Local Variability Vector for Text-independent Speaker Verification

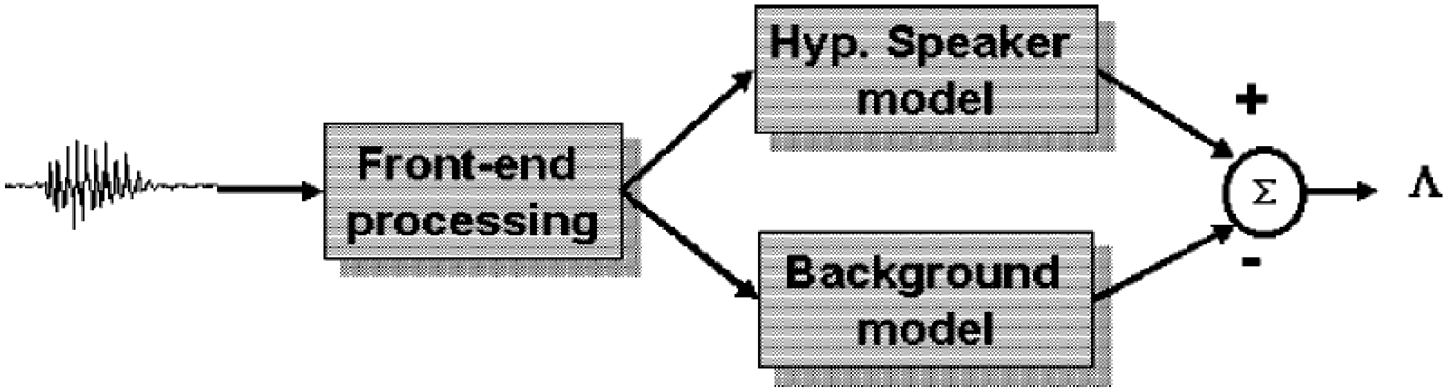
Liping Chen[1], *Kong Aik Lee*[2], *Bin Ma*[2],
Wu Guo[1], *Haizhou Li*[2], and *Li Rong Dai*[1]

[1]National Eng Lab for Speech and Language Information Processing, USTC, China

[2]Institute for Infocomm Research, A*STAR, Singapore

Introduction

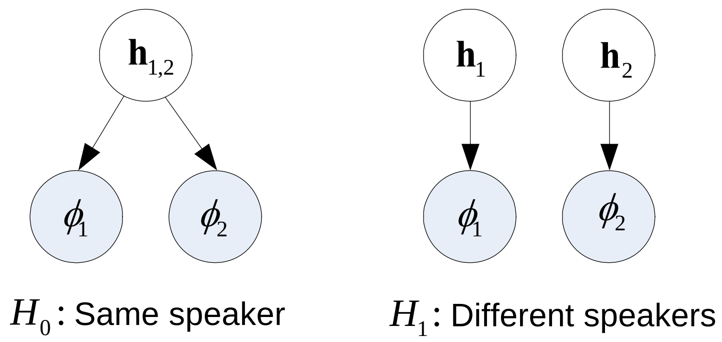
- Speaker recognition – to use a person voice as a mean to authenticate his/her identify (text-independent).
- Classical GMM-UBM paradigm [Reynolds et al, 2000]: UBM, MAP, speaker model, log-likelihood ratio.



D. A. Reynolds, T. F. Quatieri, and R. B. Dumn, "Speaker verification using adapted Gaussian mixture model," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

Introduction (cont'd)

- The i-vector PLDA paradigm:
  - Speech utterances are represented as fixed-length low dimensional vectors – the so-called i-vector (or identity vectors).
  - No speaker model – both the enrollment and test utterances are represented as i-vectors.
  - The log-likelihood ratio is computed as the hypothesis test whether two i-vectors are from the same or different speakers.
  - PLDA model facilitates the hypothesis test and channel compensation.



Motivation

- An i-vector represents the speaker and channel variability contained in an utterance.
- Local information associated with individual dimensions of the acoustic space are conflated to a single i-vector.
- We propose a local variability model to capture the local variability associated with individual dimension of the acoustic space.
- A speech utterance is represented by a set of *local variability vectors* instead of a single i-vector.
- Approach: changing the tying scheme in the total variability model.

I-vector extraction

- Given a speech utterance, we assume that it was generated from a speaker and channel dependent GMM.
- The mean super-vector lies in a low-dimensional subspace $\mathbf{T}$:

Mean super-vector

$$\mathbf{m}_r = \boldsymbol{\mu} + \mathbf{T}\mathbf{w}_r$$

Latent variable

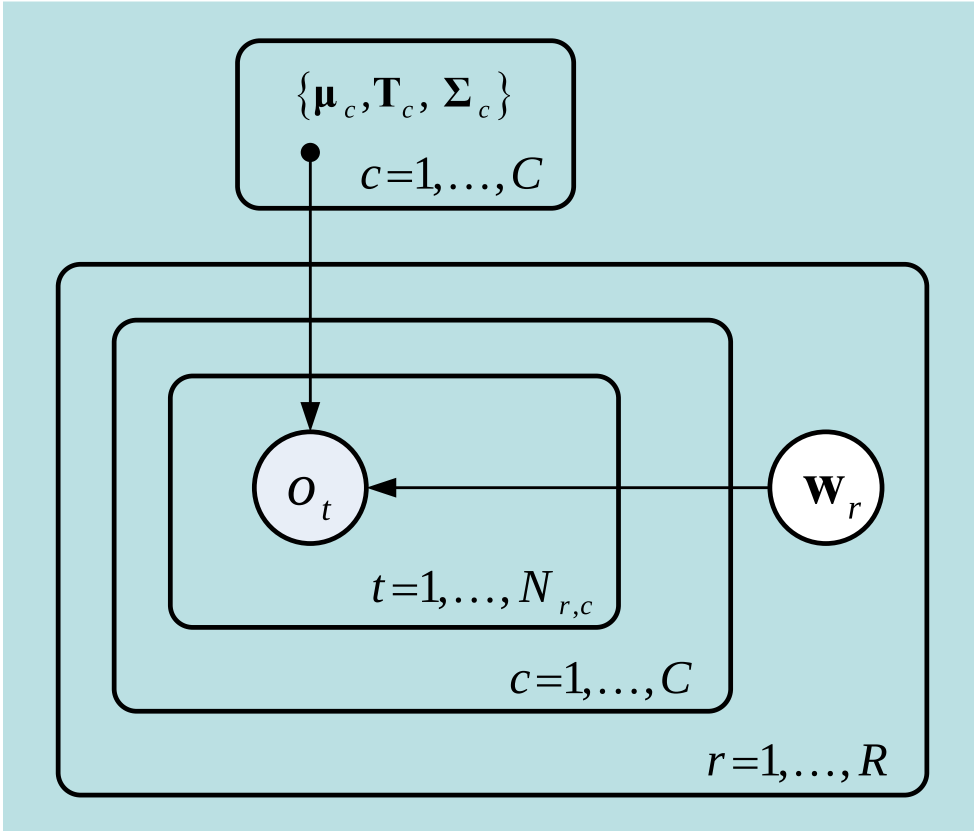
Total variability matrix

- An i-vector is the posterior mean of the latent variable $\mathbf{w}_r$.

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

Total variability model

- R – number of utterances
- C – number of Gaussians
- $N_{r,c}$ – number of frames associated with the *c*-th Gaussian component
- The latent variable $\mathbf{w}_r$ is tied (or shared) across
  - Frames
  - Mixtures



Total variability model (cont'd)

- Likelihood function

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

- Posterior estimation

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

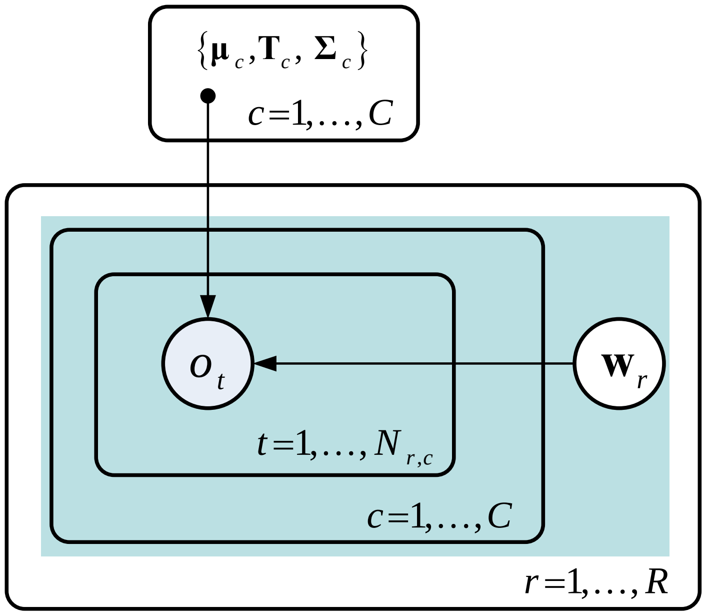
i-vector (posterior mean)

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

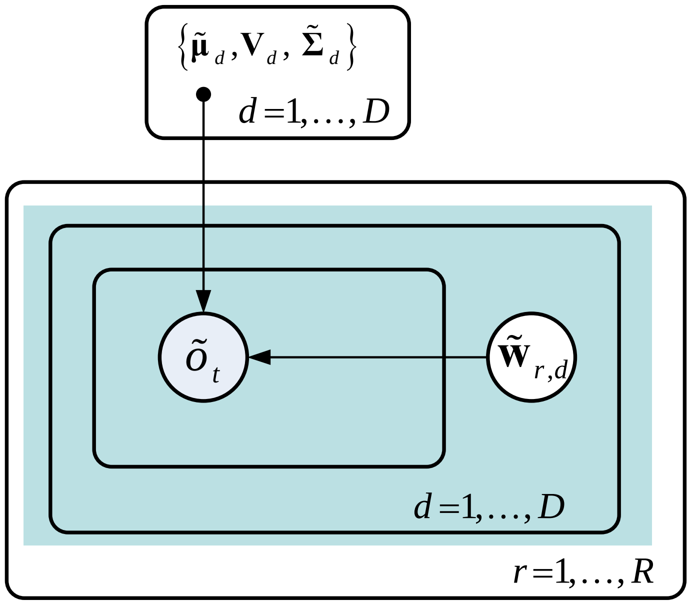
Posterior covariance

Local variability model

- We propose to remove the tying of latent variable across dimensions of the acoustic space while retaining the tying across frames and mixtures.



TVM



LVM

Local variability model (cont'd)

- Objective: to model the local variability specific to each dimension of the acoustic space.
- This formulation leads to dimension-centric variability modeling referred to as the local variability model (LVM).
- Likelihood function:

$$l_{\mathrm{CLVM}}(\theta)=\prod_{r=1}^{R}\prod_{d=1}^{D}\int\prod_{c=1}^{C}\prod_{t=1}^{N_{r,c}}\mathcal{N}\left(o_{r,c,t,d}\mid \mathbf{V}_{d,c}\mathbf{w}_{r,d},\sigma_{d,c}\right)\mathcal{N}\left(\mathbf{w}_{r,d}\mid 0,\mathbf{I}\right)d\mathbf{w}_{r,d}$$

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

Local variability model (cont'd)

- A speech utterance is represented by a set of *local variability vectors* instead of a single i-vector.
- Posterior inference (E-step):

$$\mathbf{y}_{r,d} = E\left\{\mathbf{w}_{r,d} \middle| \mathcal{O}_r\right\} = \mathbf{L}_{r,d}^{-1}\mathbf{V}_d^{\mathrm{T}}\tilde{\mathbf{F}}_{r,d} \quad \text{for } d = 1, 2, \ldots, D$$

$$\mathbf{L}_{r,d}^{-1} = \left(\mathbf{I} + \mathbf{V}_d^{\mathrm{T}}\mathbf{\Gamma}_r\mathbf{V}_d\right)^{-1}$$

- Parameter estimation (M-step):

$$\mathbf{v}_d^c = \mathbf{\Phi}_d^c\left(\sum_r \gamma_{r,c}\mathbf{K}_{r,d}\right)^{-1} \text{ for } c = 1, 2, \ldots, C, \quad d = 1, 2, \ldots, D$$

$$\mathbf{\Phi}_d = \sum_r \tilde{\mathbf{F}}_{r,d} E\left[\mathbf{w}_{r,d}^{\mathrm{T}}\right] \qquad \mathbf{K}_{r,d} = E\left[\mathbf{w}_{r,d}\mathbf{w}_{r,d}^{\mathrm{T}}\right]$$

Channel compensation and scoring with PLDA

- Local variability vectors are concatenated and taken as input to PLDA.
- PLDA is essentially a Gaussian distribution with a structured covariance for speaker and channel variability modeling:

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \boldsymbol{\mu}, \underbrace{\mathbf{F}\mathbf{F}^{\mathrm{T}}}_{\text{Intra-speaker variability}} + \underbrace{\mathbf{G}\mathbf{G}^{\mathrm{T}}}_{\text{Channel variability}} + \boldsymbol{\Sigma}\right)$$

- PLDA scoring:

$$l(\mathbf{y}_{\mathrm{t}}, \mathbf{y}_e) = \log \frac{p(\mathbf{y}_{\mathrm{t}}, \mathbf{y}_e)}{p(\mathbf{y}_{\mathrm{t}})\, p(\mathbf{y}_e)}$$

H_0: Same speaker

H_1: Different speakers

Experimental setup

Component	Configuration	DEV Set
UBM	• 512 Gaussian mixtures	NIST SRE’ 04
i-vector	• Total variability matrix of rank 400. • PLDA with F and G of rank 200 and 50 with a full covariance matrix	NIST SRE’04, 05 and 06 telephone data
Local variability vector	• 57 x 20-dim local vectors • PLDA with F and G of rank 400 and 30 with a diagonal covariance matrix	NIST SRE’04, 05 and 06 telephone data

Results – SRE'08

- Performance comparison on DET6 of *short2-short3* task of NIST SRE'08.

	Male		
	EER (%)	minDCF08	minDCF10
TVM	3.6182	0.2130	0.6820
LVM	4.7559	0.2596	0.7895
Fusion	3.3700	0.1943	0.6042
	Female		
	EER (%)	minDCF08	minDCF10
TVM	5.3908	0.2767	0.9972
LVM	6.6144	0.3367	0.9950
fusion	5.4505	0.2707	0.9961

Results – SRE’10

- Performance comparison on CC5 of *core-core* task in NIST SRE’10.

	Male		
	EER (%)	minDCF08	minDCF10
TVM	3.0836	0.1253	0.3654
LVM	3.7590	0.1453	0.5439
Fusion	2.5136	0.1212	0.3626
	Female		
	EER (%)	minDCF08	minDCF10
TVM	2.6743	0.1458	0.3239
LVM	4.3068	0.2317	0.6119
fusion	2.5399	0.1488	0.3521

Conclusion

- We proposed the local variability model (LVM) pivoted on the idea of extracting the local variability associated with each dimension of the acoustic features.
- We derived the posterior inference and the EM steps for parameter learning.
- Experimental results suggest that the proposed *local variability vector* models the speaker information that is absent in the *i-vector*.

ISCSLP 2014, Singapore

THANKS