LOCAL VARIABILITY MODELING FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

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

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 $\mathcal{O}_r$ with a fixed-length low dimensional vector ϕ_r.
- The vector ϕ_r is estimated as the posterior mean, i.e., the maximum a-posterior (MAP) estimate, of the latent variable $\mathbf{w}$ given $\mathcal{O}_r$. Ref. [4]

Mean supervector of $\mathcal{O}_r$ | $\mathbf{m}_r = \boldsymbol{\mu} + \mathbf{T}\mathbf{w}_r$ | Latent variable

i-vector | Total variability space

$$\phi_r = E[\mathbf{w}_r \mid \mathcal{O}_r] = \arg\max_{\mathbf{w}_r} p(\mathcal{O}_r \mid \boldsymbol{\mu} + \mathbf{T}\mathbf{w}_r)\mathcal{N}(\mathbf{w}_r \mid 0, \mathbf{I})$$

- Total variability model
  - i-vector tying: across features and mixtures shown in graphical model as in Fig. 1.
  - Likelihood function:

$$l_{\mathrm{TVM}}(\theta) = \prod_{r=1}^{R}\int\left(\prod_{c=1}^{C}\prod_{t=1}^{N_{r,c}}\mathcal{N}(o_t \mid \boldsymbol{\mu}_c + \mathbf{T}_c\mathbf{w}_r, \boldsymbol{\Sigma}_c)\right)\mathcal{N}(\mathbf{w}_r \mid 0, \mathbf{I})\, d\mathbf{w}_r.$$

The *t*-th frame aligned to the *c*-th mixture | Latent variable Shared by all mixtures and frames

  - Posterior estimation

$$\phi_r = E\{\mathbf{w}_r \mid \mathcal{O}_r\} = \mathbf{L}_r^{-1}\left(\sum_{c=1}^{C}\mathbf{T}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{F}_{r,c}\right)$$

$$\mathbf{L}_r^{-1} = \left(\mathbf{I} + \sum_{c=1}^{C}N_{r,c}\mathbf{T}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{T}_c\right)^{-1}$$

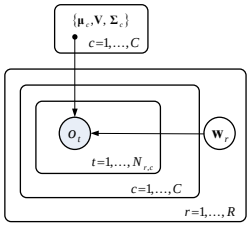

Fig. 1 the graphical model of total variability model (TVM).

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.
- Supervectors $\mathbf{m}_{r,c} = \boldsymbol{\mu}_c + \mathbf{V}\mathbf{w}_{r,c}$

Tied local variability space | Latent variable of the *c*-th mixture

- Likelihood function

$$l_{\mathrm{LVM}}(\theta) = \prod_{r=1}^{R}\prod_{c=1}^{C}\int\left(\prod_{t=1}^{N_{r,c}}\mathcal{N}(o_t \mid \boldsymbol{\mu}_c + \mathbf{V}\mathbf{w}_{r,c}, \boldsymbol{\Sigma}_c)\right)\mathcal{N}(\mathbf{w}_{r,c} \mid 0, \mathbf{I})\, d\mathbf{w}_{r,c}$$

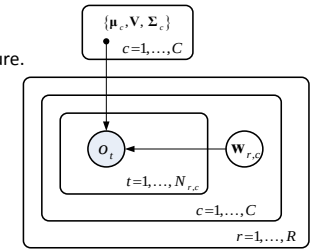

Fig. 2 the graphical model of local variability model (LVM).

- Posterior estimation
  - local latent variables

$$\tau_{r,c} = E\{\mathbf{w}_{r,c} \mid \mathcal{O}_r\} = \mathbf{L}_{r,c}^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{F}_{r,c}$$

$$\mathbf{L}_{r,c}^{-1} = \left(\mathbf{I} + N_{r,c}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{V}\right)^{-1}$$

  - Local variability vectors
    - Concatenating the *j*-th dimension of the *C* mixtures as a local vector $\rho_j = [\tau_1(j), \tau_2(j), \ldots, \tau_C(j)]^{\mathrm{T}}$
  - Model parameters

$$\mathbf{v}_i = \mathbf{a}_i\left\{\sum_{c=1}^{C}\left(\boldsymbol{\Sigma}_c^{-1}\right)_{ii}\sum_{r=1}^{R}N_{r,c}E\{\mathbf{w}_{r,c}\mathbf{w}_{r,c}^{\mathrm{T}}\}\right\}^{-1}$$

where $\mathbf{A} = \sum_{r=1}^{R}\sum_{c=1}^{C}\boldsymbol{\Sigma}_c^{-1}\mathbf{F}_{r,c}\mathbf{w}_{r,c}^{\mathrm{T}}$

- Parallel PLDA

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

  - Score calculation $l(\rho_j^{\mathrm{e}}, \rho_j^{\mathrm{t}}) = \log\dfrac{p(\rho_j^{\mathrm{t}}, \rho_j^{\mathrm{e}})}{p(\rho_j^{\mathrm{t}})\,p(\rho_j^{\mathrm{e}})}$ for $j=1,2,\ldots,J$
  - Score fusion: $s(\rho^{\mathrm{e}}, \rho^{\mathrm{t}}) = \sum_{j=1}^{J} l(\rho_j^{\mathrm{e}}, \rho_j^{\mathrm{t}})$

4. Experiments

- NIST SRE'08 (short2-short3 DET6)
- NIST SRE'10 (core-core DET5)
- For both tasks:
  - Test and train segments are telephone speech collected under clean environment
  - MFCC 57 (including $\Delta$and $\Delta\Delta$)
  - i-vector (400): UBM 512 with full covariance
  - LVM: UBM 512 with diagonal covariance; *J*=57
  - Performance criteria: EER, minDCF10 and minDCF08
- Methods: i-vector + PLDA, LVM + parallel PLDA (LVM), LVM + supervector PLDA(LVM$^+$) Ref [14]

Table 1 Performance comparison of i-vector and local variability model (LVM) on DET6 of short2-short3 task in NIST SRE'08.

	Male			Female		
	EER(%)	minDCF08	minDCF10	EER(%)	minDCF08	minDCF10
i-vector	3.6617	0.2034	0.6660	5.3716	0.2716	0.9967
LVM	5.9424	0.3251	0.9542	8.2940	0.4370	0.9884
LVM$^+$	6.3045	0.3380	0.8295	7.5206	0.3897	0.9873

Table 2 Performance comparison of i-vector and local variability model (LVM) on CC5 of core-core tests in NIST SRE'10.

	Male			Female		
	EER(%)	minDCF08	minDCF10	EER(%)	minDCF08	minDCF10
i-vector	3.2807	0.1224	0.3711	2.8001	0.1402	0.3465
LVM	4.4325	0.2113	0.6006	6.7150	0.2982	0.6648
LVM$^+$	5.3712	0.2917	0.8463	7.9109	0.3775	0.7972

- Observation:
  - local variability vectors extracted using the LVM are effective for speaker characterization even though there is still a considerable gap compared to the baseline i-vector PLDA system

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:

[4] 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.
[14] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification," in Proc. INTERSPEECH, 2012, paper 198.