1998. [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/papers/gaussian.html

[12] J. Villalba and N. Brümmer, “Bayesian two-covariance model integrating out the speaker space distribution,” Technical Report. Online: http://sites.google.com/site/nikobrummer/bay2covmuB.pdf, Oct. 2010.

[13] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, P. Kenny, L. Burget, and J. Cernocky, “Full-Covariance UBM and Heavy-Tailed PLDA in I-Vector Speaker Verification,” in *ICASSP 2011*, Prague, Czech Republic, 2011.

[14] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, “An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech,” in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[15] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Interspeech 2011*, Florence, Italy, 2011.