DEPARTMENT OF
ELECTRICAL & COMPUTER ENGINEERING
A. JAMES CLARK SCHOOL of ENGINEERING

Speech Communication Lab

University of Maryland

SRE-10

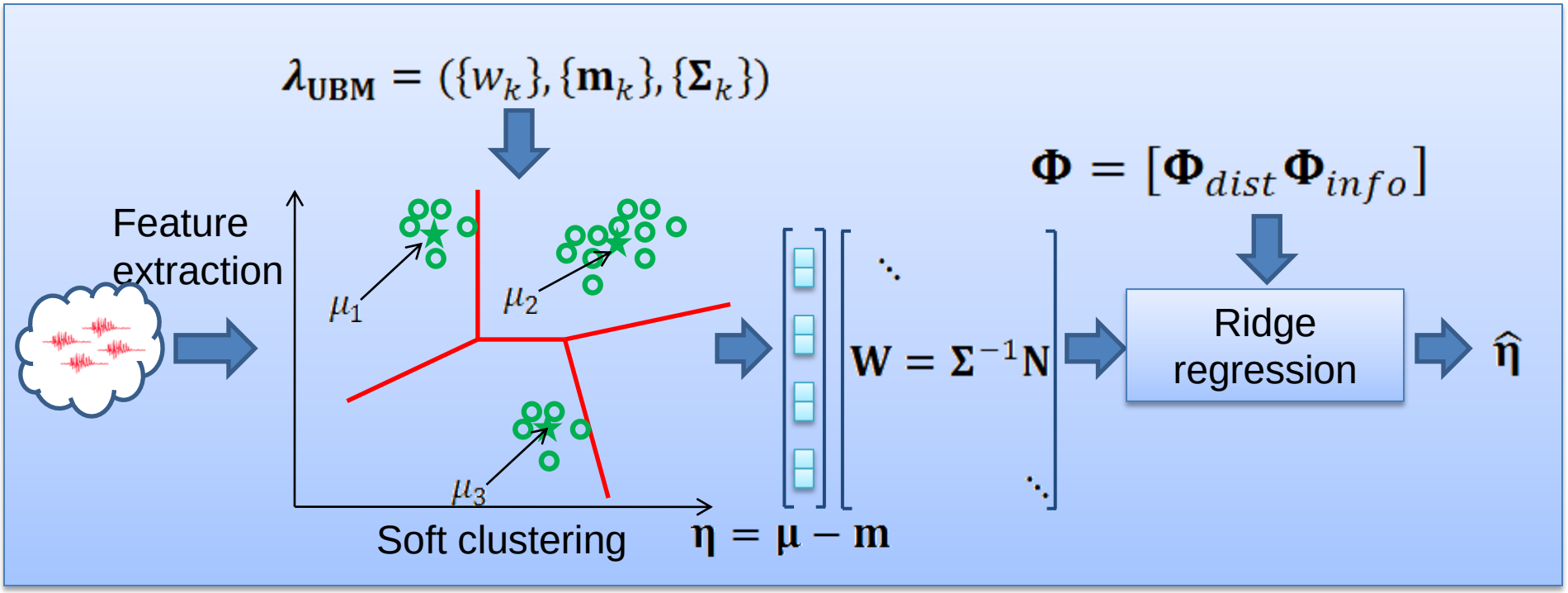
Daniel Garcia-Romero
Xinhui Zhou
Carol Espy-Wilson

The Institute for Systems Research

Overview

- Single system submission
- Core-Core condition
- Based on Signal Coding using Overcomplete Dictionaries [1]
- Multiple encoding and scoring types

Paradigm



Utterance represented by fixed-length vector $\boldsymbol{\eta}$ and a weighting matrix $\mathbf{W}$

Signal coding

Mathematical description

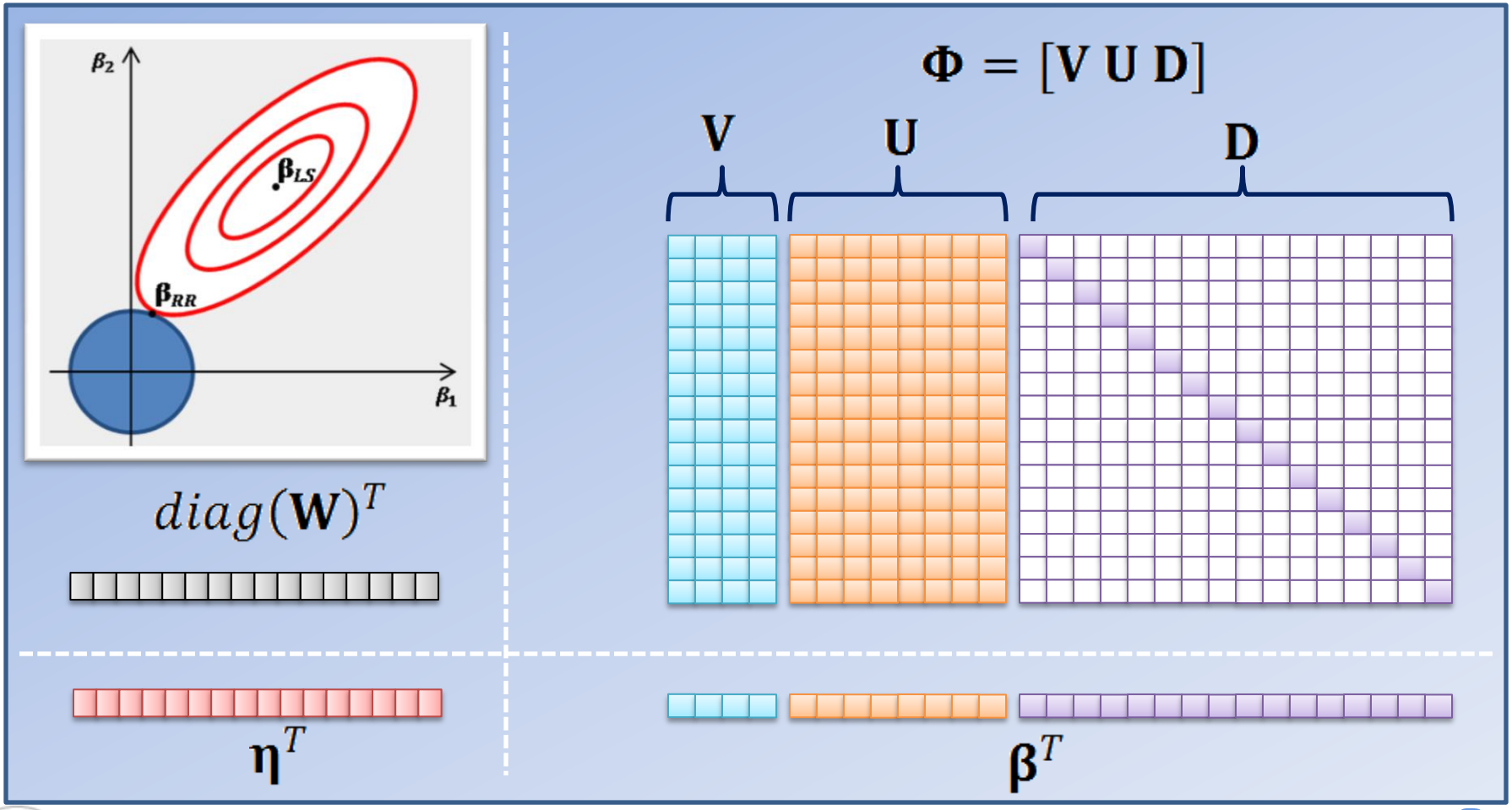
- Linear-Gaussian observation model:
  - Likelihood: $\boldsymbol{\eta} = \boldsymbol{\Phi\beta} + \boldsymbol{\epsilon}, \;\; with \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}\mathbf{N}^{-1})$
  - Prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - Posterior: $\boldsymbol{\beta}|\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{C}_{\boldsymbol{\beta}})$

    $with \; \boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{C}_{\boldsymbol{\beta}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N} \; and \; \mathbf{C}_{\boldsymbol{\beta}} = (\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}\boldsymbol{\Phi})^{-1}$
  - Mode of posterior: $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ because Gaussian

$$\boldsymbol{\beta}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \mathbf{W}^{\frac{1}{2}}(\boldsymbol{\eta} - \boldsymbol{\Phi\beta}) \right\|_2^2 + \frac{1}{2} \|\boldsymbol{\beta}\|_2^2$$

Signal Coding (SC)

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\Phi}\boldsymbol{\beta}_{\mathrm{MAP}}$$



Dictionary Learning (DL)

- Given R utterances $(\boldsymbol{\eta}_r, \mathbf{W}_r)$:

$$\min_{\boldsymbol{\Phi},\{\boldsymbol{\beta}_r\}} \sum_{r=1}^{R} \left\| \mathbf{W}_r^{\frac{1}{2}} (\boldsymbol{\eta}_r - \boldsymbol{\Phi}\boldsymbol{\beta}_r) \right\|_2^2 + \|\boldsymbol{\beta}_r\|_2^2 .$$

- Block-coordinate descent:
  - Alternating optimization

    Signal coding $\rightleftharpoons$ Dictionary Update (DU)
  - SC is performed keeping $\boldsymbol{\Phi}$ fixed
  - DU is performed by ML keeping the $\{\boldsymbol{\beta}_r\}$ fixed

Scoring

- Cosine similarity with multiple metrics:

$$\boldsymbol{score} = \frac{\langle \hat{\boldsymbol{\eta}}_{mod}, \hat{\boldsymbol{\eta}}_{test} \rangle_{\mathbf{W}_\#}}{\langle \hat{\boldsymbol{\eta}}_{mod}, \hat{\boldsymbol{\eta}}_{mod} \rangle_{\mathbf{W}_\#}^{1/2} \langle \hat{\boldsymbol{\eta}}_{test}, \hat{\boldsymbol{\eta}}_{test} \rangle_{\mathbf{W}_\#}^{1/2}}$$

$$with \quad \mathbf{W}_\# = \boldsymbol{\Sigma}^{-1}\mathbf{N}_{test} \quad or \quad \mathbf{W}_\# = \boldsymbol{\Sigma}^{-1}\mathbf{N}_{UBM}$$

- Linear scoring [2]:

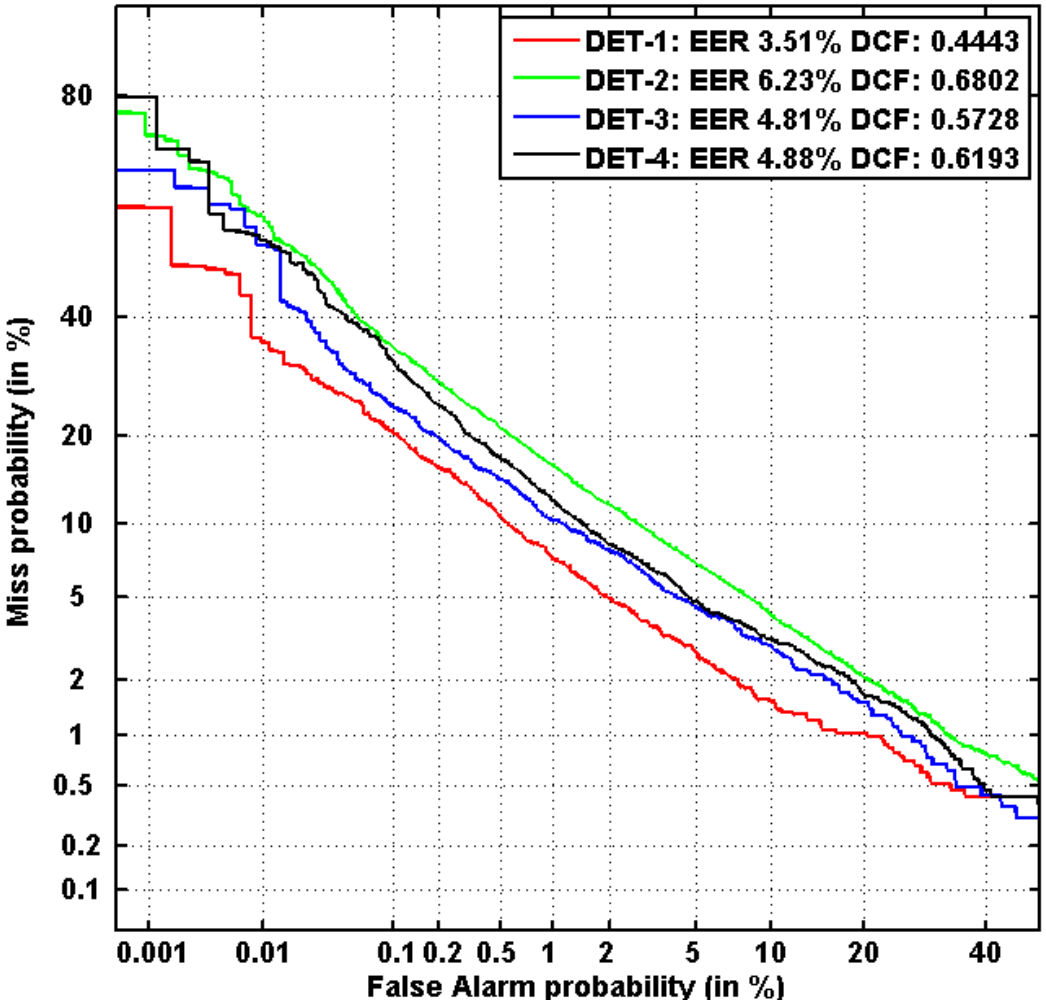
$$\boldsymbol{score} = \frac{1}{N_{TOT}} \langle \hat{\boldsymbol{\eta}}_{mod}, \boldsymbol{\eta}_{test} - \mathbf{U}\mathbf{x}_{test} \rangle_{\mathbf{W}_{test}}$$

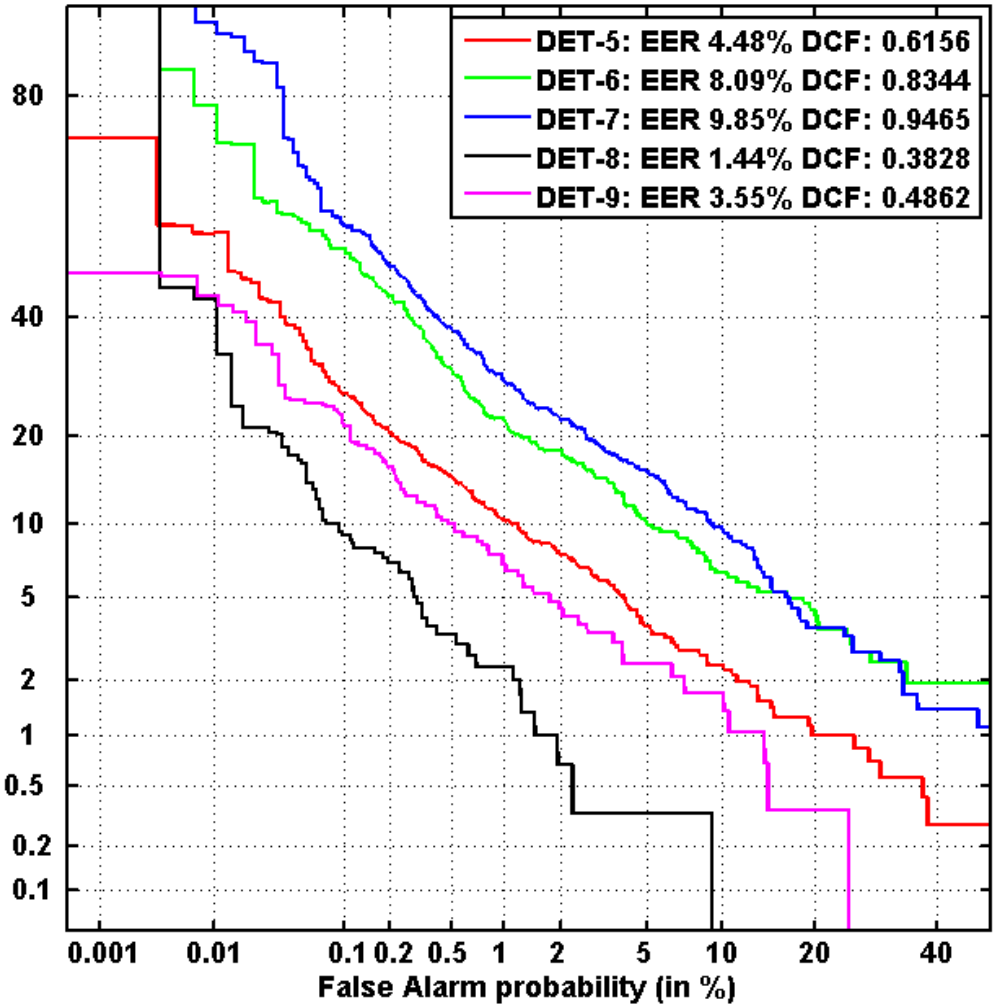
System configuration

- Feature extraction*:
  - 38 MFCCs (c1-c19 + Delta) every 10ms with 20ms window. RASTA and standardized.
  - VAD based on ASR + Energy heuristics.
- UBM*: 2048 mixtures from SRE-04,05 and 06
- Dictionary based on SRE-04, 05 and 06 and SRE08 follow-up data
  - Independent training of U and V matrices. D fixed with rel=16
  - Submitted $\mathbf{\Phi} = [\mathbf{U}_{mic}\mathbf{U}_{tel}\ \mathbf{D}]$, 50, 100, and 2048*38 dim
  - Late $\mathbf{\Phi} = [\mathbf{U}_{mic}\mathbf{U}_{tel}\mathbf{V}\ \mathbf{D}]$ with 50, 100, 300 and 2048*38 dim

* Feature extraction and UBM training done with MIT-LL software

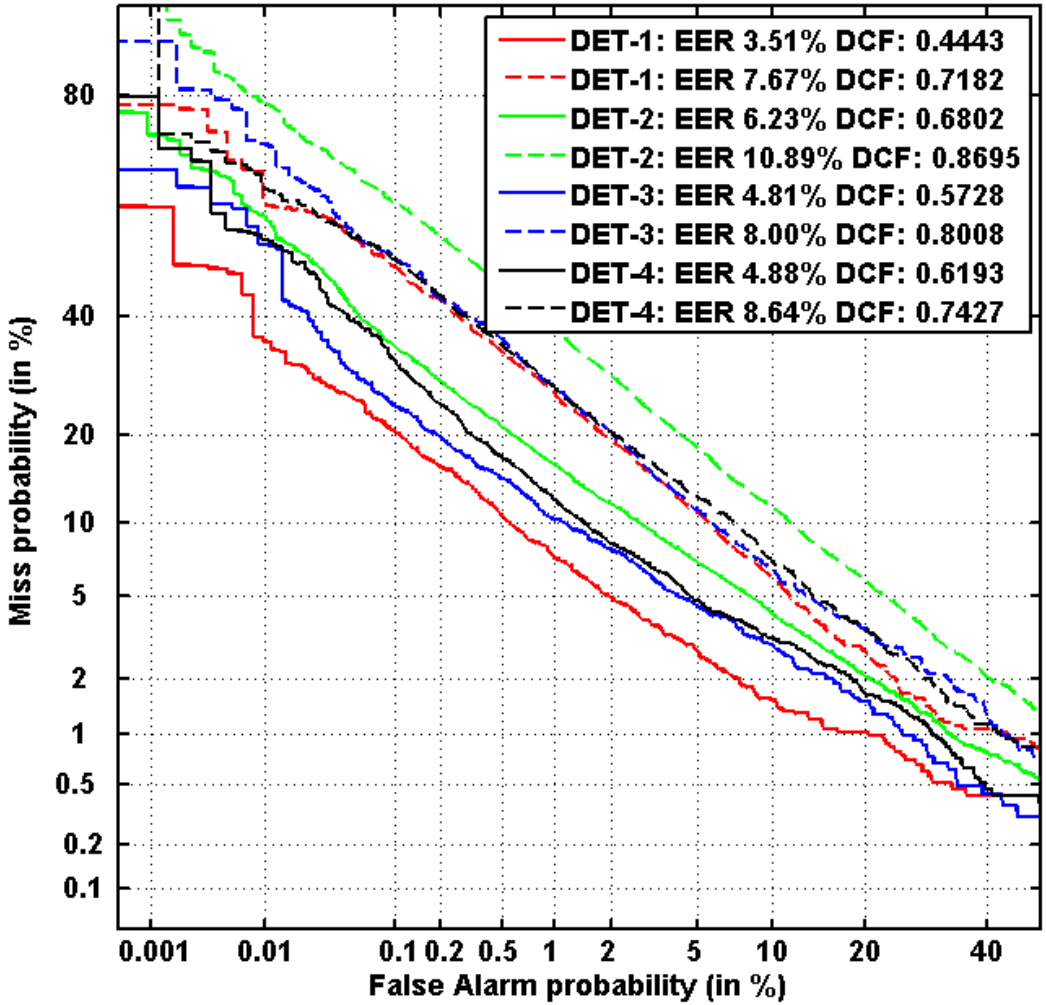
Results (primary submission)





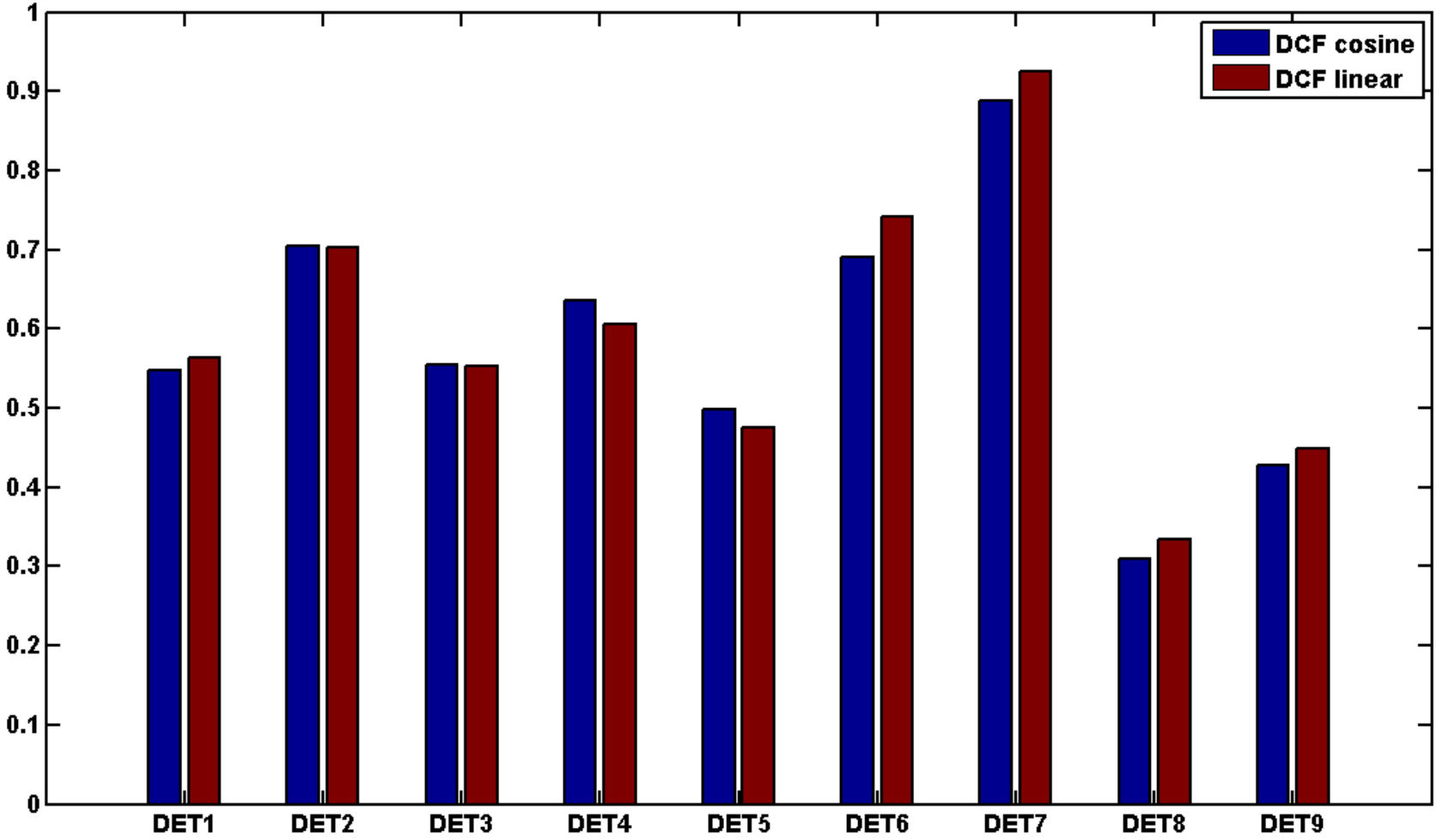
Results for cosine similarity scoring and $\mathbf{\Phi} = [\mathbf{U}_{mic}\mathbf{U}_{tel}\ \mathbf{D}]$

Analysis Results (ZT-norm)



Great improvement by ZT-norm: mode than 50% in some cases

Analysis Results (Scoring)



Cosine similarity slightly better but not significant

Conclusions

- Established a connection between JFA and signal coding in overcomplete dictionaries
- Mixed results between cosine similarity and linear scoring
- ZT-norm is essential
- High vocal effort quite detrimental to performance
- Set a baseline for comparing with future research:
  - L1-regularized regression
  - Discriminative dictionary learning
- Thanks to MIT-LL for providing binaries for feature extraction and UBM training.

References

[1] D. Garcia-Romero and C. Espy-Wilson, “Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries”, Odyssey 2010.

[2] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in *ICASSP*, 2009, pp. 4057-4060.