Minimum Divergence Estimation of Speaker Prior in Multi-session PLDA Scoring

Liping Chen[1,2], Kong Aik Lee[2], Bin Ma[2], Wu Guo[1], Haizhou Li[2], and Li Rong Dai[1]
[1]National Engineering Lab for Speech and Language Information Processing, USTC, China
[2]Institute for Infocomm Research (I2R), A*STAR, Singapore
E-mail: clp2011@mail.ustc.edu.cn, kalee@i2r.a-star.edu.sg

1. Introduction

- Current scoring method in PLDA is based on the hypothesis test whether the enrollment and test utterances are from the same or different speakers.
- In multi-session tasks, e.g. NIST SRE'12, the enrollment i-vectors are highly correlated as they might be extracted from simultaneous multi-channel recordings, shorter duration cuts or exact replicas of other utterances.
- We propose:
  - The idea of speaker adaptation in PLDA scoring.
  - The use of minimum divergence estimation of the prior distribution of speaker factor in multi-session scoring.

2. I-vector followed by PLDA

- An i-vector represents a variable-length utterance with a fixed-length low dimensional vector, estimated as the posterior mean of a latent variable [1]:

Mean supervector ← $\mathbf{m}_r = \mathbf{m} + \mathbf{T}\mathbf{w}_r$ → Total variability space

i-vector ← $\phi_r = E[\mathbf{w}_r \mid \mathcal{O}_r] = \arg\max_{\mathbf{w}_r} p(\mathcal{O}_r \mid \mathbf{m} + \mathbf{T}\mathbf{w}_r)\mathcal{N}(\mathbf{w}_r \mid 0, \mathbf{I})$

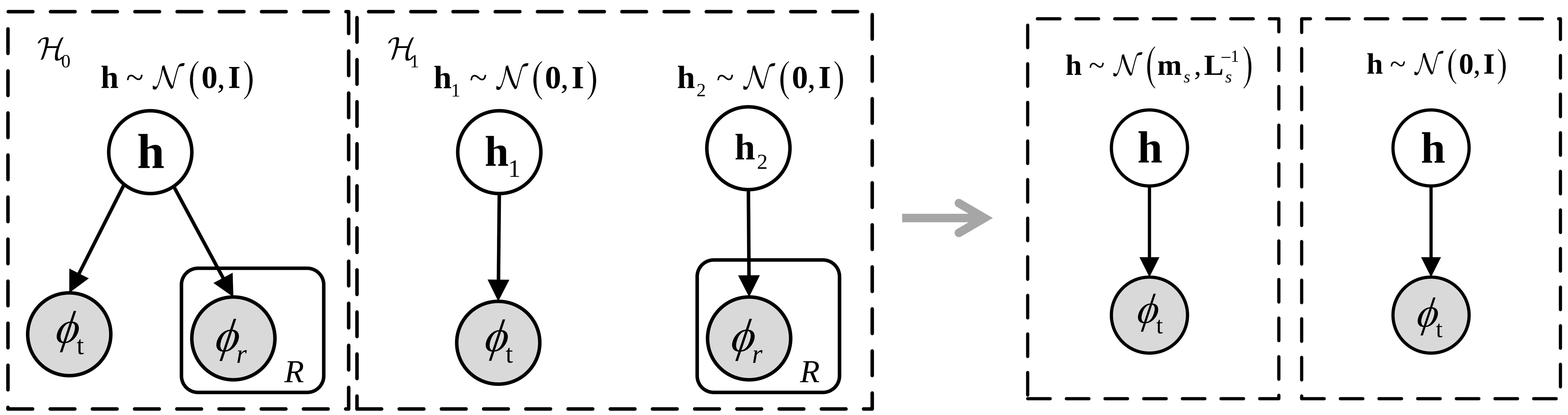
- A PLDA model is a Gaussian density with a structured covariance matrix [2]:

$$p(\phi) = \mathcal{N}\left(\phi \mid \boldsymbol{\mu}, \mathbf{F}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)$$

3. Speaker adaptation in PLDA scoring

- In conventional PLDA scoring, the score is calculated as the log likelihood ratio between two hypotheses (By-the-book scoring method):

$$\mathcal{H}_0 : \phi_{\mathrm{t}} \text{ and } \{\phi_{s,r=1,\dots,R}\} \text{ are from the same speaker}$$
$$\mathcal{H}_1 : \phi_{\mathrm{t}} \text{ and } \{\phi_{s,r=1,\dots,R}\} \text{ are from different speakers}$$



- The score could also be calculated as the log-likelihood ratio between the **speaker-dependent PLDA model** and the **universal PLDA model**:

$$l(\phi_{\mathrm{t}},\phi_{s,r=1,\dots,R}) = \log\frac{p(\phi_{\mathrm{t}},\phi_{s,r=1,\dots,R} \mid \mathcal{H}_0)}{p(\phi_{\mathrm{t}},\phi_{s,r=1,\dots,R} \mid \mathcal{H}_1)} = \log\frac{p(\phi_{\mathrm{t}},\phi_{s,r=1,\dots,R})}{p(\phi_{\mathrm{t}})\,p(\phi_{s,r=1,\dots,R})}$$
$$= \log\frac{p(\phi_{\mathrm{t}} \mid \phi_{s,r=1,\dots,R})\,p(\phi_{s,r=1,\dots,R})}{p(\phi_{\mathrm{t}})\,p(\phi_{s,r=1,\dots,R})} = \log\frac{p(\phi_{\mathrm{t}} \mid \phi_{s,r=1,\dots,R})}{p(\phi_{\mathrm{t}})}$$

$$p(\phi_{\mathrm{t}} \mid \phi_{s,r=1,\dots,R}) = \mathcal{N}\left(\phi_{\mathrm{t}} \mid \boldsymbol{\mu} + \mathbf{F}\mathbf{m}_s, \mathbf{F}\mathbf{L}_s^{-1}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)$$
$$\mathbf{m}_s = \mathbf{L}_s^{-1} \cdot \sum_{r=1}^{R} \mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\left(\phi_{s,r} - \boldsymbol{\mu}\right)$$
$$\mathbf{L}_s^{-1} = \left[\mathbf{I} + R\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{F}\right]^{-1}$$

4. Minimum Divergence Estimation of Speaker Prior

- For each enrollment session from the speaker s, we compute the mean and covariance of the posterior distribution:

$$\mathbf{m}_{s,r} = \mathbf{L}^{-1}\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}\right)^{-1}\left(\phi_{s,r}-\boldsymbol{\mu}\right)$$

$$\mathbf{L}^{-1} = \left[\mathbf{I}+\mathbf{F}^{\mathrm{T}}\left(\mathbf{G}\mathbf{G}^{\mathrm{T}}+\boldsymbol{\Sigma}\right)^{-1}\mathbf{F}\right]^{-1}$$

- We seek for another Gaussian distribution (the prior) that best represents the R posterior distributions.
- The Kullback-Leibler (KL) divergence [3] between the prior from the R posteriors , defined as follows:

$$D\left(\theta_{\mathrm{MD}}\right) = \sum_{r=1}^{R} E\left\{\log\frac{\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{m}_{s,r},\mathbf{L}^{-1}\right)}{\mathcal{N}\left(\mathbf{h}\,|\,\mathbf{y}_{s},\mathbf{P}_{s}^{-1}\right)}\right\}$$

- The minimum divergence estimates could be expressed in closed form, as follows

$$\mathcal{N}\left(\mathbf{h}|\mathbf{y}_{s},\mathbf{P}_{s}^{-1}\right) \Rightarrow \mathbf{y}_{s} = \frac{1}{R}\sum_{r=1}^{R}\mathbf{m}_{s,r}, \mathbf{P}_{s}^{-1} = \mathbf{L}^{-1}+\mathbf{S}$$

$$\mathbf{S} = \frac{1}{R}\cdot\sum_{r=1}^{R}\left(\mathbf{m}_{s,r}-\mathbf{y}_{s}\right)\left(\mathbf{m}_{s,r}-\mathbf{y}_{s}\right)^{\mathrm{T}}$$

5. Experiment

- NIST SRE’12 (Core task, CC2): one to over a hundred training segments per speaker, probably with content overlap among different segments for the same speaker.
- NIST SRE’10 (8conv-core task, CC5): 8 training segments per speaker
- For both tasks:
  - Test segments are telephone speech collected under clean environment
  - MFCC 57, UBM 512, i-vector 400
- Observations:
  - By-the-book approach does not perform better than the other two approaches.
  - Comparing to Mean only, the benefit of MinDiv is not significant on SRE’10 while the results on SRE’12 show a clear benefit where the number of enrolling segments for different speakers varies and the contents of the enrolling segments for a speaker are highly correlated.

Table 1 Comparison of three speaker adaptation approaches on CC5 of NIST SRE’10 8conv-core task

	EER (%)	minDCF10	minDCF12	
By-the-book	0.8493	0.2476	0.1915	Male
Mean	0.5194	0.1667	0.1446	
MinDiv	0.7607	0.7607	0.1623	
By-the-book	2.9370	0.3289	0.2625	Female
Mean	2.1379	0.3116	0.2546	
MinDiv	2.4747	0.3720	0.3142	

Table2 Comparison of three speaker adaptation approaches on CC2 of NIST SRE’12 core task.

	EER (%)	minDCF10	minDCF12	
By-the-book	6.8953	0.6015	0.5394	Male
Mean	3.9395	0.4765	0.4065	
MinDiv	3.5746	0.4238	0.3624	
By-the-book	6.4646	0.6338	0.5621	Female
Mean	3.2145	0.5382	0.4440	
MinDiv	3.0597	0.5235	0.4292	

6. Conclusion

- This paper presented an initial work on solving the multi-session PLDA scoring from the perspective of model adaptation.
- Based on the idea of model adaptation, we propose an adaptation method through a minimum divergence estimate of speaker prior.

References

[1] N.Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.

[2] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” *in Proc. International Conference on Computer Vision*, 2007.

[3] N. Brummer, “EM for Probabilistic LDA,” Technical Report, Feb. 2010, Available at https://sites.google.com/site/nikobrummer/.