Analysis of I-vector Length Normalization in Speaker Recognition Systems

Daniel Garcia-Romero and Carol Y. Espy-Wilson

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD
dgromero@umd.edu, espy@umd.edu

Abstract

We present a method to boost the performance of probabilistic generative models that work with i-vector representations. The proposed approach deals with the non-Gaussian behavior of i-vectors by performing a simple length normalization. This non-linear transformation allows the use of probabilistic models with Gaussian assumptions that yield equivalent performance to that of more complicated systems based on Heavy-Tailed assumptions. Significant performance improvements are demonstrated on the telephone portion of NIST SRE 2010.

Index Terms: speaker recognition, i-vectors, length normalization

1. Introduction

The recently developed paradigm of i-vector extraction [1] provides an elegant way to obtain a low dimensional fixed-length representation of a speech utterance that preserves the speaker-specific information. A Factor Analysis (FA) model is used to learn a low-dimensional subspace from a large collection of data. A speech utterance is then projected into this subspace and its coordinates vector is denoted as *i-vector* [1]. The low dimensional nature of this representation is very appealing and has opened the door for new ways to explore one of the key problems in speaker recognition, that is, how to decompose a speech signal into a speaker-specific component and an undesired variability component (often referred to as the channel component). Although both non-probabilistic [1,2] and probabilistic approaches are possible towards this goal, the focus of this paper is on the use of *probabilistic generative models* of i-vector distributions for speaker verification.

A common theme among the probabilistic approaches is that they ignore the process by which i-vectors were extracted (i.e., FA model) and instead pretend they were generated by a prescribed generative model. The distinguishing factor between these approaches is the set of assumptions embedded in the model. The two most commonly used assumptions are that (i) the speaker and channel components are statistically independent and (ii) they have Gaussian distributions. Examples following these assumptions are the Probabilistic Linear Discriminant Analysis (PLDA) model proposed in [3] and the two-covariance (2-cov) model introduced in [4]. The main advantage of these methods is that the speaker detection likelihood ratios can be obtained in closed-form.

Alternatively, a Heavy-Tailed PLDA model (HT-PLDA) was presented in [5] where the Gaussian priors where replaced by Student's t distribution. Two main motivations were behind this approach. First, to allow for larger deviations from the mean (e.g., severe channel distortions). Second, to increase the robustness to outliers in the ML estimation of the model parameters. Since no closed-form solution of the speaker detection likelihood ratio is obtained when using the heavy-tailed priors, variational Bayes was used to approximate it [6]. The results presented in [5] showed superior performance of the HT-PLDA model over the Gaussian prior based alternative. Hence, providing strong empirical evidence towards non-Gaussian behavior of speaker and channel effects.

In this paper we pursue an alternative approach to deal with the non-Gaussian behavior of the i-vectors. That is, we keep the Gaussian assumptions in the model, but perform a non-linear transformation of the i-vectors to reduce the non-Gaussian behavior (i.e., i-vector Gaussianization). In the rest of the paper we present a formal mathematical description of the whole speaker recognition system, propose a non-linear i-vector transformation and analyze its behavior on the telephone portion of NIST SRE 2010.

2. Recognition system

In this section we provide a formal description of the three fundamental components of our speaker recognition system. Namely, the i-vector extraction, modeling and likelihood ratio computation.

2.1. I-vector extraction

An i-vector extractor [1] is a system that maps a sequence of vectors (typically cepstral coefficients) obtained from a speech utterance, $\mathcal{O} = \{o_t\}_{t=1}^{N}$ with $o_t \in \mathbb{R}^F$, to a fixed-length vector $\eta \in \mathbb{R}^D$. In order to do that, a K-component Gaussian Mixture Model (GMM), $\lambda = (\{w_k\}, \{m_k\}, \{\Sigma_k\})$, denoted as Universal Background Model (UBM) is used to collect Baum-Welch statistics from the utterance. Subsequently, a supervector $\theta = [\theta_1^T, \dots, \theta_K^T]^T \in \mathbb{R}^{FK}$ is constructed by appending together the first-order statistics for each mixture component and is assumed to obey an affine linear model (i.e., factor analysis model) of the form

$$\theta = m + \mathrm{T}x, \tag{1}$$

where the supervector $m \in \mathbb{R}^{FK}$ comes from the UBM, the columns of the low-rank matrix $\mathrm{T} \in \mathbb{R}^{FK \times D}$ span the subspace where most of the speaker-specific information lives (along with channel variability) and x is a standard-normally distributed latent variable. For each speech utterance, an i-vector η is obtained as the MAP point estimate of x. The subspace spanned by T is obtained from a large collection of data representative of the task at hand by ML estimation [7].

2.2. Generative i-vector models

Once an i-vector is obtained from an utterance, we ignore the extraction mechanism and regard it as an observation from a probabilistic generative model. In the following, the two probabilistic models used in our experiments are described.

2.2.1. *Gaussian PLDA (G-PLDA)*

Assuming R utterances for a speaker, the collection of corresponding i-vectors is denoted as $\{\eta_r: r = 1, \dots, R\}$. The G-PLDA model introduced in [3] then assumes that each i-vector can be decomposed as

$$\eta_r \ = m + \Phi\beta + \Gamma\alpha_r + \ \epsilon_r. \tag{2}$$

In the jargon of speaker recognition, the model comprises two parts: (i) the speaker-specific part $s = m + \Phi\beta$ which describes the between-speaker variability and does not depend on the particular utterance; (ii) the channel component $c_r = \Gamma\alpha_r + \ \epsilon_r$ which is utterance dependent and describes the within-speaker variability. In particular, m is a global offset; the columns of Φ provide a basis for the speaker-specific subspace (eigenvoices); β is a latent identity vector having a standard normal distribution; the columns of Γ provide a basis for the channel subspace (eigenchannels); α_r is a latent vector having a standard normal distribution; and ϵ_r is a residual term assumed to be Gaussian with zero mean and diagonal covariance Σ. Moreover, all latent variables are assumed statistically independent. Since the i-vectors we are dealing with in this work are of sufficiently low dimension (i.e., 400 for our experiments), we follow the modification proposed in [5] and assume that Σ is a full covariance matrix and remove the eigenchannels from eq. (2). Thus, the modified G-PLDA model used in this paper follows:

$$\eta_r \ = m + \Phi\beta + \ \epsilon_r. \tag{3}$$

The ML point estimates of the model parameters $\{m, \Phi, \Sigma\}$ are obtained from a large collection of development data using an EM algorithm as in [3].

2.2.2. *Heavy-Tailed PLDA (HT-PLDA)*

The HT-PLDA model was first introduced in [5]. While the general formulation includes a channel subspace, the HT-PLDA model used in our experiments will be based on eq. (3) but with priors on β and ϵ_r following multivariate Student's *t* distributions rather than Gaussian. Precisely, β is assumed to have zero mean, identity scale matrix and n_β degrees of freedom. Also, ϵ_r is assumed to have zero mean, full scale matrix Σ and n_{ϵ_r} degrees of freedom. It is important to note the number of degrees of freedom parameter controls the behavior of the tails of the distribution. That is, the smaller the number of degrees of freedom the heavier the tails. On the other hand, as the number of degrees of freedom increases the Student's *t* distribution converges to a Gaussian distribution [6]. In this way, one can think of the G-PLDA model as a particularization of the HT-PLDA model where the number of degrees of freedom grows to infinity.

2.3. Verification score

For a speaker verification task, given the two i-vectors η_1 and η_2 involved in a trial, we are interested in testing two alternative hypotheses: $\mathcal{H}_s$ that both η_1 and η_2 share the same speaker identity latent variable β, or $\mathcal{H}_d$ that the i-vectors were generated using different identity variables β_1 and β_2. The verification score can now be computed as the log-likelihood ratio for this hypothesis test as

$$score = \ \log\frac{p(\eta_1, \eta_2|\mathcal{H}_s)}{p(\eta_1|\mathcal{H}_d)\ p(\eta_2|\mathcal{H}_d)}. \tag{4}$$

For the G-PLDA case, this log-likelihood ratio is easily computed in closed-form solution since the marginal likelihoods (i.e., the evidence) are Gaussian. That is,

$$\begin{aligned} score = \ & \log\mathcal{N}\left(\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}; \begin{bmatrix}m\\ m\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{ac}\\ \Sigma_{ac} & \Sigma_{tot}\end{bmatrix}\right) \\ & - \log\mathcal{N}\left(\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}; \begin{bmatrix}m\\ m\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right), \end{aligned} \tag{5}$$

where $\Sigma_{tot} = \Phi\Phi^{\mathrm{T}} + \Sigma$ and $\Sigma_{ac} = \Phi\Phi^{\mathrm{T}}$. Moreover, by setting $m = 0$ (since it is a global offset that can be precomputed and removed from all the i-vectors) and expanding we get

$$score = \eta_1^T \mathrm{Q}\ \eta_1 + \eta_2^T \mathrm{Q}\ \eta_2 + 2\eta_1^T \mathrm{P}\ \eta_2 + const, \tag{6}$$

with

$$\begin{aligned} \mathrm{Q} &= \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}, \\ \mathrm{P} &= \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}. \end{aligned} \tag{7}$$

Even though not immediately apparent, it can be shown that both P and Q have rank equal to the rank of Φ. This is a novel observation that was not made in [3] and opens the door for a fast computation of eq. (6). That is, based on the symmetry of P and assuming that $\Phi \in \mathbb{R}^{D\times K}$with $K < D$, we can obtain the decomposition

$$\begin{aligned} \mathrm{P} &= [\mathrm{U}_K|\mathrm{U}_{D-K}]diag([\lambda_1, \dots, \lambda_K, 0\ ,0, \dots])[\mathrm{U}_K|\mathrm{U}_{D-K}]^T \\ &= \mathrm{U}_K diag([\lambda_1, \dots, \lambda_K])U_K^T \end{aligned} \tag{8}$$

where the K columns of U_K are orthonormal, the vector $[\lambda_1, \dots, \lambda_K]^T$contains the non-zero eigenvalues of P and the operator $diag(\cdot)$ places the entries of its argument in the diagonal of a matrix. Letting $\Lambda = diag([\lambda_1, \dots, \lambda_K])$, we can compute

$$score = \tilde{\eta}_1^T \widetilde{\mathrm{Q}}\ \tilde{\eta}_1 + \tilde{\eta}_2^T \widetilde{\mathrm{Q}}\ \tilde{\eta}_2 + 2\tilde{\eta}_1^T \Lambda\tilde{\eta}_2 + const, \tag{9}$$

where $\widetilde{\mathrm{Q}} = \mathrm{U}_K^{\mathrm{T}}\mathrm{Q}\mathrm{U}_K$and $\tilde{\eta}_i = \mathrm{U}_K^T\eta_i$. Note that $\widetilde{\mathrm{Q}}$ and $\tilde{\eta}_1^T \widetilde{\mathrm{Q}}\ \tilde{\eta}_1$can be precomputed (assuming $\tilde{\eta}_1$ to be the enrolled model) and at verification time, after projecting the test i-vector η_2, all the remaining computations are performed in a lower dimensional space. The computational advantage becomes more significant as the ratio K/D decreases.

For the HT-PLDA model, the log-likelihood ratio in equation (4) does not have a closed form solution. In [5], a variational lower bound is used as a proxy for each of the marginal likelihoods (i.e., evidence) involved in the log-likelihood ratio. We refer the reader to the very clear exposition in [5] for the details.

3. I-vector transformation

The results presented in [5,8] showed superior performance of the HT-PLDA model over G-PLDA for the telephone conditions of NIST SRE 2010. This provides strong empirical evidence of non-Gaussian behavior of speaker and channel effects in i-vector representations. However, due to the simplicity and computational efficiency of G-PLDA we are interested in keeping the Gaussian assumptions in the model and performing a transformation of the i-vectors to reduce the non-Gaussian behavior. A successful i-vector transformation should close the gap in performance between HT-PLDA and G-PLDA.

3.1. Radial Gaussianization

Besides the Gaussian prior assumption, the statistical independence between speaker and channel factors is also questionable. As noted in [5], the success of cosine scoring [2] suggests that there is a principal axis of channel variation that

is dependent on the speaker identity. Thus, if we drop the independence assumption and keep the multivariate Student's *t* distribution assumption for the prior on the latent variables, the generative i-vector model can be expressed as

$$\eta = m + \Omega z, \tag{10}$$

where the latent variable z now represents both the speaker and channel factors and follows a Student's *t* distribution. In this way, η is nothing more than an affine transformation of a multivariate Student's *t* distribution which belongs to the family of Elliptically Symmetric Densities (ESD) [9]. Thus, a transformation of the i-vectors—that renders the Gaussian and statistical independence assumptions appropriate—needs to be able to transform members of the ESD family into a Gaussian distribution. As pointed out in [9], linear transforms have no effect on the dependencies beyond second order for ESD. Thus, if z follows a multivariate Student's *t* distribution, we need to resort to non-linear transformations to accomplish our goal. Fortunately, an effective technique denoted as Radial Gaussianization (RG) was proposed in [9]. This technique follows a two step process. First, the ESD is transformed into a Spherically Symmetric Density (SSD) by a linear whitening transformation learned from data samples. Second, a non-linear histogram warping of the length distribution of the whitened variable η_{wht} is performed (this second step stems from the fact that the length of vectors drawn from a standard Gaussian distribution follows a Chi distribution with degrees of freedom equal to the dimension of the vector). In particular, the length transformation function is given by

$$g(\|\eta_{wht}\|) = F_{\chi}^{-1} F_r(\|\eta_{wht}\|). \tag{11}$$

This is nothing more than the function composition of the inverse cumulative Chi distribution with the cumulative distribution of the length random variable $r = \|\eta_{wht}\|$. In practice, F_r needs to be estimated from data.

3.2. Length normalization

The need to estimate the cumulative distribution of the length random variable $r = \|\eta_{wht}\|$ poses a potential problem with respect to the evaluation rules of NIST SRE. That is, if the distribution of lengths of the evaluation data is to be accurately modeled, the entire collection must be used. This strategy violates the restriction to only use the two utterances involved in a trial to produce a verification score. For this reason, we propose to simplify the second step of the RG process and simply scale the lengths of each i-vector η_{wht} to unit length.

4. Experiments

In this section we present an experimental validation of the benefits of i-vector transformation in speaker verification performance. The following section provides details about the experimental setup used throughout all the experiments.

4.1. Experimental setup

The NIST SRE 2010 data from the extended-core telephone-telephone condition (i.e., condition 5) was used. Throughout the experiments we refer to this set as the evaluation data. Verification performance is reported in terms of Equal Error Rate (EER) as well as the new Detection Cost Function (DCF) defined by ($C_{MISS} = 1$, $C_{FA} = 1$ and $P_{tar} = 0.001$).

For all our experiments, we have used the i-vectors provided by Brno University of Technology (BUT) [8]. They are extracted using a 20ms short-time Gaussianized MFCCs plus delta and double-delta. A full-covariance gender-independent UBM with 2048 mixtures was trained from NIST SRE 04 and 05 telephone data. A gender-dependent i-vector extractor was trained from telephone data from: NIST SRE 04, 05, 06, Switchboard and Fisher. The dimension of the i-vectors is 400. Both the G-PLDA and HT-PLDA model parameters were estimated from the same data used in the i-vector extractor (excluding data from Fisher database since it was found in [8] to deteriorate the verification performance). We refer to this set as development data. The number of eigenvoices in G-PLDA was set to 120. Also, in order to reduce the computational cost of the HT-PLDA system, a LDA dimensionality reduction to 120 dimensions was used prior to any other processing of the i-vectors. The number of eigenvoices was also set to 120 as in the G-PLDA case. No score normalization is used in the reported results since it did not help improve the performance.

4.2. I-vector length analysis

As mentioned before, the length of vectors drawn from a standard Gaussian distribution follows a Chi distribution with number of degrees of freedom (DOF) equal to the dimension of the vector (i.e., 400). In principle, an i-vector extractor is supposed to generate i-vectors that behave in this way (especially if a Minimum Divergence [7] step is used). Figure 1 depicts the probability density function of a Chi distribution with 400 DOF. Also, histograms of the i-vector length distribution for development and evaluation data (separated by gender) are presented. Three important observations are in order. First, neither the development data not the evaluation data match the Chi distribution. Second, the behavior for both genders is similar. Third, and most important, there is a significant mismatch between the length distributions of the development and evaluation i-vectors. This is not surprising since the i-vector extractor is trained on the development set and therefore fits this data set better than the evaluation set.

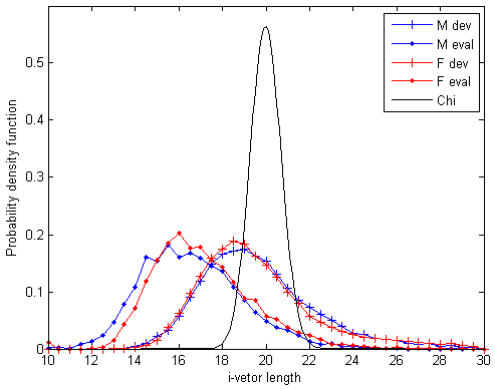

Figure 1. (Color) Histograms of the i-vector length distribution for development and evaluation data separated by gender: male (M) and female (F). Also the probability density function of a Chi distribution with 400 degrees of freedom is depicted.

However, although not surprising, this behavior is undesirable—especially if we are interested in using a simple G-PLDA model—since the mismatch can be considered as a strong source of heavy-tailed behavior. To further investigate the effects of this mismatch, we used the HT-PLDA system and checked how the ML estimates of the degrees of freedom parameters n_β and n_ϵ behaved as we transformed both the

development and evaluation i-vectors by RG and length (L) normalization. The results are summarized in Table 1. We can observe that the behavior between male and female speakers is consistent. More interestingly, both RG and L transformations increase the value of n_β and decrease the value of n_ϵ when compared to the original i-vectors. This indicates that the transformations make the HT-PLDA more like a partially-HT model where the eigenvoices have lighter tails (i.e., more Gaussian) and the residual shows a stronger heavy-tailed behavior. Also, the L normalization seems to induce a more extreme behavior.

Transformation type	Eigenvoices (n_β)		Residual (n_ϵ)	
	Male	Female	Male	Female
dev_eval	11.09	12.39	17.10	17.42
RG-dev_RG-eval	25.35	27.30	13.24	14.81
L-dev_L-eval	48.07	54.71	9.21	10.42

Table 1. Value of the degrees of freedom parameters for i-vector transformations in the HT-PLDA system.

4.3. Speaker verification results

Table 2 summarizes the results for multiple combinations of transformations of development and evaluation sets. For the G-PLDA system both the RG and L transformation provide an impressive improvement over the unprocessed i-vectors both in EER and in DCF. Also, the simpler length normalization achieves equivalent performance to the RG technique with the advantage of being NIST SRE compliant.

System codes	Male scores		Female scores	
	EER(%)	minDCF	EER(%)	minDCF
U_U_G	3.08	0.4193	3.41	0.4008
U_RG_G	1.44	0.3032	2.15	0.3503
U_L_G	1.29	0.3084	1.97	0.3511
L_L_G	**1.27**	**0.3019**	2.02	0.3562
RG_RG_G	1.37	0.3066	2.16	0.3393
U_U_HT	1.48	0.3357	2.21	0.3410
L_L_HT	1.28	0.3036	**1.95**	**0.3297**
RG_RG_HT	**1.27**	0.3143	**1.95**	0.3339

Table 2. Verification results for extended-core condition 5 of NIST SRE 2010. G-PLDA and HT-PLDA systems are evaluated with various combinations of i-vector transformations. Both systems use 120 eigenvoices and full-covariance residual. In the case of HT-PLDA system, an initial LDA dimensionality reduction to 120 dimensions was used to decrease the computational cost. The top 5 rows correspond to G-PLDA system and the lower 3 to HT-PLDA system. The system codes correspond to: dev_eval_system. For example, the first row indicates that both the dev and eval data were not transformed and the system was G-PLDA.

Another interesting observation is that as long as the evaluation data is transformed, keeping the development i-vectors in their original form does not affect the performance much (rows 2 and 3). Thus, the key step is the transformation of the evaluation data i-vectors. In the case of length normalization, this can be explained by taking a look at the scoring equation (6) and noting that a global scaling of the length of all the evaluation i-vectors only produces a global scaling of the scores (i.e., it does not alter the relative position of the scores). Hence, once the length normalization has been applied, instead of unit-length we can select the target length to match the mode of the development data distribution. In this way, we have greatly eliminated the mismatch and the results should reflect that. Thus, the choice of unit length is an arbitrary one and we can think that effectively the length normalization is mapping the length of all the evaluation data to the mode of the development data length histogram.

Regarding the HT-PLDA system, first we can note that the performance gap between G-PLDA and HT-PLDA is greatly reduced (if not completely eliminated). Also, although HT-PLDA is able to successfully cope with the development and evaluation mismatch induced by the i-vector extraction procedure, a small improvement is observed after transforming the i-vectors.

5. Conclusions

We have presented a method to boost the performance of probabilistic generative models that work with i-vector representations. By performing a simple length normalization of the i-vectors, the performance of a G-PLDA system was able to match that of a more complicated HT-PLDA one. Also, we have identified the mismatched induced by the i-vector extraction mechanism as a major source of non-Gaussian behavior. Additionally a computational improvement for G-PLDA scoring was pointed out by noting the low-rank nature of the matrices involved in the log-likelihood ratio computation.

6. Acknowledgments

We thank BUT for providing the i-vectors and Carlos Vaquero for the HT-PLDA system. Special thanks to Lukas Burget, Niko Brummer and Patrick Kenny for helpful discussions. The work of Lukas has helped in the i-vector analysis in section 4.2. Also, the discussion with Niko Brummer has helped in the architecture selection of the G-PLDA system. This work has been supported by NSF grant #0917104.

7. Bibliography

[1] N. Dehak, P. 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, May 2010.

[2] N. Dehak et al., "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in *Interspeech 2009*, Brighton, UK, 2009.

[3] S. J. D. Prince, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, 2007 , pp. 1 - 8.

[4] N. Brummer and E. De Villiers, "The Speaker Partitioning Problem," in *Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

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

[6] C. M. Bishop, Pattern Recognition and Machine Learning , 1st ed.: Springer, 2006.

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

[8] P. Matejka et al., "Full-Covariance UBM and Heavy-Tailed PLDA in I-Vector Speaker Verification," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[9] S. Lyu and E. P. Simoncelli, "Nonlinear Extraction of Independent Components of Natural Images using Radial Gaussianization," *Neural Computation*, vol. 21, no. 6, June 2009.