FULL-COVARIANCE UBM AND HEAVY-TAILED PLDA IN I-VECTOR SPEAKER VERIFICATION

Pavel Matejka[1], Ondrej Glembek[1], Fabio Castaldo[2], M.J. Alam[3,4], Oldrich Plchot[1], Patrick Kenny[3],Lukas Burget[1], Jan Cernocky[1]
[1]Brno University of Technology, Czech Republic, [3]Centre de Recherche Informatique de Montreal (CRIM), Montreal, Canada,
[2]Loquendo, Italy, [4]INRS-EMT, Montreal, Canada

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

- Single best system in post-analysis of ABC (Agnitio+BUT+CRIM) NIST SRE 2010 submission was Full covariance UBM with the state-of-the-art scheme - iVector + PLDA
- Do we really need full-covariance matrices?
- Let us take a look at some analysis.

iVector + PLDA

- **iVector extractor** – model similar to JFA, where GMM mean supervector

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{i}$$

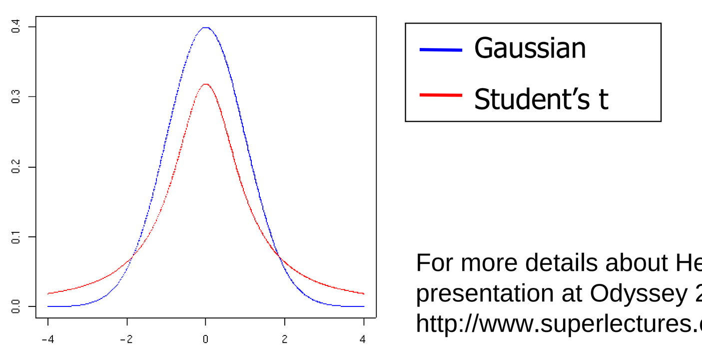
  is constrained to leave in single subspace **T** spanning both speaker and channel variability ➔ no need for speaker labels to train **T**
- **iVector** – point estimate of **i** – can now be extracted for every recording as its low-dimensional, fixed-length representation (typically 400 dimensions)
- contains information about both speaker and channel
- are assumed to be normal distributed
- Natural choice is simplified JFA model with only single Gaussian. Such model is known as PLDA and is described by familiar equation:

$$\mathbf{i} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \epsilon$$

- PLDA has nice interpretation in face verification where it was introduced by Simon J.D. Prince
- Each face image **i** can be constructed by adding
  - mean face **m**
  - linear combination of basis **V** corresponding to between-individual variability (moving from **m** in these directions gives us images that look like different people)
  - linear combination of basis **U** corresponding to within-individual variability (moving from **m** in these directions gives us images that looks like different pictures of the same person)
  - residual noise vector ϵ

Picture taken from: S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," ICCV, 2007

- **Gausian PLDA** – assume standard normal prior for iVectors
- **Heavy tailed PLDA** – assume student's-t distribution prior for iVectors



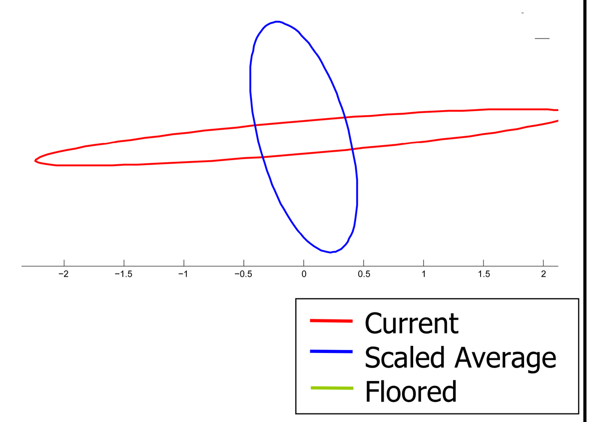
For more details about Heavy tailed PLDA see Keynote presentation at Odyssey 2010 from Patrick Kenny: http://www.superlectures.com/odyssey/lecture.php?lang=en&id=15

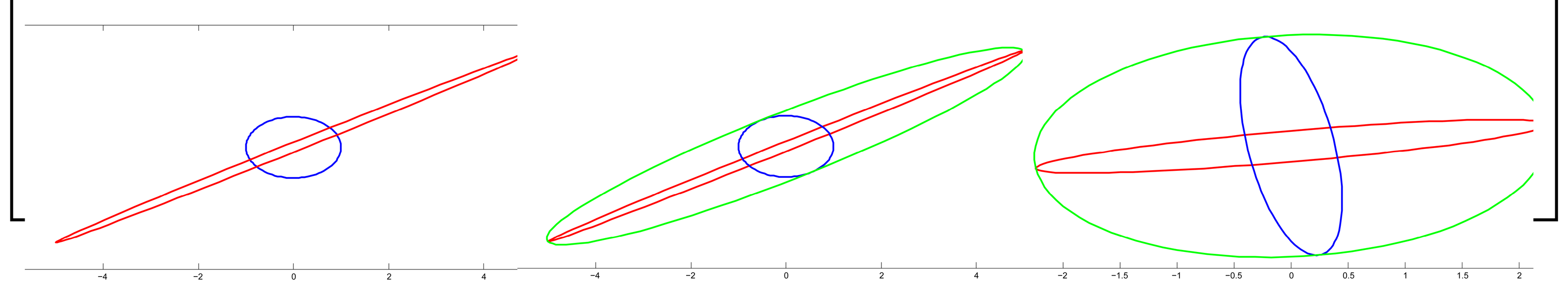
Motivations for full covariance GMM:

- Better description of feature space while preserving reasonable size of GMM mean supervector
- Higher computational complexity ➔ investigation into possible simplifications
- Full covariance Gaussians are more sensitive to very low values of off-diagonal elements -> variance flooring:

Function: $\tilde{\mathbf{S}} = \text{floor}(\mathbf{S}, \mathbf{F})$

1. $\mathbf{F} = \mathbf{L}\mathbf{L}^T$ (Cholesky decomposition)
2. $\mathbf{T} \leftarrow \mathbf{L}^{-1}\mathbf{S}(\mathbf{L}^{-1})^T$ (normalization of target matrix)
3. $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ (Eigenvalue Decomposition - diagonalization of target matrix)
4. Set diagonal matrix $\tilde{\mathbf{D}}$ to $\mathbf{D}$ floored to 1, i.e. $\tilde{d}_{ii} = max(d_{ii}, 1)$
5. $\tilde{\mathbf{T}} \leftarrow \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T$ (making the matrix full again)
6. $\tilde{\mathbf{S}} \leftarrow \mathbf{L}\tilde{\mathbf{T}}\mathbf{L}^T$ (de-normalization)





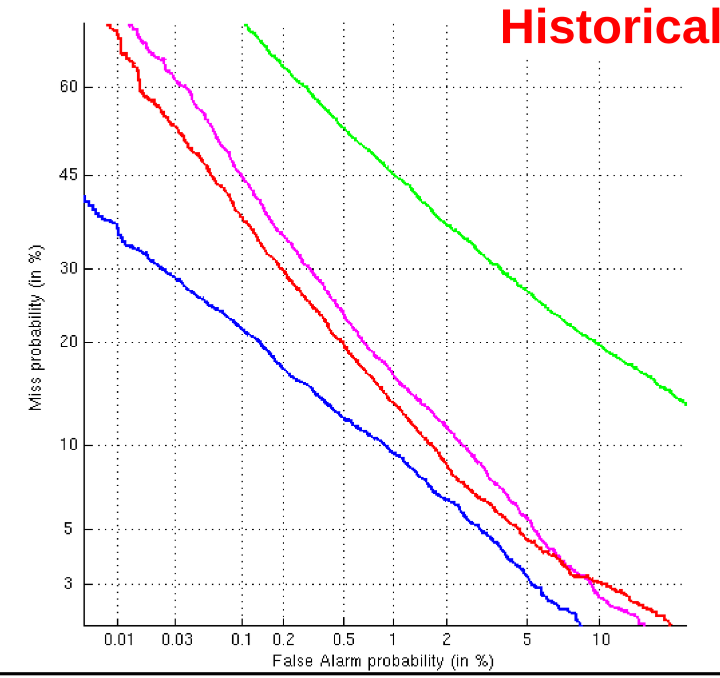
Experimental Setup

Features: MFCC 19+E, Delta + double delta

Short time cepstral mean and variance normalization over 300frames,

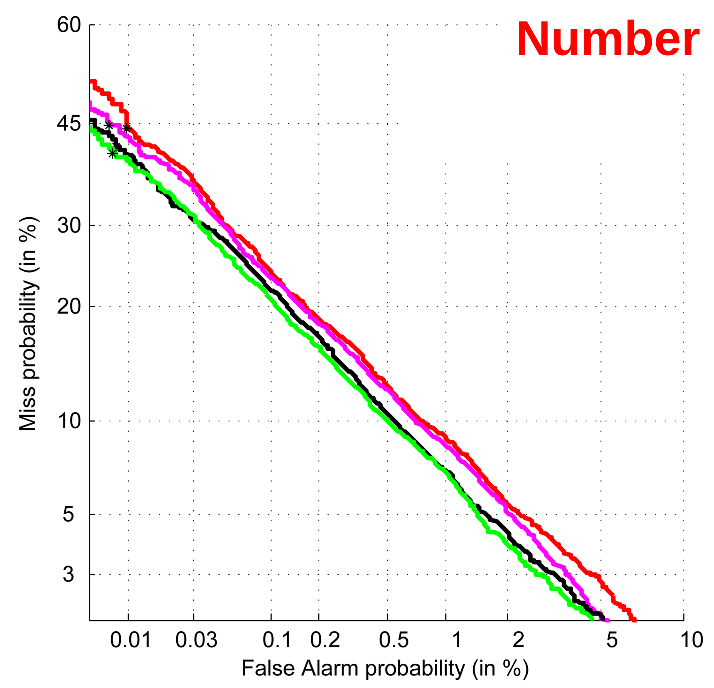
Dataset: NIST SRE 2010, Extended core condition 5 – tel-tel, Female only

Historical way to iVector + PLDA



iVector+PLDA system:
- Implementation simpler than for JFA
- Allows for extremely fast verification
- Provides significant improvements especially in important low False Alarm region

Number Of Parameters



System/#param	UBM	Ivector
2048G Diag i400	0,24M	49,1M
4096G Diag i400	0,49M	98,3M
2048G Diag i800	0,24M	98,3M
2048G Full i400	3,80M	49,1M

Different statistic normalization

Zero order statistics: $N_{\mathcal{X}}^{(c)} = \sum_t \gamma_t^{(c)}$

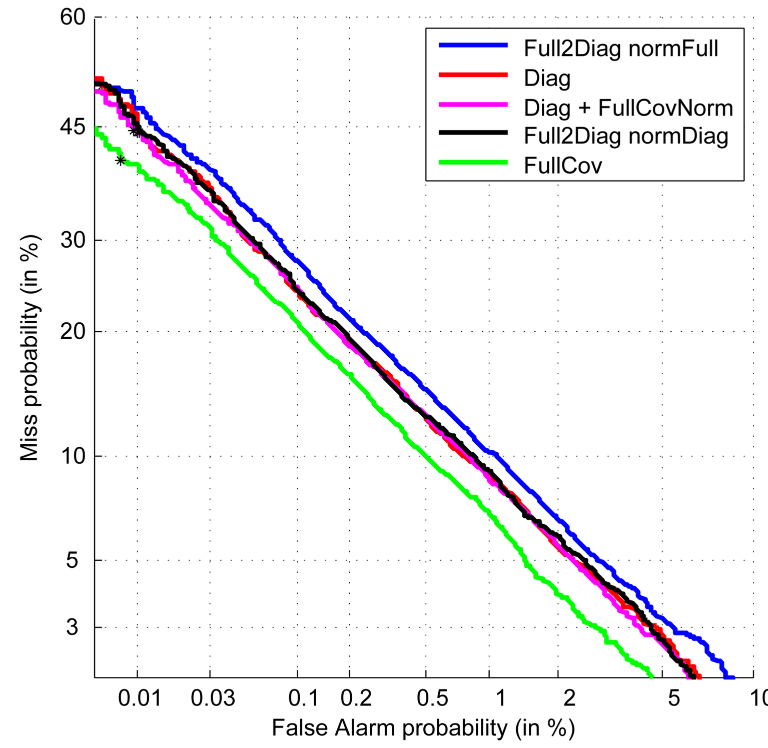
First order statistics: $\mathbf{f}_{\mathcal{X}}^{(c)} = \sum_t \gamma_t^{(c)} \mathbf{o}_t$

Centering around UBM:

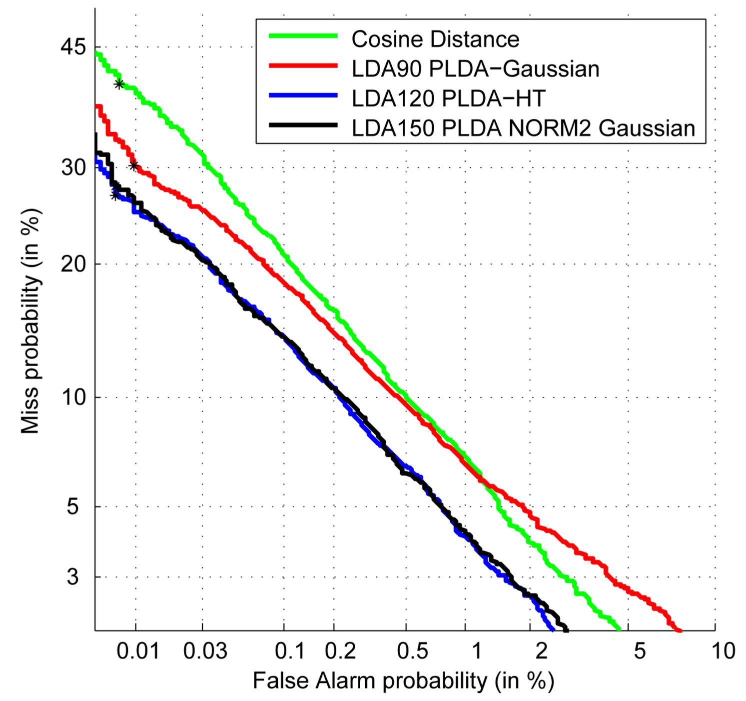
$$\mathbf{f}_{\mathcal{X}}^{(c)} \leftarrow \mathbf{f}_{\mathcal{X}}^{(c)} - N_{\mathcal{X}}^{(c)} \mathbf{m}^{(c)}$$
$$\mathbf{m}^{(c)} \leftarrow \mathbf{0}.$$

Normalization:

$$\mathbf{f}_{\mathcal{X}}^{(c)} \leftarrow \boldsymbol{\Sigma}^{(c)-\frac{1}{2}} \mathbf{f}_{\mathcal{X}}^{(c)}$$
$$\mathbf{T}^{(c)} \leftarrow \boldsymbol{\Sigma}^{(c)-\frac{1}{2}} \mathbf{T}^{(c)}$$
$$\boldsymbol{\Sigma}^{(c)} \leftarrow \mathbf{I},$$

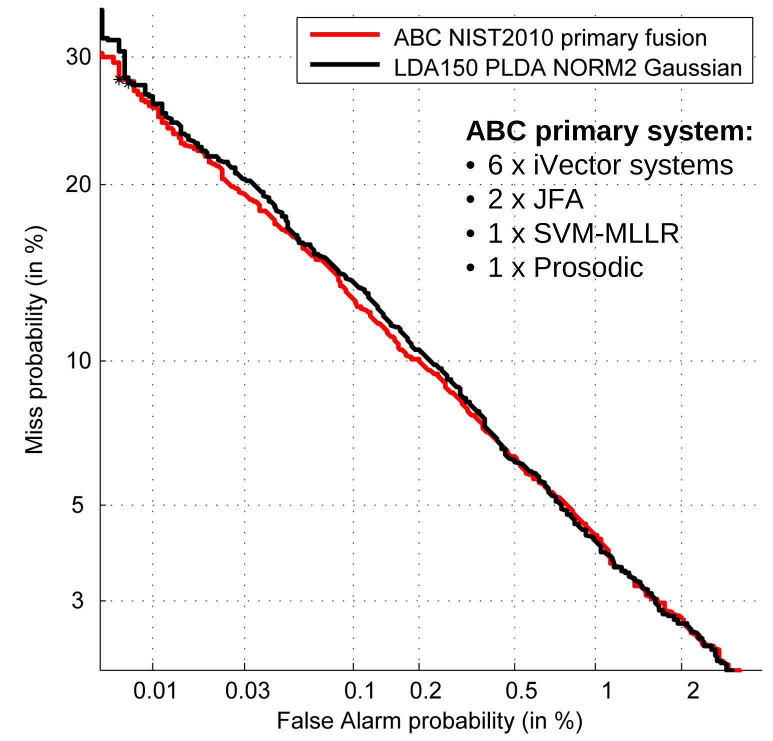


Different scoring



Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems".

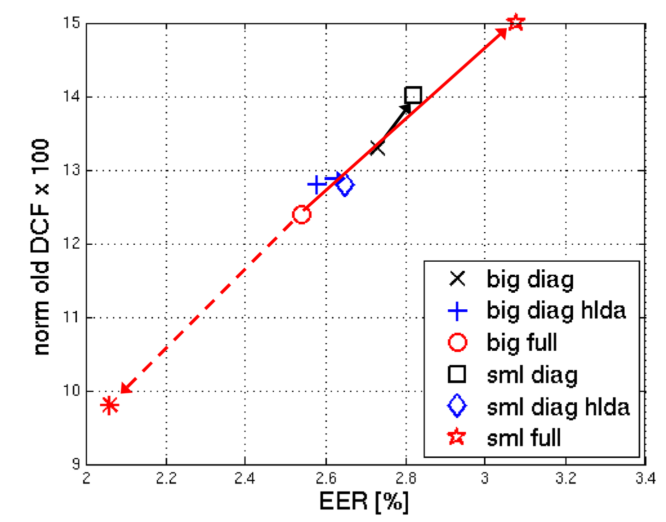
State-of-the-art comparison

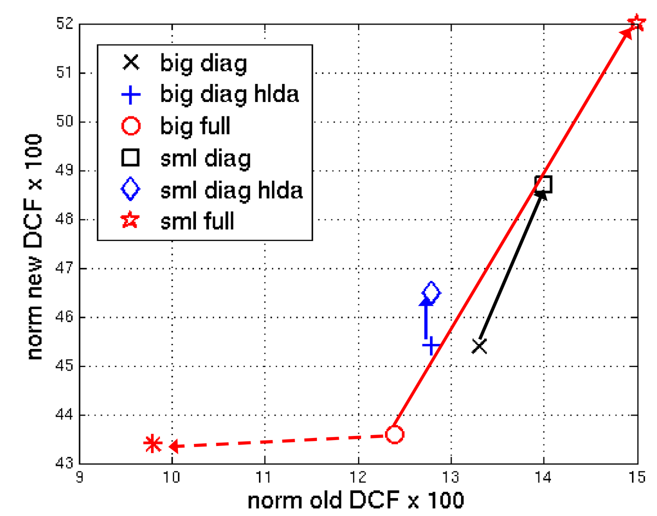


ABC primary system:
- 6 x iVector systems
- 2 x JFA
- 1 x SVM-MLLR
- 1 x Prosodic

Amount of training data

- Full covariance | Diagonal cov. | Diagonal cov + HLDA
- iVector 400, LDA 150, Norm2, Gaussian PLDA
- big = NIST SRE 2004 + 2005 = 310 hours
- sml = 3 hours subset of big set





CONLUSION

- Full covariance UBM gives the best results
- With unity length normalization of iVector you can use Gauss PLDA
- Diagonal covariance UBM with MLLT/HLDA goes very close and have benefit of fast evaluation of Gaussians