The CRIM System for the 2006 NIST SRE

Patrick Kenny, Shou-Chun Yin

Centre de recherche informatique de Montréal

The ‘Fundamental Equation’ of Speaker Recognition

$S = \text{speaker supervector}$

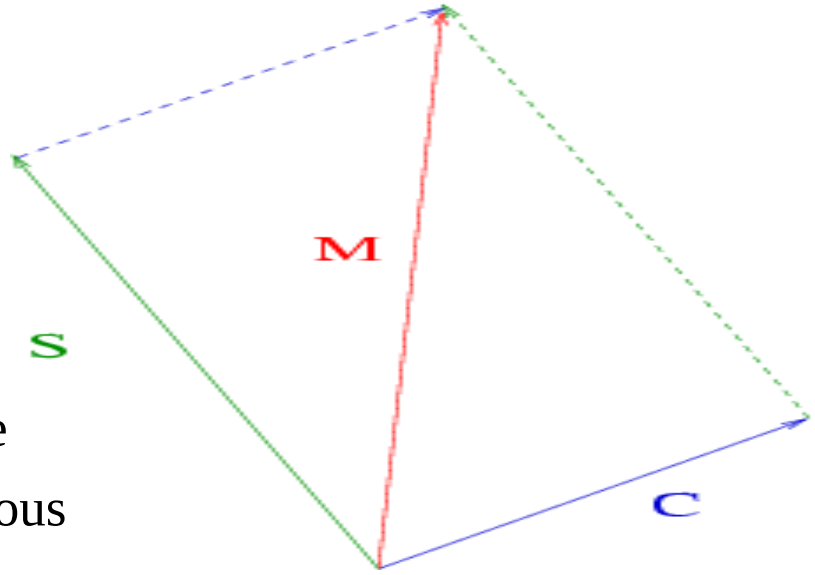
$C = \text{channel supervector}$

$M = \text{speaker} + \text{channel supervector}$

$$M = S + C$$

Feature Mapping (supervised) : C is discrete

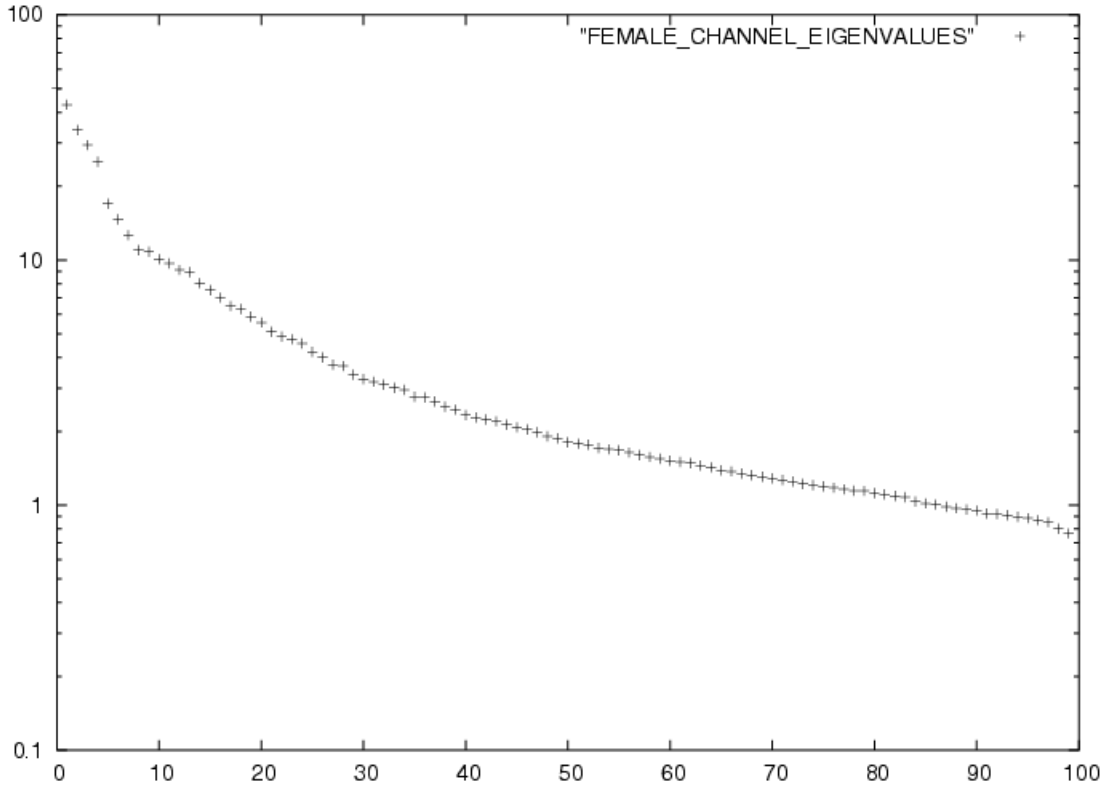
New Methods (unsupervised) : C is continuous



Covariance modeling is the key to handling session variability

- Is Cov(C, C) of full rank or of low rank?
  - Is a small number of latent variables enough?
- Discriminative or generative?
  - SVMs or GMMs?
- GMM Model domain or feature domain?

Session Variability
Cov(C, C) = uu*



100 channel eigenvalues sorted in decreasing order

Discriminative Approaches

- SVM-GMM-NAP (e.g. MIT)
  - Channel space = range of Cov(C, C)
  - Decomposition M = S + C:
    - Speaker supervector S extracted from M by orthogonal projection onto channel space
  - Kernel is essentially Euclidean
    - Length scales defined by GMM variances
- WCCN (e.g. ICSI)
  - Kernel is Mahalanobis inner product defined by Cov(C, C) (full rank achieved by adding a multiple of identity matrix)
  - Two supervectors which differ by a principal eigenvector of Cov(C, C) are close in Mahalanobis sense

The Basic Generative Approach: Eigenchannel MAP

- Eigenchannels or channel factors are just the eigenvectors of Cov(C, C)
  - Channel space defined just as in SVM-GMM-NAP
- Given a speaker GMM and a test utterance, compensate for channel effects by adapting to the test utterance using Cov(C, C) as a prior
  - Naïve implementation very expensive with t-norm (but BUT doesn't seem to need it!)
  - Fast approximations exist (CRIM, QUT, SDV)

Why Joint Factor Analysis?

- M = S + C
  - We are really interested in the speaker supervector S
  - The channel supervector C is just noise
- Channel effects should be suppressed at enrollment time
- Classical MAP does not achieve this
  - QUT improves on classical MAP with Gauss-Siedel iterative method
  - MIT improves on classical MAP by orthogonally projecting M
- Joint factor analysis aims to achieve direct MAP estimation of S from enrollment data

What assumptions can be made about Cov(S, S)?

Classical MAP : d diagonal

$$\mathrm{Cov}(S, S) = d^2$$

$$\text{e.g. } d^2 = \frac{1}{16}\sum$$

Eigenvoice MAP : v low rank

$$\mathrm{Cov}(S, S) = vv^*$$

Factor Analysis MAP : d diagonal, v low rank

$$\mathrm{Cov}(S, S) = vv^* + d^2$$

Compare eigenchannel MAP : u low rank

$$\mathrm{Cov}(C, C) = uu^*$$

Hidden Variable Formulations

Classical MAP : $S = m + dz$

Eigenvoice MAP : $S = m + vy$

Factor Analysis MAP : $S = m + vy + dz$

Eigenchannel MAP : $C = ux$

Joint Factor Analysis : $M = (m + vy + dz) + ux$

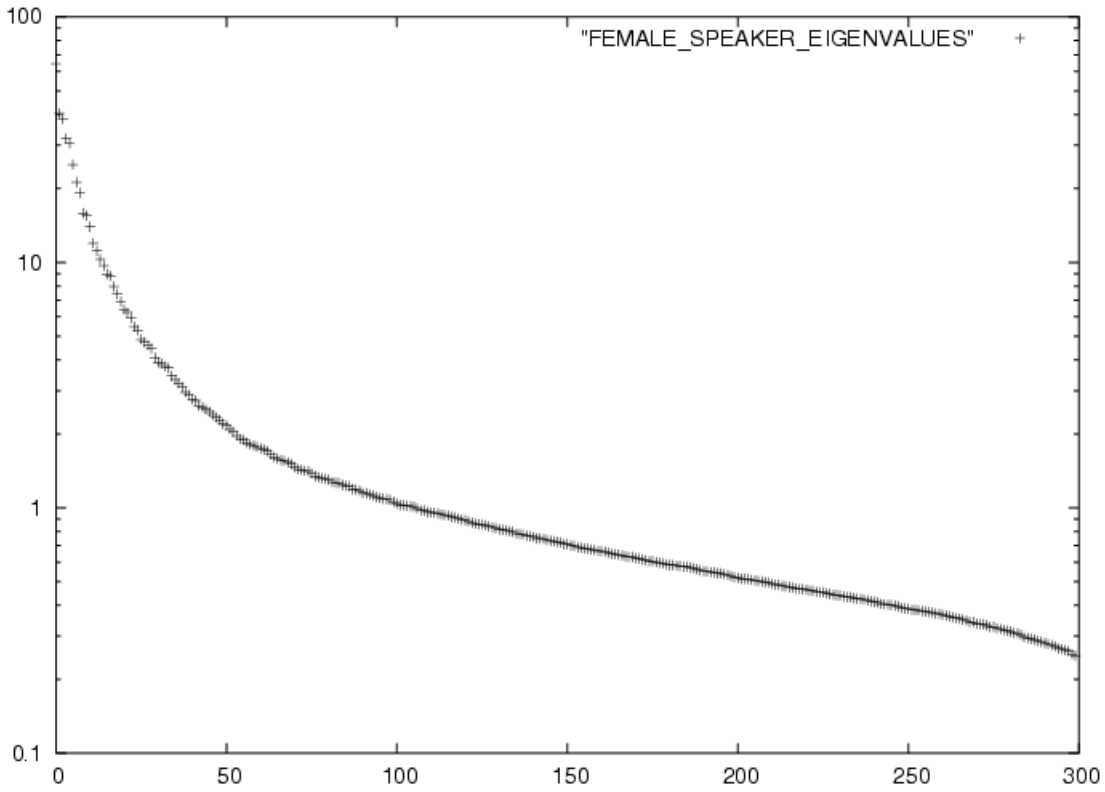
x, y, z standard normal random vectors (factors)

m = speaker-independent supervector

range of $uu^* =$ channel space

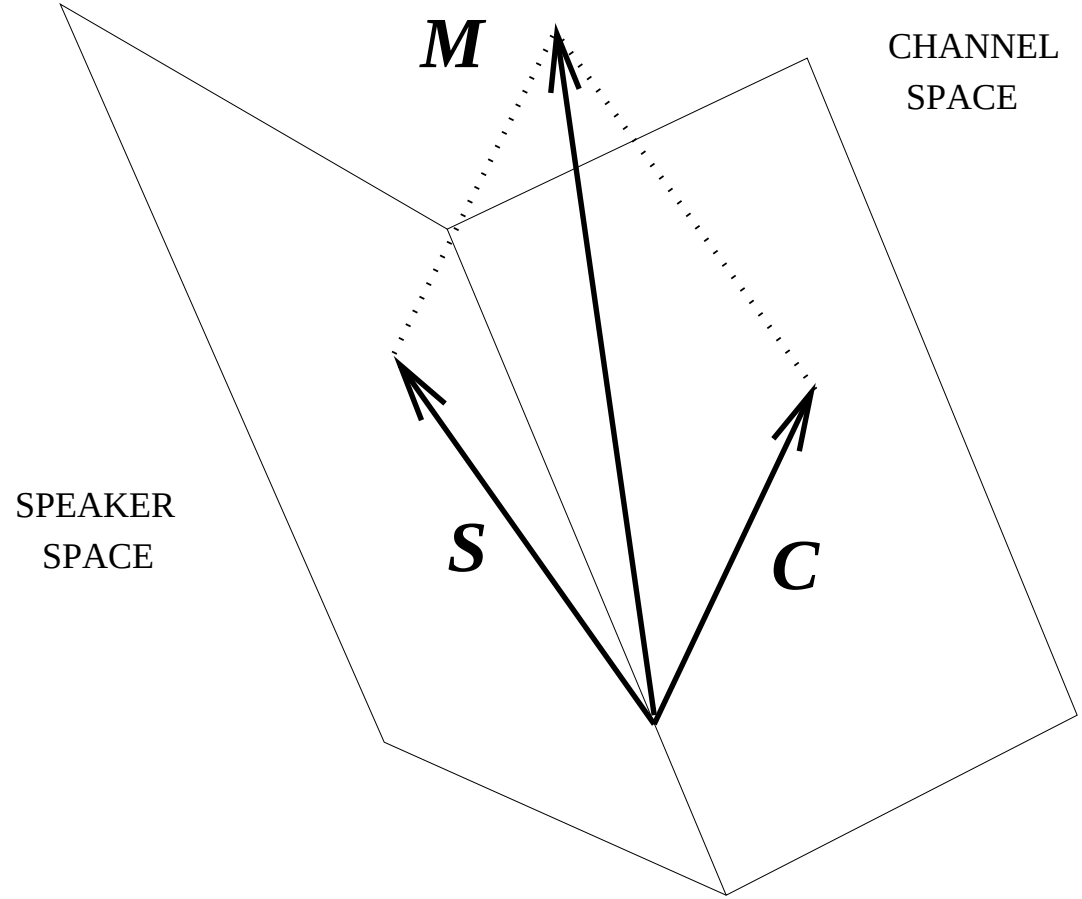
range of $vv^* =$ speaker space

Speaker Variability
Cov(S, S) = vv*



300 speaker eigenvalues sorted in decreasing order

The Speaker Space and the Channel Space



$$M = S + C$$

Disentangling Speaker and Channel Effects

M = speaker and channel supervector for a given utterance

$$M = S + C$$
$$= (m + vy + dz) + ux$$

Extract the Baum - Welch statistics from the utterance and calculate the MAP estimates $\hat{y}, \hat{z}$ and $\hat{x}$ of y, z and x

Speaker supervector $= m + v\hat{y} + d\hat{z}$

Channel supervector $= u\hat{x}$

LPT says : don't just throw away the channel supervector!

How to determine whether two speakers are the same using FA

For a given speaker, the speaker supervector S is given by

$$S = m + vy + dz$$

For a given recording of the speaker, the speaker and channel dependent supervector M is given by

$$M = S + ux$$

One way of deciding whether or not 2 recordings have been uttered by the same speaker :

Evaluate the likelihood of the data in 2 ways

(i) Assume the speaker factors y and z are the same for both recordings

(ii) Assume that they are different

System Configuration

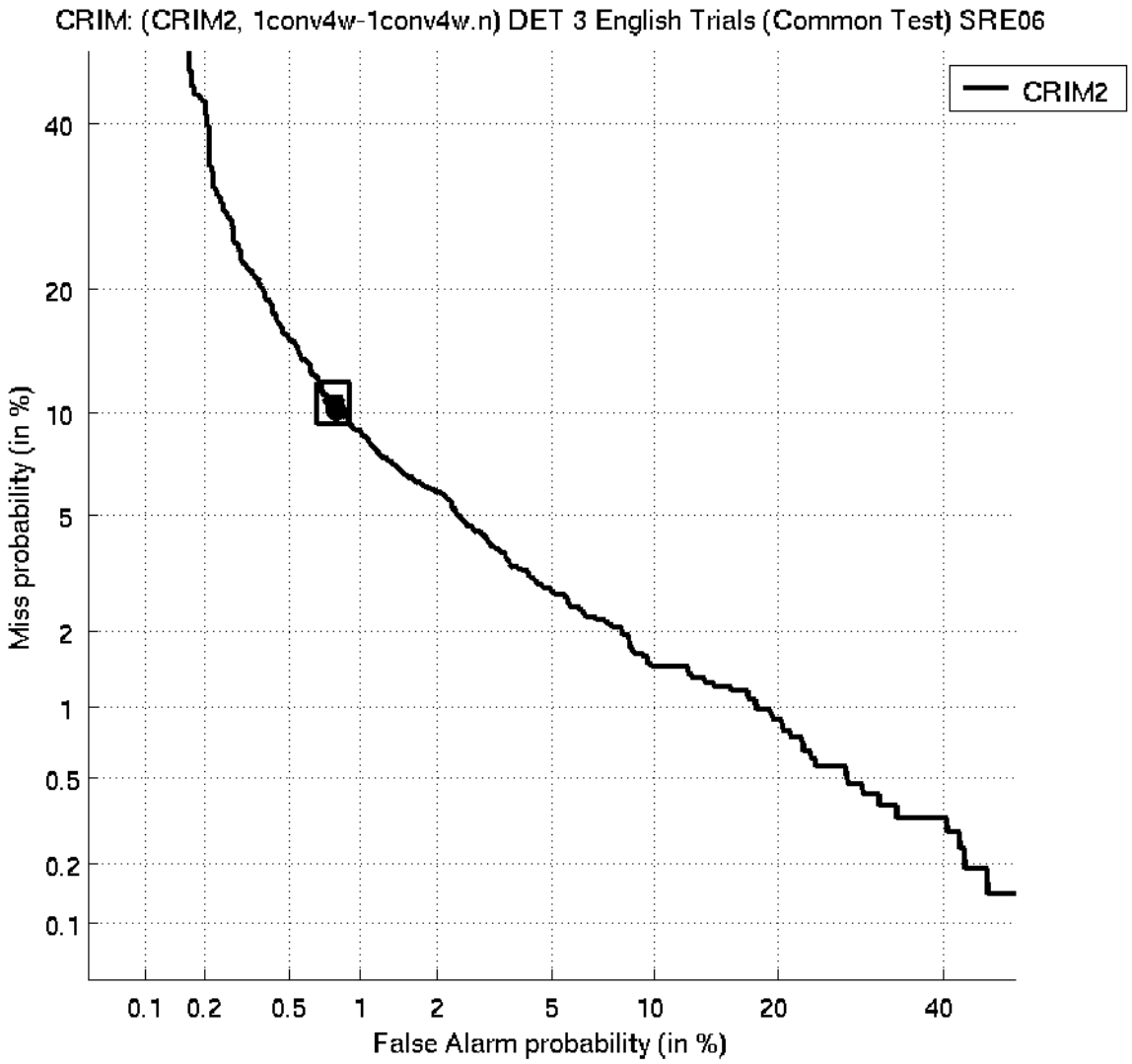
- Front end
  - 12 MFCC's+ log energy + first derivatives
  - Feature warping (short term Gaussianization)
  - No Feature mapping
  - Silence detection with .ctm files
- Factor Analysis
  - Gender dependent
  - 2K Gaussians in UBM
  - 300 speaker factors in CRIM_1, CRIM_2; 0 in CRIM_3
  - 75 channel factors
  - ***NIST 2005 data added to training***
- Score normalization and calibration
  - ZT-norm
  - Gender dependent logistic regression (FoCal)

Development Results on Fisher Data

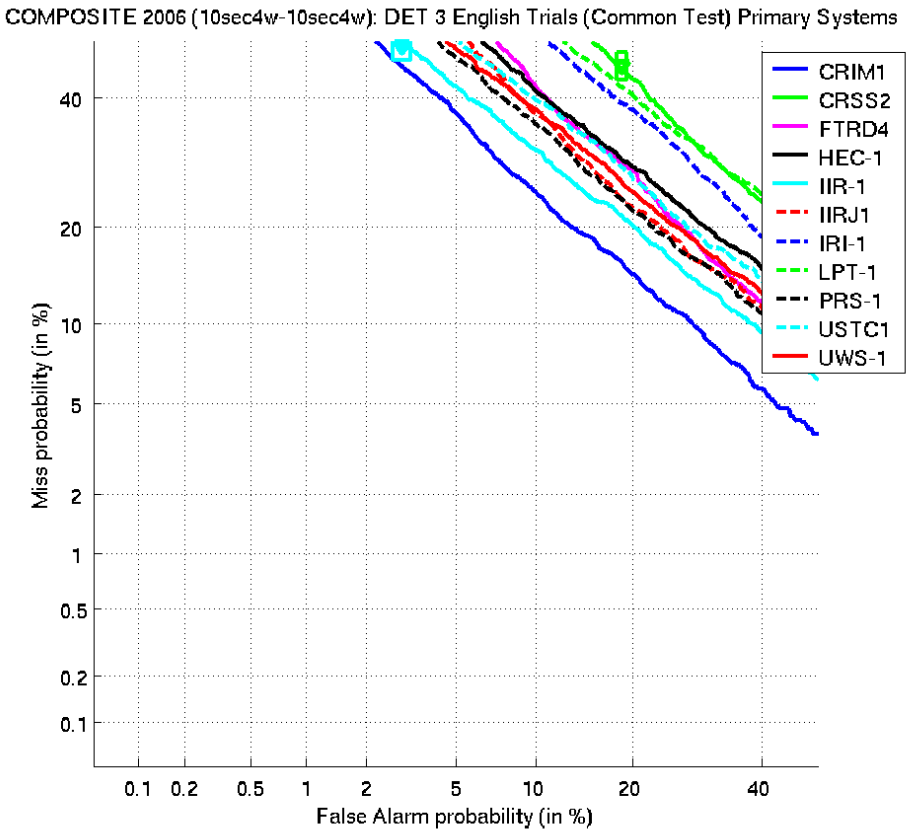
- Whole Fisher conversation sides
- For each gender:
  - 1000 target speakers
  - 1000 target trials
  - 10 000 non-target trials

	EER	DCF
MALES	1.8%	0.005
FEMALES	2.0%	0.004

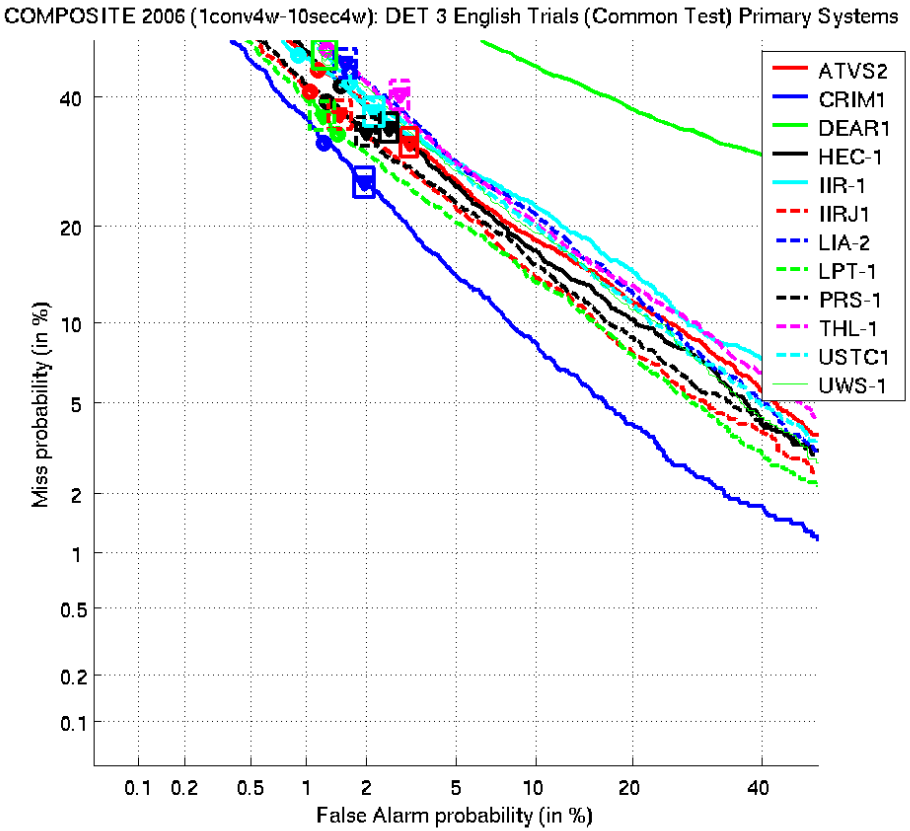
Core condition (English)



10sec4w-10sec4w (English)



1conv4w-10sec4w (English)



8conv4w-1conv4w (English)