Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition

Mohammed Senoussaoui,[1,3] *Patrick Kenny,*[1]
Niko Brümmer,[2] *Edward de Villiers*[2] *and Pierre Dumouchel*[1,3]

[1]CRIM, Canada, {mohammed.senoussaoui,patrick.kenny,pierre.dumouchel}@crim.ca
[2]AGNITIO Labs, South Africa, {nbrummer,edevilliers}@agnitio.es
[3]École de Technologie Suprieure (ÉTS), Canada

Abstract

The Speaker Recognition community that participates in NIST evaluations has concentrated on designing gender- and channel-conditioned systems. In the real word, this conditioning is not feasible. Our main purpose in this work is to propose a *mixture of Probabilistic Linear Discriminant Analysis models* (PLDA) as a solution for making systems independent of speaker gender. In order to show the effectiveness of the mixture model, we first experiment on 2010 NIST telephone speech (det5), where we prove that there is no loss of accuracy compared with a baseline gender-dependent model. We also test with success the mixture model on a more realistic situation where there are cross-gender trials. Furthermore, we report results on microphone speech for the det1, det2, det3 and det4 tasks to confirm the effectiveness of the mixture model.

Index Terms: i-vectors, speaker recognition, mixture model

1. Introduction

The series of NIST Speaker Recognition Evaluations [1] has had a strong influence on research in text-independent speaker recognition for more than a decade. In these evaluations, the canonical *speaker detection* task has always prescribed trials where (i) genders are not mixed and (ii) the genders of the speakers involved are given. Although this task definition has allowed researchers to concentrate on certain core aspects of the technology, it has also encouraged system designs that cannot function *as is* in more realistic environments where males and females may be mixed and where no gender labels are given. In this paper we propose a system design for a speaker detector that can function without gender labels and we demonstrate that we do not lose accuracy when we compare it with a gender-dependent baseline where such labels are provided.

The *i-vector* is a low-dimensional representation of an entire speech segment [2, 3, 4]. In recent work we showed that i-vectors respond well to generative modelling [5, 6]. In those papers, we used separate male and female i-vector extractors, followed by separate male and female generative modelling stages, choosing between them by using the gender labels provided by NIST. In this paper, we use a gender-independent *i-vector extractor* and combine male and female models in a mixture, where the gender label is treated as a hidden variable.

Handling *cross-gender* trials is tricky for traditional speaker verification systems, including those based on joint factor analysis, because some sort of score normalization such as zt-norm is essential. (For each trial, z-norm and t-norm imposter cohorts have to be selected for both the hypothesized speaker and the test segment using a gender detector.) However as shown in [5, 7], score normalization heuristics are not generally needed for a PLDA based speaker verification system. We will show how this makes it possible to handle cross gender trials by a straightforward application of the rules of probability, using a mixture of male and female PLDA models (but no explicit gender detection) to perform speaker recognition.

The remainder of this paper is organized as follows. In the next section we present the generative model used for speaker detection. We will focus on the mixture solution proposed to deal with the gender dependent problem. The third section reports experimental results on NIST 2010 telephone and microphone speech. The last section presents concluding remarks.

2. Generative Model for Speaker Detection

Here we define speaker detection and present our generative model for this problem.

2.1. Speaker detection

In a *speaker detection trial*, two speech segments are given, each assumed to have been produced by a single speaker and the question is asked whether the segments were produced by the same speaker, or by two different speakers. We use the convention that the same-speaker hypothesis is called a *target* trial and the different speaker hypothesis a *non-target* trial. The speaker detector processes the given speech segments and outputs a scalar *score*, where a more positive value favours the target hypothesis and a more negative value favours the non-target hypothesis.

Note that by definition, target trials cannot have mixed gender, but non-target trials may be *male*, *female* or *mixed*. In the NIST evaluations, there are no trials of mixed gender and male/female labels are provided for each trial. In this paper, we are interested in the case where there may be mixed non-target trials and where no gender labels are provided.

In what follows, we assume that every speech segment is mapped to an i-vector, which lives in D-dimensional Euclidean space. (We used $D = 800$.) This mapping is done with an *i-vector extractor*, as explained in [4, 5, 6]. In those papers, we used separate male and female extractors. In this work, we train a single extractor on pooled male and female data and then apply it to all evaluation speech segments regardless of gender.

The consequence of the single i-vector mapping is that the input data of every detection trial is just a pair of i-vectors. (If one follows tradition, denoting the two speech segments of the detection trial as *enrollment* and *test*, then the two i-vectors may also labelled as such. However, the i-vector recipe is symmetric in its inputs, so that there is no real difference between enrollment and test. Below we simply number the two segments as 1 and 2.)

2.2. Generative Model

We construct a speaker detector by using a generative PLDA model for the pair of i-vectors in a trial. The model ignores the real-world complexities of speech production, transmission and processing and instead pretends that the i-vectors were produced by simple random processes. The model pretends the pair of i-vectors, denoted $\mathbf{z}_1, \mathbf{z}_2$, is produced as follows:

$$\mathbf{z}_1 = \mathbf{V}\mathbf{y}_1 + \mathbf{x}_1 \qquad\qquad \mathbf{z}_2 = \mathbf{V}\mathbf{y}_2 + \mathbf{x}_2 \tag{1}$$

where the *hidden speaker variables*, $\mathbf{y}_1, \mathbf{y}_2$, are d-dimensional vectors sampled from a continuous multivariate *between speaker* distribution. For target trials: $y_1 = y_2$, while for non-target trials: y_1 and y_2 are sampled independently. The *hidden channel*[1] *variables*, $\mathbf{x}_1, \mathbf{x}_2$ are D-dimensional and sampled independently from a continuous multivariate *within speaker* distribution. Usually $d \leq D$, but in our experiments we use $d = D$. The between and within speaker distributions are either normal or heavy-tailed [5]. The D-by-d matrix $\mathbf{V}$ is a fixed hyperparameter.

In this model, $\mathbf{z}_1, \mathbf{z}_2$ are the *observed* variables. There are two types of *hidden* variables: (i) the continuous *nuisance* variables: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$ and (ii) the variable of interest to be inferred, i.e. the *trial-type*, which can have the discrete values *target* ($\mathcal{T}$) or *non-target* ($\mathcal{N}$).

2.2.1. Gender modelling

Now we add another pair of discrete hidden variables, namely g_1, g_2, which respectively represent the genders of the speakers that produced $\mathbf{z}_1$ and $\mathbf{z}_2$. These variables take the values *male* ($\mathcal{M}$) or *female* ($\mathcal{F}$). For a target trial, $g_1 = g_2$, while and for a non-target trial they may be different.

The generative model needs *priors* for all hidden variables. The priors for the continuous hidden variables are the above-mentioned within and between speaker distributions. The prior for the trial type will not be needed in this paper. The priors for the gender labels are trial-type dependent and are defined as:

$$\begin{aligned} P_{\mathcal{M}} &= P(\mathcal{MM}|\mathcal{T}) & P_{\mathcal{F}} &= P(\mathcal{FF}|\mathcal{T}) \\ Q_{\mathcal{MM}} &= P(\mathcal{MM}|\mathcal{N}) & Q_{\mathcal{FF}} &= P(\mathcal{FF}|\mathcal{N}) \\ Q_{\mathcal{MF}} &= P(\mathcal{MF}|\mathcal{N}) & Q_{\mathcal{FM}} &= P(\mathcal{FM}|\mathcal{N}) \end{aligned}$$

where e.g. $\mathcal{MF}$ denotes the event: $g_1 = \mathcal{M}$ and $g_2 = \mathcal{F}$. Note $P_{\mathcal{M}} + P_{\mathcal{F}} = 1$ and $Q_{\mathcal{MM}} + Q_{\mathcal{FF}} + Q_{\mathcal{MF}} + Q_{\mathcal{FM}} = 1$. The gender priors must be supplied by the user of the speaker detector (we will use equiprobable priors in this work). In the limiting case of given gender labels, these priors will have values of 0 or 1.

2.3. Scoring

For scoring this model in a gender-independent way, we marginalize over the gender hidden variables. For this purpose we assume that we have available the following gender-dependent *likelihoods*: For targets, we need $P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{MM}, \mathcal{T})$, $P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{FF}, \mathcal{T})$ and for non-targets we need:

$$P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{MM}, \mathcal{N}) = P(\mathbf{z}_1|\mathcal{M})P(\mathbf{z}_2|\mathcal{M}) \tag{2}$$

$$P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{FF}, \mathcal{N}) = P(\mathbf{z}_1|\mathcal{F})P(\mathbf{z}_2|\mathcal{F}) \tag{3}$$

When using normal distributions for the continuous hidden variables, all these likelihoods can be computed in closed form [6], or when using heavy-tailed distributions, they can be approximated [5]. These likelihoods can be arranged into the following likelihood-ratios:

$$R_{\mathcal{M}} = \frac{P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{MM}, \mathcal{T})}{P(\mathbf{z}_1|\mathcal{M})P(\mathbf{z}_2|\mathcal{M})} \tag{4}$$

$$R_{\mathcal{F}} = \frac{P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{FF}, \mathcal{T})}{P(\mathbf{z}_1|\mathcal{F})P(\mathbf{z}_2|\mathcal{F})} \tag{5}$$

$$G_i = \frac{P(\mathbf{z}_i|\mathcal{M})}{P(\mathbf{z}_i|\mathcal{F})} \tag{6}$$

Note that $\log R_{\mathcal{M}}$ and $\log R_{\mathcal{F}}$ are the usual gender-dependent speaker detection scores for the male and female models respectively. $\log G_i$ can be used as a gender discrimination score (in experiments with the normal model variant on the telephone data of NIST SRE2010,[2] G_i gives an equal error-rate (EER) of below 2% when discriminating gender).

The *gender-independent likelihood-ratio score*, denoted $\bar{R}$, is obtained by marginalizing (summing) over the gender variables:

$$\begin{aligned} \bar{R} &= \frac{P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{T})}{P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{N})} \\ &= \frac{P_{\mathcal{M}}P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{MM}, \mathcal{T}) + P_{\mathcal{F}}P(\mathbf{z}_1, \mathbf{z}_2|\mathcal{FF}, \mathcal{T})}{\sum_{g_1,g_2} Q_{g_1g_2} P(z_1|g_1)P(z_2|g_2)} \end{aligned} \tag{7}$$

where there are four terms in the denominator. Finally, $\bar{R}$ can be expressed in terms of the above-defined likelihood-ratios and gender priors:

$$\bar{R} = \frac{P_{\mathcal{M}}}{Q_{\mathcal{MM}}} S_{\mathcal{M}} R_{\mathcal{M}} + \frac{P_{\mathcal{F}}}{Q_{\mathcal{FF}}} S_{\mathcal{F}} R_{\mathcal{F}} \tag{8}$$

where

$$S_{\mathcal{M}} = \frac{Q_{\mathcal{MM}} G_1 G_2}{Q_{\mathcal{MM}} G_1 G_2 + Q_{\mathcal{FM}} G_2 + Q_{\mathcal{MF}} G_1 + Q_{\mathcal{FF}}} \tag{9}$$

$$S_{\mathcal{F}} = \frac{Q_{\mathcal{FF}}}{Q_{\mathcal{FF}} + Q_{\mathcal{MF}} G_1 + Q_{\mathcal{FM}} G_2 + Q_{\mathcal{MM}} G_1 G_2} \tag{10}$$

2.3.1. Caveat

The independence assumption in (2) and (3) holds when the model parameters are assumed known at scoring time [5, 6]. This would not be the case in a more fully Bayesian treatment, where the uncertainty in the estimates of the model parameters is taken into account during scoring [8].

3. Experiments

We performed experiments on the *coreext-coreext* condition of the NIST extended list. We use the *Equal Error Rate* (EER) and the (new and old) normalized minimum *Detection Cost Function* [2] (DCF) of NIST as metrics. In order to show the effectiveness of the mixture gender independent model, we first report

[1] The synecdoche *channel* is understood to represent everything that causes different i-vectors of the same speaker to be different.

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

results on telephone speech (det5). We also carried out experiments on *cross-gender* trials by creating our own list. That list contains the same *target* trials as the NIST extended list det5 (3465), and the same number of *non-target* trials (175873), but all those *non-target* trials are *cross-gender* (i.e. male as model segment and female as test segment or vice versa, since the scoring is symmetric in our generative model). The reason for using the *cross-gender* list is to test our mixture model in a more realistic scenario. We also perform experiments on microphone speech and we will present results for det1 to det4 conditions.

3.1. Feature extraction

3.1.1. Universal Background Model

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. Speech parameters are represented by a 60-dimensional vector of Mel Frequency Cepstral Coefficients (MFCC) i.e. static MFCC, first and second derivative of MFCC.

3.1.2. 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 05, 06 and 08 interview development microphone data).

3.1.3. Linear Discriminant Analysis (LDA) in i-vector space

LDA is a well known supervised method, widely used for dimensionality reduction in classification problems. In our work, LDA is applied to map i-vectors from the 800-dimension space to a reduced space spanned by the *eigenvectors* corresponding to the biggest *eigenvalues* of the following generalized eigenvalue problem:

$$S_b u = \lambda S_w u \tag{11}$$

where S_b and S_w represent respectively the between class and the within class scatter matrices. Given a set of S speakers with speaker s having n_s utterances and $\sum_{s=1}^{S} n_s = N$, the scatter matrices S_b and S_w are given by the following formulas:

$$S_b = \sum_{s=1}^{S} n_s (\bar{\mathbf{z}}_s - \bar{\mathbf{z}})(\bar{\mathbf{z}}_s - \bar{\mathbf{z}})^t \tag{12}$$

$$S_w = \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{z}_i^s - \bar{\mathbf{z}}_s)(\mathbf{z}_i^s - \bar{\mathbf{z}}_s)^t \tag{13}$$

In practice, the estimation of S_b on only telephone data and S_w on telephone and microphone data works well. An optimal reduced dimension equal to 200 is also determined empirically.

3.1.4. Length normalization of i-vectors

Recently, in [9],some experimental results on speaker verification using i-vectors as features and a PLDA model (without score normalization) based normal assumption [5], have shown that the normalization of the length of i-vectors to one, after LDA projection, gives comparable results to those obtained by the same PLDA model based heavy-tailed prior distribution [5]. High dimensional data can be Gaussianized by whitening and projecting onto the unit sphere. For an entertaining discussion of this curious fact see.[3] The advantage behind this normalization is the ability to use the Gaussian assumption in the PLDA model rather than the heavy-tailed assumption, which is more complicated and time consuming.

3.1.5. PLDA model training

Three PLDA models were trained for our experiments: two gender-dependent (*GD*) models and a gender-independent (*GI*) model. All models were trained on the same data sets as the *i-vector extractor* (i.e. telephone and microphone speech) except Fisher data. For all three models, the fixed hyperparameter $\mathbf{V}$ is a full rank matrix of dimension $d = 200$. The *mixture* (*Mix*) model is implemented by combining the *GD* models (i.e. male and female models) as shown in section 2.2.

3.2. Test on telephone speech

3.2.1. NIST list

In order to show the effectiveness of the mixture model, we perform the first series of experiments on only telephone speech. We report male and female results of *mixture*, gender-independent and gender-dependent systems as scored on NIST's extended list (det5). Observing these results, we can easily see that the use of the mixture model (see columns 1 and 4 of Table 1) gives almost the same results as the baseline gender-dependent model (see columns 3 and 6 of Table 1). On the other hand, we see an improvement of *mixture* model results compared with gender-independent model (see columns 2 and 5 of Table 1). In the *male* case, the improvement is about **9.5%** in *EER* and the *old MinDCF* decreases from **0.112** to **0.096**.

Table 1: *Male and female det5 results for mixture, gender dependent and gender independent models, measured by EER and normalized minimum DCF.*

	MALE			*FEMALE*		
	Mix	*GI*	*GD*	*Mix*	*GI*	*GD*
EER(%)	1.81	2.00	1.81	2.46	2.75	2.47
OldDCF	0.096	0.112	0.096	0.124	0.133	0.124
NewDCF	0.322	0.386	0.320	0.388	0.415	0.387

3.2.2. Cross gender list

To realize cross-gender tests we proceed as follows: First, we score the cross-gender list that we have created. Then, we use θ_{old} and θ_{new} to refer to thresholds used to obtain respectively *Old* and *New* minimums of *DCF* already calculated on the scored *det5* list of NIST (pooled males and females) using the *mixture* model. The idea is to use θ_{old} and θ_{new} to calculate actual DCFs of the scores of the cross-gender list. Since, both lists share the same *target* trials, and also 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.

As we expected, the *old actual DCF* decreases from *old minimum DCF* = 0.119 to **0.078** and the *new actual DCF* decreases from *new minimum DCF* = 0.381 to **0.349** (see Table 2).

[3] `http://ontopo.wordpress.com/2009/03/10/reasoning-in-higher-dimensions-measure/`

Table 2: *Pooled male and female det5 results as measured by EER, Normalized minimum DCFs using mixture model and pooled gender results as measured by EER, Normalized actual DCFs scored using mixture model on cross-gender list.*

	NIST list	*CrossG list*
EER(%)	2.24	**0.40**
OldDCF	0.119	**0.078**
NewDCF	0.381	**0.349**

3.3. Test on microphone speech

In the previous section we reported results that demonstrate the effectiveness of the mixture model, at least with telephone speech. Given the encouraging results obtained in the telephone speech case, we decided to go further and test the mixture model on microphone data. First, we report in Table 3 results on *interview-interview* det2 task of the NIST extended list.

Table 3: *Male and female det2 results on mixture, gender dependent and gender independent models, measured by EER and normalized minimum DCF.*

	MALE			*FEMALE*		
	Mix	*GI*	*GD*	*Mix*	*GI*	*GD*
EER(%)	2.03	2.11	2.02	3.87	**3.80**	3.86
OldDCF	0.097	0.098	0.097	0.190	**0.187**	0.190
NewDCF	0.365	0.397	0.363	0.541	**0.536**	0.543

We can draw two principal conclusions from these results. First, by comparing *Mix* (see columns 2 and 5 of Table 3) results and *GD* results (see columns 4 and 7 of Table 3), we can see that there is no loss of accuracy. Thus, the mixture model successfully handles microphone speech in the same way as telephone speech. Second, the comparison between *Mix* results and *GI* (see columns 3 and 6 of Table 3) shows a small anomaly (probably due to experimental error), since *GI* results for female are a bit better than *Mix* and *GD* results (see boldface entries of Table 3).

Finally, we tested the mixture model on the other main NIST SRE tasks. In order to facilitate comparison, we report pooled gender results rather than separate gender of det1, det3 and det4 tasks in Table 4. Again, those results confirm that there is no loss of accuracy when using the mixture model as shown in the det5 and det2 cases. The boldface entries in Table 4 show that we again have the same small anomaly as in the det2 case. However, it is clear that it is not a general problem, since it doesn't appear in the whole *GI* column.

4. Conclusion

In this paper we have shown how to build a speaker recognition system which is blind with respect to both, the male/female and telephone/microphone distinctions. Using a mixture of male and female PLDA models enables us to obtain good results on the NIST 2010 test data *without* taking advantage of the gender information provided in the evaluation protocol (Table 1). We have also shown that our approach works well on cross gender trials (Table 2), a problem which is not encountered in the NIST evaluations but which is important in real word applications. The system that we have presented falls short of the ideal of being fully blind in just one respect; namely that we have to set the decision thresholds for the various conditions (det1,det2 etc) in a way which takes account of the telephone/microphone distinction. It remains to find a way of performing score calibration which remedies this defect.

Table 4: *Pooled male and female det1, det3 and det4 results of mixture, gender dependent and gender independent models, measured by EER and normalized minimum DCF.*

		Mix	*GI*	*GD*
Det1	*EER(%)*	1.58	**1.44**	1.58
	OldDCF	0.070	0.071	0.070
	NewDCF	0.246	0.262	0.246
Det3	*EER(%)*	2.68	**2.57**	2.68
	OldDCF	0.125	**0.124**	0.126
	NewDCF	0.397	0.439	0.402
Det4	*EER(%)*	2.90	3.05	2.90
	OldDCF	0.129	0.133	0.128
	NewDCF	0.384	0.403	0.385

5. Acknowledgements

The authors would like to thank the organizers and other participants of the 2010 BOSARIS Workshop,[4] where this work was done.

6. References

[1] 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.

[2] 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.

[3] 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 *Proceedings of Interspeech*, Brighton, UK, Sep. 2009, pp. 1559–1562.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," 2010, to appear in IEEE Transactions on Audio, Speech and Language Processing.

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

[6] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[7] M. Senoussaoui, P. Kenny, P. Dumouchel, and F. Castaldo, "Well-calibrated heavy tailed Bayesian speaker verification for microphone speech," in *Proceedings of ICASSP*, Prague, Czech Republic, May 2011.

[8] J. Villalba and N. Brümmer, "Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[9] D. Garcia-Romero and C. Y. Espy-Wilso, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[4] http://speech.fit.vutbr.cz/en/workshops/bosaris-2010