Outline

Outline

Motivation

- **Motivation of i-vectors**:
  - Insufficiency of joint factor analysis (JFA) in distinguishing between speaker and channel information, as channel factors were shown to contain speaker information.
  - Better to use a two-step approach: (1) use low-dimensional vectors (called i-vectors) that comprise both speaker and channel information to represent utterances; and (2) model the channel and variabilities of the i-vectors during scoring.
- **Motivation of Heavy-tailed PLDA**:
  - JFA assumes that the speaker and channel components follow Gaussian distributions.
  - The Gaussian assumption prohibits large deviations from the mean.
  - But speaker effects (e.g., non-native speakers) and channel effect (gross channel distortion) could cause large deviations.
  - Use heavy-tailed distributions instead of Gaussians for modeling the speaker and channel components in i-vectors [Kenny, 2010].

Generative model with heavy-tailed priors

- Assuming that we have H_i i-vectors $\mathcal{X}_i = \{\mathbf{x}_{ij}, j = 1, \ldots, H_i\}$ from speaker i, the generative model is

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

  where $\mathbf{V}$ and $\mathbf{G}$ represent the the speaker and channel subspaces, respectively.

- In heavy-tailed PLDA,

$$\begin{aligned}
\mathbf{h}_i &\sim \mathcal{N}(\mathbf{0}, u_1^{-1}\mathbf{I}) \quad & u_1 &\sim \mathcal{G}(n_1/2, n_1/2) \\
\mathbf{r}_{ij} &\sim \mathcal{N}(\mathbf{0}, u_{2j}^{-1}\mathbf{I}) \quad & u_{2j} &\sim \mathcal{G}(n_2/2, n_2/2) \\
\boldsymbol{\epsilon}_{ij} &\sim \mathcal{N}(\mathbf{0}, (v_j\boldsymbol{\Lambda})^{-1}) \quad & v_j &\sim \mathcal{G}(\nu_j/2, \nu_j/2)
\end{aligned}$$

- By integrating out the hyperparameters (u_1, u_{2j}, and v_j), one can show [Eq. 2.161 of Bishop (2006)] that the priors of $\mathbf{h}_i$, $\mathbf{r}_{ij}$, and $\boldsymbol{\epsilon}_{ij}$ follow Student's t. So, $\mathbf{x}_{ij}$ also follows Student's t.

Performance on NIST 2008 SRE

- Telephone speech, without score normalization

	Gaussian	heavy-tailed
short2-short3	3.6% / 0.014	**2.2% / 0.010**
8conv-short3	3.7% / 0.009	1.3% / **0.005**
10sec-10sec	16.4% / 0.070	**10.9% / 0.053**

- Microphone speech, with score normalization

	partially heavy-tailed	fully heavy-tailed
det1	**3.3% / 0.017**	3.4% / **0.017**
det4	**2.8% / 0.016**	3.1% / 0.018
det5	4.0% / **0.020**	**3.8% / 0.020**

[Kenny, 2010]

From HT-PLDA to Gaussian PLDA

- In 2011, [Garcia-Romero and Espy-Wilson, 2011] discovered that Gaussian PLDA performs as good as HT-PLDA provided that i-vectors have been subjected to the following pre-processing steps:

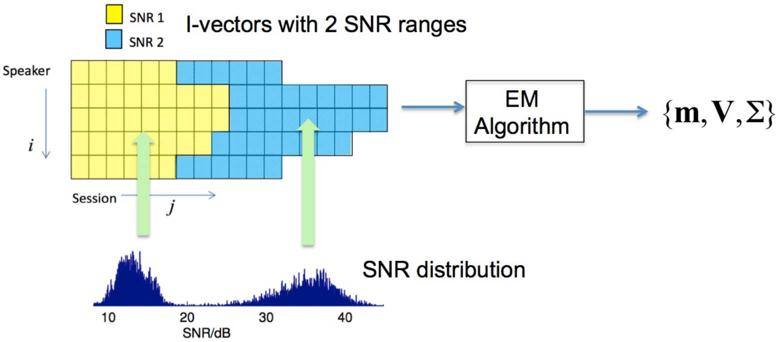
$$\text{Whitening} + \text{Length normalization}$$

- These steps have the effect of making the i-vectors more Gaussian.
- As Gaussian PLDA is computationally much simpler than HT-PLDA, the former has been extensively used in speaker verification.

Outline

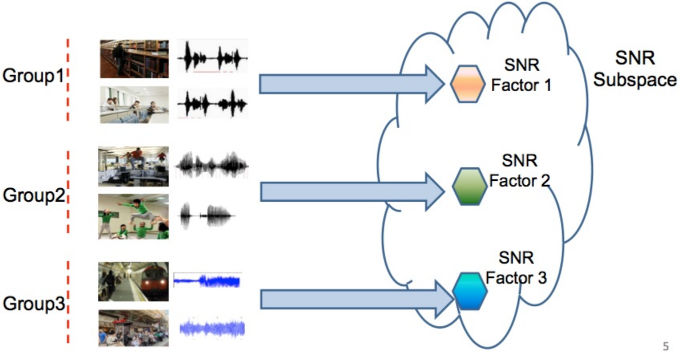
Motivation

- While i-vector extraction followed by PLDA is very effective in addressing channel variability
- Performance degrades rapidly in the presence of background noise with a wide range of signal-to-noise ratios (SNR)
- Classical approach: Multi-condition training where i-vectors from various background noise level are pooled to train a PLDA model.



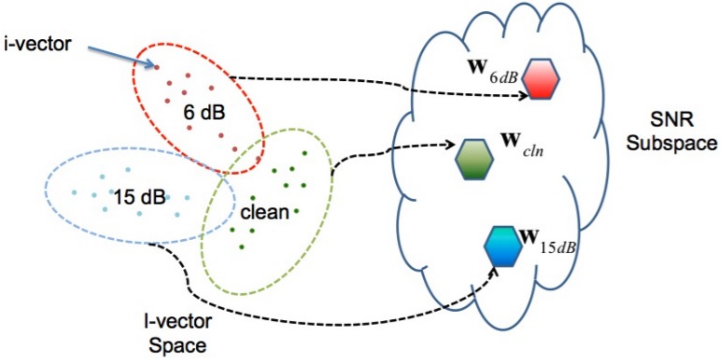
Motivation

- We argue that the variation caused by SNR variability can be modeled by an **SNR subspace** and utterances falling within a narrow SNR range should share the same set of SNR factors.
- SNR-specific information were separated from speaker-specific information through marginalizing out the SNR factors during scoring



Motivation

- I-vectors derived from utterances of similar SNR will be mapped to a small region in the SNR subspace.



SNR-Invariant PLDA

- Classical PLDA: $\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \boldsymbol{\epsilon}_{ij}$
- By adding an SNR factor to the conventional PLDA, we have **SNR-invariant PLDA** [Li and Mak, 2015]:

$$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \boldsymbol{\epsilon}_{ij}^k, \quad k = 1, \ldots, K$$

  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:

$$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{G}\mathbf{r}_{ij} + \boldsymbol{\epsilon}_{ij},$$

  where $\mathbf{G}$ defines the channel subspace and $\mathbf{r}_{ij}$ represents the channel factors.

SNR-invariant PLDA

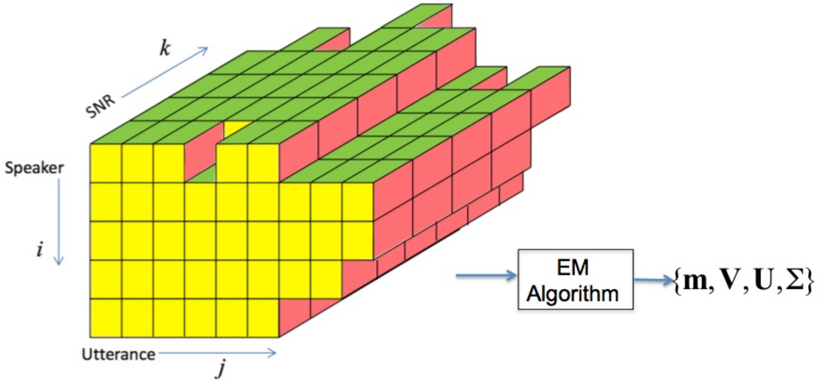
Generative model:

$$\mathbf{x}_{ij}^{k} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \epsilon_{ij}^{k}, \qquad k{=}1,\ldots,K$$

- $\mathbf{h}_i$ is speaker factors with prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- $\mathbf{x}_{ij}^{k}$ is the j-th i-vector from speaker i in the k-th SNR subgroup
- $\mathbf{V}$ is the eigenvoice matrix
- $\mathbf{U}$ defines the SNR subspace
- $\mathbf{w}_k$ is SNR factor with prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- ϵ_{ij}^{k} is a residual term with prior distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$; $\boldsymbol{\Sigma}$ is a full covariance matrix aiming to model the channel variability

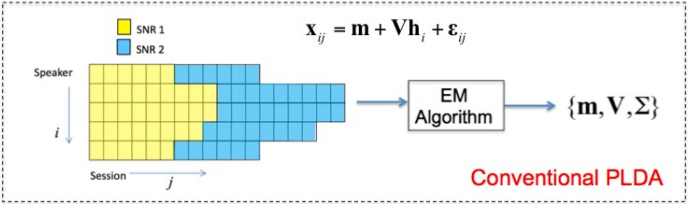
SNR-invariant PLDA

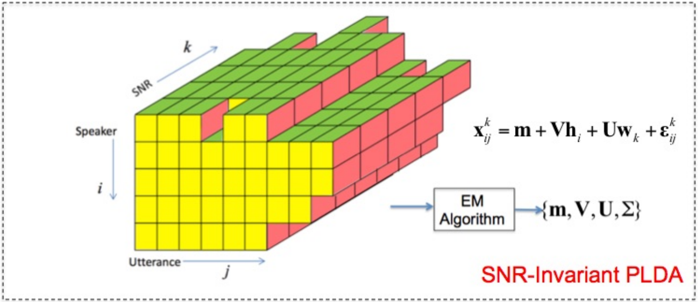
- Training utterances are divided into K groups, accroding to their SNR



PLDA vs. SNR-invariant PLDA

- Comparing the use of training i-vectors with conventional PLDA





PLDA vs. SNR-invariant PLDA

- Comparing generative models:

PLDA	SNR-Invariant PLDA
$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \epsilon_{ij}$	$\mathbf{x}_{ij}^k = \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k + \epsilon_{ij}^k$
$\mathbf{x} \sim \mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \mathbf{\Sigma})$	$\mathbf{x} \sim \mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \mathbf{U}\mathbf{U}^\mathsf{T} + \mathbf{\Sigma})$
$\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{\Sigma}\}$	$\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \mathbf{\Sigma}\}$

Auxiliary function for SNR-invariant PLDA

- The parameters $\boldsymbol{\theta} = \{\mathbf{m}, \mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$ can be learned from a training set $\mathcal{X}$ using maximum likelihood estimation.
- $\mathcal{X} = \{\mathbf{x}_{ij}^k; i = 1, \ldots, S; j = 1, \ldots, H_i(k); k = 1, \ldots, K\}$
  - S: No. of training speakers
  - K: No. of SNR groups
  - $H_i(k)$: No. of utterances from speaker i in the k-th SNR group.
- Given an initial value $\boldsymbol{\theta}$, we aim to find a new estimate $\hat{\boldsymbol{\theta}}$ that maximizes the auxiliary function:

$$
\begin{aligned}
\mathbf{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{h},\mathbf{w}}\Big[\sum\nolimits_{ikj} \ln\Big(p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \mathbf{w}_k, \hat{\boldsymbol{\theta}})p(\mathbf{h}_i, \mathbf{w}_k)\Big)\Big|\mathcal{X}, \boldsymbol{\theta}\Big] \\
&= \mathbb{E}_{\mathbf{h},\mathbf{w}}\Big[\sum\nolimits_{ikj} \Big(\ln \mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Sigma}) \\
&\qquad\qquad + \ln p(\mathbf{h}_i, \mathbf{w}_k)\Big)\Big|\mathcal{X}, \boldsymbol{\theta}\Big]
\end{aligned}
$$

Posterior distributions of latent variables

- We show 3 ways to compute the posteriors:
  1. Computing the posterior of $\mathbf{h}_i$ and $\mathbf{w}_k$ separately.
  2. Computing the posterior $\mathbf{h}_i$ while fixing $\mathbf{w}_k$ using the Gauss-Seidel method.
  3. Computing the joint posterior of $\mathbf{h}_i$ and $\mathbf{w}_k$ using variational Bayes.

Method 1: Computing posteriors separately

- Given i-vectors $\mathbf{x}_{ij}^{k}$, the posterior density of $\mathbf{h}_i$ has the form:

$$
\begin{aligned}
p(\mathbf{h}_i|\mathbf{x}_{ij}^{k},\boldsymbol{\theta}) &\propto p(\mathbf{x}_{ij}^{k}|\mathbf{h}_i,\boldsymbol{\theta})p(\mathbf{h}_i) \\
&= \int p(\mathbf{x}_{ij}^{k},\mathbf{w}_k|\mathbf{h}_i,\boldsymbol{\theta})p(\mathbf{h}_i)\mathrm{d}\mathbf{w}_k \\
&= \int p(\mathbf{x}_{ij}^{k}|\mathbf{h}_i,\mathbf{w}_k,\boldsymbol{\theta})p(\mathbf{w}_k)p(\mathbf{h}_i)\mathrm{d}\mathbf{w}_k \\
&= \int \mathcal{N}(\mathbf{x}_{ij}^{k}|\mathbf{m}+\mathbf{V}\mathbf{h}_i+\mathbf{U}\mathbf{w}_k,\boldsymbol{\Sigma})\mathcal{N}(\mathbf{w}_k|\mathbf{0},\mathbf{I})\mathcal{N}(\mathbf{h}_i|\mathbf{0},\mathbf{I})\mathrm{d}\mathbf{w}_k \\
&= \mathcal{N}(\mathbf{x}_{ij}^{k}|\mathbf{m}+\mathbf{V}\mathbf{h}_i,\boldsymbol{\Phi})\mathcal{N}(\mathbf{h}_i|\mathbf{0},\mathbf{I}) \\
&\propto \exp\left\{\mathbf{h}_i^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\boldsymbol{\Phi}^{-1}(\mathbf{x}_{ij}^{k}-\mathbf{m})-\frac{1}{2}\mathbf{h}_i^{\mathsf{T}}(\mathbf{I}+\mathbf{V}^{\mathsf{T}}\boldsymbol{\Phi}^{-1}\mathbf{V})\mathbf{h}_i\right\}
\end{aligned}
$$

where $\boldsymbol{\Phi}=\mathbf{U}\mathbf{U}^{\mathsf{T}}+\boldsymbol{\Sigma}$.

Method 1: Computing posteriors separately

- If all of the i-vectors of speaker i, say $\mathcal{X}_i$, are given,

$$p(\mathbf{h}_i|\mathbf{x}_{ij}^k\ \forall j \text{ and } k, \boldsymbol{\theta}) \propto \prod_{k=1}^{K}\prod_{j=1}^{H_i(k)} p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \boldsymbol{\theta})p(\mathbf{h}_i)$$

$$\propto \exp\left\{\mathbf{h}_i^\mathsf{T}\mathbf{V}^\mathsf{T}\boldsymbol{\Phi}^{-1}\sum_{k=1}^{K}\sum_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m}) - \frac{1}{2}\mathbf{h}_i^\mathsf{T}\left(\mathbf{I} + \sum_{k=1}^{K} H_i(k)\mathbf{V}^\mathsf{T}\boldsymbol{\Phi}^{-1}\mathbf{V}\right)\mathbf{h}_i\right\}$$

- This is a Gaussian with mean and 2nd-order (uncentralized) moment

$$\langle\mathbf{h}_i|\mathcal{X}_i\rangle = \left(\mathbf{I} + \sum\nolimits_{k=1}^{K} H_i(k)\mathbf{V}^\mathsf{T}\boldsymbol{\Phi}^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^\mathsf{T}\boldsymbol{\Phi}^{-1}\sum\nolimits_{k=1}^{K}\sum\nolimits_{j=1}^{H_i(k)}(\mathbf{x}_{ij}^k - \mathbf{m})$$

$$\langle\mathbf{h}_i\mathbf{h}_i^\mathsf{T}|\mathcal{X}_i\rangle = \left(\mathbf{I} + \sum\nolimits_{k=1}^{K} H_i(k)\mathbf{V}^\mathsf{T}\boldsymbol{\Phi}^{-1}\mathbf{V}\right)^{-1} + \langle\mathbf{h}_i|\mathcal{X}_i\rangle\langle\mathbf{h}_i|\mathcal{X}_i\rangle^\mathsf{T}, \tag{1}$$

$$\boxed{\mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_h, \mathbf{C}_h) \propto \exp\left\{-\frac{1}{2}(\mathbf{h} - \boldsymbol{\mu}_h)^\mathsf{T}\mathbf{C}_h^{-1}(\mathbf{h} - \boldsymbol{\mu}_h)\right\} \propto \exp\left\{\mathbf{h}^\mathsf{T}\mathbf{C}_h^{-1}\boldsymbol{\mu}_h - \frac{1}{2}\mathbf{h}^\mathsf{T}\mathbf{C}_h^{-1}\mathbf{h}\right\}}$$

Method 1: Computing posteriors separately

- Similarly, to compute the posterior expectations of $\mathbf{w}_k$, we marginalize over $\mathbf{h}_i$'s. Thus, the posterior density of $\mathbf{w}_k$ is

$$
\begin{aligned}
p(\mathbf{w}_k|\mathbf{x}_{ij}^k, \boldsymbol{\theta}) &\propto \int p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \mathbf{w}_k, \boldsymbol{\theta}) p(\mathbf{h}_i) p(\mathbf{w}_k) \mathrm{d}\mathbf{h}_i \\
&= \int \mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Sigma}) \mathcal{N}(\mathbf{h}_i|\mathbf{0}, \mathbf{I}) \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I}) \mathrm{d}\mathbf{h}_i \\
&= \mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m} + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Psi}) \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I}) \\
&\propto \exp\left\{ \mathbf{w}_k^\mathsf{T} \mathbf{U}^\mathsf{T} \boldsymbol{\Psi}^{-1} (\mathbf{x}_{ij}^k - \mathbf{m}) - \frac{1}{2} \mathbf{w}_k^\mathsf{T} (\mathbf{I} + \mathbf{U}^\mathsf{T} \boldsymbol{\Psi}^{-1} \mathbf{U}) \mathbf{w}_k \right\}
\end{aligned}
$$

Method 1: Computing posteriors separately

- Given all of the i-vectors ($\mathcal{X}^k$) from the k-th SNR group, we can compute the posterior expectations as follows:

$$\langle \mathbf{w}_k | \mathcal{X}^k \rangle = \left(\mathbf{I} + \sum_{i=1}^{S} H_i(k) \mathbf{U}^{\mathsf{T}} \boldsymbol{\Psi}^{-1} \mathbf{U} \right)^{-1} \mathbf{U}^{\mathsf{T}} \boldsymbol{\Psi}^{-1} \sum_{i=1}^{S} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m})$$

$$\langle \mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} | \mathcal{X}^k \rangle = \left(\mathbf{I} + \sum_{i=1}^{S} H_i(k) \mathbf{U}^{\mathsf{T}} \boldsymbol{\Psi}^{-1} \mathbf{U} \right)^{-1} + \langle \mathbf{w}_k | \mathcal{X}^k \rangle \langle \mathbf{w}_k | \mathcal{X}^k \rangle^{\mathsf{T}} \tag{2}$$

where $\boldsymbol{\Psi} = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}$

Method 2: Computing posteriors by Gauss-Seidel method

- Another approach to computing $p(\mathbf{h}_i|\mathcal{X}_i)$ is to assume that $\mathbf{w}_k$'s are fixed for all $k = 1, \dots, K$.
- This is called the Gauss-Seidel method [Barrett et al., 1994]
- We fix $\mathbf{w}_k$ to its posterior mean: $\mathbf{w}_k^* \equiv \langle \mathbf{w}_k | \mathcal{X}^k \rangle$
- The posterior density of $\mathbf{h}_i$ becomes:

$$
\begin{aligned}
p(\mathbf{h}_i|\mathcal{X}_i, \mathbf{w}_k^*, \boldsymbol{\theta}) &\propto \prod_{k=1}^{K} \prod_{j=1}^{H_i(k)} p(\mathbf{x}_{ij}^k|\mathbf{h}_i, \mathbf{w}_k^*, \boldsymbol{\theta}) p(\mathbf{h}_i) \\
&= \prod_{k=1}^{K} \prod_{j=1}^{H_i(k)} \mathcal{N}(\mathbf{x}_{ij}^k|\mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k^*, \boldsymbol{\Sigma}) \mathcal{N}(\mathbf{h}_i|\mathbf{0}, \mathbf{I}) \\
&\propto \exp\left\{ \mathbf{h}_i^\mathsf{T} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{U}\mathbf{w}_k^*) - \right. \\
&\qquad \left. \frac{1}{2} \mathbf{h}_i^\mathsf{T} \left(\mathbf{I} + \sum_{k=1}^{K} H_i(k) \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{V} \right) \mathbf{h}_i \right\}
\end{aligned}
$$

Method 2: Computing posteriors by Gauss-Seidel method

- Comparing this posterior density with a standard Gaussian, we have

$$\begin{aligned}\langle \mathbf{h}_i | \mathcal{X}_i \rangle &= \left(\mathbf{L}_i^{(1)}\right)^{-1} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \sum_{k=1}^{K} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{U}\mathbf{w}_k^*) \\ \langle \mathbf{h}_i \mathbf{h}_i^{\mathsf{T}} | \mathcal{X}_i \rangle &= \left(\mathbf{L}_i^{(1)}\right)^{-1} + \langle \mathbf{h}_i | \mathcal{X}_i \rangle \langle \mathbf{h}_i | \mathcal{X}_i \rangle^{\mathsf{T}},\end{aligned} \tag{3}$$

  where $\mathbf{L}_i^{(1)} \equiv \mathbf{I} + \sum_{k=1}^{K} H_i(k) \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V}$
- Note that these formulations is similar to the JFA model estimation in [Vogt and Sridharan, 2008].

Method 2: Computing posteriors by Gauss-Seidel method

- Apply the same approach to computing the posterior density of $\mathbf{w}_k$, we have

$$\langle \mathbf{w}_k | \mathcal{X}^k \rangle = \left(\mathbf{L}_k^{(2)}\right)^{-1} \mathbf{U}^\mathsf{T} \mathbf{\Sigma}^{-1} \sum_{i=1}^{S} \sum_{j=1}^{H_i(k)} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{V}\mathbf{h}_i^*)$$
$$\langle \mathbf{w}_k \mathbf{w}_k^\mathsf{T} | \mathcal{X}^k \rangle = \left(\mathbf{L}_k^{(2)}\right)^{-1} + \langle \mathbf{w}_k | \mathcal{X}^k \rangle \langle \mathbf{w}_k | \mathcal{X}^k \rangle^\mathsf{T} \quad (4)$$

where $\mathbf{L}_k^{(2)} = \mathbf{I} + \sum_{i=1}^{S} H_i(k) \mathbf{U}^\mathsf{T} \mathbf{\Sigma}^{-1} \mathbf{U}$ and $\mathbf{h}_i^* \equiv \langle \mathbf{h}_i | \mathcal{X}_i \rangle$

Method 3: Computing posteriors by variational Bayes

- Denote $\underline{\mathbf{w}} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ and $\underline{\mathbf{h}} = [\mathbf{h}_1, \ldots, \mathbf{h}_S]$
- In variational Bayes [Bishop, 2006, Kenny, 2010], we factorize the joint posterior as follows:

$$\ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}} | \mathcal{X}) \approx \ln q(\underline{\mathbf{h}}) + \ln q(\underline{\mathbf{w}}) = \sum_{i=1}^{S} \ln q(\mathbf{h}_i) + \sum_{k=1}^{K} \ln q(\mathbf{w}_k)$$

where

$$\begin{aligned} \ln q(\underline{\mathbf{h}}) &= \mathbb{E}_{\underline{\mathbf{w}}}\{\ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}}, \mathcal{X})\} + \text{const} \\ \ln q(\underline{\mathbf{w}}) &= \mathbb{E}_{\underline{\mathbf{h}}}\{\ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}}, \mathcal{X})\} + \text{const} \end{aligned}$$

where $\mathbb{E}_{\underline{\mathbf{w}}}$ means taking expectation with respect to $\underline{\mathbf{w}}$.

Method 3: Computing posteriors by variational Bayes

- Consider $\ln q(\underline{\mathbf{h}})$:

$$\begin{aligned}
&\ln q(\underline{\mathbf{h}}) = \mathbb{E}_{\underline{\mathbf{w}}}\{\ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}}, \mathcal{X})\} + \text{const} \\
&= \langle \ln p(\mathcal{X}|\underline{\mathbf{h}}, \underline{\mathbf{w}})\rangle_{\underline{\mathbf{w}}} + \langle \ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}})\rangle_{\underline{\mathbf{w}}} + \text{const} \\
&= \sum_{ijr} \left\langle \ln \mathcal{N}(\mathbf{x}_{ij}^r | \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_r, \boldsymbol{\Sigma}) \right\rangle_{\mathbf{w}_r} \\
&\qquad + \sum_i \langle \ln \mathcal{N}(\mathbf{h}_i|\mathbf{0}, \mathbf{I})\rangle_{\underline{\mathbf{w}}} + \sum_r \langle \ln \mathcal{N}(\mathbf{w}_r|\mathbf{0}, \mathbf{I})\rangle_{\underline{\mathbf{w}}} + \text{const} \\
&= -\frac{1}{2} \sum_{ijr} (\mathbf{x}_{ij}^r - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{U}\mathbf{w}_r^*)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ij}^r - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{U}\mathbf{w}_r^*) \\
&\qquad - \frac{1}{2} \sum_i \mathbf{h}_i^\mathsf{T} \mathbf{h}_i + \text{const} \qquad (5) \\
&= \sum_i \left[\mathbf{h}_i^\mathsf{T} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \sum_{jr} (\mathbf{x}_{ij}^r - \mathbf{m} - \mathbf{U}\mathbf{w}_r^*) - \frac{1}{2} \mathbf{h}_i^\mathsf{T} \left(\mathbf{I} + \sum_{jr} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{V} \right) \mathbf{h}_i^\mathsf{T} \right] + \text{const}
\end{aligned}$$

- $q(\mathbf{h}_i)$ a Gaussian with mean and precision identical to Eq. 3:

$$\begin{aligned}
\langle \mathbf{h}_i | \mathcal{X}_i \rangle &= \left(\mathbf{L}_i^{(1)} \right)^{-1} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \sum_{jr} (\mathbf{x}_{ij}^r - \mathbf{m} - \mathbf{U}\mathbf{w}_r^*) \\
\mathbf{L}_i^{(1)} &= \mathbf{I} + \sum_{jr} \mathbf{V}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{V}
\end{aligned} \qquad (6)$$

Method 3: Computing posteriors by variational Bayes

$$
\begin{aligned}
\ln q(\underline{\mathbf{w}}) &= \langle \ln p(\mathcal{X}|\underline{\mathbf{h}}, \underline{\mathbf{w}}) \rangle_{\underline{\mathbf{h}}} + \langle \ln p(\underline{\mathbf{h}}, \underline{\mathbf{w}}) \rangle_{\underline{\mathbf{h}}} + \text{const} \\
&= \sum_{ijk} \left\langle \ln \mathcal{N}(\mathbf{x}_{ij}^k | \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \mathbf{\Sigma}) \right\rangle_{\mathbf{h}_i} \\
&\qquad + \sum_i \langle \ln \mathcal{N}(\mathbf{h}_i | \mathbf{0}, \mathbf{I}) \rangle_{\mathbf{h}_i} + \sum_k \langle \ln \mathcal{N}(\mathbf{w}_k | \mathbf{0}, \mathbf{I}) \rangle_{\underline{\mathbf{h}}} + \text{const} \\
&= -\frac{1}{2} \sum_{ijk} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{V}\mathbf{h}_i^* - \mathbf{U}\mathbf{w}_k)^{\mathsf{T}} \mathbf{\Sigma}^{-1} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{V}\mathbf{h}_i^* - \mathbf{U}\mathbf{w}_k) \\
&\qquad - \frac{1}{2} \sum_k \mathbf{w}_k^{\mathsf{T}} \mathbf{w}_k + \text{const} \\
&= \sum_k \left[\mathbf{w}_k^{\mathsf{T}} \mathbf{U}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \sum_{ij} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{V}\mathbf{h}_i^*) - \frac{1}{2} \mathbf{w}_k^{\mathsf{T}} \left(\mathbf{I} + \sum_{ij} \mathbf{U}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \mathbf{U} \right) \mathbf{w}_k^{\mathsf{T}} \right] + \text{const}
\end{aligned}
$$

- $q(\mathbf{w}_k)$ is a Gaussian with mean and precision identical to Eq. 4:

$$
\begin{aligned}
\langle \mathbf{w}_k | \mathcal{X}^k \rangle &= \left(\mathbf{L}_k^{(2)} \right)^{-1} \mathbf{U}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \sum_{ij} (\mathbf{x}_{ij}^k - \mathbf{m} - \mathbf{V}\mathbf{h}_i^*) \\
\mathbf{L}_k^{(2)} &= \mathbf{I} + \sum_{ij} \mathbf{U}^{\mathsf{T}} \mathbf{\Sigma}^{-1} \mathbf{U}
\end{aligned}
\tag{7}
$$

Note : $\langle \ln \mathcal{N}(\mathbf{h}_i | \mathbf{0}, \mathbf{I}) \rangle_{\mathbf{h}_i}$ is the differential entropy of normal distribution and is independent of $\mathbf{w}_k$, see [Norwich, 1993](Ch 8).

Computing posterior moment

- The exact posterior moment $\langle \mathbf{w}_k \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle$ will be complicated because $\mathbf{h}_i$ and $\mathbf{w}_k$ are correlated in the posterior.
- If Gauss-Seidel's method is used, we may approximate the posterior moments by (Kenny 2010, p.6)

$$\langle \mathbf{w}_k \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \approx \langle \mathbf{w}_k | \mathcal{X}^k \rangle (\mathbf{h}_i^*)^\mathsf{T}$$
$$\langle \mathbf{h}_i \mathbf{w}_k^\mathsf{T} | \mathcal{X} \rangle \approx \langle \mathbf{h}_i | \mathcal{X}_i \rangle (\mathbf{w}_k^*)^\mathsf{T}$$

  where $\mathbf{h}_i^*$ and $\mathbf{w}_k^*$ are the most up-to-date posterior means in the EM iterations.
- Alternatively, we may compute the exact joint posterior.[3] But it will be computationally intensive.

[3] http://www.eie.polyu.edu.hk/∼mwmak/papers/si-plda.pdf

Computing posterior moment

- A better approach is to use variational Bayes:

$$p(\mathbf{h}_i, \mathbf{w}_k | \mathcal{X}) \approx q(\mathbf{h}_i) q(\mathbf{w}_k) \tag{8}$$

- Note that as both $q(\mathbf{h}_i)$ and $q(\mathbf{w}_k)$ are Gaussians. Based on the law of total expectation,[4] the factorization in Eq. 8 gives

$$\langle \mathbf{w}_k \mathbf{h}_i^\mathsf{T} | \mathcal{X} \rangle \approx \langle \mathbf{w}_k | \mathcal{X}^k \rangle \langle \mathbf{h}_i | \mathcal{X}_i \rangle^\mathsf{T}$$
$$\langle \mathbf{h}_i \mathbf{w}_k^\mathsf{T} | \mathcal{X} \rangle \approx \langle \mathbf{h}_i | \mathcal{X}_i \rangle \langle \mathbf{w}_k | \mathcal{X}^k \rangle^\mathsf{T}$$

[4] https://en.wikipedia.org/wiki/Product_distribution

Maximization Step

- In the M-step, we maximize the auxiliary function:

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\underline{\mathbf{h}},\underline{\mathbf{w}}} \left\{ \sum\nolimits_{ijk} \ln \mathcal{N}\left(\mathbf{x}_{ij}^{k} \middle| \mathbf{m} + \mathbf{V}\mathbf{h}_i + \mathbf{U}\mathbf{w}_k, \boldsymbol{\Sigma} \right) p(\mathbf{h}_i, \mathbf{w}_k) \middle| \mathcal{X}, \boldsymbol{\theta} \right\}$$

$$= \sum_{ijk} \mathbb{E}_{\underline{\mathbf{h}},\underline{\mathbf{w}}} \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \left(\mathbf{x}_{ij}^{k} - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{U}\mathbf{w}_k \right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \right.$$

$$\left. \times \left(\mathbf{x}_{ij}^{k} - \mathbf{m} - \mathbf{V}\mathbf{h}_i - \mathbf{U}\mathbf{w}_k \right) + \ln p(\mathbf{h}_i, \mathbf{w}_k) \middle| \mathcal{X}, \boldsymbol{\theta} \right\}$$

- As $p(\mathbf{h}_i, \mathbf{w}_k)$ is independent of the model parameters $\mathbf{V}$, $\mathbf{U}$, and $\boldsymbol{\Sigma}$, they could be taken out of $Q(\boldsymbol{\theta})$ in the M-step [Prince and Elder, 2007].

Maximization Step

- Differentiating $Q(\boldsymbol{\theta})$ with respect to $\mathbf{V}$, $\mathbf{U}$, and $\boldsymbol{\Sigma}$ and set the results to 0, we obtain

$$
\begin{aligned}
\mathbf{V} &= \left\{ \sum_{ijk} \left[(\mathbf{x}_{ij}^{k} - \mathbf{m}) \langle \mathbf{h}_i | \mathcal{X}_i \rangle - \mathbf{U} \langle \mathbf{w}_k \mathbf{h}_i^{\mathsf{T}} | \mathcal{X} \rangle \right] \right\} \left[\sum_{ijk} \langle \mathbf{h}_i \mathbf{h}_i^{\mathsf{T}} | \mathcal{X} \rangle \right]^{-1} \\
\mathbf{U} &= \left\{ \sum_{ijk} \left[(\mathbf{x}_{ij}^{k} - \mathbf{m}) \langle \mathbf{w}_k | \mathcal{X}^k \rangle - \mathbf{V} \langle \mathbf{h}_i \mathbf{w}_k^{\mathsf{T}} | \mathcal{X} \rangle \right] \right\} \left[\sum_{ijk} \langle \mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} | \mathcal{X} \rangle \right]^{-1} \\
\boldsymbol{\Sigma} &= \frac{1}{N} \sum_{ijk} \Big[(\mathbf{x}_{ij}^{k} - \mathbf{m})(\mathbf{x}_{ij}^{k} - \mathbf{m})^{\mathsf{T}} \\
&\qquad\qquad - \mathbf{V} \langle \mathbf{h}_i | \mathcal{X}_i \rangle (\mathbf{x}_{ij}^{k} - \mathbf{m})^{\mathsf{T}} - \mathbf{U} \langle \mathbf{w}_k | \mathcal{X}^k \rangle (\mathbf{x}_{ij}^{k} - \mathbf{m})^{\mathsf{T}} \Big]
\end{aligned}
$$

Likelihood Ratio Scores

- Given target-speaker's i-vector $\mathbf{x}_s$ and test-speaker's i-vector $\mathbf{x}_t$
- If $\mathbf{x}_s$ and $\mathbf{x}_t$ are from the same speaker, they should **share** the same speaker factor $\mathbf{h}$:

$$\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \mathbf{U} & \mathbf{0} \\ \mathbf{V} & \mathbf{0} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{w}_s \\ \mathbf{w}_t \end{bmatrix} + \begin{bmatrix} \epsilon_s \\ \epsilon_t \end{bmatrix}$$

$$\Longrightarrow \hat{\mathbf{x}}_{st} = \hat{\mathbf{m}} + \hat{\mathbf{A}}\hat{\mathbf{z}}_{st} + \hat{\epsilon}_{st}.$$

- Same-speaker likelihood:

$$\begin{aligned} p(\hat{\mathbf{x}}_{st}|\text{same-speaker}) &= \int p(\hat{\mathbf{x}}_{st}|\hat{\mathbf{z}}_{st})p(\hat{\mathbf{z}}_{st})d\hat{\mathbf{z}}_{st} \\ &= \int \mathcal{N}(\hat{\mathbf{x}}_{st}|\hat{\mathbf{m}} + \hat{\mathbf{A}}\hat{\mathbf{z}}_{st}, \hat{\mathbf{\Sigma}})\mathcal{N}(\hat{\mathbf{z}}_{st}|\mathbf{0}, \mathbf{I})d\hat{\mathbf{z}}_{st} \\ &= \mathcal{N}(\hat{\mathbf{x}}_{st}|\hat{\mathbf{m}}, \hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathsf{T}} + \hat{\mathbf{\Sigma}}) \\ &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{tot} & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Sigma}_{tot} \end{bmatrix}\right) \end{aligned}$$

where $\hat{\mathbf{\Sigma}} = \text{diag}\{\mathbf{\Sigma}, \mathbf{\Sigma}\}$, $\mathbf{\Sigma}_{tot} = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{U}\mathbf{U}^{\mathsf{T}} + \mathbf{\Sigma}$ and $\mathbf{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}}$

Likelihood Ratio Scores

- If $\mathbf{x}_s$ and $\mathbf{x}_t$ are from different speakers, they should have their own speaker factor $(\mathbf{h}_s, \mathbf{h}_t)$:

$$\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_t \\ \mathbf{w}_s \\ \mathbf{w}_t \end{bmatrix} + \begin{bmatrix} \epsilon_s \\ \epsilon_t \end{bmatrix}$$

$$\Longrightarrow \hat{\mathbf{x}}_{st} = \hat{\mathbf{m}} + \bar{\mathbf{A}}\bar{\mathbf{z}}_{st} + \hat{\epsilon}_{st}$$

- Different-speaker likelihood:

$$\begin{aligned} p(\hat{\mathbf{x}}_{st}|\text{diff-speaker}) &= \int p(\hat{\mathbf{x}}_{st}|\bar{\mathbf{z}}_{st}) p(\bar{\mathbf{z}}_{st}) d\bar{\mathbf{z}}_{st} \\ &= \int \mathcal{N}(\hat{\mathbf{x}}_{st}|\hat{\mathbf{m}} + \bar{\mathbf{A}}\bar{\mathbf{z}}_{st}, \hat{\boldsymbol{\Sigma}}) \mathcal{N}(\bar{\mathbf{z}}_{st}|\mathbf{0}, \mathbf{I}) d\bar{\mathbf{z}}_{st} \\ &= \mathcal{N}(\hat{\mathbf{x}}_{st}|\hat{\mathbf{m}}, \bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T} + \hat{\boldsymbol{\Sigma}}) \\ &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{tot} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{tot} \end{bmatrix} \right) \end{aligned}$$

Likelihood Ratio Scores

- Log-likelihood ratio score (assuming i-vectors have been mean subtracted, $\mathbf{x} \leftarrow \mathbf{x} - \mathbf{m}$)

$$
\begin{aligned}
S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t) &= \log \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{Same-speaker})}{p(\mathbf{x}_s, \mathbf{x}_t | \text{Diff-speaker})} \\
&= \log \frac{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{tot} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_{tot} \end{bmatrix} \right)}{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{tot} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{tot} \end{bmatrix} \right)} \\
&= \frac{1}{2}[\mathbf{x}_s^{\mathsf{T}} \mathbf{Q} \mathbf{x}_s + 2\mathbf{x}_s^{\mathsf{T}} \mathbf{P} \mathbf{x}_t + \mathbf{x}_t^{\mathsf{T}} \mathbf{Q} \mathbf{x}_t] + \text{const}
\end{aligned}
\tag{9}
$$

where

$$
\begin{aligned}
\mathbf{Q} &= \boldsymbol{\Sigma}_{tot}^{-1} - (\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1} \\
\mathbf{P} &= \boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac}(\boldsymbol{\Sigma}_{tot} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{tot}^{-1}\boldsymbol{\Sigma}_{ac})^{-1}
\end{aligned}
$$

Likelihood Ratio Scores

- The LLR in Eq. 9 assumes that the SNR of both target-speaker's utterance and test utterance are unknown.
- If both SNRs (ℓ_s, ℓ_t) are known, we may compute the score as follows:

$$S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t) = \log \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{Same-speaker}, \ell_s, \ell_t)}{p(\mathbf{x}_s, \mathbf{x}_t | \text{Diff-speaker}, \ell_s, \ell_t)}$$

Likelihood Ratio Scores

- Given i-vector $\mathbf{x}$ and utterance SNR ℓ, the likelihood of $\mathbf{x}$ is

$$\begin{aligned}
p(\mathbf{x}|\ell) &= \int_{\mathbf{h}}\int_{\mathbf{w}} p(\mathbf{x}|\mathbf{h},\mathbf{w},\ell)p(\mathbf{h},\mathbf{w}|\ell)d\mathbf{h}d\mathbf{w} \\
&= \int_{\mathbf{h}}\int_{\mathbf{w}} p(\mathbf{x}|\mathbf{h},\mathbf{w},\ell)p(\mathbf{h}|\mathbf{w},\ell)p(\mathbf{w}|\ell)d\mathbf{h}d\mathbf{w} \\
&= \int_{\mathbf{h}}\int_{\mathbf{w}} p(\mathbf{x}|\mathbf{h},\mathbf{w},\ell)p(\mathbf{h})d\mathbf{h}p(\mathbf{w}|\ell)d\mathbf{w}
\end{aligned}$$

  where we have assumed that $\mathbf{h}$ is *a priori* independent of $\mathbf{w}$ and ℓ.

- Note that if $\ell \in k$-th SNR group, we have $\mathbf{w} = \mathbf{w}_k^* \equiv \langle \mathbf{w}_k | \mathcal{X}^k \rangle$

$$\begin{aligned}
p(\mathbf{x}|\ell \in k\text{-th SNR group}) &= \int_{\mathbf{h}} p(\mathbf{x}|\mathbf{h},\mathbf{w}_k^*)p(\mathbf{h})d\mathbf{h} \\
&= \int_{\mathbf{h}} \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{V}\mathbf{h}+\mathbf{U}\mathbf{w}_k^*,\boldsymbol{\Sigma})\mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I})d\mathbf{h} \\
&= \mathcal{N}(\mathbf{x}|\mathbf{m}+\mathbf{U}\mathbf{w}_k^*,\mathbf{V}\mathbf{V}^{\mathsf{T}}+\boldsymbol{\Sigma})
\end{aligned}$$

Likelihood Ratio Scores

$$
\begin{aligned}
S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t) &= \log \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{Same-speaker}, \ell_s, \ell_t)}{p(\mathbf{x}_s, \mathbf{x}_t | \text{Diff-speaker}, \ell_s, \ell_t)} \\
&= \log \frac{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} + \mathbf{U}\mathbf{w}^*_{k_s} \\ \mathbf{m} + \mathbf{U}\mathbf{w}^*_{k_t} \end{bmatrix}, \begin{bmatrix} \mathbf{\Psi} & \mathbf{\Sigma}_{ac} \\ \mathbf{\Sigma}_{ac} & \mathbf{\Psi} \end{bmatrix} \right)}{\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} + \mathbf{U}\mathbf{w}^*_{k_s} \\ \mathbf{m} + \mathbf{U}\mathbf{w}^*_{k_t} \end{bmatrix}, \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi} \end{bmatrix} \right)} \\
&= \frac{1}{2}[\bar{\mathbf{x}}_s^{\mathsf{T}} \mathbf{Q} \bar{\mathbf{x}}_s + 2\bar{\mathbf{x}}_s^{\mathsf{T}} \mathbf{P} \bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t^{\mathsf{T}} \mathbf{Q} \bar{\mathbf{x}}_t] + \mathrm{const}
\end{aligned} \tag{10}
$$

where

$$
\begin{aligned}
\bar{\mathbf{x}}_s &= \mathbf{x}_s - \mathbf{m} - \mathbf{U}\mathbf{w}^*_{k_s} \\
\bar{\mathbf{x}}_t &= \mathbf{x}_t - \mathbf{m} - \mathbf{U}\mathbf{w}^*_{k_t} \\
\mathbf{Q} &= \mathbf{\Psi}^{-1} - (\mathbf{\Psi} - \mathbf{\Sigma}_{ac}\mathbf{\Psi}^{-1}\mathbf{\Sigma}_{ac})^{-1} \\
\mathbf{P} &= \mathbf{\Psi}^{-1}\mathbf{\Sigma}_{ac}(\mathbf{\Psi} - \mathbf{\Sigma}_{ac}\mathbf{\Psi}^{-1}\mathbf{\Sigma}_{ac})^{-1} \\
\mathbf{\Psi} &= \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{\Sigma}; \quad \mathbf{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^{\mathsf{T}}
\end{aligned}
$$

Compare with scoring in JFA

- Scoring in JFA is based on the **sequential mode** [Kenny et al., 2007b]:

$$S_{\text{JFA-LR}}(\mathcal{O}_s, \mathcal{O}_t) = \frac{P_{\Lambda(s)}(\mathcal{O}_t)}{P_{\Lambda}(\mathcal{O}_t)}$$

  where $\Lambda(s)$ denotes the adapted speaker model based on enrollment speech $\mathcal{O}_s$ from speaker s.
- Computing $P_{\Lambda(s)}(\mathcal{O}_t)$ requires the posterior density of speaker factors [$\mathbf{y}(s)$ and $\mathbf{z}(s)$ in Kenny 2007], which are posteriorly correlated.
- The scoring function in Eq. 9 is based on the **batch mode**.
- Batch mode is similar to speaker comparison in which no model adaptation is performed. So, the posterior correlation between speaker factors and SNR factors does not occur in Eq. 9.

Scoring based on sequential mode

- The batch-mode scoring (Eq. 10) requires inverting a big matrix if the target speaker has a large number of enrollment utterances.
- The sequential-mode scoring can mitigate this problem.
- For notational simplicity, we assume that the target speaker only have one enrollment utterance with i-vector $\mathbf{x}_s$:

$$\begin{aligned}
S_{\mathrm{LR}}(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t) &= \frac{p(\mathbf{x}_s, \mathbf{x}_t | \ell_s, \ell_t)}{p(\mathbf{x}_s|\ell_s)p(\mathbf{x}_t|\ell_t)} \\
&= \frac{p(\mathbf{x}_t|\mathbf{x}_s, \ell_s, \ell_t)p(\mathbf{x}_s|\ell_s)}{p(\mathbf{x}_s|\ell_s)p(\mathbf{x}_t|\ell_t)} \\
&= \frac{p(\mathbf{x}_t|\mathbf{x}_s, \ell_s, \ell_t)}{p(\mathbf{x}_t|\ell_t)}
\end{aligned}$$

Scoring based on sequential mode

- For simplicity, we omit ℓ_s and ℓ_t from now on.

$$p(\mathbf{x}_t|\mathbf{x}_s) = \int\int p(\mathbf{x}_t|\mathbf{h},\mathbf{w})p(\mathbf{h},\mathbf{w}|\mathbf{x}_s)d\mathbf{h}d\mathbf{w}$$

- As $\mathbf{h}$ and $\mathbf{w}$ are posteriorly dependent, we use variational Bayes to approximate the joint posterior:

$$p(\mathbf{x}_t|\mathbf{x}_s) \approx \int\int p(\mathbf{x}_t|\mathbf{h},\mathbf{w})q(\mathbf{h})q(\mathbf{w})d\mathbf{h}d\mathbf{w}$$
$$= \int_{\mathbf{h}}\int_{\mathbf{w}} \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\mathbf{h}+\mathbf{U}\mathbf{w},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_{\mathbf{h}_s},\boldsymbol{\Sigma}_{\mathbf{h}_s})\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}_s},\boldsymbol{\Sigma}_{\mathbf{w}_s})d\mathbf{h}d\mathbf{w} \quad (11)$$

where $\boldsymbol{\mu}_{\mathbf{h}_s}$, $\boldsymbol{\Sigma}_{\mathbf{h}_s}$, $\boldsymbol{\mu}_{\mathbf{w}_s}$, and $\boldsymbol{\Sigma}_{\mathbf{w}_s}$ are posterior means and posterior covariances.

Scoring based on sequential mode

- Eq. 11 is a convolution of Gaussians

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{x}_s) &\approx \int_{\mathbf{w}}\int_{\mathbf{h}} \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\mathbf{h}+\mathbf{U}\mathbf{w},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_{\mathbf{h}_s},\boldsymbol{\Sigma}_{\mathbf{h}_s})d\mathbf{h}\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}_s},\boldsymbol{\Sigma}_{\mathbf{w}_s})d\mathbf{w} \\
&= \int_{\mathbf{w}} \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\boldsymbol{\mu}_{\mathbf{h}_s}+\mathbf{U}\mathbf{w},\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{h}_s}\mathbf{V}^{\mathsf{T}}+\boldsymbol{\Sigma})\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}_s},\boldsymbol{\Sigma}_{\mathbf{w}_s})d\mathbf{w} \\
&= \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\boldsymbol{\mu}_{\mathbf{h}_s}+\mathbf{U}\boldsymbol{\mu}_{\mathbf{w}_s},\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{h}_s}\mathbf{V}^{\mathsf{T}}+\mathbf{U}\boldsymbol{\Sigma}_{\mathbf{w}_s}\mathbf{U}^{\mathsf{T}}+\boldsymbol{\Sigma})
\end{aligned}
$$

- If ℓ_s falls on the k-th SNR group, we may replace $\boldsymbol{\mu}_{\mathbf{w}_s}$ by $\mathbf{w}_k^* \equiv \langle \mathbf{w}_k|\mathcal{X}^k\rangle$ and assume that $\boldsymbol{\Sigma}_{\mathbf{w}_k^*} \to \mathbf{0}$:

$$
p(\mathbf{x}_t|\mathbf{x}_s) = \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\boldsymbol{\mu}_{\mathbf{h}_s}+\mathbf{U}\mathbf{w}_k^*,\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{h}_s}\mathbf{V}^{\mathsf{T}}+\boldsymbol{\Sigma})
$$

- $p(\mathbf{x}_t)$ is a marginal density

$$
\begin{aligned}
p(\mathbf{x}_t) &= \int p(\mathbf{x}_t|\mathbf{h},\mathbf{w})p(\mathbf{h},\mathbf{w})d\mathbf{h}d\mathbf{w} \\
&= \int \mathcal{N}(\mathbf{x}_t|\mathbf{m}+\mathbf{V}\mathbf{h}+\mathbf{U}\mathbf{w},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{I})d\mathbf{h}d\mathbf{w} \\
&= \mathcal{N}(\mathbf{x}_t|\mathbf{m},\mathbf{V}\mathbf{V}^{\mathsf{T}}+\mathbf{U}\mathbf{U}^{\mathsf{T}}+\boldsymbol{\Sigma})
\end{aligned}
$$

Scoring based on sequential mode

- The posteriors means and covariances can be obtained from Eq. 6 and Eq. 7 by considering a single utterance from target-speaker s:

$$\boldsymbol{\mu}_{\mathbf{h}_s} = \langle \mathbf{h}_s | \mathbf{x}_s \rangle = \boldsymbol{\Sigma}_{\mathbf{h}_s} \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_s - \mathbf{m} - \mathbf{U} \boldsymbol{\mu}_{\mathbf{w}_s})$$
$$\boldsymbol{\mu}_{\mathbf{w}_s} = \langle \mathbf{w}_s | \mathbf{x}_s \rangle = \boldsymbol{\Sigma}_{\mathbf{w}_s} \mathbf{U}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_s - \mathbf{m} - \mathbf{V} \boldsymbol{\mu}_{\mathbf{h}_s})$$
$$\boldsymbol{\Sigma}_{\mathbf{h}_s} = (\mathbf{I} + \mathbf{V}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{V})^{-1}$$
$$\boldsymbol{\Sigma}_{\mathbf{w}_s} = (\mathbf{I} + \mathbf{U}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{U})^{-1}$$

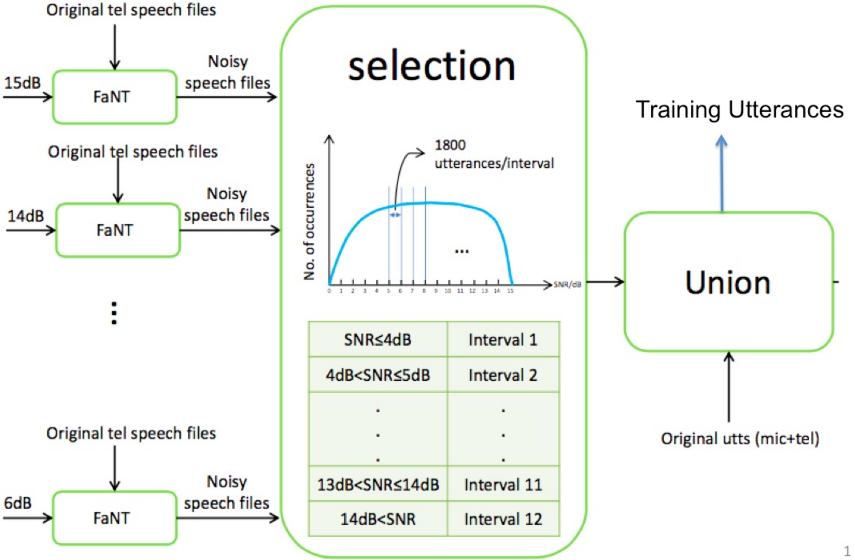
- Note that $\boldsymbol{\mu}_{\mathbf{h}_s}$ and $\boldsymbol{\mu}_{\mathbf{w}_s}$ depend on each other, meaning that they should be found iteratively.

Experiments on SRE12

- **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 $\Longrightarrow$ 60-Dim acoustic vectors
- **UBM**: Gender-dependent, mic+tel, 1024 mixtures
- **Total Variability Matrix**: Gender-dependent, mic+tel, 500 total factors
- **I-Vector Preprocessing**: Whitening by WCCN then length normalization followed by non-parametric feature analysis (NFA)[5] and WCCN (500-dim $\rightarrow$ 200-dim)

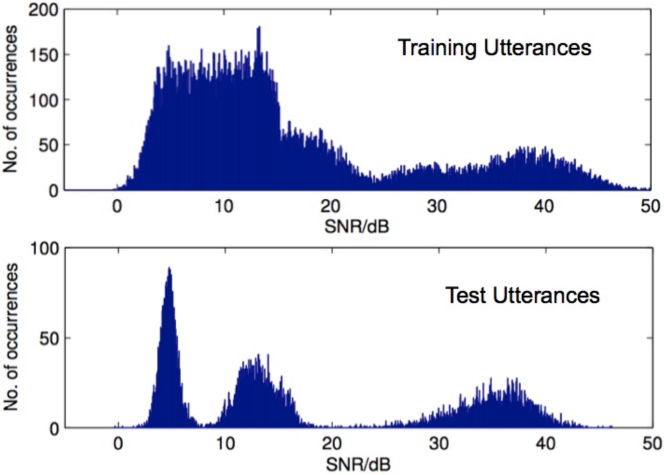
[5] Z. Li, D. Lin, and X. Tang, "*Nonparametric discriminant analysis for face recognition*," IEEE Trans. on PAMI, 2009.

Prepare training speech files



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 $\mathbf{w}_k$)

Performance on SRE12

Mixture of PLDA (*Mak, Interspeech14*)

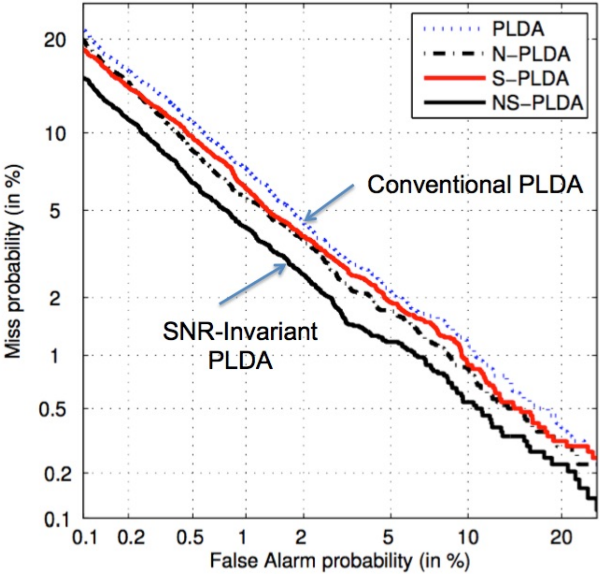
CC4

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 $\mathbf{w}_k$)

Performance on SRE12

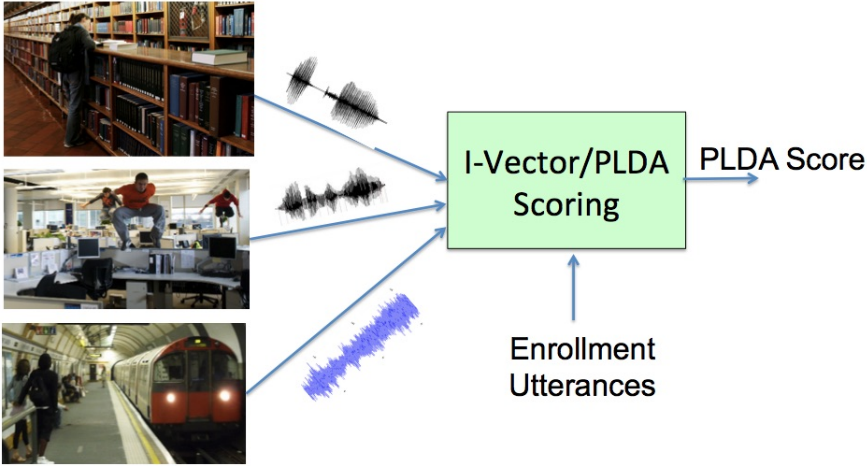


CC4, Female

Outline

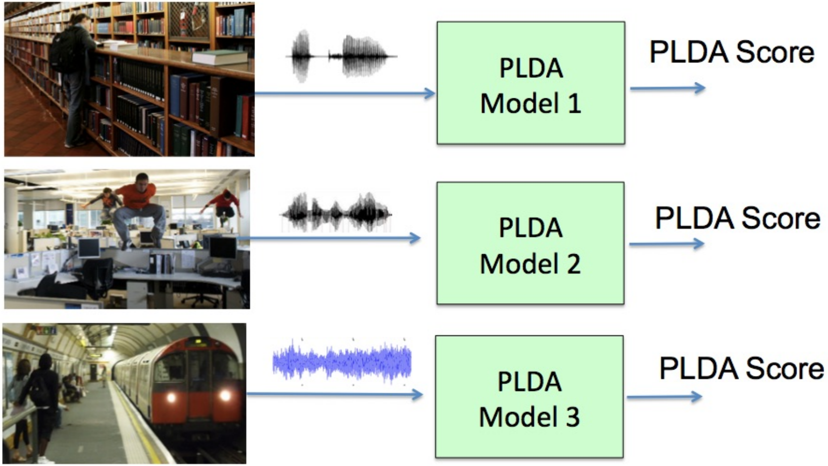
Mixture of PLDA: Motivation

- Conventional i-vector/PLDA systems use a single PLDA model to handle all SNR conditions

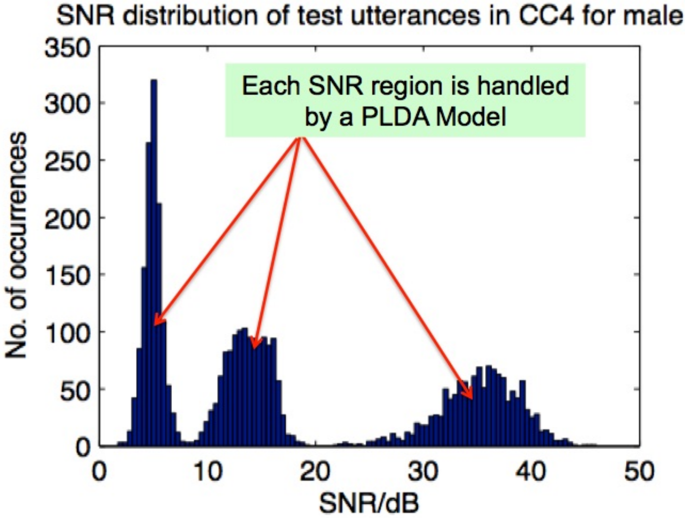


Mixture of PLDA: Motivation

- We argue that a PLDA model should focus on a small range of SNR.

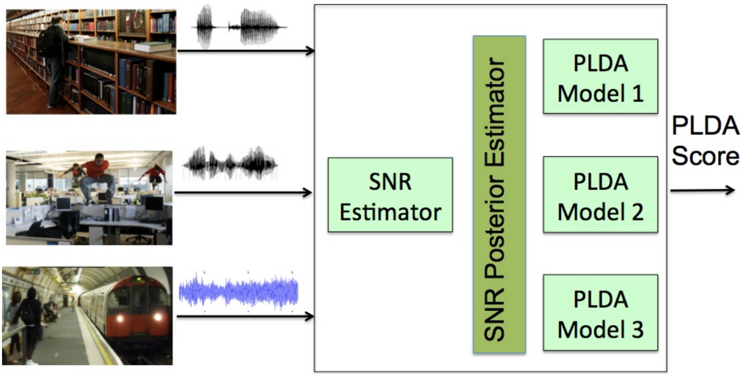


Distribution of SNR



Proposed solution

- The full spectrum of SNRs is handled by a mixture of PLDA in which the posteriors of the indicator variables depend on the utterance's SNR.
- Verification scores depend not only on the same-speaker and different-speaker likelihoods but also on the posterior probabilities of SNR.



Mixture of PLDA [Mak et al., 2016]

- Model parameters:

$$\begin{aligned}\boldsymbol{\theta} &= \{\underline{\pi}, \underline{\mu}, \underline{\sigma}, \underline{\mathbf{m}}, \underline{\mathbf{V}}, \underline{\boldsymbol{\Sigma}}\} \\ &= \{\underbrace{\pi_k, \mu_k, \sigma_k}_{\text{Modeling SNR}}, \underbrace{\mathbf{m}_k, \mathbf{V}_k, \boldsymbol{\Sigma}_k}_{\text{Speaker subspaces}}\}_{k=1}^{K}\end{aligned}$$

- Generative model:

$$\mathbf{x}_{ij} \sim \sum_{k=1}^{K} P(y_{ijk} = 1|\ell_{ij})\mathcal{N}(\mathbf{x}_{ij}|\mathbf{m}_k, \mathbf{V}_k\mathbf{V}_k^{\mathsf{T}} + \boldsymbol{\Sigma}_k)$$

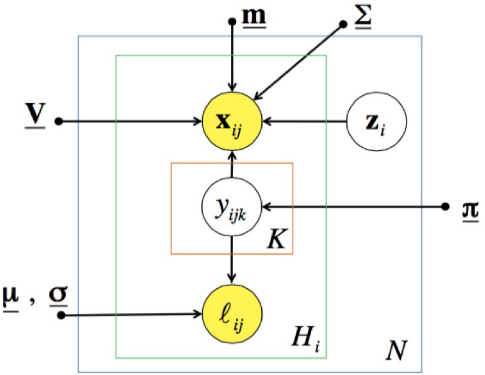
where

$$P(y_{ijk} = 1|\ell_{ij}) = \frac{\pi_k \mathcal{N}(\ell_{ij}|\mu_k, \sigma_k^2)}{\sum_{k'=1}^{K} \pi_{k'}\mathcal{N}(\ell_{ij}|\mu_{k'}, \sigma_{k'}^2))}$$

and ℓ_{ij} is the SNR of the utterance j from speaker i.

Mixture of PLDA

- Graphical model:



$\mathbf{x}_{ij}$: i-vector of the j-th utterance from the i-th speaker

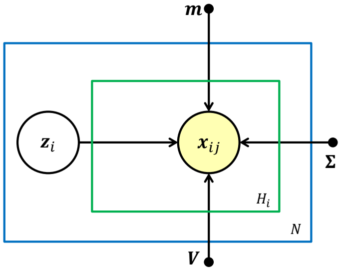
ℓ_{ij} : SNR of the j-th utterance from the i-th speaker

$$\underline{\boldsymbol{\pi}} = \{\boldsymbol{\pi}_k\}_{k=1}^{K}$$

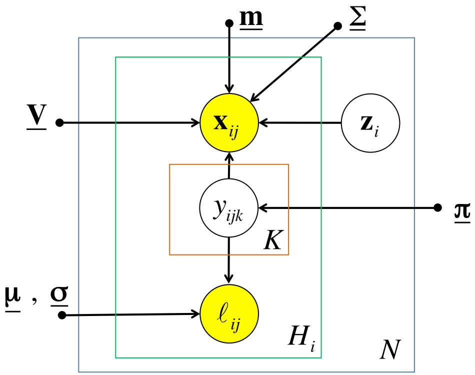
$$\underline{\mathbf{V}} = \{\mathbf{V}_k\}_{k=1}^{K}$$

$$\boldsymbol{\theta} = \{\underbrace{\pi_k, \mu_k, \sigma_k}_{\text{For modeling SNR of utts.}}, \underbrace{\mathbf{m}_k, \mathbf{V}_k, \boldsymbol{\Sigma}_k}_{\text{For modeling SNR-dependent i-vectors}}\}_{k=1}^{K}$$

- Graphical models:



PLDA



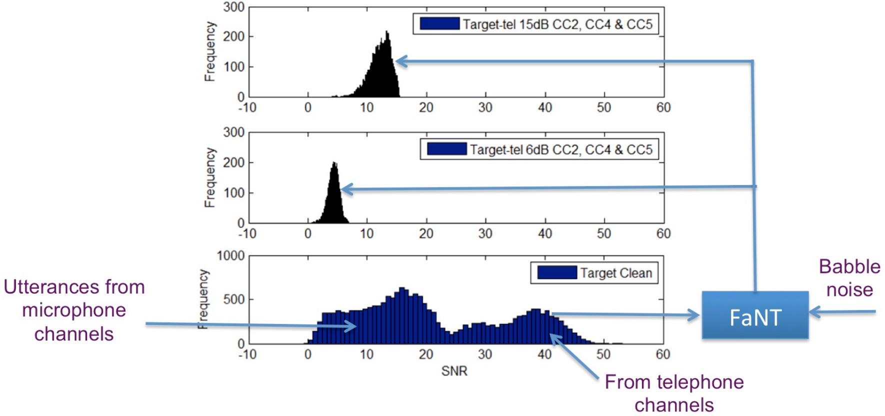
Mixture of PLDA

Experiments on SRE12

- **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 $\Longrightarrow$ 60-Dim acoustic vectors
- **UBM**: Gender-dependent, mic+tel, 1024 mixtures
- **Total Variability Matrix**: Gender-dependent, mic+tel, 500 total factors
- **I-Vector Preprocessing**: Whitening by WCCN then length normalization followed by LDA and WCCN (500-dim $\rightarrow$ 200-dim)

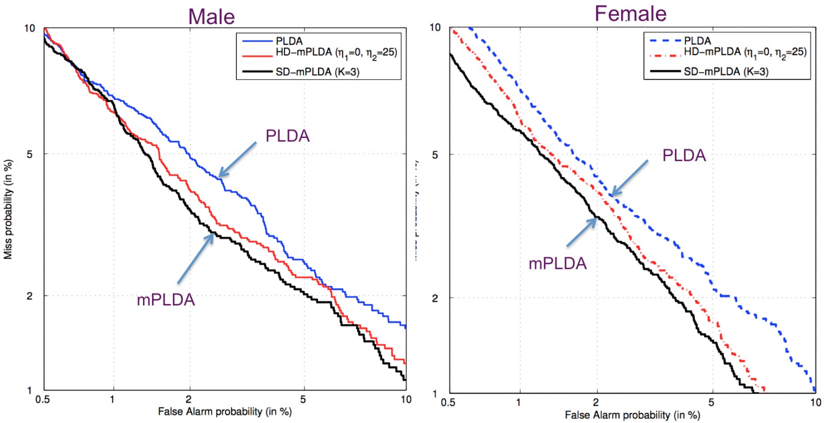
Experiments on SRE12

- In NIST 2012 SRE, training utterances from telephone channels are clean, but some of the test utterances are noisy.
- We used the FaNT tool to add babble noise to the clean training utterances

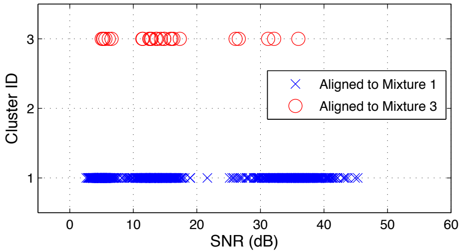


DET Performance

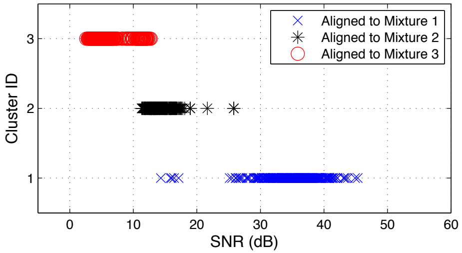
- Train on tel+mic speech and test on noisy tel speech (CC4).



I-vector Cluster Alignment



Without using SNR

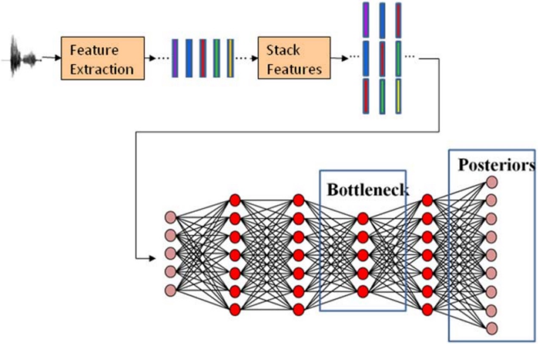


With SNR as guidance

Outline

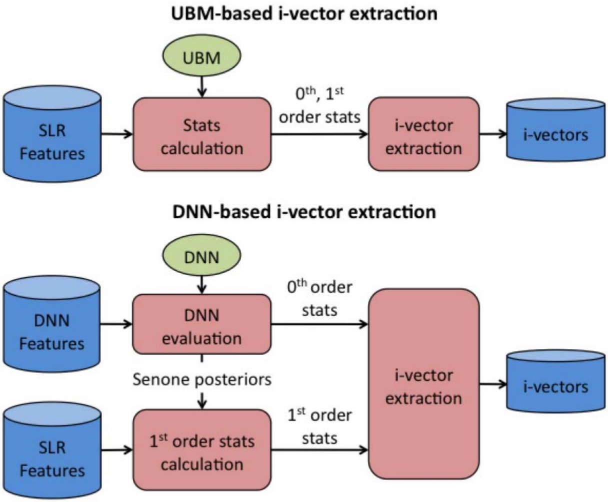
DNN for speaker recognition

- **Main idea**: replacing the universal background model (UBM) with a phonetically-aware DNN for computing the frame posterior probabilities.
- The most successful application of DNN to speaker recognition [Lei et al., 2014, Ferrer et al., 2016, Richardson et al., 2015]



Source: Richardson, F., et al. *IEEE Signal Processing Letters*, 2015

DNN I-vector extraction



Source: Ferrer, L. et al. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

UBM I-vectors

- Factor analysis model for UBM i-vectors:

$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_c^{(b)} + \mathbf{T}_c \mathbf{w} \qquad\qquad c = 1, \ldots, C$$

- Given the MFCC vectors of an utterance $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, its i-vector is the posterior mean of $\mathbf{w}$

$$\mathbf{x} \equiv \langle \mathbf{w} | \mathcal{O} \rangle = \mathbf{L}^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\boldsymbol{\Sigma}_c^{(b)} \right)^{-1} \sum_{t=1}^{T} \gamma_c(\mathbf{o}_t)(\mathbf{o}_t - \boldsymbol{\mu}_c^{(b)})$$

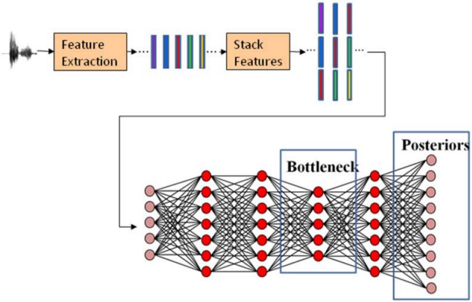
where

$$\mathbf{L} = \mathbf{I} + \sum_{c=1}^{C} \sum_{t=1}^{T} \gamma_c(\mathbf{o}_t) \mathbf{T}_c^{\mathsf{T}} (\boldsymbol{\Sigma}_c^{(b)})^{-1} \mathbf{T}_c$$

$$\gamma_c(\mathbf{o}_t) \equiv \Pr(\text{Mixture} = c | \mathbf{o}_t) = \frac{\lambda_c^{(b)} \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_c^{(b)}, \boldsymbol{\Sigma}_c^{(b)})}{\sum_{j=1}^{C} \lambda_j^{(b)} \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_j^{(b)}, \boldsymbol{\Sigma}_j^{(b)})}$$

DNN I-vectors

- Replace $\gamma_c(\mathbf{o}_t)$ by DNN output, $\gamma_c^{\mathrm{DNN}}(\mathbf{a}_t)$
- The DNN is trained to produce posterior probabilities of senones, given multiple frames of acoustic features, $\mathbf{a}_t$, as input.



- Acoustic features for speech recognition in the DNN are not necessary the same as the features for the i-vector extractor.

DNN I-vectors

- Given MFCC or bottleneck (BN) feature vectors $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, the DNN i-vector is[6]

$$\mathbf{x} \equiv \langle \mathbf{w} | \mathcal{O} \rangle = \mathbf{L}^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\boldsymbol{\Sigma}_c^{\mathrm{DNN}} \right)^{-1} \sum_{t=1}^{T} \gamma_c^{\mathrm{DNN}}(\mathbf{a}_t)(\mathbf{o}_t - \boldsymbol{\mu}_c^{\mathrm{DNN}})$$

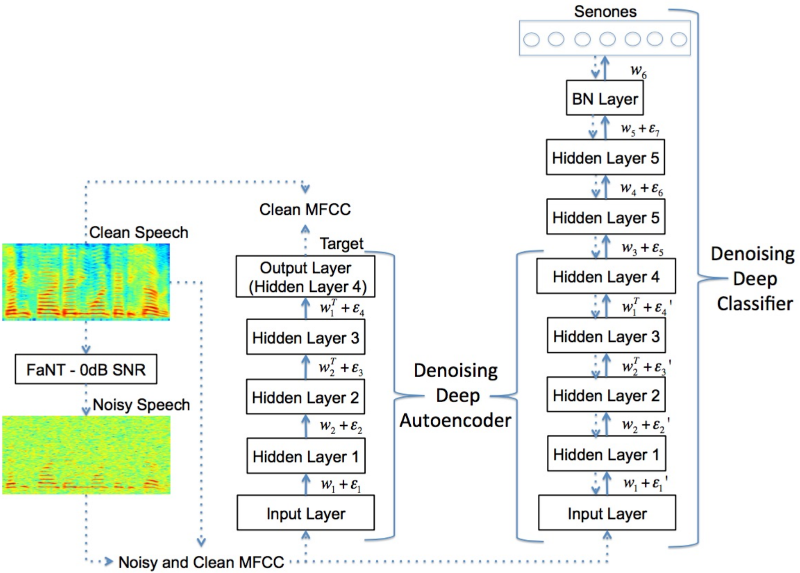
where

$$\boldsymbol{\mu}_c^{\mathrm{DNN}} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_c^{\mathrm{DNN}}(\mathbf{a}_{it}) \mathbf{o}_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_c^{\mathrm{DNN}}(\mathbf{a}_{it})}$$

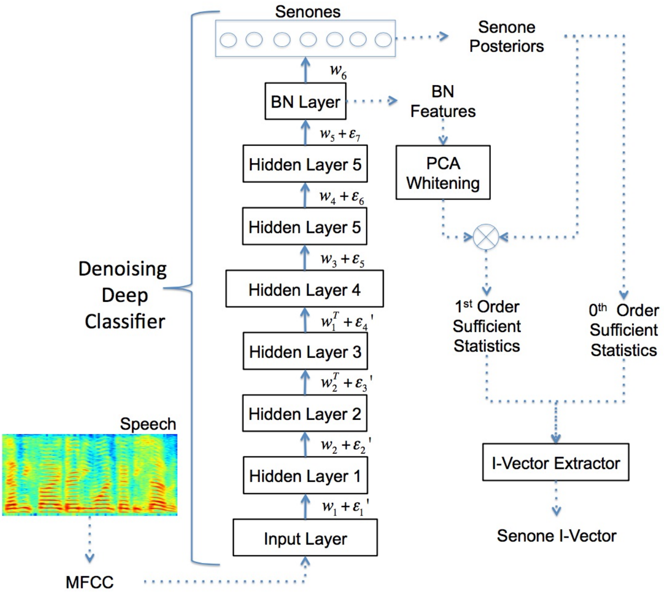
$$\boldsymbol{\Sigma}_c^{\mathrm{DNN}} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_c^{\mathrm{DNN}}(\mathbf{a}_{it}) \mathbf{o}_{it} \mathbf{o}_{it}^{\mathsf{T}}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_c^{\mathrm{DNN}}(\mathbf{a}_{it})} - \boldsymbol{\mu}_c^{\mathrm{DNN}} (\boldsymbol{\mu}_c^{\mathrm{DNN}})^{\mathsf{T}}$$

[6] For a full set of formulae, see http://www.eie.polyu.edu.hk/∼mwmak/papers/FA-Ivector.pdf

Denoising deep classifier [Tan et al., 2016]



DNN I-vectors from denoising deep classifier



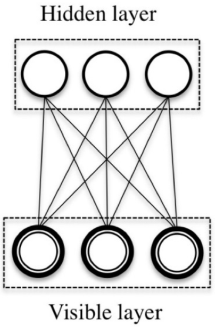
Performance on NIST 2012 SRE

- Performance on CC4 with test utterances contaminated with babble noise.

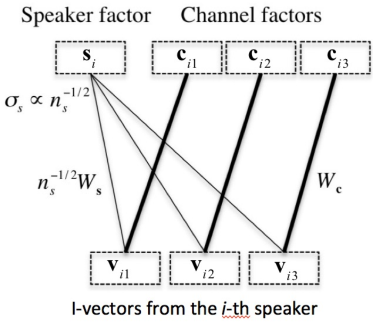
		15dB		6dB		0dB	
Acoustic Features	Posteriors from	EER	minDCF	EER	minDCF	EER	minDCF
MFCC	GMM (1024 mixtures)	3.366	0.322	3.243	0.353	5.353	0.631
MFCC	GMM (2048 mixtures)	4.215	0.352	3.819	0.379	5.332	0.646
BN Features	GMM (1024 mixtures)	3.269	0.263	3.493	0.368	4.608	0.551
BN Features	DNN (2000 senones)	**2.448**	**0.236**	**2.774**	**0.311**	**4.503**	**0.544**

Outline

- **Main idea**:
  - Use i-vectors as input to the Gaussian visible layer of an RBM
  - Divide RBM weights into two parts: speaker and channel
  - Consider RBM weights as analogue to PLDA's loading matrices
  - Divide the Gaussian hidden layer into two parts: speaker and channel
  - Hidden nodes are considered as latent factors



RBM



RBM–PLDA

PLDA vs. PLDA–RBM

- **PLDA (omitting global mean)**:

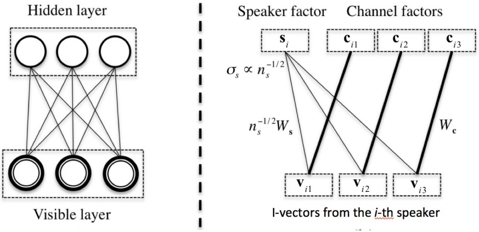
$$\mathbf{v} = \mathbf{V}\mathbf{s} + \mathbf{U}\mathbf{c} + \epsilon$$

where $\mathbf{v}$ is an i-vector, $\mathbf{s}$ and $\mathbf{c}$ are speaker and channel factors.

- **RBM-PLDA**:

$$\mathbf{v}_n = \sigma_v \left[\mathbf{W}_s \frac{\mathbf{s}_{st}}{\sigma_s} + \mathbf{W}_c \frac{\mathbf{c}_{st}}{\sigma_c} \right]$$

where $\mathbf{v}_n$ is the expected value of visible layer in the negative phase of CD-1 sampling, $\mathbf{s}_{st}$ and $\mathbf{c}_{st}$ are the states of Gaussian hidden nodes of the RBM.



Scoring in PLDA–RBM

- Given two i-vectors $\mathbf{v}_i, i = 1, 2$, compute $\mathbf{s}_i = \mathbf{W}_s^\mathsf{T} \mathbf{v}_i$.
- The log-likelihood ratio is

$$LLR = -\frac{1}{2}(\mathbf{s}_1 - \mathbf{s}_2)^\mathsf{T}(\mathbf{s}_1 - \mathbf{s}_2) + \text{const}$$

- If $\|\mathbf{s}_i\| = 1$, the model is similar to cosine distance scoring.

Results on NIST 2010 SRE

- NIST'10, female, core-extended: