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

Mohammed Senoussaoui, Patrick Kenny, Niko Brümmer, Edward de Villiers, Pierre Dumouchel

Interspeech 2011

August 28, 2011
The NIST speaker recognition evaluations have stimulated the development of a gender-dependent approach to speaker recognition:

- gender labels are given
- there are no cross-gender trials

Real-world deployment of a gender-dependent system is not straightforward and typically involves making a premature hard-decision based on the output of a gender detector (error rate $\sim 2\%$)

In the i-vector/PLDA approach, a mixture of gender-dependent models can be used to calculate likelihood ratios for speaker verification in a way which avoids the need for explicit gender detection.
We will see that this approach

- enables us to ignore the gender labels supplied by NIST without any loss in performance
- handles cross-gender trials (not evaluated by NIST) properly
- works for both telephone and microphone speech across multiple operating points

In the absence of gender information, the likelihood ratio for a speaker verification trial is evaluated by a simple probability calculation

This works because score normalization heuristics are not needed the i-vector/PLDA approach
An example to show why premature hard decisions are bad

- Suppose you have a traditional gender-dependent GMM/UBM system with $t$-norm and you find that a $\sim 2\%$ error rate in gender detection is unsatisfactory.
- In a given non-target verification trial, you are comparing a female speaker model with a test speaker who happens to be female.
- Your gender detector tells you that the speaker in the test segment is male.
- So you select the male impostor cohort for $t$-norm (another type of hard decision) and find that the test segment score is very high.
- Score normalization will lead you to conclude that the trial is a target trial.
i-vector extraction

- Speech segments are represented by low dimensional i-vectors
  - essentially, the hidden variables in the Joint Factor Analysis model if the distinction between speaker and channel variability is dropped
- Gender-independent UBM, gender-independent i-vector extractor (trained on microphone as well as telephone speech)
- Dimensionality reduction (from 800 to 200) via ordinary linear discriminant analysis to handle microphone speech as well as telephone speech
- Length normalization to Gaussianize the i-vector distribution so that heavy-tailed modeling is not needed [Ramos Interspeech 2010]
In its simplest form, Probabilistic Linear Discriminant Analysis (PLDA) assumes that i-vectors are distributed according to

\[ i = m + Vy + \epsilon \]

where

- the *speaker variable* \( y \) is Gaussian distributed and its value is *common to all recordings of a given speaker*
- the mean vector \( m \), the matrix \( V \) and the noise covariance matrix are usually taken to be *gender-dependent* (this is generally optimal for NIST conditions)

Probability calculations with this model involve Gaussian integrals which can be evaluated in closed form [Kenny Odyssey 2010]
The likelihood ratio for speaker verification

Given a pair of i-vectors \( D = (i_1, i_2) \) we have to evaluate the ratio

\[
\frac{P(D|H_1)}{P(D|H_0)}
\]

where

- \( H_1 \) : the speaker variables \( y_1 \) and \( y_2 \) are the same (target trial)
- \( H_0 \) : the speaker variables are different (non-target trial)

For the denominator, \( P(D|H_0) = P(i_1)P(i_2) \); the numerator is just another Gaussian integral.
Suppose now we have a mixture consisting of a male PLDA model $M$ and a female model $F$

Then $P(D|H_0) = P(i_1)P(i_2)$ where

$$P(i_1) = 0.5P(i_1|M) + 0.5P(i_1|F)$$

$$P(i_2) = 0.5P(i_2|M) + 0.5P(i_2|F)$$

and

$$P(D|H_1) = 0.5P(D|H_1, M) + 0.5P(D|H_1, F)$$
So the calculations are no more difficult than in the gender-dependent case.

There is no need to rely on gender detection to perform speaker recognition in the absence of gender labels but it is worth mentioning that the ratio

\[
\frac{P(i|M)}{P(i|F)}
\]

can serve as a very good gender detector (EER < 2%).
We use the trial lists from the extended core condition of the 2010 NIST speaker recognition evaluation.

We report results on all of the principal subconditions (microphone speech as well as telephone speech) obtained with gender-independent PLDA, gender-dependent PLDA and mixture PLDA.

The operating points tested are the equal error rate (EER), the "new DCF" (the detection cost function introduced in 2010) and the "old DCF".

We will also report the results of an experiment involving cross gender trials.
Mixture modeling (Mix) gives the same results as gender-dependent modeling (GD) which is substantially better than gender-independent modeling (GI)

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>GD</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>1.81%</td>
<td>1.81%</td>
<td>2.00%</td>
</tr>
<tr>
<td>old DCF</td>
<td>0.096</td>
<td>0.096</td>
<td>0.112</td>
</tr>
<tr>
<td>new DCF</td>
<td>0.322</td>
<td>0.320</td>
<td>0.386</td>
</tr>
</tbody>
</table>

- Male trials
Likewise for females

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>GD</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>2.46%</td>
<td>2.47%</td>
<td>2.75%</td>
</tr>
<tr>
<td>old DCF</td>
<td>0.124</td>
<td>0.124</td>
<td>0.133</td>
</tr>
<tr>
<td>new DCF</td>
<td>0.388</td>
<td>0.387</td>
<td>0.415</td>
</tr>
</tbody>
</table>
Cross gender trials (telephone/telephone)

It is not clear how to design a trial list which includes cross-sex trials as there are no benchmark results in the literature.

The error rates you obtain will depend on the proportions of cross-sex trials to same-sex trials among the non-target trials.

Retaining the target trials in the NIST extended list and replacing the non-target trials by cross gender trials reduces the EER from 2.24% to 0.4% (using the PLDA mixture model).

Quoting a minimum DCF in these circumstances is not really meaningful.
The cross gender trial list (CG) is created by replacing the non-target trials in the NIST extended list by cross gender trials.

The actual DCFs for the CG list are computed by using the decision thresholds which are optimal for the NIST extended list (which contains no cross gender trials).

As expected, the actual DCFs for CG are less than the minimum DCFs for NIST.
Again, mixture modeling (Mix) works just as well as gender-dependent modeling (GD) which is better than gender-independent modeling (GI)

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>GD</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>2.03%</td>
<td>2.02%</td>
<td>2.11%</td>
</tr>
<tr>
<td>old DCF</td>
<td>0.097</td>
<td>0.097</td>
<td>0.098</td>
</tr>
<tr>
<td>new DCF</td>
<td>0.365</td>
<td>0.363</td>
<td>0.397</td>
</tr>
</tbody>
</table>

- Male trials
For females, mixture modeling (Mix) works just as well as gender-dependent modeling (GD)

But the error rates are much higher than for males and the results obtained with gender-independent modeling (GI) are slightly anomalous

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>GD</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>3.87%</td>
<td>3.86%</td>
<td>3.80%</td>
</tr>
<tr>
<td>old DCF</td>
<td>0.190</td>
<td>0.190</td>
<td>0.187</td>
</tr>
<tr>
<td>new DCF</td>
<td>0.541</td>
<td>0.543</td>
<td>0.536</td>
</tr>
</tbody>
</table>
The other microphone conditions

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>GD</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>det1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>1.58%</td>
<td>1.58%</td>
</tr>
<tr>
<td></td>
<td>old DCF</td>
<td>0.070</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>new DCF</td>
<td>0.246</td>
<td>0.246</td>
</tr>
<tr>
<td>det3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>2.68%</td>
<td>2.68%</td>
</tr>
<tr>
<td></td>
<td>old DCF</td>
<td>0.125</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>new DCF</td>
<td>0.397</td>
<td>0.402</td>
</tr>
<tr>
<td>det4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>2.90%</td>
<td>2.90%</td>
</tr>
<tr>
<td></td>
<td>old DCF</td>
<td>0.129</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>new DCF</td>
<td>0.384</td>
<td>0.385</td>
</tr>
</tbody>
</table>

- All trials, male and female
Conclusion

- Gender-dependent PLDA is more effective than gender-independent PLDA for same-sex trials
  - The difference is substantial in the case of telephone speech
  - But for microphone speech the results are ambiguous in the case of female speakers
- Taking advantage of the fact that PLDA does not benefit from score normalization heuristics, it is easy to use a mixture of male and female PLDA models for speaker verification
  - A slight modification of the likelihood ratio calculation is all that is required
- This makes it possible to do speaker recognition without gender labels or a gender detector
Across all conditions and operating points tested, the results obtained without using gender labels are essentially the same as those obtained with gender labels.

The proposed method behaves properly on cross-sex trials.