GENDER INDEPENDENT DISCRIMINATIVE SPEAKER RECOGNITION IN I–VECTOR SPACE

Sandro Cumani [*], *Ondřej Glembek* [o], *Niko Brümmer* [+], *Edward de Villiers* [+], *Pietro Laface* [*]

[*] Politecnico di Torino, Italy, Sandro.Cumani, Pietro.Laface@polito.it
[o] Brno University of Technology, Czech Republic, glembek@fit.vutbr.cz
[+] AGNITIO, South Africa, edwarddsp@gmail.com, niko.brummer@gmail.com

ABSTRACT

Speaker recognition systems attain their best accuracy when trained with gender dependent features and tested with known gender trials. In real applications, however, gender labels are often not given. In this work we illustrate the design of a system that does not make use of the gender labels both in training and in test, i.e. a completely Gender Independent (GI) system. It relies on discriminative training, where the trials are i–vector pairs, and the discrimination is between the hypothesis that the pair of feature vectors in the trial belong to the same speaker or to different speakers. We demonstrate that this pairwise discriminative training can be interpreted as a procedure that estimates the parameters of the best (second order) approximation of the log–likelihood ratio score function, and that a pairwise SVM can be used for training a gender independent system. Our results show that a pairwise GI SVM, saving memory and execution time, achieves on the last NIST evaluations state–of–the–art performance, comparable to a Gender Dependent(GD) system.

Index Terms— Speaker Recognition, I-vector, PLDA, Discriminative Training, SVM.

1. INTRODUCTION

State–of–the–art text–independent speaker recognition systems are often designed to achieve best performance when the gender label is known both at training and test time. The development of gender dependent systems is normal practice in NIST Speaker Recognition Evaluations [1], where the sites participating in the evaluations have access to the gender labels of the development and test segments. Gender information, however, is not available in a number of real applications. The speaker gender can be estimated from the trial data, but this preliminary classification is a potential source of accuracy degradation. A gender independent system has two benefits: a larger amount of training data can be used for off-line estimation of the Universal Background Model, and of channel and speaker sub-spaces, and its models require less memory and computation during testing.

In [2] a solution for the Probabilistic Linear Discriminant Analysis (PLDA) [3, 4] based on *i–vectors* [5] has been proposed to deal with the gender dependent problem. It uses a mixture of male and female PLDA models, but does not need the gender information in testing to obtain a gender-independent likelihood-ratio score. Two speaker detection and two gender discrimination scores, one for the male and one for the female models, are computed. These scores are then appropriately combined also taking into account the gender priors.

In this work we address the problem of designing a fully gender independent speaker recognition system. Our system, rather than using mixtures of generative models, relies instead on discriminative training in i-vector space and ignores the gender labels both in training and in test. In particular we show that in the pairwise discriminative framework [6, 7], derived from PLDA, it is possible to train a single GI SVM system whose performance is comparable to that of gender dependent SVM systems trained on the same data. Moreover, we give a new interpretation of pairwise discriminative training which helps to explain why a pairwise SVM system reaches the performance of state–of–the–art generative models both for GD and GI training.

The paper is organized as follows: Section 3 describes the i-vectors and we briefly recall the main PLDA models that have been proposed for GD and GI speaker recognition. Section 4 presents the pairwise discriminative training approach, focusing on a novel interpretation of this technique. Section 5 presents the experimental results, and in Section 6 we draw our conclusions.

2. I-VECTORS

I-vectors [8] provide an elegant way of reducing large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information.

The main idea is that the speaker- and channel-dependent Gaussian Mixture Model (GMM) supervector $\mathbf{s}$ can be modeled as:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the Universal Background Model (UBM) GMM mean supervector, $\mathbf{T}$ is a low-rank matrix representing M bases spanning subspace with important variability in the

mean supervector space, and $\mathbf{w}$ is a latent variable of size M with standard normal distribution. For an observation $\mathcal{X}$ the i-vector ϕ is the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{w}$, i.e. the mean $\mathbf{w}_{\mathcal{X}}$ of the posterior distribution $\mathrm{p}(\mathbf{w}|\mathcal{X})$.

3. GENERATIVE MODELS

I-vectors have been able to achieve good performance by means of simple LDA and cosine distance scoring [8]. Since their introduction, the speaker recognition community has focused on models for computing speaker verification scores directly from these low–dimensional features. Generative models based on Probabilistic Linear Discriminant Analysis [3, 4] have been proposed that achieve better performance than simple cosine scoring. In this section we briefly recall the general PLDA framework and its application to gender independent scoring [2].

3.1. PLDA

PLDA makes use of the following latent variable model for i-vectors:

$$\phi = m + U_1 y + U_2 x + \epsilon$$

where ϕ is an i-vector, y is a speaker factor, x a channel factor, and ϵ a residual noise. The model parameters are a mean vector m, a matrix U_1 whose columns are referred to as eigenvoices, and a matrix U_2 whose columns are referred to as eigenchannels. These matrices constrain the speaker and channel factors to be low–dimensional. The generation of an i-vector requires choosing a random speaker factor y according to speaker prior distribution $p(y)$ and a random channel factor x according to channel prior distribution $p(x)$. The i-vector is then the sum of $U_1 y + U_2 x$ and the residual noise ϵ generated according to distribution prior $p(\epsilon)$.

The model parameters are estimated in order to maximize the posterior probability for the observed i-vectors, assuming that i-vectors extracted from the same speaker share the same speaker factor, i.e. the same value for the latent variable y. The conditional likelihoods of two i-vectors can then be computed to obtain the speaker verification log–likelihood ratio score between the "same–speaker" hypothesis H_s and "different–speaker" hypothesis H_d:

$$s = \log \frac{P(\phi_1, \phi_2|H_s)}{P(\phi_1, \phi_2|H_d)} \quad (2)$$

where ϕ_1, ϕ_2 are two i-vectors (or two sets of i-vectors) that must be scored.

3.2. PLDA models

The simplest PLDA model (GPLDA) assumes a Gaussian distribution for the prior parameters. Under the Gaussian assumption, the log-likelihood ratio in (2) can be expressed in a closed form as a quadratic function of the i-vectors under consideration [3]. However, in the same reference it is shown that ML estimation of the PLDA parameters under Gaussian assumption fails to produce accurate models for i-vectors.

Thus, heavy tailed distributions for the model priors have been proposed leading to the Heavy-Tailed PLDA model, which however, is computationally expensive.

A simpler approach keeps the Gaussian distribution assumption, but incorporates a preprocessing step where the vector dimensionality is further reduced by means of LDA, and length normalization is then applied to the resulting patterns [9]. The performance of the two approaches is comparable, the latter being much faster both in training and testing.

The mixture of Gaussian PLDA models [2] is a recently proposed application of the GPLDA approach for GI speaker recognition. In this approach, two GD Gaussian PLDA models are trained using GI i–vectors, but a GI score is obtained from the two GD models' detection scores appropriately combined with the gender discrimination scores and the gender priors.

4. PAIRWISE DISCRIMINATIVE TRAINING

The PLDA framework has contributed to the development of a new flavour of discriminative techniques for speaker verification. Based on the two covariance PLDA formulation [10], a framework for speaker trial verification based on SVM and Logistic Regression has been introduced in [6, 7]. In the following we will refer to such a framework as pairwise discriminative training, where a speaker trial is a pair of speech segments, and the hypothesis to test is whether the two segments were spoken by the same speaker. In particular, we will discuss and present results for the pairwise SVM approach, where an SVM hyperplane is trained to directly answer this question.

4.1. From PLDA to discriminative training

Under the assumption of Gaussian distribution for the priors $p(y)$, $p(x)$ and $p(\epsilon)$, PLDA scoring can be reformulated as a quadratic function of the pair of i–vectors in a given trial. The formal expression of the log–likelihood ratio for the i–vector pair ϕ_1 and ϕ_2 is given by

$$\begin{aligned} s(\phi_1, \phi_2) &= \phi_1^T \Lambda \phi_2 + \phi_2^T \Lambda \phi_1 + \phi_1^T \Gamma \phi_1 + \phi_2^T \Gamma \phi_2 \\ &+ (\phi_1 + \phi_2)^T c + k \end{aligned} \quad (3)$$

where c, k, Λ, Γ are functions of the parameters of the PLDA model [6].

In [6] it has been shown that this expression can be reformulated as a dot–product in a feature space expanded from the original feature space consisting of i–vector pairs. As a consequence, a pairwise SVM system has been proposed: a classifier able to discriminate directly between the *same speaker*

class and *different speakers* class. The feature vectors for the SVM are pairs of i–vectors, and non–linear classification is obtained through the mapping

$$\varphi(\phi_1,\phi_2)=\begin{bmatrix}\mathrm{vec}(\phi_1\phi_2^T+\phi_2\phi_1^T)\\ \mathrm{vec}(\phi_1\phi_1^T+\phi_2\phi_2^T)\\ \phi_1+\phi_2\\ 1\end{bmatrix}=\begin{bmatrix}\varphi_\Lambda(\phi_1,\phi_2)\\ \varphi_\Gamma(\phi_1,\phi_2)\\ \varphi_c(\phi_1,\phi_2)\\ \varphi_k(\phi_1,\phi_2)\end{bmatrix} \quad (4)$$

where vec is the operator that stacks the columns of a matrix into a single vector.

4.2. A novel interpretation of pairwise SVM

The derivations in the previous section show that pairwise SVM and GPLDA are closely related. However, in [6] we showed that pairwise SVM achieves much better performance than GPLDA, unless normalization of the i–vectors is performed to better fit the Gaussian assumptions [9].

In this section we introduce a novel interpretation for pairwise SVM training which allows us to explain why this discriminative model achieves the same performance as Heavy–Tailed PLDA and why a GI system can be trained by simply pooling male and female trials, without performance loss with respect to a GD system.

Rather than making assumptions about the probability distributions that models the i–vector generation process, we focus directly on the log–likelihood ratio score as a function of two i–vectors, defined in (2):

$$s=\log\frac{P(\phi_1,\phi_2|H_s)}{P(\phi_1,\phi_2|H_d)}=s(\phi_1,\phi_2)=s(\Phi) \quad (5)$$

where $\Phi=[\phi_1^T\phi_2^T]^T$.

Under the assumption that $s(\Phi)$ admits a Taylor expansion around a point $\hat{\Phi}$, we can write $s(\Phi)$ as

$$s(\Phi)=\sum_{k=0}^{+\infty}\frac{\left(\left(\Phi-\hat{\Phi}\right)\cdot\nabla\right)^k s|_{\hat{\Phi}}}{k!} \quad (6)$$

where ∇ is the differential operator $\nabla=\left(\frac{\partial}{\partial\Phi_1},\ldots,\frac{\partial}{\partial\Phi_d}\right)$.

If we consider the second order approximation of such function around $\hat{\Phi}=0$ we obtain

$$s(\Phi)=s(\hat{\Phi})+(\Phi\cdot\nabla s|_{\hat{\Phi}})+\Phi^T(H_s|_{\hat{\Phi}})\Phi \quad (7)$$

where H_s is the Hessian of function s.

Assuming that $s(\phi_1,\phi_2)$ is an analytic function symmetric in its two arguments, that is $s(\phi_1,\phi_2)=s(\phi_2,\phi_1)$, we can set

$$H_s|_{\hat{\Phi}} = \begin{bmatrix}\Gamma & \Lambda\\ \Lambda & \Gamma\end{bmatrix} \quad (8)$$

By also setting $s(\hat{\Phi})=k$ and $\nabla s|_{\hat{\Phi}}=c$ we can rewrite (7) as:

$$\begin{aligned}s(\phi_1,\phi_2) &= k+(\phi_1+\phi_2)^Tc+\phi_1^T\Lambda\phi_2^T\\ &+ \phi_2^T\Lambda\phi_1+\phi_1^T\Gamma\phi_1+\phi_2^T\Gamma\phi_2\end{aligned} \quad (9)$$

It is worth noting that the choice of $\hat{\Phi}=0$ is not restrictive, since any other choice would lead to a formally equivalent expression for s. The log–likelihood ratio given in (3) for GPLDA and the one obtained in (9) by Taylor expansion have exactly the same expression. Although the two expressions are formally equivalent, an important difference has to be highlighted. The parameters estimated in GPLDA are constrained, due to the positive definiteness constraints of the covariance matrices of the PLDA model. In pairwise discriminative training, on the contrary, no parameter constraints are imposed, except for the ones arising from the regularization of the optimization function. Thus, pairwise discriminative training can be interpreted as a procedure that estimates the parameters of the best (second order) approximation of the log–likelihood ratio score function.

4.3. GI Pairwise SVM

The interpretation of pairwise discriminative training illustrated in the previous section provides the rationale for a straightforward approach to gender independent pairwise SVM training. The generative models of Heavy-Tailed or Mixtures of PLDA differ only in the formal expression of their log–likelihood ratio score function. In pairwise SVM training we directly optimize the best second order approximation of such functions. A gender independent SVM can therefore be implemented by training a single system with pooled gender i–vectors, without even the need for gender labels in the training stage. Some care might, however, be required in case of very unbalanced male and female training sets.

5. EXPERIMENTS

5.1. Experimental setup

In our experiments, we used cepstral features, extracted using a 25 ms Hamming window. 19 Mel frequency cepstral coefficients together with log-energy were calculated every 10 ms. This 20–dimensional feature vector was subjected to short time mean and variance normalization using a 3 s sliding window. Delta and double delta coefficients were then calculated using a 5–frame window giving 60-dimensional feature vectors. Segmentation was based on the BUT Hungarian phoneme recognizer and relative average energy thresholding. Also, short segments were pruned out, after which the speech segments were merged together.

One gender-independent UBM was represented as a full covariance 2048-component GMM. It was trained using LDC

Table 1. Results for the SRE2008 tests in terms of % EER and minDCF08 with 400 and 600 dimension i–vectors

Gender	Female		Male	
System	EER	minDCF08	EER %	minDCF08
400 GD	2.65 %	0.081	1.26	0.053
400 PGI	2.69 %	0.077	1.34	0.056
400 GI	2.54 %	0.078	1.29	0.056
600 GD	2.64 %	0.078	1.74	0.055
600 PGI	2.59 %	0.076	1.42	0.052
600 GI	2.39 %	0.067	1.18	0.044

releases of NIST 2004–2005 SRE telephone data. Both gender dependent and gender independent i–vector extractors were trained on the following telephone data: NIST SRE 2004–2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2.

Both 400 and 600 dimensional i–vectors were extracted.

5.2. Pairwise SVM

We compare the performance of three types of pairwise SVM systems: a fully GD system (GD), where both i–vector extraction and SVM training is gender dependent, a partially gender independent system (PGI) where the i–vectors are gender independent, whereas SVM is trained using GD trials, and finally a totally gender independent (GI) system, where both i–vectors and SVM are trained without using gender labels. For GD and PGI systems gender labels are provided at test time, while for the GI system no gender information is used to score trials. Pairwise SVMs are trained according to [6], applying Within–Class Covariance Normalization to the i–vectors.

Results are reported in Tables 1 and 2 for the tel–tel condition in the NIST 2008 and for the extended telephone condition in the NIST 2010 evaluations, respectively, and for 400 and 600 dimension i–vectors. The recognition accuracy is given in terms of Equal Error Rate (EER) and Minimum Detection Cost Functions defined by NIST for SRE 2008 (minDCF08) and SRE 2010 (minDCF10) [1].

Considering the performance of the 400-dimension i–vector systems on SRE08, the GI system has results comparable to the GD and to the partially gender independent system (PGI). Surprising results were obtained with the 600-dimension i–vector system on the same data: the minDCF is similar to the 400 i–vector GI and the PGI systems, but the EER is worse. The 600 GI system, on the contrary shows a substantial improvement mostly for male speakers.

On the more meaningful extended telephone tests of SRE10 the GI systems, both with 400 and 600 i–vectors, are comparable to the PGI systems and not far from the GD systems.

Overall these experiments show that the performance of a fully gender independent pairwise discriminative SVM system is comparable to the one of a more expensive GD model.

Table 2. EER and minDCFs for the SRE2010 tests with 400 and 600 dimension i–vectors

Gender	Female			Male		
System	EER	DCF08	DCF10	EER	DCF08	DCF10
400 GD	2.21 %	0.109	0.360	1.73 %	0.081	0.303
400 PGI	2.49 %	0.115	0.369	1.84 %	0.084	0.298
400 GI	2.51 %	0.115	0.382	1.82 %	0.087	0.309
600 GD	2.32 %	0.106	0.342	1.76 %	0.077	0.290
600 PGI	2.59 %	0.103	0.358	1.82 %	0.082	0.274
600 GI	2.51 %	0.108	0.383	1.80 %	0.078	0.307

6. CONCLUSIONS

A novel interpretation of pairwise discriminative training for speaker recognition has been presented, based on the best second order approximation of the the log-likelihood ratio score, which explains why discriminative training achieves state–of–the–art performance in speaker verification. A fully Gender Independent discriminative system has been trained which achieves, using GI i–vectors, the same performance of similar Gender Dependent systems. Its accuracy is just slightly worse than a fully GD system where i–vector extraction is also gender dependent.

7. REFERENCES

[1] NIST, "The NIST year 2008 and 2010 Speaker Recognition Evaluation plans," http://www.itl.nist.gov/iad/mig/tests/sre.

[2] M. Senoussaoui, P. Kenny, N. Brümmer, E. de Villiers, and P. Dumouchel, "Mixture of PLDA models in i–vector space for gender–independent speaker recognition," in *Proc. of Interspeech 2011*, 2011, pp. 25–28.

[3] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," in *Keynote presentation, Odyssey 2010*, 2010.

[4] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for inferences about identity," in *11th International Conference on Computer Vision*, 2007, pp. 1–8.

[5] N. Dehak, P. Kenny, R. Dehak, and P. Dumoucheland P. Ouellet, "Front–end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[6] S. Cumani, N. Brümmer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i-vector space," in *Proc. of ICASSP 2011*, 2011, pp. 4852–4855.

[7] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification," in *Proc. of ICASSP 2011*, 2011, pp. 4833–4836.

[8] 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 *Proc. of Interspeech 2009*, 2009, pp. 1559–1562.

[9] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i–vector length normalization in speaker recognition systems," in *Proc. of Interspeech 2011*, 2011, pp. 249–252.

[10] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey 2010*, 2010, pp. 194–201.