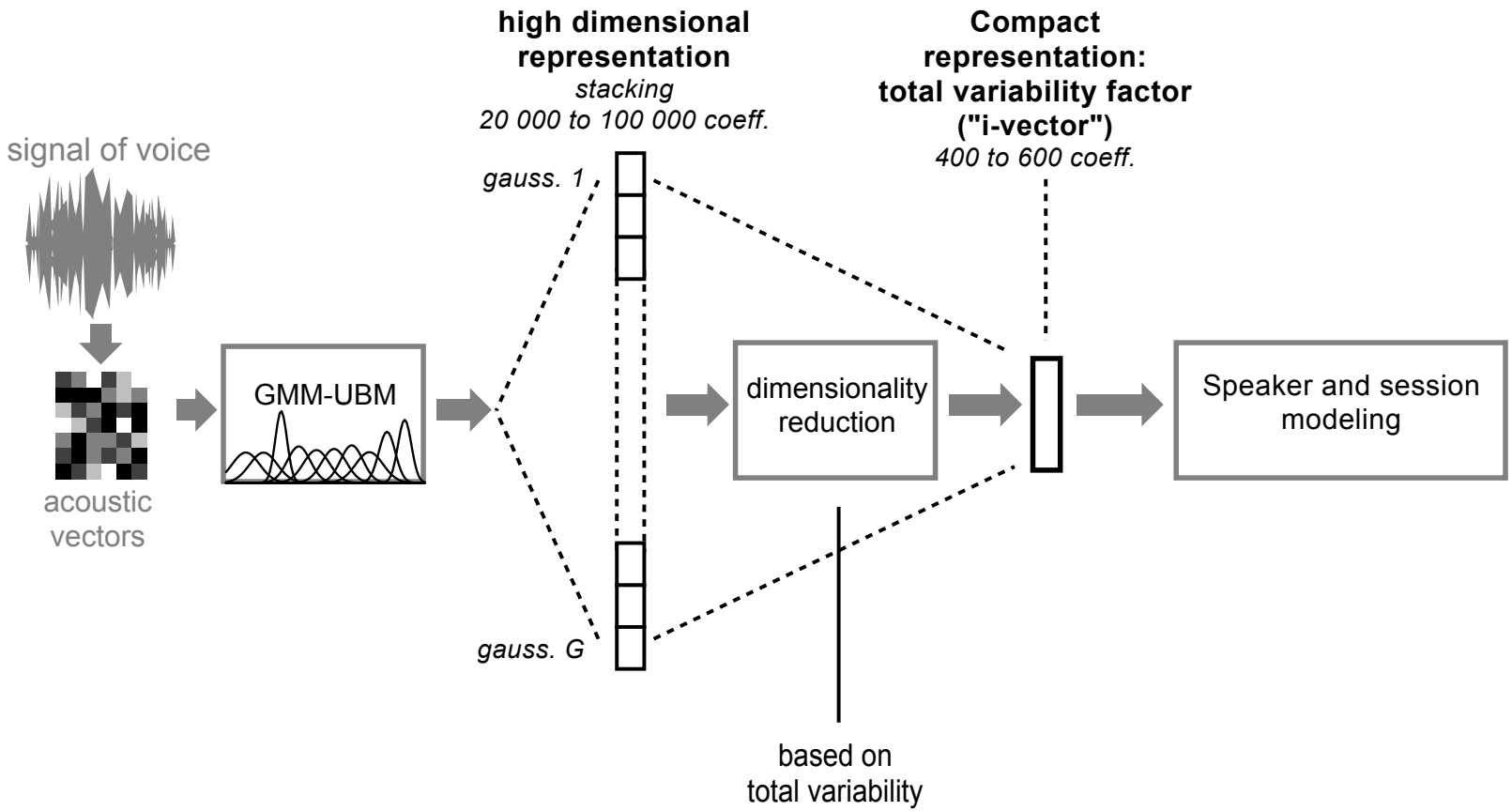
On the importance of pre- and post-conditioning procedures for speaker recognition systems based on total variability factors

P.M. Bousquet, J.F. Bonastre

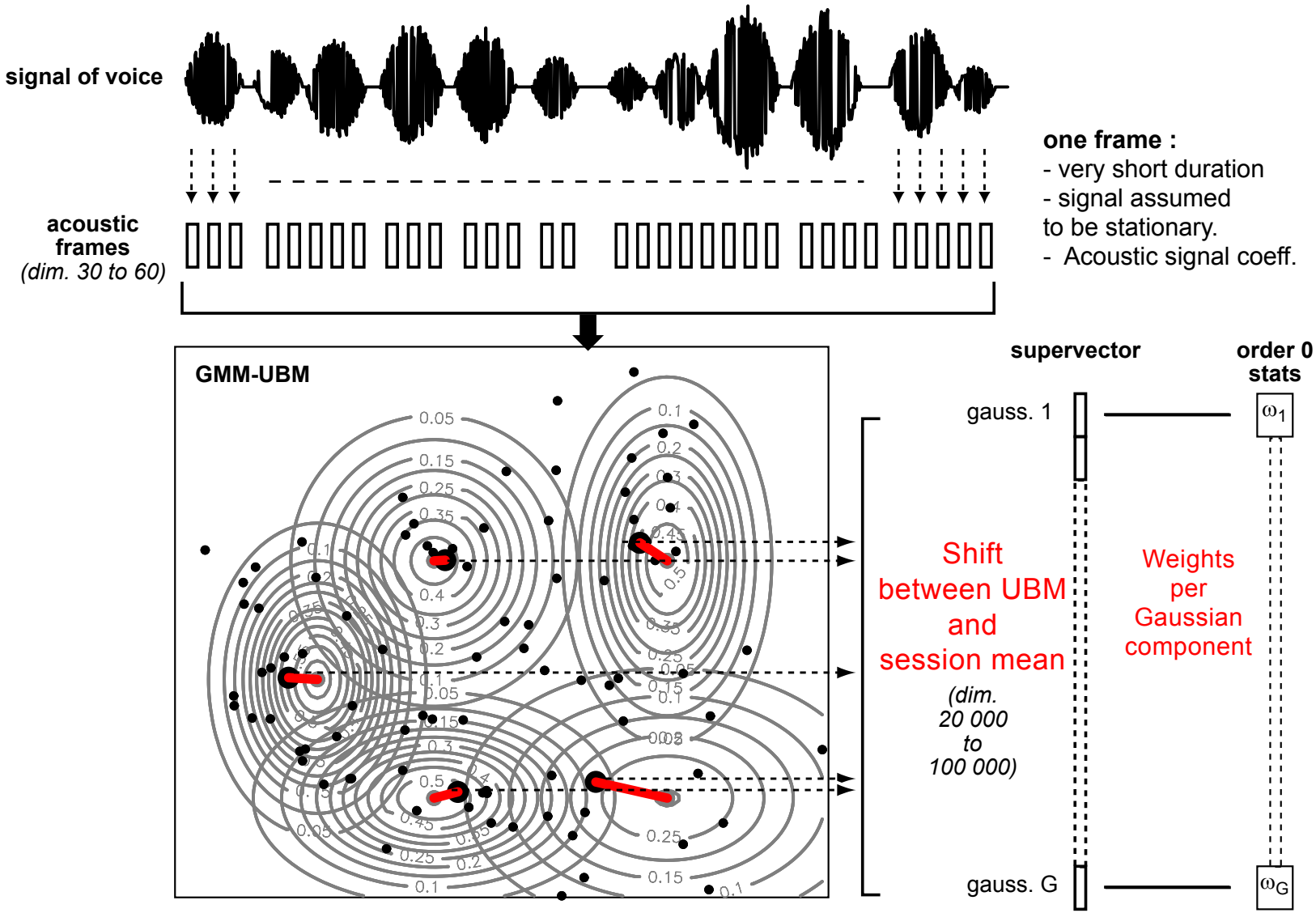
LIA
University of Avignon

The different steps of the i-vector solution



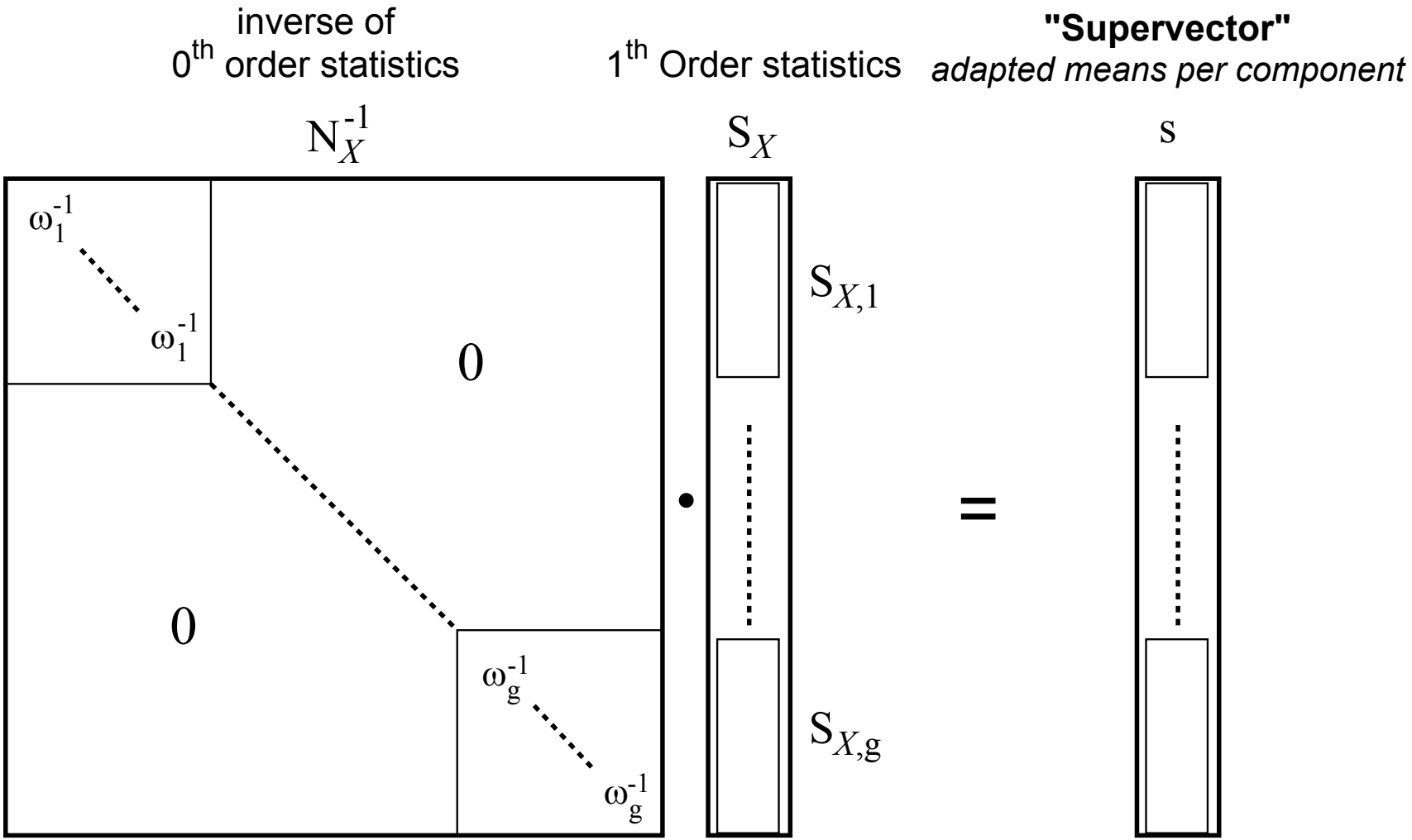
GMM-UBM high-dim. representation

i. With 0th and 1th order statistics



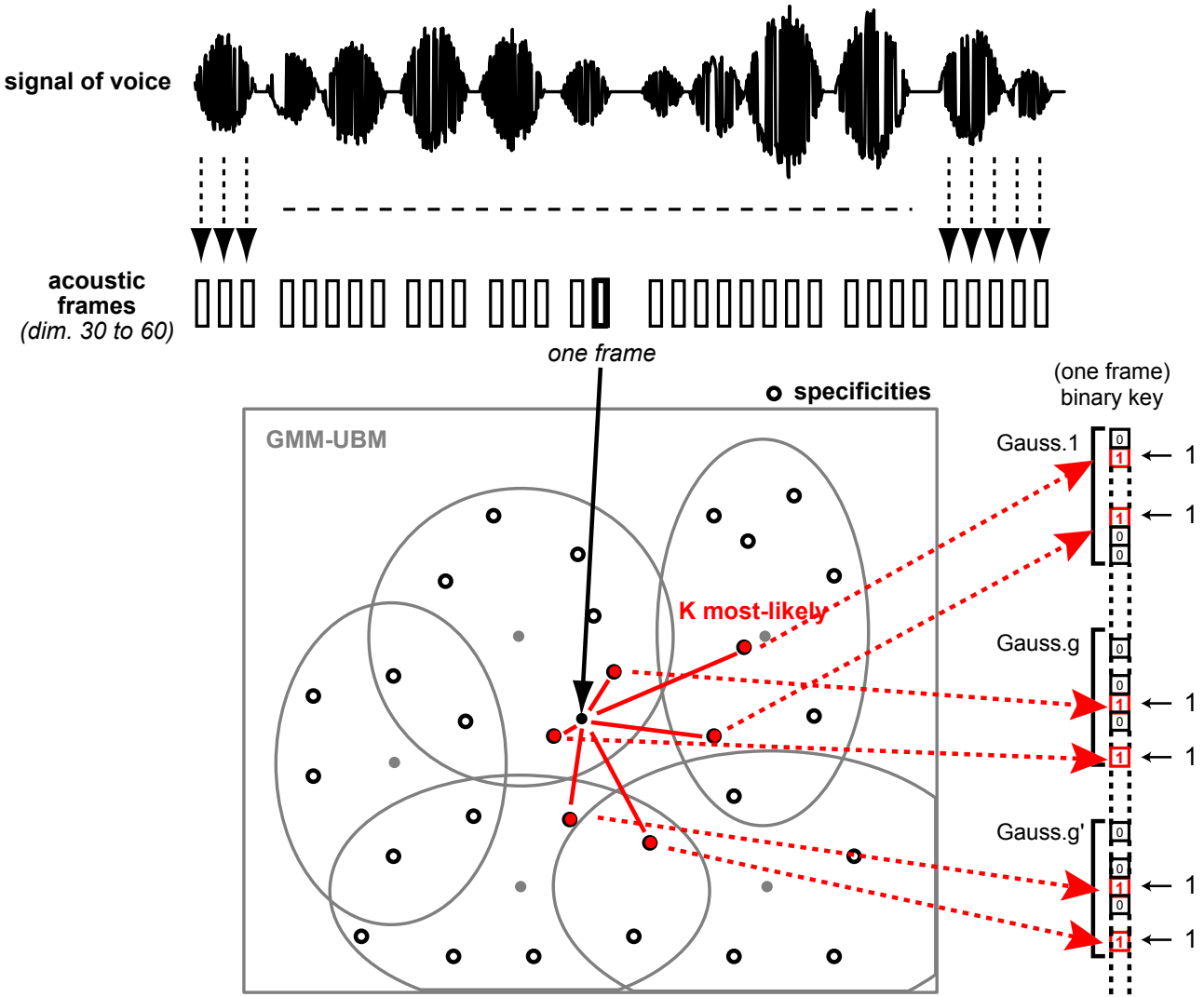
GMM-UBM high-dim. representation

i. With 0th and 1th order statistics



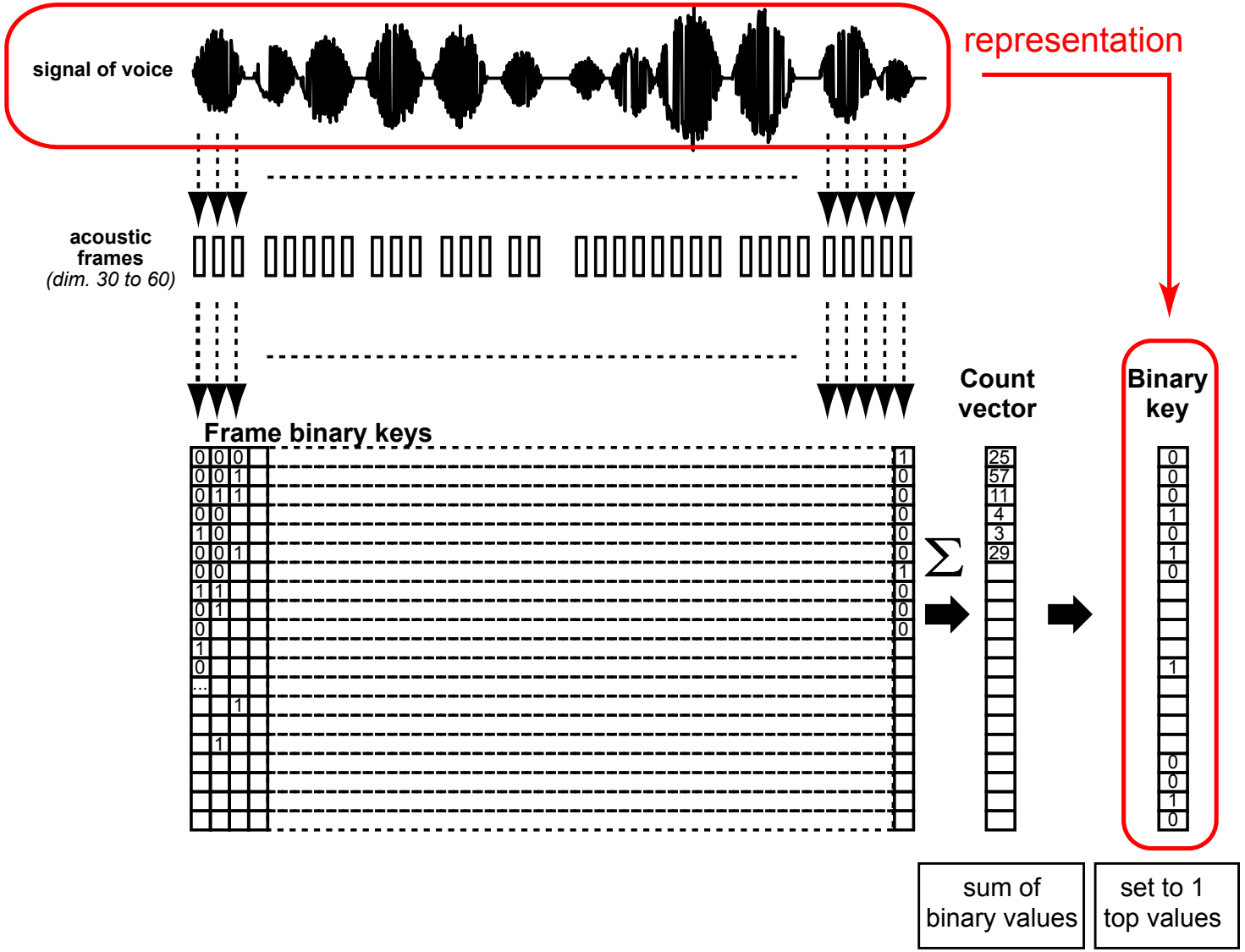
GMM-UBM high-dim. representation

ii. With speaker binary keys



GMM-UBM high-dim. representation

ii. With speaker binary keys



Dimensionality reduction (I-vector extractor)

i. With 0th and 1th order statistics

- A sequence $\mathcal{X}$ of acoustic vectors $\rightarrow$ supervector $\mathbf{s}$

Factor analysis : a probabilistic dimensionality reduction technique

- Supervector $\mathbf{s}$: dimension GF, where G number of GMM components and F feature dimension: **20 000 to 100 000 coefficients**.
- Factor Analysis Total Variability:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

- Factor $\mathbf{w}$ (**i-vector**), dimension $p << GF$: **400 to 600 coefficients**.

- $\mathbf{m}$ is the mean super-vector of the Universal Background Model (UBM),
- $\mathbf{T}$ is the low-rank variability matrix $GF \times p$ ($p << GF$)
- Low-dimensional factor $\mathbf{w}$ is assumed to have a standard normal prior distribution $\mathcal{N}(0, \mathbf{I})$.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing.

Dimensionality reduction (I-vector extractor)

i. With 0th and 1th order statistics

Probabilistic determination of $\mathbf{T}$

by using EM-ML (*Expectation-Maximization - Maximum Likelihood*) procedure.

$$\mathbf{w} = \left(\mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}N_{\mathcal{X}}\mathbf{T}\right)^{-1}\mathbf{T}^t\boldsymbol{\Sigma}^{-1}N_{\mathcal{X}}\left(\mathbf{s}-\mathbf{m}\right) \quad (2)$$

where

- $\boldsymbol{\Sigma}$ is the UBM diagonal covariance matrix,
- $N_{\mathcal{X}}$ is the $GF \times GF$ diagonal matrix composed of F blocks of $N_{\mathcal{X}}^{(g)}\mathbf{I}$ $(g = 1, ..., G)$ where $N_{\mathcal{X}}^{(g)}$ are the zero-order statistics estimated on the g-th Gaussian component of the UBM observing the set of feature vectors in the sequence $\mathcal{X}$.

Dimensionality reduction (I-vector extractor)
ii. With speaker binary keys

- Specificities: categorical variables (selected $\rightarrow$ 1, otherwise $\rightarrow$ 0).
  - *Multiple Correspondence Analysis* (MCA): equivalent-to-PCA dimensionality reduction technique for categorical variables.
  - Special case of binary variables (all variables have only two levels): MCA is equivalent to PCA on the binary coded vectors. Eigenvectors provided by both techniques are identical.
- Binary key $\rightarrow$ dimensionality reduction $\rightarrow$ "i-vector".

Speaker and session modeling
Gaussian-PLDA

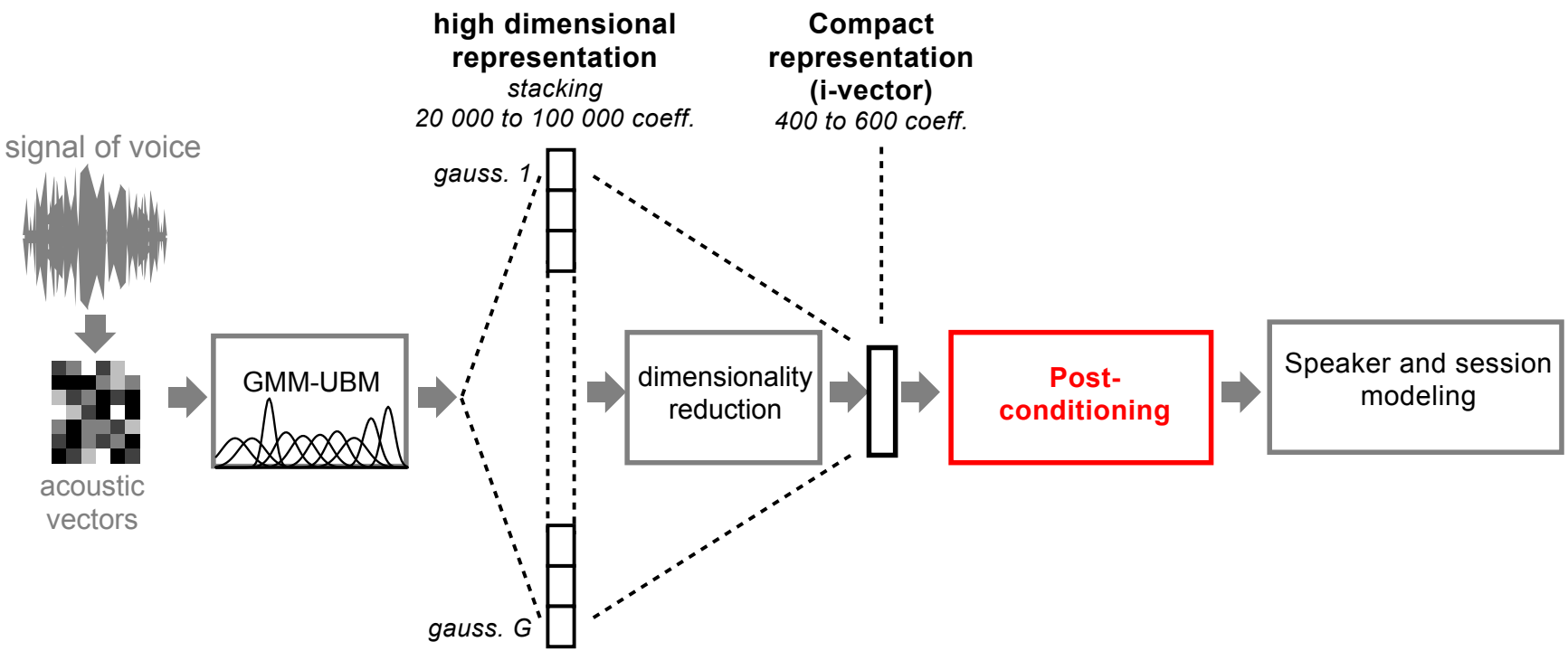
Given an i-vector $\mathbf{w}$

$$\mathbf{w} = \mu + \mathbf{\Phi}\mathbf{y}_s + \varepsilon \qquad \begin{cases} \mathbf{y}_s \sim \mathcal{N}(0, \mathbf{I}_r) \text{ where } r \leqslant p \text{ and } \mathbf{\Phi} \text{ is a } p \times r \text{ matrix} \\ \varepsilon \sim \mathcal{N}(0, \mathbf{\Lambda}) \text{ where } \mathbf{\Lambda} \text{ is a } p \times p \text{ matrix (full)} \\ \mathbf{y}_s \text{ and } \varepsilon \text{ statistically independent} \end{cases}$$

- Decision

$$\begin{aligned} score(\mathbf{w}_1, \mathbf{w}_2) &= \log \frac{P(\mathbf{w}_1, \mathbf{w}_2 | H0 : \text{same speaker})}{P(\mathbf{w}_1, \mathbf{w}_2 | H1 : \text{not the same})} \\ &= \log \frac{\mathcal{N}\left(\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \mid \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Lambda} & \mathbf{\Phi}\mathbf{\Phi}^t \\ \mathbf{\Phi}\mathbf{\Phi}^t & \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Lambda} \end{bmatrix} \right)}{\mathcal{N}\left(\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \mid \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Lambda} \end{bmatrix} \right)} \end{aligned} \tag{3}$$

Post-conditioning



Post-conditioning

$$\mathbf{w} \leftarrow \frac{\mathbf{A}^{-\frac{1}{2}}(\mathbf{w}-\mu)}{\left\|\mathbf{A}^{-\frac{1}{2}}(\mathbf{w}-\mu)\right\|} \tag{4}$$

where μ and $\mathbf{A}$ denote the mean vector and a covariance matrix of a given i-vector corpus.

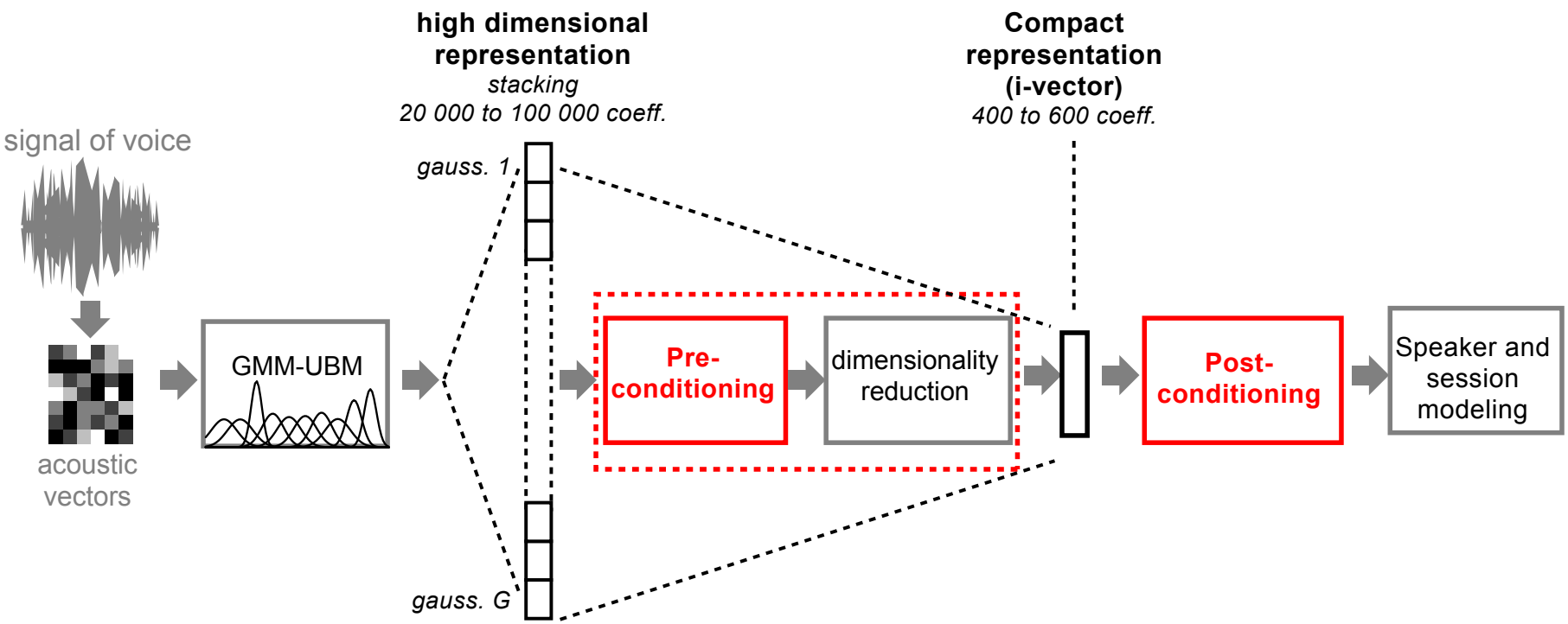
- $\mathbf{A} = \mathbf{\Sigma}$ (total covariance matrix) $\Longrightarrow$ **LΣ**-conditioning
- $\mathbf{A} = \mathbf{W}$ (within-speaker covariance matrix) $\Longrightarrow$ **LW**-conditioning

... eventually iterated.

Parameters are computed for the i-vectors present in the training corpus and applied to test i-vectors.

- Improve Gaussianity of data and many other properties. See:
  - Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In International Conference on Speech Communication and Technology, pages 249-252.
  - Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., and Plchot, O. (2012). Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. In Speaker and Language Recognition Workshop (IEEE Odyssey).
  - Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2011). Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In International Conference on Speech Communication and Technology, pages 485-488.

Pre-conditioning



Pre-conditioning

With 0th and 1th order statistics

$$\mathbf{w} = \left(\mathbf{I} + \mathbf{T}^t \mathbf{\Sigma}^{-1} N_{\mathcal{X}} \mathbf{T}\right)^{-1} \mathbf{T}^t \mathbf{\Sigma}^{-1} N_{\mathcal{X}} \left(\mathbf{s} - \mathbf{m}\right) \tag{5}$$

As $\mathbf{N}_{\mathcal{X}}$ and $\mathbf{\Sigma}$ are diagonal, this equation can be rewritten:

$$\mathbf{w} = \left[\left(\mathbf{I} + \widetilde{\mathbf{T}}^t \mathbf{N}_{\mathcal{X}} \widetilde{\mathbf{T}}\right)^{-1} \widetilde{\mathbf{T}}^t \mathbf{N}_{\mathcal{X}}\right] \mathbf{\Sigma}^{-\frac{1}{2}} \left(s - \mu\right) \tag{6}$$

where $\widetilde{\mathbf{T}} = \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{T}$.

$\Longrightarrow$ Implicit pre-conditioning procedure

- $\mathbf{\Sigma}^{-\frac{1}{2}} \left(s - \mu\right)$
  standardized version of supervector s (according to the world mean and covariance matrix).
- $\left(\mathbf{I} + \widetilde{\mathbf{T}}^t \mathbf{N}_{\mathcal{X}} \widetilde{\mathbf{T}}\right)^{-1} \widetilde{\mathbf{T}}^t \mathbf{N}_{\mathcal{X}}$
  segment-dependent projection matrix (normalization depending on the amount of informations per Gaussian component, expressed in $\mathbf{N}_{\mathcal{X}}$).

Pre-conditioning

i. With 0th and 1th order statistics

To assess the importance of the pre-conditioning procedure:

- replacing F.A. extraction with the simple *Probabilistic Principal Component Analysis* PPCA on **non standardized** supervectors.
  - A **unique projection matrix** is estimated by using Gaussian EM-ML procedure then applied to any supervectors, without taking into account the amount of informations per Gaussian component.

Bishop, C. M. (2006). Pattern recognition and machine learning, volume 4. Springer.
Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology).

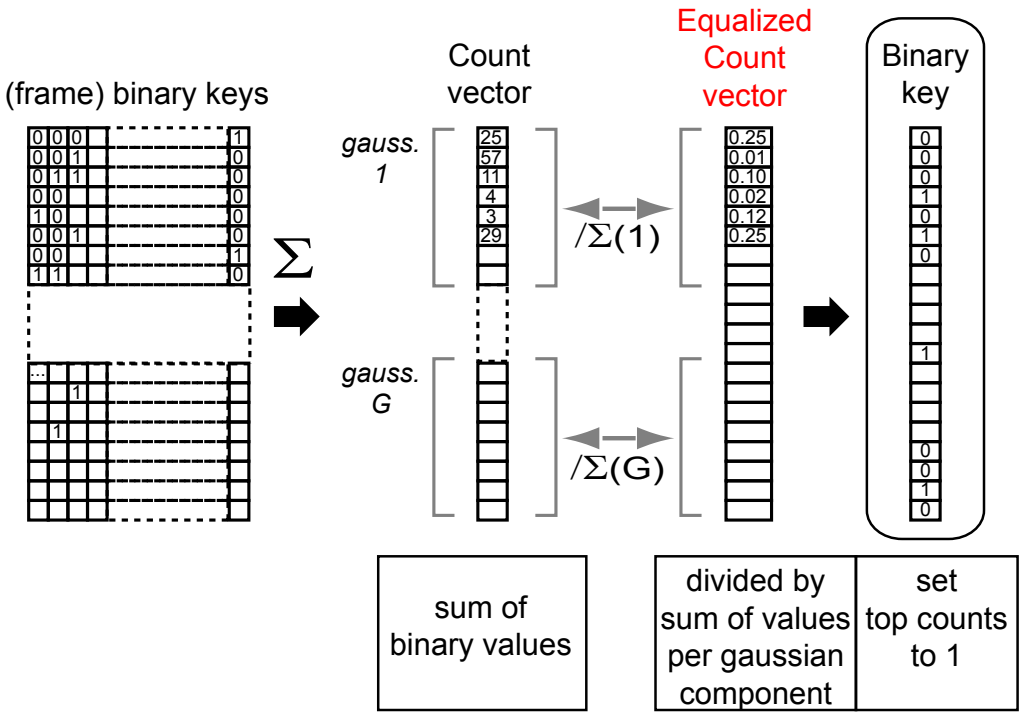
Pre-conditioning

ii. With speaker binary keys (*Equalization per Gaussian Component*)

Let c denote the Gq-dimensional count vector of a segment.
Let $c_{g,k}$ denote the value of c for the kth specificity of the gth GMM component.

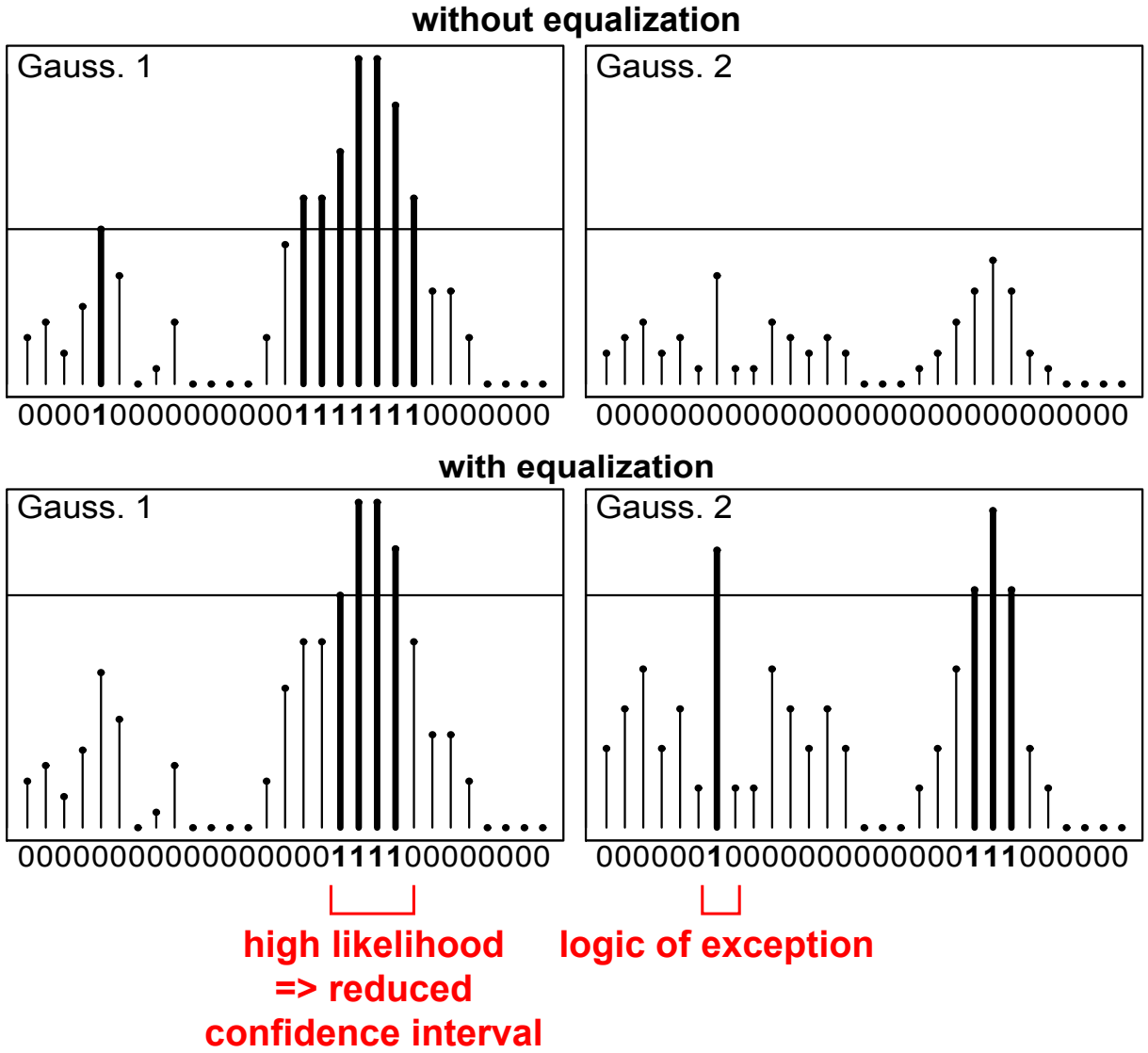
$$\widehat{c}_{g,k} = \frac{c_{g,k}}{\sum_{k=1}^{q} c_{g,k}} \tag{7}$$

The count vector c becomes a local *frequency* vector $\widehat{c}$ (the sum of $\widehat{c}$ values for a Gaussian component is equal to 1).

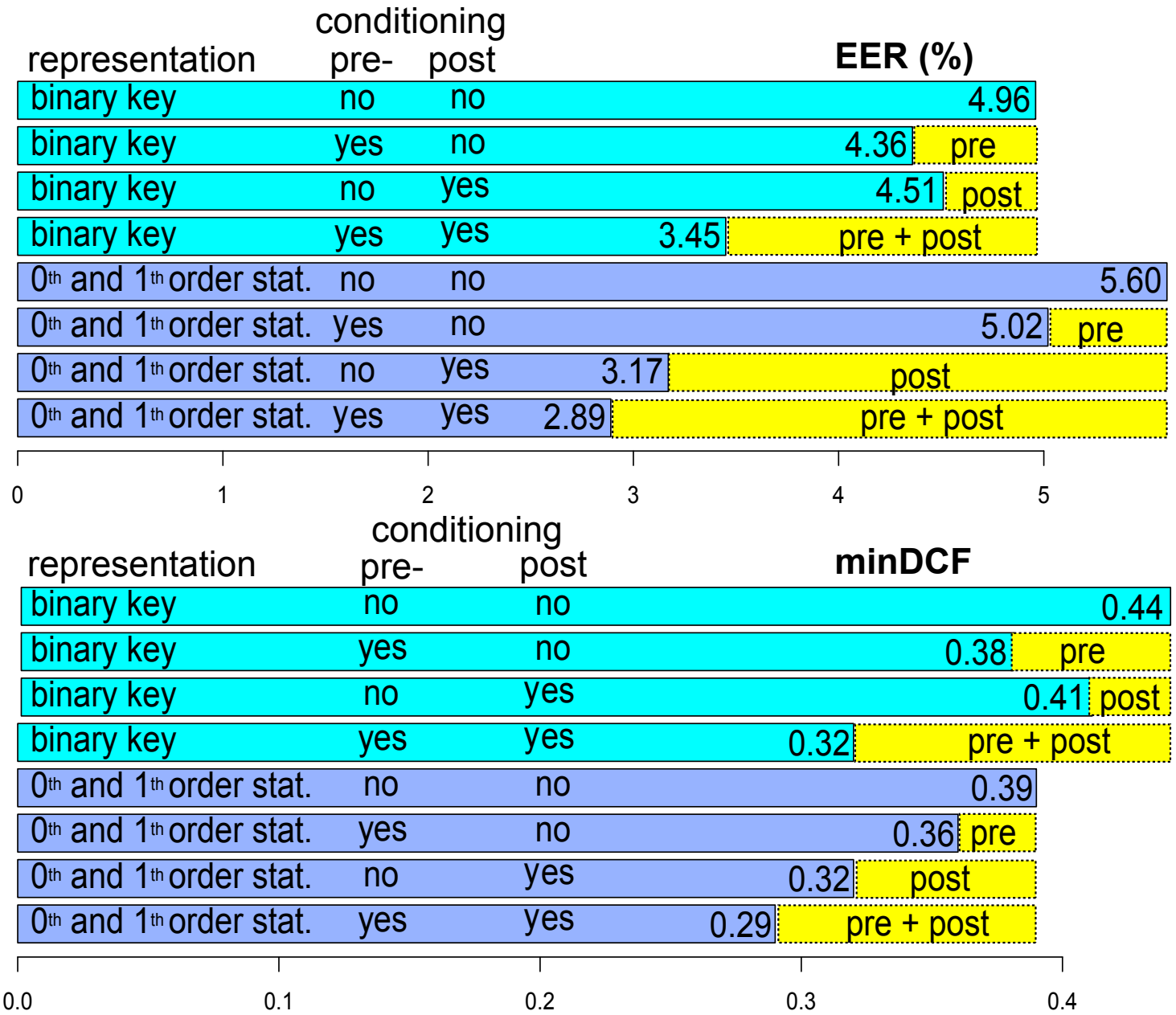


Pre-conditioning

ii. With speaker binary keys



Results


conditioning
representation
pre-
post
EER (%)
binary key no no 4.96
binary key yes no 4.36 pre
binary key no yes 4.51 post
binary key yes yes 3.45 pre + post
0th and 1th order stat. no no 5.60
0th and 1th order stat. yes no 5.02 pre
0th and 1th order stat. no yes 3.17 post
0th and 1th order stat. yes yes 2.89 pre + post
0 1 2 3 4 5
conditioning
representation
pre-
post
minDCF
binary key no no 0.44
binary key yes no 0.38 pre
binary key no yes 0.41 post
binary key yes yes 0.32 pre + post
0th and 1th order stat. no no 0.39
0th and 1th order stat. yes no 0.36 pre
0th and 1th order stat. no yes 0.32 post
0th and 1th order stat. yes yes 0.29 pre + post
0.0 0.1 0.2 0.3 0.4

Conclusion

- Gaussian-PLDA model for total variability factors (*i-vectors*) achieves the best performance in speaker recognition, but these results are only obtained if transformations are applied to i-vectors, before modeling and after their generation by the dimensionality reduction technique.
- Today, the benefits of the **post-conditioning** procedures (comprised of standardization and length-normalization) are well known.
- The evaluation presented in this paper highlights the importance of **pre-conditioning** procedures, applied to high dimensional representations of utterance before their reduction.
  - With the binary key approach, the pre-conditioning procedure that we have designed for this representation turns out to be the most decisive of the two procedures, in terms of system accuracy.
  - With the Baum-Welch 0th and 1th statistics, we shows that FA dimensionality reduction includes an implicit pre-conditioning. Combining the two procedures (pre- and post-) is essential to achieve the state-of-the-art performance.
- This evaluation shows that the strategy of extending the GMM-UBM based speaker recognition systems through an additional stage of total variability factor is only relevant if **conditioning** procedures are applied **before** and **after** the dimensionality reduction stage.

Thank you ...