1. Introduction

- i-vector is the state-of-the-art for the text-independent speaker verification.
- As a length-fixed and dimension-reduced vector, the low dimension of i-vector makes it possible to apply simple classifiers for speaker verification tasks.
- Contributions:
  - Analyze the i-vector as a tied vector for session variability across all the frames and mixtures.
  - Propose a kind of local variability vector which contains the session variability contained in every single mixture through local variability modeling.
  - Derive the posterior inference and sort out the maximum likelihood estimate of the model parameters using the expectation-maximization (EM) algorithm.

2. i-vector

- Represents a variable-length utterance with a fixed-length low dimensional vector $\hat{\phi}$.
- The vector $\hat{\phi}$ is estimated as the posterior mean, i.e., the maximum a-posteriori (MAP) estimate, of the latent variable $w$ given $\theta_i$.
- Total variability model
  - i-vector tying: across features and mixtures shown in graphical model as in Fig. 1.
  - Likelihood function:
    $$p(\theta | \phi) = \arg \max p(\theta | \phi) \cdot p(\phi) \cdot p(\theta | w, \Sigma) \cdot p(w) / \prod_{c} \int p(\theta | w, \Sigma) \cdot p(w) dw$$
- Posterior estimation
  - local latent variables $\tau_c = E[w_i | \phi] = \mu + \Sigma \nu_{i,c}$
  - Model parameters $\nu_{i,c} = \sum_{j=1}^{N_j} E[w_i | C_j]$ $\Sigma = \sum_{j=1}^{N_j} E[w_i | C_j] E[w_i | C_j]^T$.

3. Local variability models and vectors

- Local variability model (Fig. 2)
  - Each mixture is assigned with a local vector.
  - Latent variable are tied across the frames which are aligned to the same mixture.
  - Latent variable are untied across mixtures.
  - Loading matrices of different mixtures are tied.
- Likelihood function
  $$l_{iv}(\theta) = \prod_{i=1}^{C} \prod_{c} \int p(\theta | w, \Sigma) \cdot p(w) dw$$
- Posterior estimation
  - local latent variables $\tau_c = E[w_i | \phi] = \mu + \Sigma \nu_{i,c}$
- Model parameters $\nu_{i,c} = \sum_{j=1}^{N_j} E[w_i | C_j]$ $\Sigma = \sum_{j=1}^{N_j} E[w_i | C_j] E[w_i | C_j]^T$.

4. Experiments

- NIST SRE’08 (short2-short3 DET6)
- NIST SRE’10 (core-data DET5)
- For both tasks:
  - Test and train segments are telephone speech collected under clean environment
  - MFCC 57 (including $A$ and $\Delta A$)
  - i-vector (400): UBM 512 with full covariance.
  - LVM: UBM 512 with diagonal covariance; $j=57$
- Performance criteria: EER, minDCF08 and minDCF10
- Methods: i-vector + PLDA, LVM + parallel PLDA (LVM), LVM + supervector PLDA (LVM*)

<table>
<thead>
<tr>
<th>i-vector</th>
<th>LVM</th>
<th>LVM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER(%)</td>
<td>3.3807</td>
<td>5.9424</td>
</tr>
<tr>
<td>minDCF08</td>
<td>0.1224</td>
<td>0.3251</td>
</tr>
<tr>
<td>minDCF10</td>
<td>0.3711</td>
<td>0.9542</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER(%)</td>
<td>3.3807</td>
</tr>
<tr>
<td>minDCF08</td>
<td>0.1224</td>
</tr>
<tr>
<td>minDCF10</td>
<td>0.3711</td>
</tr>
</tbody>
</table>

Table 2 Performance comparison of i-vector and local variability model (LVM) on C05 core-data tests in NIST SRE’10.

<table>
<thead>
<tr>
<th>i-vector</th>
<th>LVM</th>
<th>LVM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER(%)</td>
<td>3.3807</td>
<td>5.9424</td>
</tr>
<tr>
<td>minDCF08</td>
<td>0.1224</td>
<td>0.3251</td>
</tr>
<tr>
<td>minDCF10</td>
<td>0.3711</td>
<td>0.9542</td>
</tr>
</tbody>
</table>

5. Conclusion

- We have proposed the local variability model (LVM) pivoted on the idea of cross-mixture tying upon a common loading matrix.
- We also derived the posterior inference and the EM steps for parameter learning.

References: