New Cosine Similarity Scorings to Implement Gender-independent Speaker Verification

Mohammed Senoussaoui[1,2], *Patrick Kenny*[2], *Pierre Dumouchel*[1] *and Najim Dehak*[3]

[1] École de technologie supérieure (ÉTS), Canada
[2] Centre de recherche informatique de Montréal, Canada
[3] MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA

{mohammed.senoussaoui, patrick.kenny}@crim.ca,
pierre.dumouchel@etsmtl.ca, najim@csail.mit.edu

Abstract

This paper is a natural extension of our previous work on gender-independent speaker verification systems [1]. In a previous paper, we presented a solution to avoid using gender information in the Probabilistic Linear Discriminant Analysis (PLDA) without any loss of accuracy compared with a gender-dependent base-line implementation. In this work, we propose two solutions to make a speaker verification system based on Cosine similarity independent of speaker gender. Our choice of the Cosine similarity is motivated by the fact that it is proved itself as a second state-of-the art - in parallel with PLDA- of i-vector based speaker verification systems. As measured by Equal Error Rate and min DCF's, performance results on the extended telephone list *coreext-coreext* condition of SRE2010[1] show no performance decrease in gender-independent Cosine similarity system compared to gender-dependent one. Tests were also successful for gender-independent propositions on a cross gender list as done in [1].

Index Terms: Speaker verification, Cosine similarity, gender-independent, i-vector.

1. Introduction

Traditionally, NIST Speaker Recognition Evaluation (SRE) tasks [2] involve lists that contain: (i) no cross-gender trials, (ii) information about trial gender. In the real word applications, it is not obvious how one can obtains such information. This is the reason why it is important to design a robust speaker verification system that can overcome the lack of information mentioned above.

The introduction of the low dimension representation, commonly called i-vector [3][4][5] greatly facilitated the task of the speaker recognition community. In fact, two principal methods emerged in speaker verification field with the arrival of i-vectors, namely Probabilistic Linear Discriminant Analysis (PLDA) with its two variants (i.e. Gaussian and heavy-tailed) [6] and Cosine similarity [4]. In our previous work we have successfully implemented a mixture of PLDA to deal with the total lack of gender information in enrollment and in test utterances. The main objective of this work is to propose solutions in order to implement a gender-independent Cosine similarity, which is the historic competitor of PLDA. Unlike to the mixture of Probabilistic Linear Discriminant Analysis proposed in [1], build a gender-independent system in the case of Cosine similarity is not straightforward. Indeed, the probabilistic nature of PLDA model allows building the gender-independent system by introducing a simple modification of the gender-dependent scoring rule. Effectively, this modification follows probability rules and prevents the explicit use of a gender detector usually based on a hard decision. In addition, PLDA model doesn't need score normalization, which facilitates the implementation of the mixture of PLDA. In this work we introduce new scoring rules in which we combine the Cosine similarity with a soft decision Gaussian gender detector in order to implement a gender-independent speaker verification system.

The rest of this paper is organized as follows. In the next section, we overview the standard gender-dependent Cosine similarity based speaker verification system. In Section 3 we present with more details our gender-independent speaker verification system based on Cosine similarity. In Section 4 we describe our feature extraction and carried out experiments, we also discuss the results. The conclusion is in Section 5.

2. Gender-dependent Cosine similarity (Gd)

The remarkable capability of i-vectors to capture the most important information characterizing a speaker with a moderate dimensionality allows the use of multiple traditional methods of filtering and normalizing data in order to improve classification accuracy. In the following paragraphs we will show the steps needed to make Cosine similarity working for a speaker verification system [4].

Suppose we are given two i-vectors e and t (e and t stand for enrollment test i-vectors respectively) in a typical speaker verification system based on Cosine similarity. These i-vectors are subject to a Linear Discriminant Analysis (LDA) projection (e.g. from 800 dimensions to 200 dimensions). Usually, a shift of LDA projected data using a sample mean - estimated on background utterances- followed by a rotation via the inverse of the Within Class Covariance (WCC) matrix are required in order to centralize data and to penalize axes that maximize within class variability. After these steps, a length (i.e. Euclidean norm) normalization of i-vectors [7] allows an accurate classification by a simple dot product.

Recently in [8], a further normalization followed LDA, sample mean, WCC, and length normalization of i-vectors using the mean and covariance matrix of some background cohorts was proposed. This proposal is to simulate the score normalization, which is an essential step for improving Cosine scoring performances. The above process of i-vector normalizing is depicted in Figure 1.

2.1. Raw Cosine scoring

Without loss of generality we can suppose that e and t were already subject to the sequence of transformations (except of

[1] http://www.itl.nist.gov/iad/mig//tests/sre/ 2010/index.html

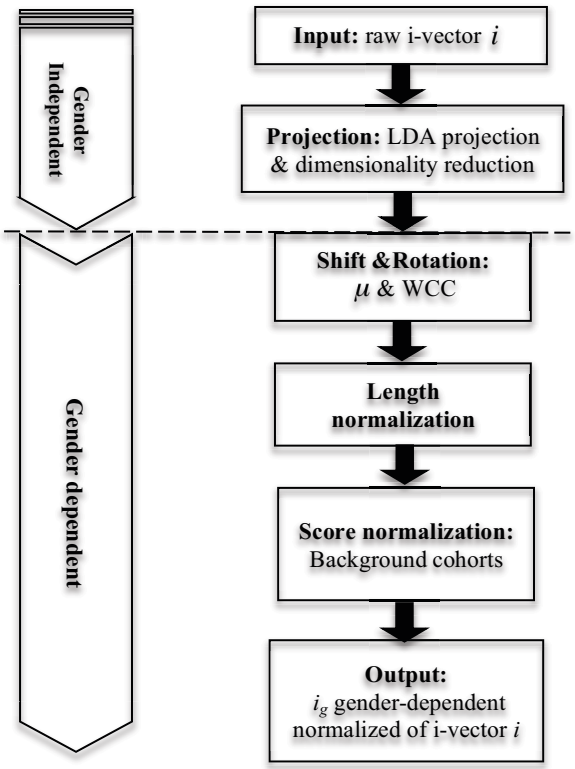

Figure 1: *Block diagram describing the extraction procedure of normalized gender-dependent i-vector i_g from a given raw i-vector. The index g is a discrete hidden variable indicates the gender of the i-vector.*

score normalization step) depicted in Figure 1, namely LDA projection, mean shifting normalization, WCC rotation. In this case, the Cosine scoring reduced to a dot product as follows:

$$raw\_score\_Gd(e,t) = e^* \cdot t \tag{1}$$

where *e* and *t* are normalized as we cited previously and their lengths are normalized as follow:

$$e = \frac{e}{\|e\|} \qquad t = \frac{t}{\|t\|} \tag{2}$$

Note that the exponent star in all our mathematic formulas will refer to the transpose operator and $\|\cdot\|$ is the Euclidean norm of a given vector.

2.2. Score normalization

Normalization of scores is an essential step to produce an efficient speaker verification system based on Cosine similarity. Traditionally, Cosine similarity scoring works well with zt-norm score normalization method. In this work we will use a proxy of zt-norm proposed in [8]. Due to the symmetric propriety of Cosine scoring, the zt-norm like scoring has a great advantage, as one could estimate all needed parameters (namely a mean vector and covariance matrix of a cohort set) from a set of background i-vectors in an offline way. The zt-norm like scoring is given by the following formula:

$$zt\_score\_Gd(e,t) = \frac{(e-\mu_{imp})^* \cdot (t-\mu_{imp})}{\|C_{imp} \cdot e\| \cdot \|C_{imp} \cdot t\|} = \left(\frac{(e-\mu_{imp})}{\|C_{imp} \cdot e\|}\right)^* \cdot \left(\frac{(t-\mu_{imp})}{\|C_{imp} \cdot t\|}\right) \tag{4}$$

where μ_{imp} and C_{imp} are respectively the mean and the Cholesky decomposition of covariance matrix of some cohort background i-vectors normalized with the same sequence of operations as *e* and *t*. In fact, if we suppose independence constraint, we can take C_{imp} as the square root of the diagonal of impostor covariance matrix. Observing zt-norm formula we can realize that we are able to apply score normalization separately on enrollment and test i-vectors.

3. Gender-Independent Cosine similarity

We use a discrete hidden variable *g* that takes its values in the set $\{F, M\}$ for female and male genders respectively. Before going into details of the proposed gender-independent system, we first describe the gaussian gender-detector based on the WCC's covariance matrix.

3.1. Gaussian gender detector

Despite the fact that it is possible to build a gender detector simply by training a two-class classifier, we propose to explore a simple idea of modeling each gender with a multi-dimensional gaussian distribution in the i-vector space. The originality of our idea is the use of the gender-dependent sample mean μ and the within class covariance matrix (see Figure 1) respectively as mean vector and covariance matrix of the gender-dependent gaussian model. Proceeding as proposed we need no extra training procedures for the gender detector. When we test this gender detector on SRE 2010 telephone data (det5 utterances) we get ~1.9% of error of identification (see Table 1 for more results).

Table 1 *Performance of the gaussian gender-detector on a subset of SRE 2010 telephone data (set5 data) as measured by error of identification (Error shown in %); the number of test observations is also provided (#Obs).*

MALE		**FEMALE**		**Mean / Total**	
Error (%)	#Obs	Error (%)	#Obs	Error (%)	#Obs
1.56	2294	2.29	2740	**1.92**	5034

3.2. Naïve Gender-independent scoring (NGi)

If we prefer a simplest (naïve) way to implement a gender independent scoring using Cosine similarity, one can estimate gender-dependent parameters depicted in Figure 1 (i.e. WCC, sample mean μ and parameters of score normalization) in a gender-independent manner by pooling all male and female data. Indeed, in this case we will not need any gender-detector to perform gender-independent scoring.

In fact, from our previous work [1] and the theoretical point of view we will explain later in the section 3.4, that is a hard task to get the naïve gender independent working especially in the presence of score normalization step. Although this proposal does not seem mature enough to deal with this problem we believe that it should serve as a benchmark.

3.3. Gender-independent scoring (Gi)

Without loss of generality, let us suppose that we can derive two gender-dependent normalized i-vectors (e_F, e_M and t_F, t_M) of each of enrollment i-vectors *e* and test i-vector *t* by applying the full procedure illustrated in Figure 1. In the context of NIST SRE in which the gender of trials is controlled (i.e. there are only same gender trials), a simple sum of male and female scores weighted with combined gender detector outputs λ_{FF} and λ_{MM} allows the calculation of the gender-independent (Gi) score:

$$\begin{aligned} zt\_score\_Gi(e,t) &= \left(p(e|F)\cdot e_F^* \cdot p(t|F)\cdot t_F\right) + \\ &\quad \left(p(e|M)\cdot e_M^* \cdot p(t|M)\cdot t_M\right) \\ &= \left(\lambda_{FF}\cdot(e_F^*\cdot t_F)\right) + \left(\lambda_{MM}\cdot(e_M^*\cdot t_M)\right) \end{aligned} \quad (4)$$

where e_F, e_M, t_F and t_M are male and female normalized i-vectors of enrolment and test raw i-vectors extracted respectively. Weight coefficients are calculated by combining gender detector outputs as follows:

$$\lambda_{FF} = p(e|F)\cdot p(t|F) \quad (5)$$

$$\lambda_{MM} = p(e|M)\cdot p(t|M) \quad (6)$$

We observe in the first line of gender-independent (Gi) score that each normalized enrollment or test i-vectors is weighted with the probability that its raw i-vector is either male or female gender (i.e. $P(.|g)$). These probabilities are the outputs of the gaussian gender-detector and are normalized in order to sum to one.

3.4. Cross Gender-independent scoring (CGi)

So far, we have proposed a gender-independent Cosine scoring suitable in situations where we will never be exposed to a cross gender trial (i.e. male for enrollment and female for test or vice versa). At first glance, this problem seems easy because the discrimination of two speakers from different genders is easier than if they were from the same gender. Indeed, this is not entirely right since one can imagine a situation of a traditional gender-dependent GMM/UBM system with t-norm scoring based on a hard decision of a gender detector to select the gender of the model and the cohorts to be used. In a given non-target verification trial, we compare a female speaker model with a test speaker who happens to be female, but our gender detector suggests us to use a male model. So we select the male impostor cohorts for t-normalization and find that the test segment score is very high. Thus, our system will wrongly conclude that the trial in question is a target trial.

In the previous paragraph we explained the importance of addressing the issue of cross gender with caution. Now, we present a Cosine scoring that takes care of cross gender trials by combining all scoring possibilities as follows:

$$\begin{aligned} zt\_score\_CGi(e,t) &= \left(\lambda_{FF}\cdot(e_F^*\cdot t_F)\right) + \left(\lambda_{FM}\cdot(e_F^*\cdot t_M)\right) + \\ &\quad \left(\lambda_{MF}\cdot(e_M^*\cdot t_F)\right) + \left(\lambda_{MM}\cdot(e_M^*\cdot t_M)\right) \end{aligned} \quad (7)$$

where λ_{FM} and λ_{MF} are cross gender weights calculated as in formulas (3) and (4):

$$\lambda_{FM} = p(e|F)\cdot p(t|M) \quad (8)$$

$$\lambda_{MF} = p(e|M)\cdot p(t|F) \quad (9)$$

Weights are normalized by their sum in order to sum-up to one.

In the subsequent section we will present experiential results which confirm that Gi and CGi propositions provide a nice solution for making Cosine similarity scoring independent from speakers gender.

4. Experimental setup

4.1. Feature extraction and normalization

4.1.1. MFCC extraction

Each 10ms, 60 Mel Frequency Cepstral Coefficients (MFCC) were extracted within a 25ms hamming window (19 MFC Coefficients + Energy + first & second Deltas) from speech signal. These features were normalized with a short time Gaussianization.

4.1.2. Universal Background Model (UBM)

We use a gender-independent GMM UBM containing 2048 Gaussians. This UBM is trained with the LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004–2005 SRE.

4.1.3. i-vector extractor

We use a gender-independent i-vector extractor of dimension 800. Its parameters are estimated on the following data: LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; Fisher data and NIST 2004 and 2005 SRE (i.e. telephone speech) and all NIST microphone data (i.e. NIST 2005, 2006 and 2008 interview development microphone data).

4.1.4. Linear Discriminant Analysis

We use a gender-independent LDA to reduce i-vector dimensionality from 800 to 200. Its parameters are estimated on the following data: LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; NIST 2004 and 2005 SRE (i.e. telephone speech) and all NIST microphone data (i.e. NIST 2005, 2006 and 2008 interview development microphone data).

4.1.5. WCC and the sample mean

Unlike from UBM, i-vector extractor and LDA are gender-dependent within class covariance matrix and sample mean μ.

These two parameters are estimated using the 200-dimensional i-vectors of NIST 2004, 2005 and 2006 telephone data. Moreover, we estimate a gender-independent WCC by pooling male and female data.

4.1.6. Score normalization parameters

The score normalization parameters, namely μ_{imp} and C_{imp}, are estimated on NIST 2004, 2005 and 2006 telephone data. Before estimating μ_{imp} and C_{imp} background i-vectors are subject to a sequence of transformations as depicted in Figure 1, namely LDA projection, mean subtraction, WCC rotation

and length normalization. Note that we used exactly the same i-vector extractor as used in our previous paper [1].

4.2. Experiments and discussions

4.2.1. Test on NIST (det5) list

In this section we present results for the three gender-independent Cosine scorings and compare them with the gender-dependent Cosine scoring. In addition, we present our results of the mixture of PLDA published in [1] in order to carry out further comparisons.

Table 2. *Performance of gender-dependent (Gd), Naïve gender-independent (NGi), gender independent (Gi), cross gender-independent (CGi) and the mixture of PLDA (*MixPLDA*) test on NIST det5 (telephone) list.*

		EER (%)	MinDCF_08	MinDCF_10
MALE	Gd	1.67	0.091	0.415
	NGi	2.60	0.167	0.699
	Gi	**1.66**	**0.090**	**0.402**
	CGi	1.67	0.091	0.405
	MixPLDA	1.81	0.096	0.322
FEMALE	Gd	2.60	0.151	0.583
	NGi	3.69	0.250	0.687
	Gi	**2.59**	0.149	0.550
	CGi	2.61	**0.148**	**0.545**
	MixPLDA	2.46	0.124	0.388

Table 2 shows results of our gender-independent propositions (i.e. Gi and CGi) that maintain the same performance of a gender-dependent (Gd) system. In addition, we can also see that the cross gender-independent (CGi) scoring outperforms all others for female data. Finally, as expected, naïve gender-independent system performances were the worst for both genders. We can also observe that Cosine similarity results are similar with PLDA ones (see grey highlighted row in Table 2), given that we did a *perfect back-to-back* comparison.

In order to provide a more general demonstration of results on different operating points, we produce a DET plot depicted in Figure 2. In fact, the effectiveness of our proposals, namely gender-independent Gi and cross-gender independent CGi, is clear from observing Figure 2, in which we can note the superposition of Gi and CGi curves with the gender-dependent Gd curve.

4.2.2. Test on cross gender list

To carry out cross-gender tests we proceeded as follows. Firstly, we score the cross-gender list that we have created by replacing the non-target trials in the NIST extended list by the cross gender trials [1]. Finally, we use θ_{08} and θ_{10} to refer to thresholds used to obtain respectively 2008 and 2010 ***minimum*** of NIST DCFs already calculated on the scored det5 list of NIST (pooled males and females) using Gi or CGi scorings. The idea is to use θ_{08} and θ_{10} to calculate ***actual*** DCFs of the cross-gender list scores. Since, both lists share the same target trials, and have the same number of non-target trials, we expect that these actual DCFs should be at least equal to or less than the minimum DCFs calculated on the NIST list. Note that, the error rates in these circumstances will depend on the proportions of cross-gender trials to same-gender trials among the non-target trials. Therefore, the minimum of DCF is not really meaningful.

Table 3. *NIST list vs. cross-gender list for gender independent (Gi), cross gender-independent (CGi) and the mixture of PLDA (*MixPLDA*). Results are for pooled gender scores.*

		EER (%)	Actual DCF_08	Actual DCF_10
NIST list	Gi	**2.18**	0.125	0.522
	CGi	2.19	**0.124**	**0.509**
	MixPLDA	2.24	0.119	0.381
Cross-gender list	Gi	**0.89**	**0.071**	0.412
	CGi	1.03	**0.071**	**0.394**
	MixPLDA	0.40	0.078	0.349

As expected there is a net gain in the actual DCFs in the cross-gender list compared to NIST list (see Table 3). Furthermore, we observe that EER for gender-independent (Gi) proposal outperforms cross-gender independent (CGi) one in both lists. However, DCF_10 of CGi is slightly better than Gi one.

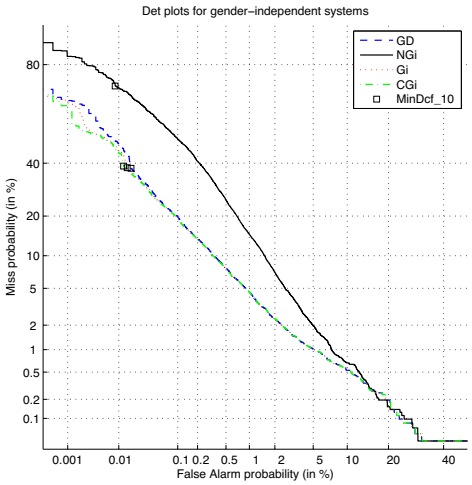

Figure 2: *Det curves for different systems. Note that these curves are obtained by pooling male and female scores.*

5. Conclusions

This paper is an extension of our pervious one [1] in which we succeeded to implement a gender-independent speaker verification system based on mixture of PLDA. We show how to build a gender-independent speaker verification system based on Cosine similarity. Using, two proposals, namely gender-independent Gi and cross-gender independent CGi, to combine Cosine scoring, enable us to get good results on extended det5 of SRE 2010 without taking advantage of gender information. In this work, our main contribution was twofold. In one hand, we have designed an efficient Gaussian gender-detector based on WCC as covariance matrix, which has spared us efforts of training extra parameters of an independent gender-detector. In the other hand, we used the outputs of this gender-detector to weight the sum of gender-dependent Cosine similarity in order to build the gender-independent system. Finally, in a *back-to-back* comparison, Cosine results were comparable with mixture of PLDA results obtained in the previous work using the same i-vectors.

6. References

[1] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers and P. Dumouchel, “Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition,” in Proceedings of Interspeech, Florence, Italy, Aug. 2011.

[2] A.F. Martin and C.S. Greenberg, “NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels,” in Proceedings of Interspeech, Brighton, UK, Sep. 2009, pp. 2579–2582.

[3] L. Burget et al., “Robust speaker recognition over varying channels,” in Johns Hopkins University CLSP Summer Workshop Report, 2008, online: http://www.clsp.jhu.edu/workshops/ws08/documents/jhu_report_main.pdf.

[4] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” in Proceedings of Interspeech, Brighton, UK, Sep. 2009, pp. 1559–1562.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” 2011, in IEEE Transactions on Audio, Speech and Language Processing, 16(5), pp. 980-988.

[6] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” in Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, Jun. 2010.

[7] D. Garcia-Romero, “Analysis of i-vector length normalization in speaker recognition systems,” in Proceedings of Interspeech, Florence, Italy, Aug. 2011.

[8] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine Similarity Scoring without Score Normalization Techniques," Proc. IEEE Odyssey Workshop, Brno, Czech Republic, June 2010.