# SNR-Invariant PLDA with Multiple Speaker Subspaces


**Na LI and Man-Wai MAK**

Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University


## Introduction

- Noise level variability can shift the i-vectors to different regions of the i-vector space, and i-vectors with similar SNRs tend to cluster together.
- This phenomenon limits the capability of SNR-invariant PLDA with a single speaker subspace.
- This paper proposes a new SNR-invariant PLDA model by introducing multiple speaker subspaces to the SNR-invariant PLDA model.
- Experiments on NIST 2012 SRE demonstrate the effectiveness of the proposed method compared with PLDA and SNR-invariant PLDA.

## Background

**Conventional PLDA:** $\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\varepsilon}_{ij}$

Pool i-vectors from various background noise levels to train a PLDA model.

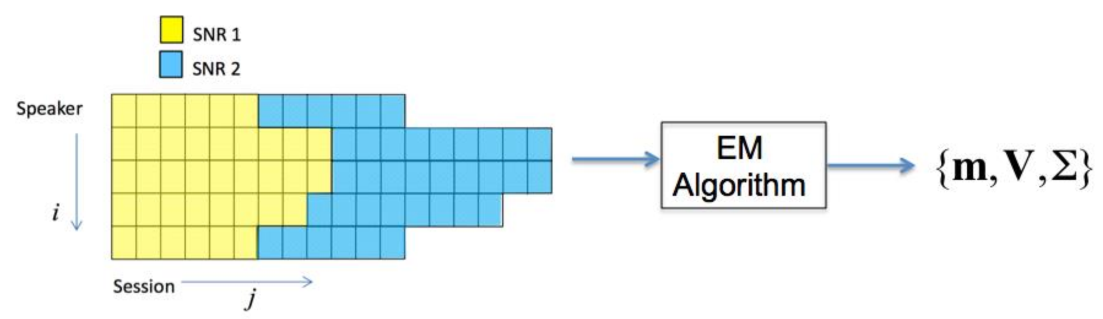


**Fig.1:** *Training process of conventional PLDA.*

**SNR-invariant PLDA:** $\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^k$

I-vectors within the same SNR group share the same SNR factor $\mathbf{w}_k$; the model is trained using the pooled data.

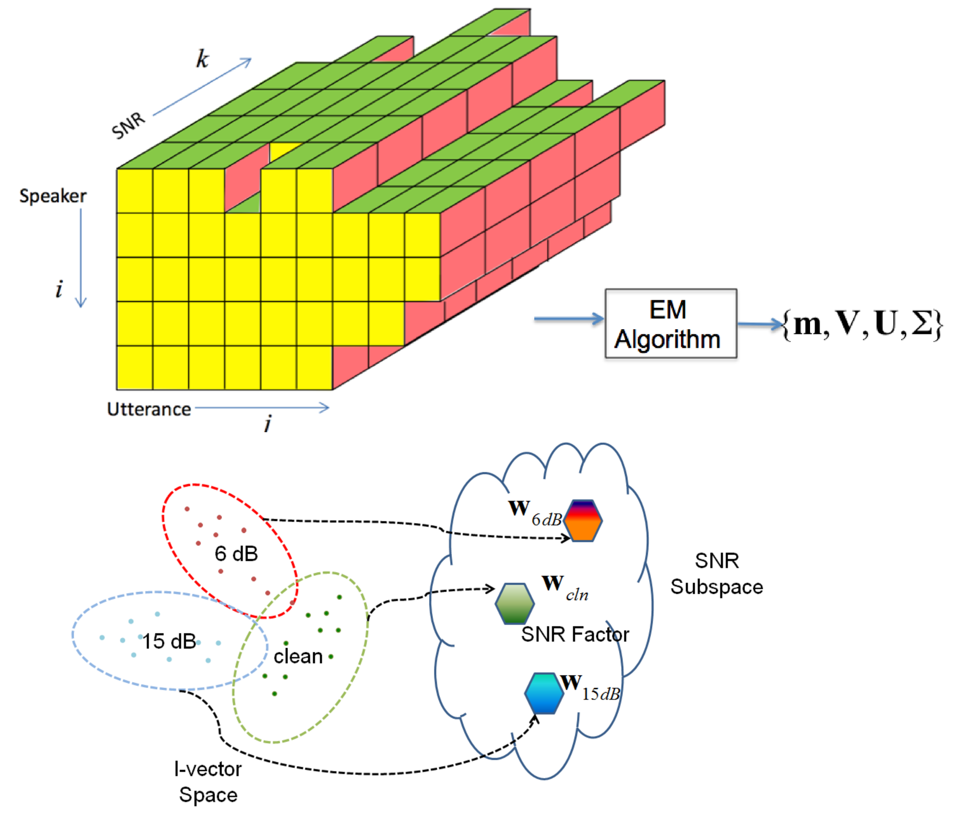


**Fig.2:** *Training process of SNR-invariant PLDA.*

## Proposed Method

Assuming that speaker variability within a narrow range of SNR occurs in a unique speaker-subspace, multiple speaker subspaces are introduced.

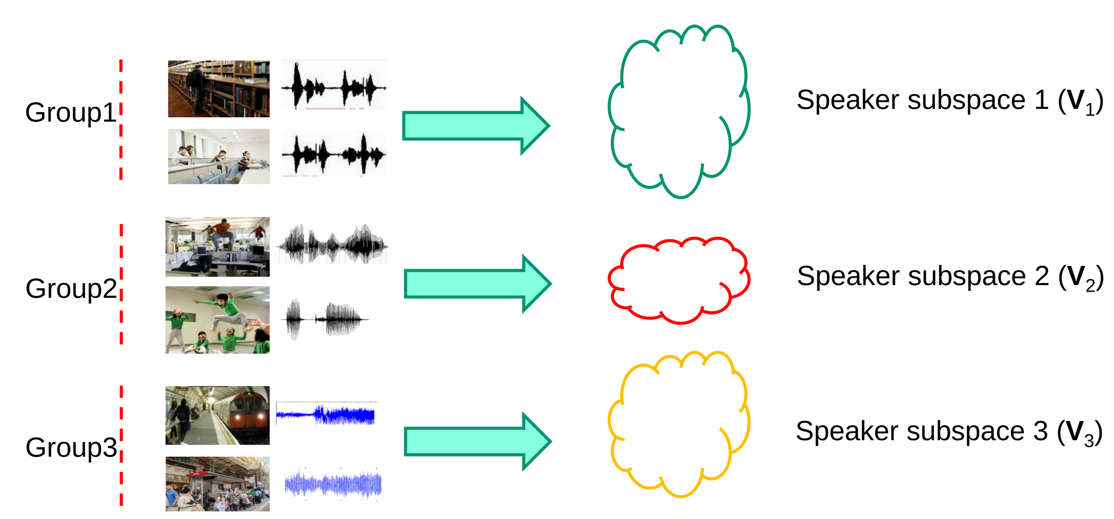


**Fig.3:** *Multiple speaker subspaces in the proposed model.*

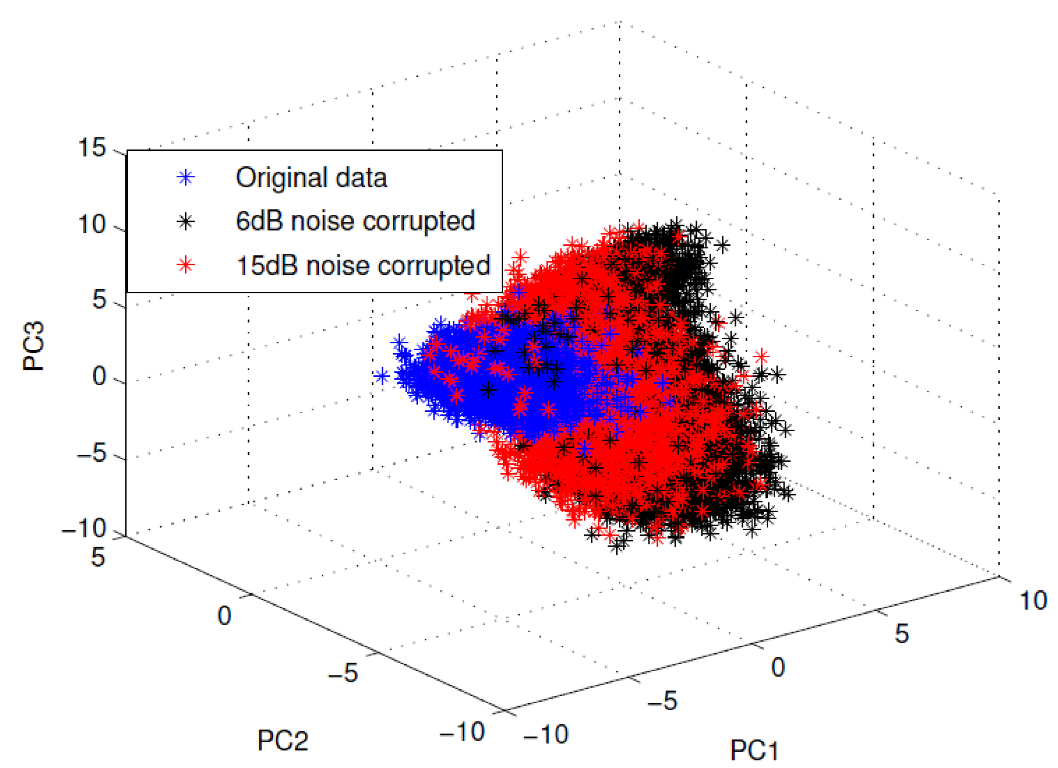


**Fig.4:** *The mean-shift effect of i-vectors caused by different levels of background noise in the corresponding utterances. This figure displays the three groups of i-vectors on the first 3 principal components.*

**SNR Subgroups:**

The training set is divided into multiple SNR subgroups according to the highest posterior probability with respect to a GMM trained using the SNRs of the training utterances.

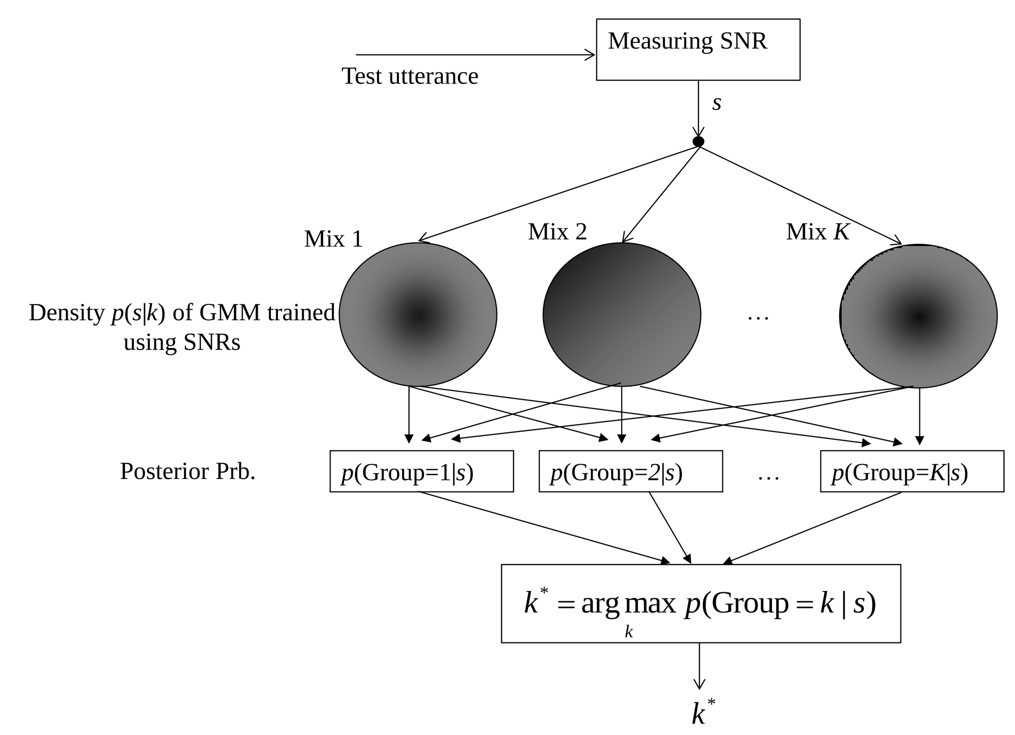


**Fig.5:** *Determination of the SNR subgroup of a test utterance.*

**The proposed SNR-invariant PLDA:**

$$\mathbf{x}_{ij}^k = \mathbf{m}_k + \mathbf{V}_k\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^k \qquad \boldsymbol{\varepsilon}_{ij}^k \sim \mathcal{N}(\boldsymbol{\varepsilon}|0, \boldsymbol{\Sigma}_k)$$

**Auxiliary Function:**

$$\mathbf{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{h},\mathbf{w}}\Big[\sum_{ikj}\Big(\ln\mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m}_k + \mathbf{V}_k\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Sigma}_k) + \ln\mathcal{N}(\mathbf{h}_i|\mathbf{0},\mathbf{I}) + \ln\mathcal{N}(\mathbf{w}_k|\mathbf{0},\mathbf{I})\Big)\Big|\mathcal{X},\boldsymbol{\theta}\Big].$$

$$\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}$$

**EM-Step:**

$$\langle\mathbf{h}_i|\mathcal{X}\rangle = (\mathbf{L}_i^1)^{-1}\sum_{k=1}^{K}\mathbf{V}_k^\top\boldsymbol{\Phi}_k^{-1}\sum_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m}_k)$$

$$\langle\mathbf{w}_k|\mathcal{X}\rangle = (\mathbf{L}_k^2)^{-1}\mathbf{U}^\top\boldsymbol{\Psi}_k^{-1}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m}_k)$$

$$\mathbf{L}_i^1 = \mathbf{I} + \sum_{k=1}^{K}H_i(k)\mathbf{V}_k^\top\boldsymbol{\Phi}_k^{-1}\mathbf{V}_k$$

$$\mathbf{L}_k^2 = \mathbf{I} + M_k\mathbf{U}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{U}$$

$$\boldsymbol{\Phi}_k = \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_k \qquad \boldsymbol{\Psi}_k = \mathbf{V}_k\mathbf{V}_k^\top + \boldsymbol{\Sigma}_k$$

$$\mathbf{V}_k = \left\{\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\left[(\mathbf{x}_{ij}^k - \mathbf{m}_k)\langle\mathbf{h}_i|\mathcal{X}\rangle^\top - \mathbf{U}\langle\mathbf{w}_k\mathbf{h}_i^\intercal|\mathcal{X}\rangle\right]\right\}\left\{\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\langle\mathbf{h}_i\mathbf{h}_i^\intercal|\mathcal{X}\rangle\right\}^{-1}$$

$$\mathbf{U} = \left\{\sum_{i=1}^{S}\sum_{k=1}^{K}\sum_{j=1}^{H_i(k)}\left[(\mathbf{x}_{ij}^k - \mathbf{m}_k)\langle\mathbf{w}_k|\mathcal{X}\rangle^\top - \mathbf{V}_k\langle\mathbf{h}_i\mathbf{w}_k^\intercal|\mathcal{X}\rangle\right]\right\}\left\{\sum_{i=1}^{S}\sum_{k=1}^{K}\sum_{j=1}^{H_i(k)}\langle\mathbf{w}_k\mathbf{w}_k^\intercal|\mathcal{X}\rangle\right\}^{-1}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{M_k}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\Big[(\mathbf{x}_{ij}^k - \mathbf{m}_k)(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top - \mathbf{V}_k\langle\mathbf{h}_i|\mathcal{X}\rangle(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top - \mathbf{U}\langle\mathbf{w}_k|\mathcal{X}\rangle(\mathbf{x}_{ij}^k - \mathbf{m}_k)^\top\Big]$$

$$\mathbf{m}_k = \frac{1}{M_k}\sum_{i=1}^{S}\sum_{j=1}^{H_i(k)}\mathbf{x}_{ij}^k$$

**Likelihood Ratio Scores:**

$$S_{\mathrm{LR}}(\mathbf{x}_s,\mathbf{x}_t) = \ln\frac{p(\mathbf{x}_s,\mathbf{x}_t|\text{same-speaker})}{p(\mathbf{x}_s,\mathbf{x}_t|\text{different-speakers})} = \ln\frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\\mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}_{k_s}\\\mathbf{m}_{k_t}\end{bmatrix},\begin{bmatrix}\mathbf{A}_{k_s} & \mathbf{B}_{k_sk_t}\\\mathbf{B}_{k_sk_t}^\top & \mathbf{A}_{k_t}\end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s\\\mathbf{x}_t\end{bmatrix}\middle|\begin{bmatrix}\mathbf{m}_{k_s}\\\mathbf{m}_{k_t}\end{bmatrix},\begin{bmatrix}\mathbf{A}_{k_s} & \mathbf{0}\\\mathbf{0} & \mathbf{A}_{k_t}\end{bmatrix}\right)}$$

where

$$\mathbf{A}_{k_s} = \mathbf{V}_{k_s}\mathbf{V}_{k_s}^\top + \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_{k_s},$$
$$\mathbf{A}_{k_t} = \mathbf{V}_{k_t}\mathbf{V}_{k_t}^\top + \mathbf{U}\mathbf{U}^\top + \boldsymbol{\Sigma}_{k_t}, \text{ and}$$
$$\mathbf{B}_{k_sk_t} = \mathbf{V}_{k_s}\mathbf{V}_{k_t}^\top.$$

## Results

- **Table1:** *Performance of PLDA, S-PLDA and Proposed multi-speaker subspace PLDA on CC4*

| Method | K | Male | | Female | |
|---|---|---|---|---|---|
| | | EER(%) | minDCF | EER(%) | minDCF |
| PLDA | - | 3.39 | 0.325 | 3.10 | 0.354 |
| S-PLDA | 3 | 3.20 | **0.300** | 2.95 | **0.327** |
| Proposed | 2 | 3.31 | 0.302 | 3.09 | 0.333 |
| | 3 | **3.06** | 0.309 | 2.88 | 0.332 |
| | 4 | 3.12 | 0.316 | **2.84** | 0.339 |

- **Table2:** *Performance of PLDA, S-PLDA and Proposed multi-speaker subspace PLDA on CC5*

| Method | K | Male | | Female | |
|---|---|---|---|---|---|
| | | EER(%) | minDCF | EER(%) | minDCF |
| PLDA | - | 2.80 | 0.303 | 2.34 | 0.331 |
| S-PLDA | 3 | 2.80 | 0.302 | 2.37 | **0.319** |
| Proposed | 2 | **2.74** | **0.276** | 2.36 | 0.350 |
| | 3 | 2.80 | 0.278 | 2.31 | 0.325 |
| | 4 | 2.79 | 0.284 | **2.26** | 0.321 |

- **Table3:** *Performance comparison of different SNR-invariant PLDA models on CC4*

| Model | Model Parameters | EER(%) | minDCF |
|---|---|---|---|
| 1 | $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$ | 3.20 | **0.300** |
| 2 | $\boldsymbol{\theta} = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}$ | 3.06 | 0.309 |
| 3 | $\boldsymbol{\theta} = \{\mathbf{m}_k, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$ | 3.30 | 0.305 |
| 4 | $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}_k\}$ | 3.15 | 0.308 |
| 5 | $\boldsymbol{\theta} = \{\mathbf{m}_k, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}_k\}$ | 3.57 | 0.319 |
| 6 | $\boldsymbol{\theta} = \{\mathbf{m}_k, \mathbf{V}_k, \mathbf{U}, \boldsymbol{\Sigma}\}$ | **2.81** | 0.332 |

## References:


- N. Li and M. W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1648–1659, 2015.
- P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in Proc. of *Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.