THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

SNR-Invariant PLDA Modeling for Robust Speaker Verification

Na Li and Man-Wai Mak

Interspeech 2015
Dresden, Germany

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR, China

The Hong Kong Polytechnic University
香港理工大學

Contents

1. Background and Motivation of Work
2. SNR-invariant PLDA modeling for Robust Speaker Verification
3. Experiments on SRE12
4. Conclusions

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Background

- I-vector/PLDA Framework

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

$\mathbf{m}$: Global mean of all i-vectors

$\mathbf{V}$: Bases of speaker subspace

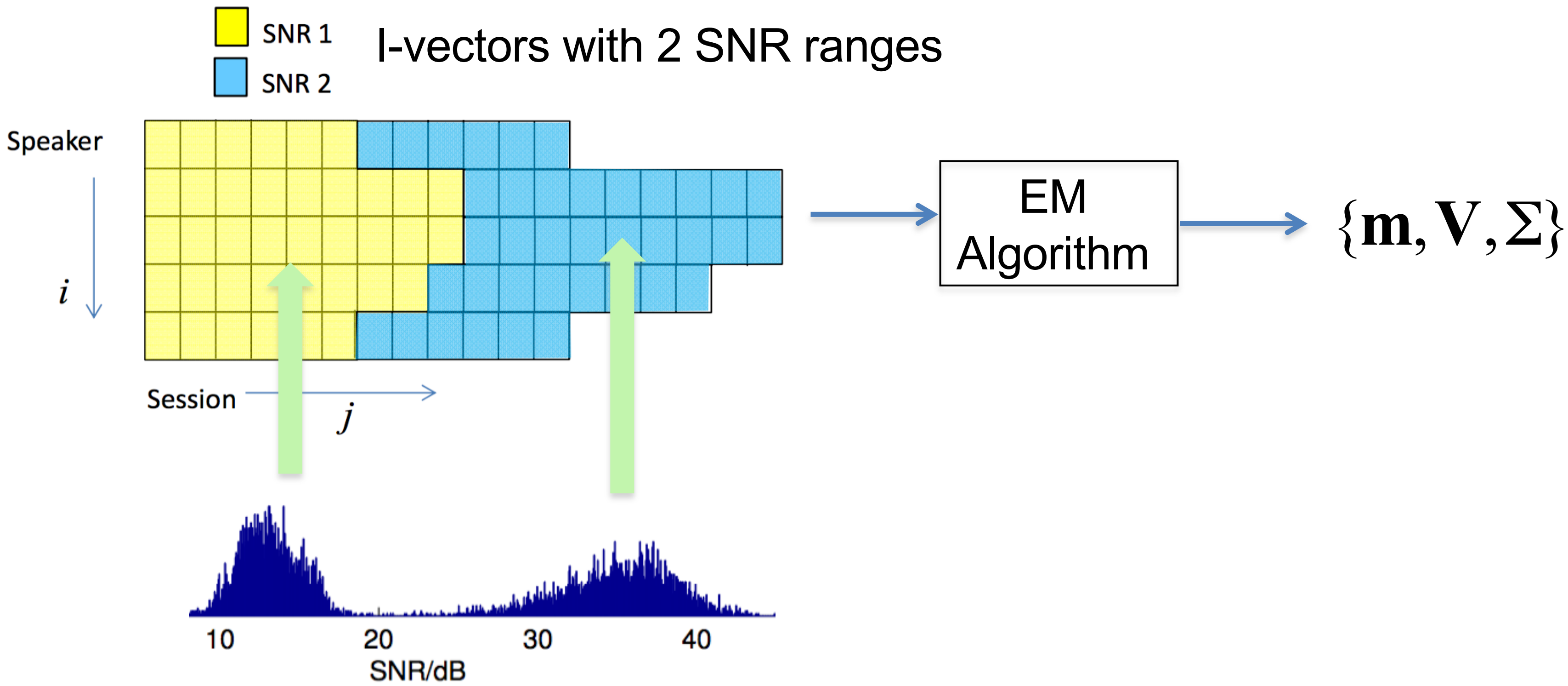
$\mathbf{h}_i$: Latent speaker factor with a standard normal distributio $\mathcal{N}(\mathbf{0}, \mathbf{I})$

$\boldsymbol{\varepsilon}_{ij}$: Residual term follows a Gaussian distribution with zero mean and full covariance $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

$\mathbf{x}_{ij}$: Length-normalized i-vector of speaker *i*

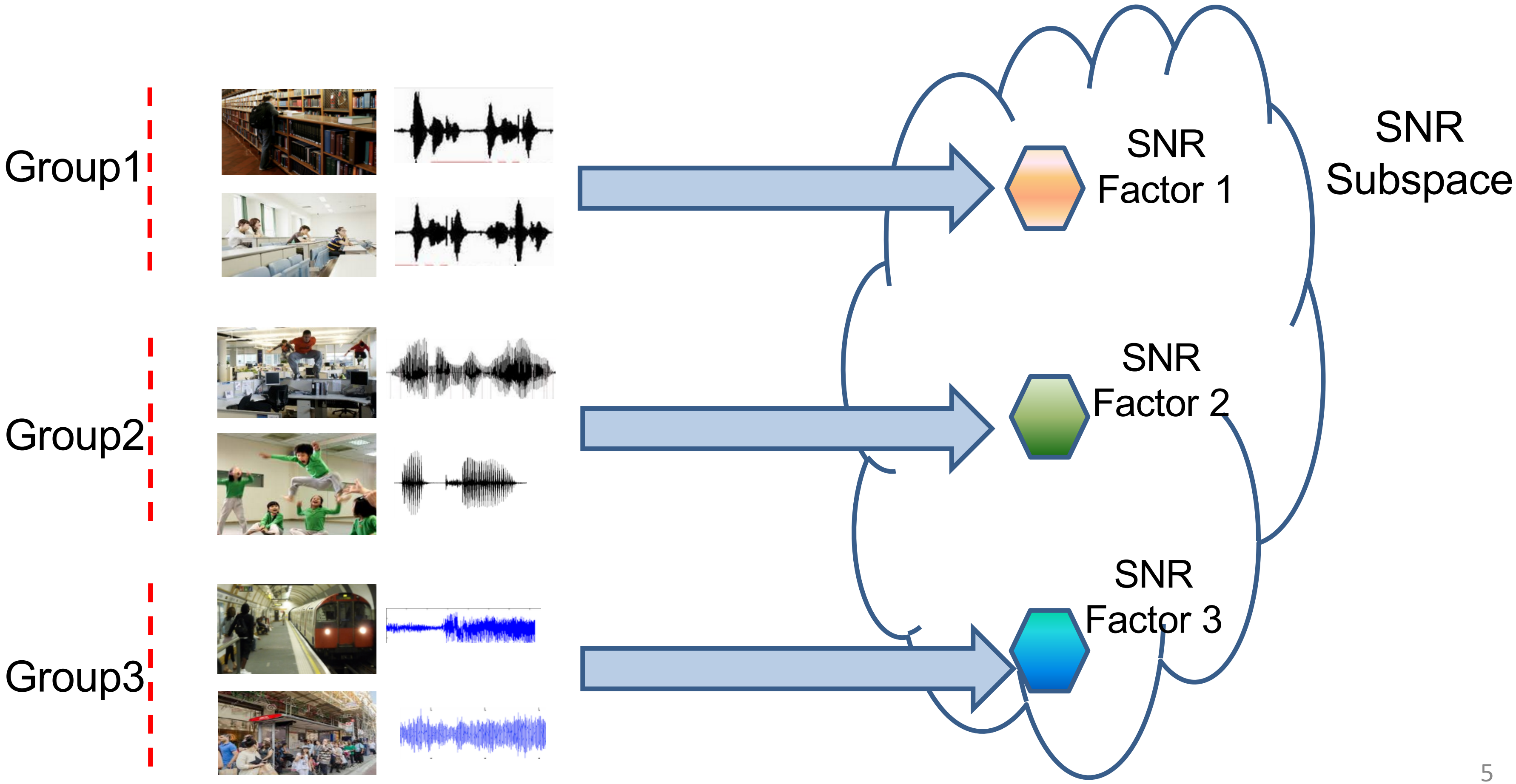
Background

- In conventional multi-condition training, we **pool** i-vectors from various background noise levels to train $\mathbf{m}$, $\mathbf{V}$ and $\mathbf{\Sigma}$.



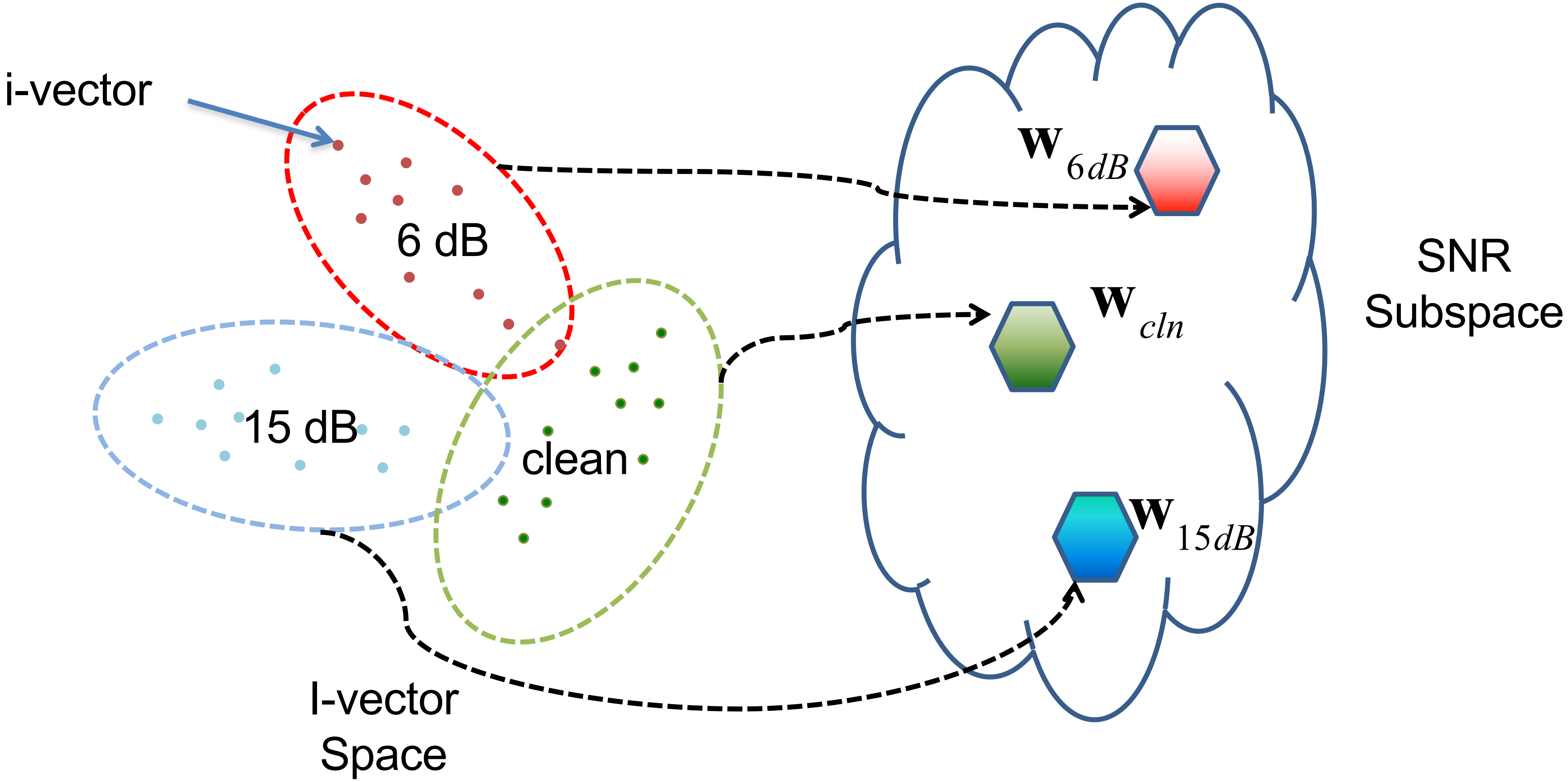
Motivation

- We argue that the variation caused by SNR can be modeled by an **SNR subspace** and utterances falling within a narrow SNR range should share the same set of **SNR factors**.

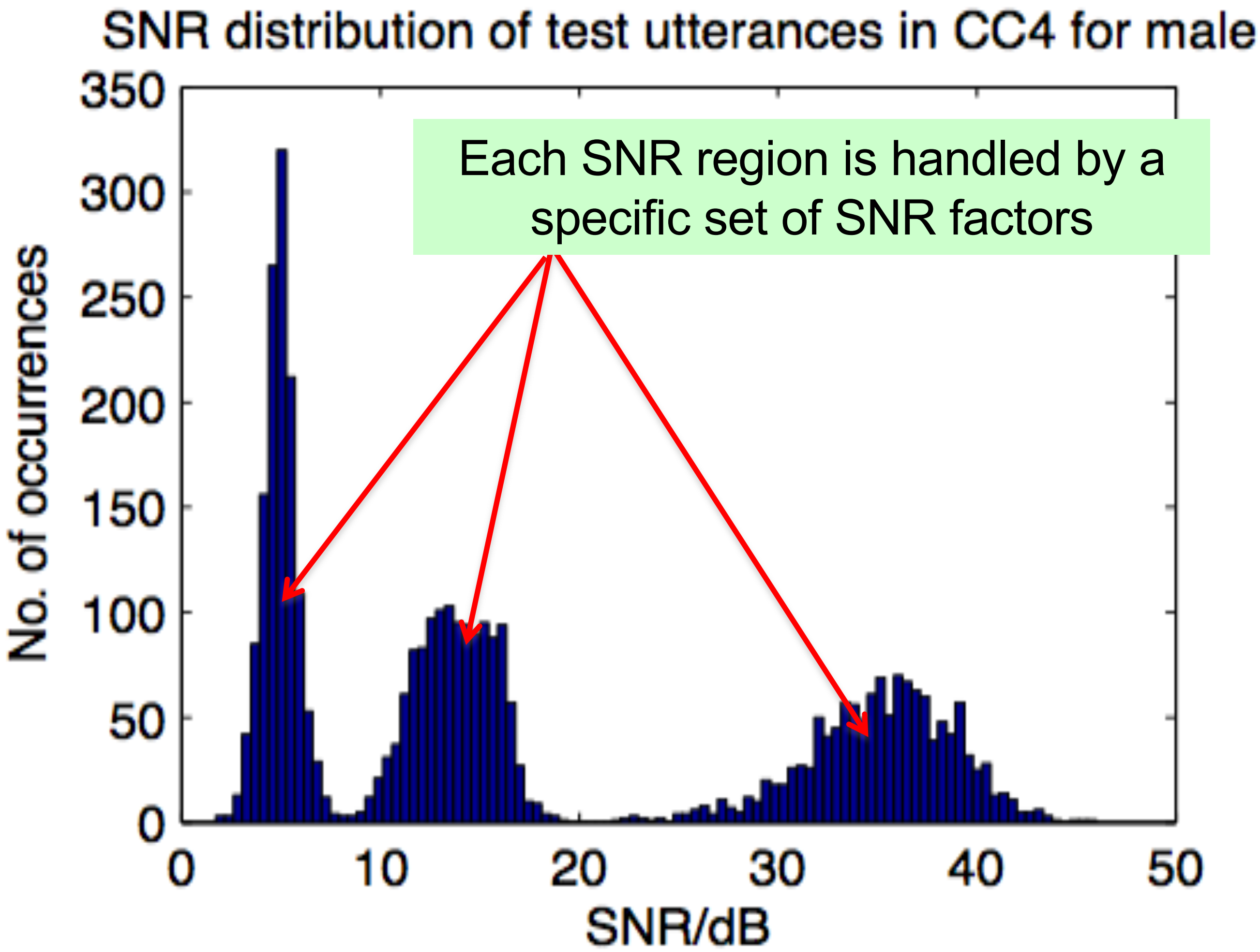


Motivation

- Method of modeling SNR information



Distribution of SNR in SRE12



THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Contents

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

SNR-invariant PLDA

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

i: Speaker index
j: Session index

- By adding an SNR factor to the conventional PLDA, we have **SNR-invariant PLDA**:

$$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^k$$

k: SNR index

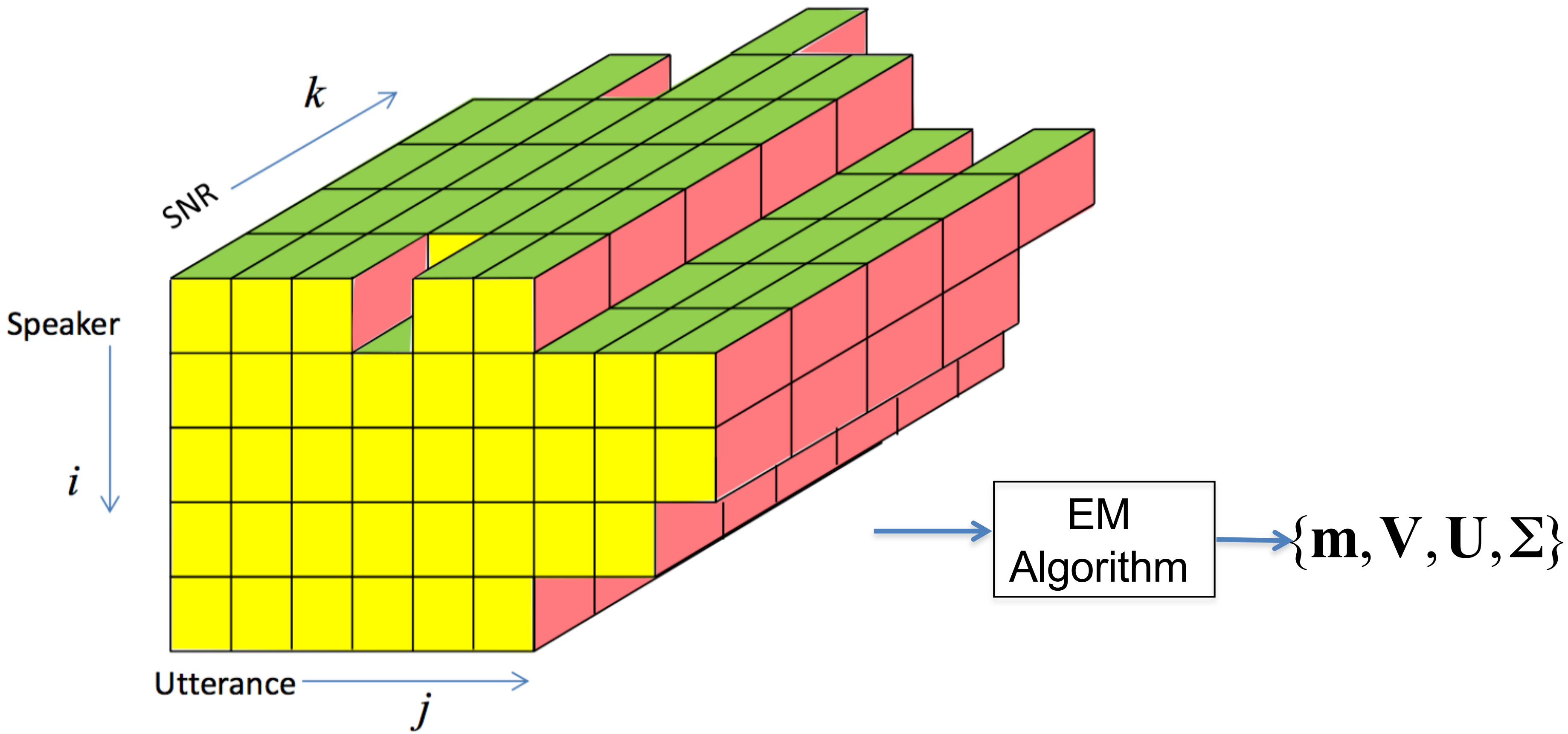
where $\mathbf{U}$ denotes the SNR subspace, $\mathbf{w}_k$ is an SNR factor, and $\mathbf{h}_i$ is the speaker (identity) factor for speaker i.

- Note that it is not the same as PLDA with channel subspace:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{R}\mathbf{r}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

SNR-invariant PLDA

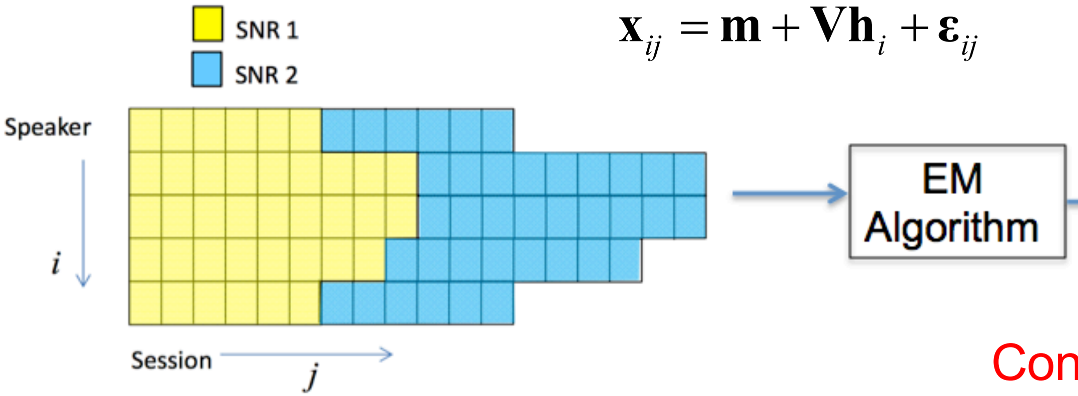
- We separate I-vectors into different **groups** according to the SNR of their utterances



$$\mathbf{x}_{ij}^{k} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^{k}$$

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

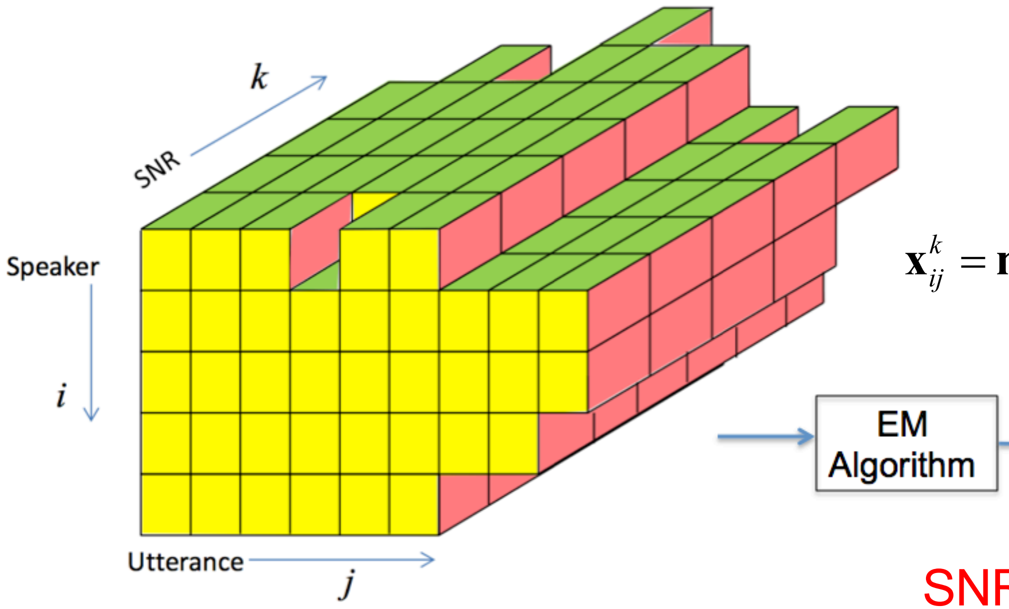
Compared with Conventional PLDA



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

$$\{\mathbf{m}, \mathbf{V}, \Sigma\}$$

Conventional PLDA



$$\mathbf{x}_{ij}^{k} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^{k}$$

$$\{\mathbf{m}, \mathbf{V}, \mathbf{U}, \Sigma\}$$

SNR-Invariant PLDA

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

PLDA vs SNR-invariant PLDA

Generative Model

PLDA	SNR-invariant PLDA
$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\varepsilon}_{ij}$ $p(\mathbf{x}) = N(\mathbf{x} \mid \mathbf{m}, \mathbf{V}\mathbf{V}^T + \Sigma)$	$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\varepsilon}_{ij}^k$ $p(\mathbf{x}) = N(\mathbf{x} \mid \mathbf{m}, \mathbf{V}\mathbf{V}^T + \mathbf{U}\mathbf{U}^T + \boldsymbol{\Sigma})$
$\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \boldsymbol{\Sigma}\}$	$\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

PLDA vs SNR-invariant PLDA

E-Step

PLDA	SNR-invariant PLDA
$\langle \mathbf{h}_i \mid X \rangle = \mathbf{L}_i^{-1} \mathbf{V}^T \mathbf{\Sigma}^{-1} \sum_{j=1}^{H_i} (\mathbf{x}_{ij} - \mathbf{m})$ $\langle \mathbf{h}_i \mathbf{h}_i^T \mid X \rangle = \mathbf{L}_i^{-1} + \langle \mathbf{h}_i \mid X \rangle \langle \mathbf{h}_i \mid X \rangle^T$ $\mathbf{L}_i = \mathbf{I} + H_i \mathbf{V}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \mathbf{V}$	$\langle \mathbf{h}_i \vert \mathcal{X} \rangle = (\mathbf{L}_i^1)^{-1} \mathbf{V}^{\top} \mathbf{\Phi}_1^{-1} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m})$ $\langle \mathbf{w}_k \vert \mathcal{X} \rangle = (\mathbf{L}_k^2)^{-1} \mathbf{U}^{\top} \mathbf{\Phi}_2^{-1} \sum_{i=1}^{S} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m})$ $\langle \mathbf{h}_i \mathbf{h}_i^{\mathsf{T}} \vert \mathcal{X} \rangle = (\mathbf{L}_i^1)^{-1} + \langle \mathbf{h}_i \vert \mathcal{X} \rangle \langle \mathbf{h}_i \vert \mathcal{X} \rangle^{\mathsf{T}}$ $\langle \mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \vert \mathcal{X} \rangle = (\mathbf{L}_k^2)^{-1} + \langle \mathbf{w}_k \vert \mathcal{X} \rangle \langle \mathbf{w}_k \vert \mathcal{X} \rangle^{\mathsf{T}}$ $\langle \mathbf{w}_k \mathbf{h}_i^{\mathsf{T}} \vert \mathcal{X} \rangle = \langle \mathbf{w}_k \vert \mathcal{X} \rangle \langle \mathbf{h}_i \vert \mathcal{X} \rangle^{\mathsf{T}}$ $\langle \mathbf{h}_i \mathbf{w}_k^{\mathsf{T}} \vert \mathcal{X} \rangle = \langle \mathbf{h}_i \vert \mathcal{X} \rangle \langle \mathbf{w}_k \vert \mathcal{X} \rangle^{\mathsf{T}}$ $\mathbf{L}_i^1 = \mathbf{I} + N_i \mathbf{V}^{\top} \mathbf{\Phi}_1^{-1} \mathbf{V} \quad \mathbf{L}_k^2 = \mathbf{I} + M_k \mathbf{U}^{\top} \mathbf{\Phi}_2^{-1} \mathbf{U}$ $\mathbf{\Phi}_1 = \mathbf{U}\mathbf{U}^{\top} + \mathbf{\Sigma} \quad \mathbf{\Phi}_2 = \mathbf{V}\mathbf{V}^{\top} + \mathbf{\Sigma}$

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

PLDA versus SNR-invariant PLDA

M-Step

<table>
<tr><th>PLDA</th><th>SNR-invariant PLDA</th></tr>
<tr><td>

$$\mathbf{V} = \left[\sum_{ij} (\mathbf{x}_{ij} - \mathbf{m}) \langle \mathbf{h}_i \mid X \rangle^T \right] \left[\sum_{ij} \langle \mathbf{h}_i \mathbf{h}_i^T \mid X \rangle \right]^{-1}$$

$$\boldsymbol{\Sigma} = \frac{\sum_{ij} \left[(\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})^T - \mathbf{V} \langle \mathbf{h}_i \mid X \rangle (\mathbf{x}_{ij} - \mathbf{m})^T \right]}{\sum_i H_i}$$

$$\mathbf{m}' = \frac{\sum_{ij} \mathbf{x}_{ij}}{\sum_i H_i}$$

</td><td>

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

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

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

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

</td></tr>
</table>

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

SNR-invariant PLDA Score

- Likelihood Ratio Scores

Given a test i-vector $\mathbf{x}_t$ and a target-speaker i-vector $\mathbf{x}_s$, the likelihood ratio score can be computed as follows:

$$L(\mathbf{x}_s, \mathbf{x}_t) = \ln \frac{P(\mathbf{x}_s, \mathbf{x}_t | \text{same-speaker})}{P(\mathbf{x}_s, \mathbf{x}_t | \text{different-speakers})}$$
$$= \text{const} + \frac{1}{2}\mathbf{x}_s^\top \mathbf{Q}\mathbf{x}_s + \frac{1}{2}\mathbf{x}_t^\top \mathbf{Q}\mathbf{x}_t + \mathbf{x}_s^\top \mathbf{P}\mathbf{x}_t$$

where

$$\mathbf{P} = \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1},$$
$$\mathbf{Q} = \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1},$$
$$\Sigma_{ac} = \mathbf{V}\mathbf{V}^\top, \text{ and } \Sigma_{tot} = \mathbf{V}\mathbf{V}^\top + \mathbf{U}\mathbf{U}^\top + \Sigma.$$

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

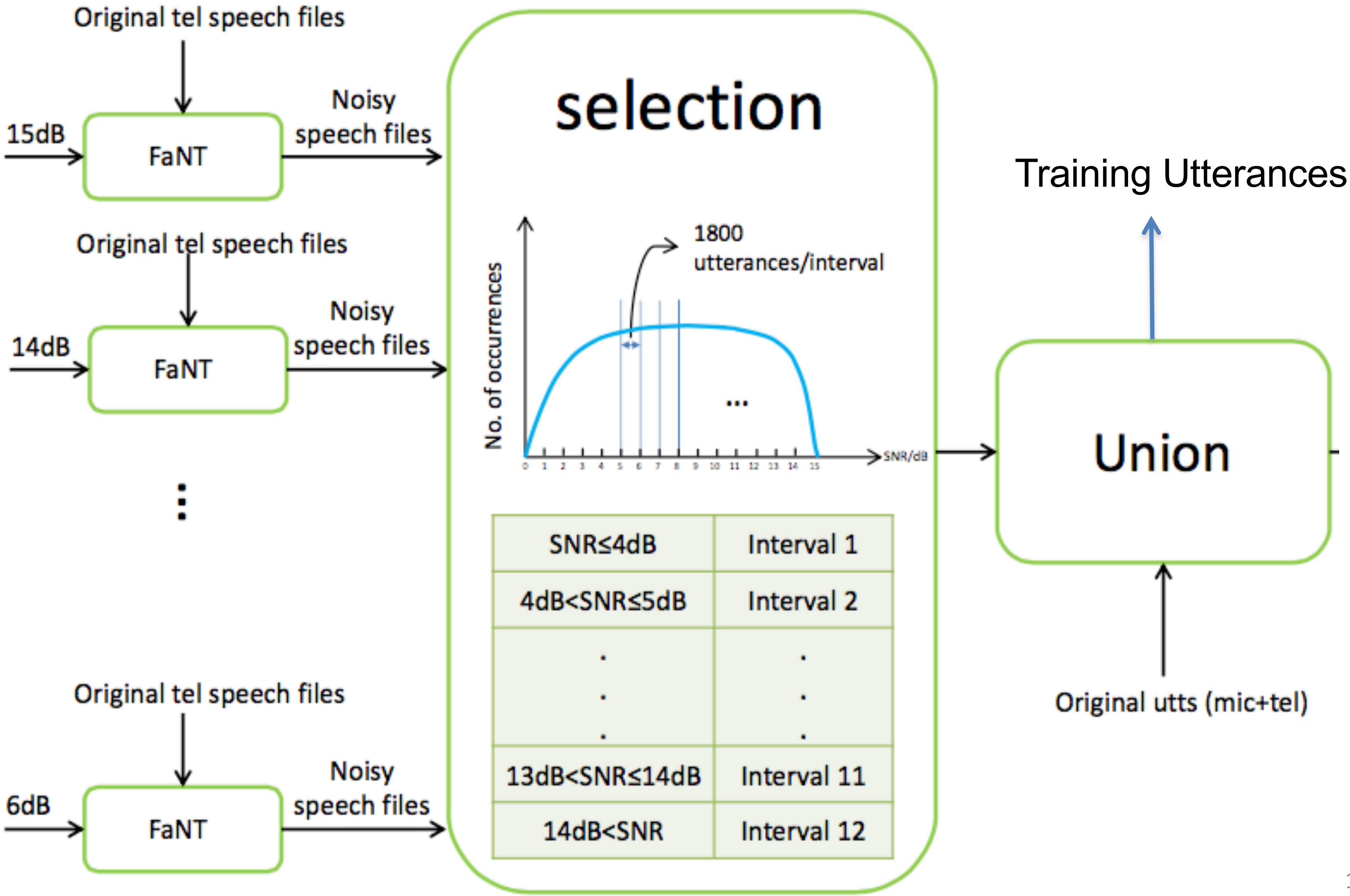
Contents

Data and Features

- **Evaluation dataset:** Common evaluation condition 1 and 4 of NIST SRE 2012 core set.
- **Parameterization:** 19 MFCCs together with energy plus their 1st and 2nd derivatives → 60-Dim
- **UBM:** gender-dependent, 1024 mixtures
- **Total Variability Matrix:** gender-dependent, 500 total factors
- **I-Vector Preprocessing:**
  - Whitening by WCCN then length normalization
  - Followed by NFA (500-dim → 200-dim)

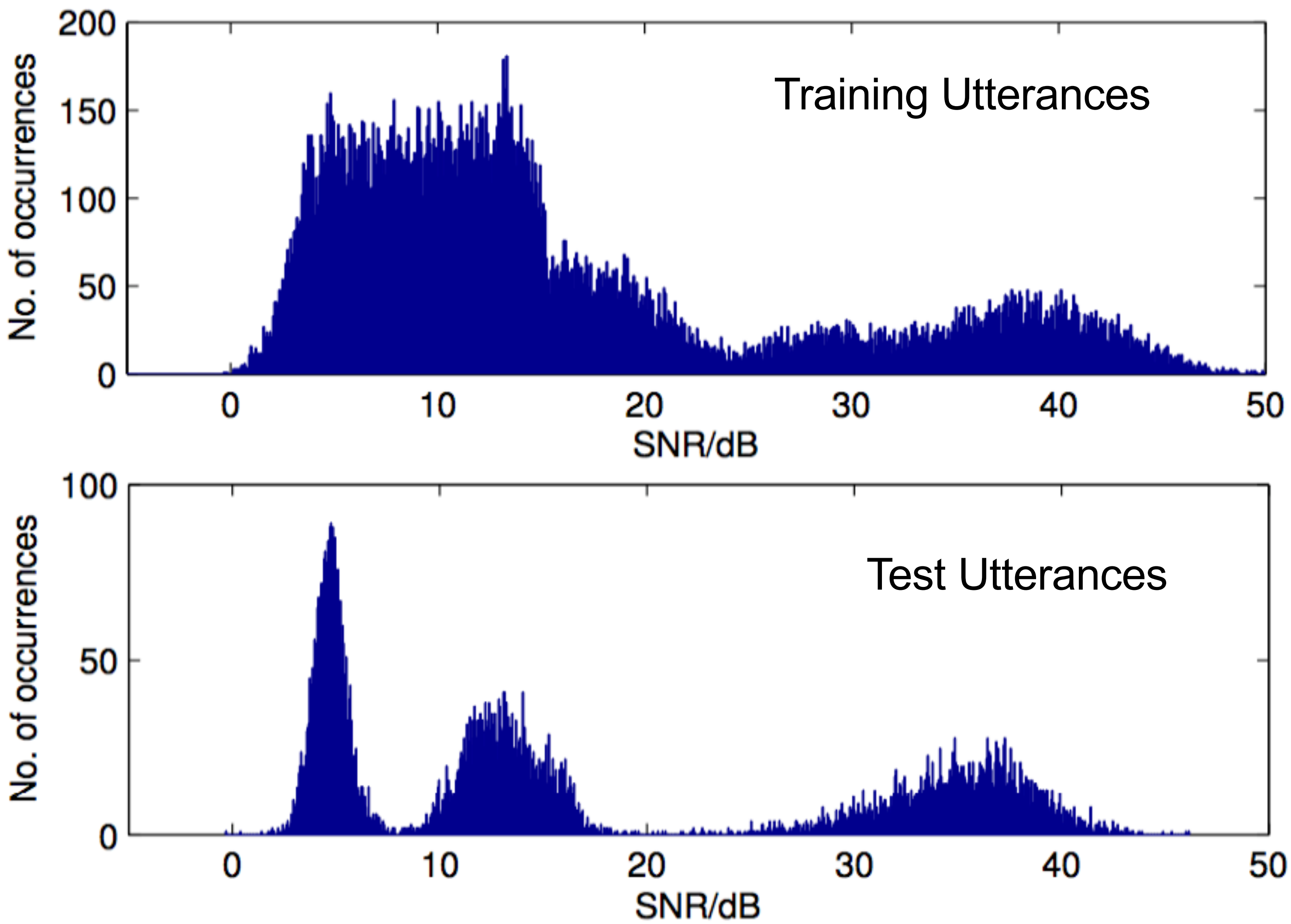
The Hong Kong Polytechnic University
香港理工大學

Finding SNR Groups



SNR Distributions

- SNR Distribution of training and test utterances in CC4



Performance on SRE12

Mixture of PLDA (*Mak, Interspeech14*)

CC1

Method	Parameters		Male		Female	
	K	Q	EER(%)	minDCF	EER(%)	minDCF
PLDA	-	-	5.42	**0.371**	7.53	0.531
mPLDA	-	-	5.28	0.415	7.70	0.539
SNR-Invariant PLDA	3	40	5.42	0.382	6.93	0.528
	5	40	**5.28**	0.381	**6.89**	**0.522**
	6	40	5.29	0.388	6.90	0.536
	8	30	5.56	0.384	7.05	0.545

No. of SNR Groups

No. of SNR factors (dim of $\mathbf{w}_k$)

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Performance on SRE12

CC4

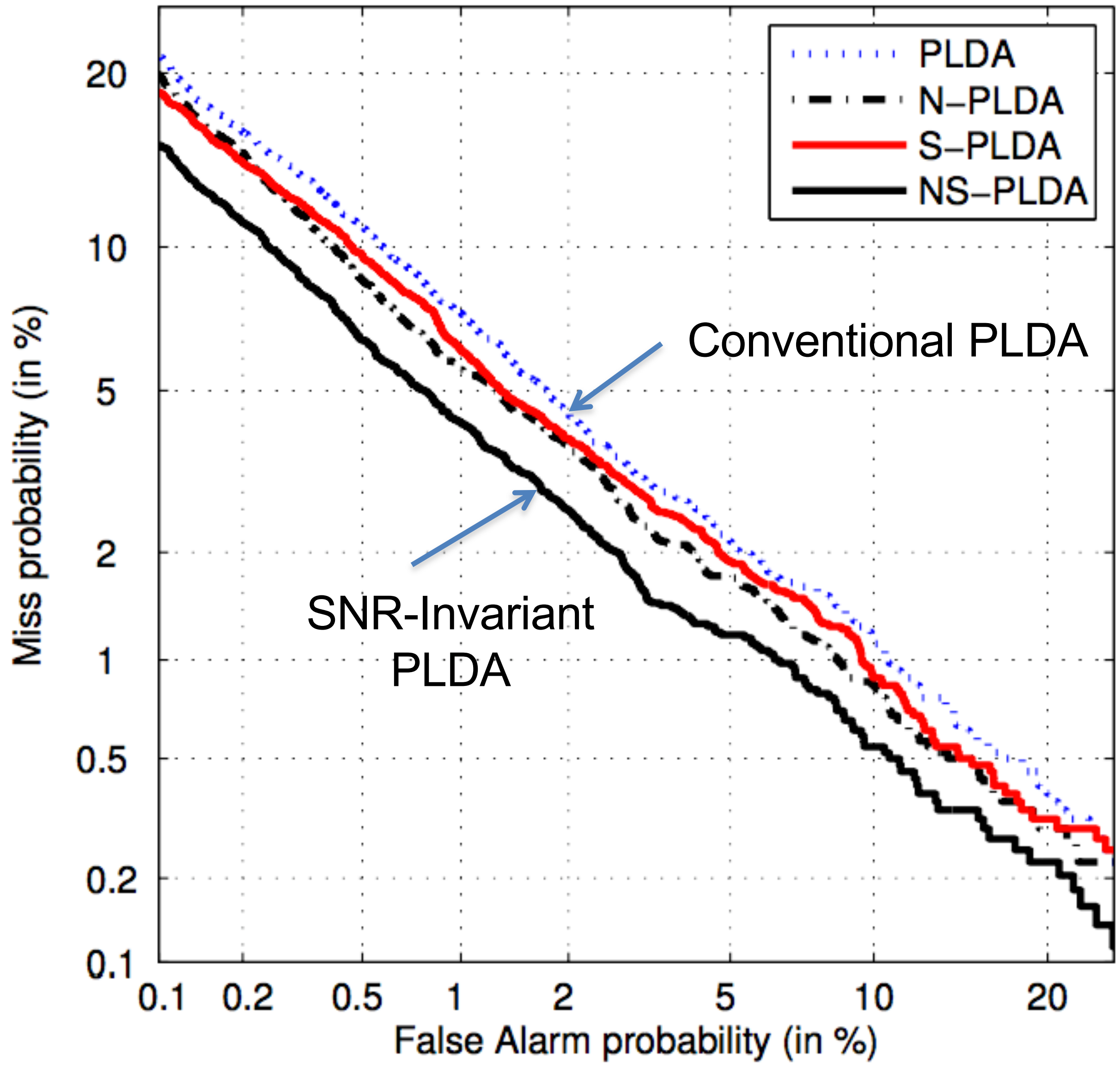
Method	Parameters		Male		Female	
	K	Q	EER(%)	minDCF	EER(%)	minDCF
PLDA	-	-	3.13	0.312	2.82	0.341
mPLDA	-	-	2.88	0.329	2.71	0.332
SNR-Invariant PLDA	3	40	2.72	0.289	2.36	0.314
	5	40	2.67	0.291	2.38	0.322
	6	40	**2.63**	**0.287**	2.43	0.319
	8	30	2.70	0.292	**2.29**	**0.313**

No. of SNR Groups

No. of SNR factors (dim of $\mathbf{w}_k$)

The Hong Kong Polytechnic University
香港理工大學

Performance on SRE12



CC4,
Female

Conclusions

- We show that while I-vectors of different SNR fall on different regions of the I-vector space, they vary within a single cluster in an SNR-subspace.
- Therefore, it is possible to model the SNR variability by adding an SNR loading matrix and SNR factors to the conventional PLDA model.