The CRIM System for the NIST 2005 SRE

Corpus-Based Models of Speaker and Channel Variability

Patrick Kenny

Objectives

- Explore new ways of modeling utterance and session variability with large corpora
- Lots of data available but almost all speakers are either landline or cellular
- See if factor analysis can be simplified
  - Eigenchannel modeling
- Core condition only
  - Post evaluation results

Utterance Variability I: Classical MAP

M = supervector for an utterance (conversation side)

m = speaker-independent supervector

d^2 a diagonal covariance matrix

Assumption: M normal with mean m, covariance d^2

Hidden variable formulation: $M = m + dz$, z standard normal

Utterance Variability II: Eigenvoice MAP

Assumption: M normal with mean m, covariance vv^*

v is a rectangular matrix of low rank

Equivalently: $M = m + vy$, y standard normal

M is a linear combination of m and the columns of v (speaker space)

Utterance Variability III: Factor Analysis MAP

Assumption : $M = m + vy + dz$, y and z standard normal

M normal with mean m, covariance $d^2 + vv^*$

(i.e. a factor analysis in the sense of Rubin & Thayer)

Components of y are *speaker factors*

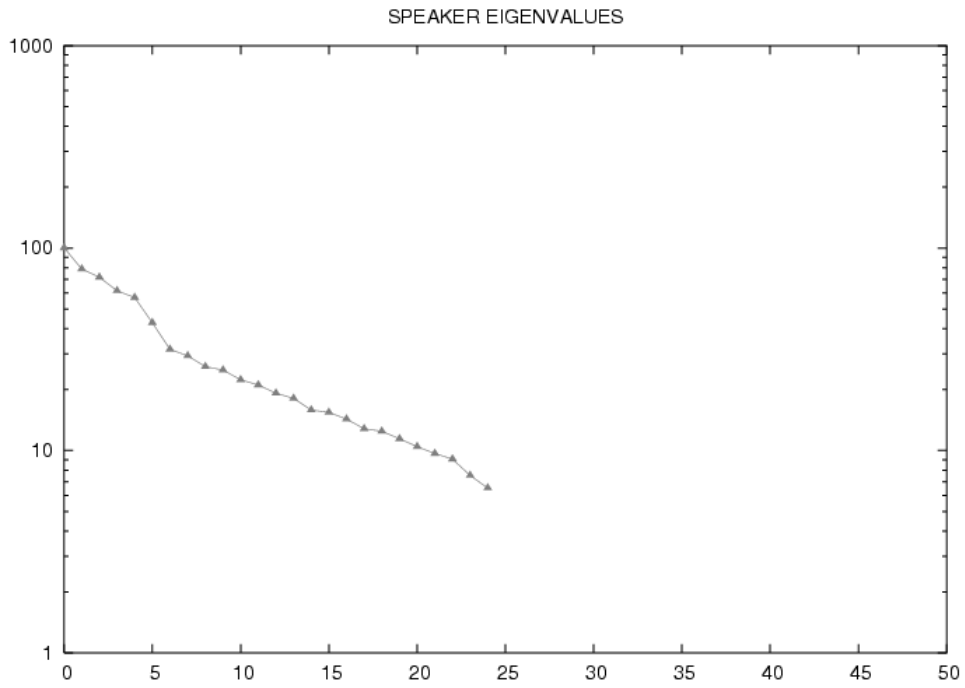
dz is a residual which compensates for the fact

that the speaker space constraint may be too stringent

Utterance Variablity (25 eigenvoices, female data)

$$\mathrm{tr}(d^2) = 77$$

$$\mathrm{tr}(vv^*) = 730$$



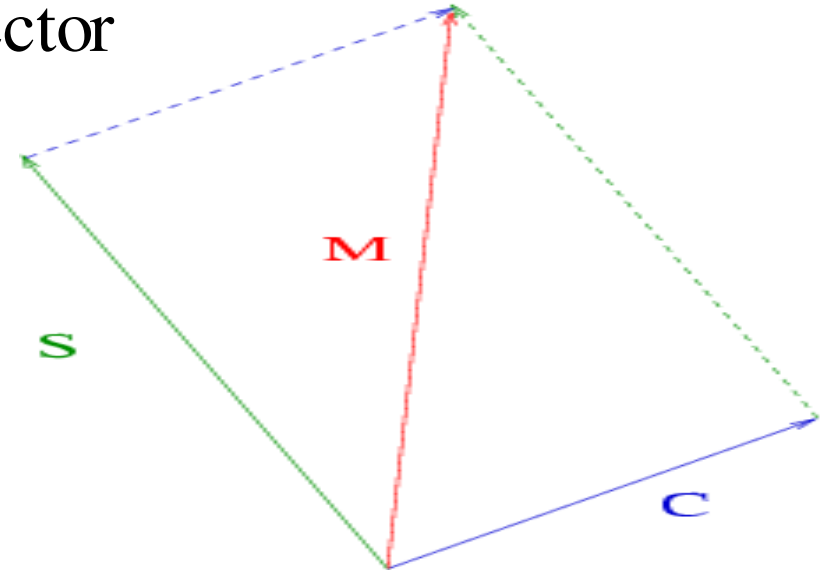
Channel Variability I: Joint Factor Analysis and Feature Mapping

$S = \text{speaker supervector}$

$C = \text{channel supervector}$

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

$$M = S + C$$

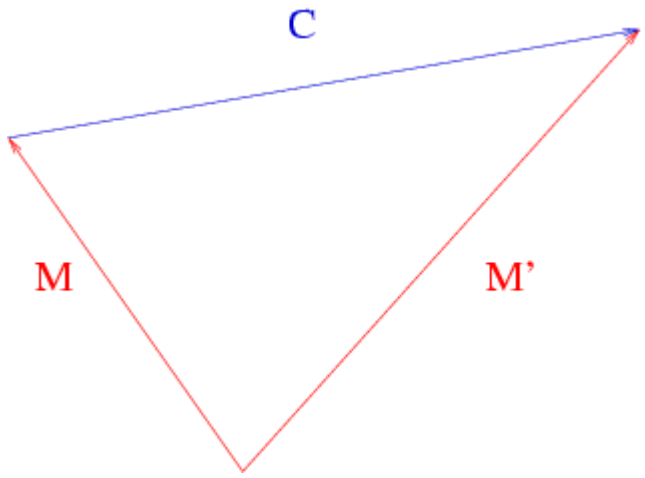


Channel Variability II: Eigenchannel Modeling and Speaker Model Synthesis

$$M, M' = \text{speaker} + \text{channel supervectors}$$

$$C = \text{channel supervector}$$

$$M' = M + C$$



Channel Variability III: Eigenchannel MAP

M and M' = supervectors for 2 utterances by the same speaker

Assumption: $M' = M + ux$, x standard normal

u rectangular, low rank

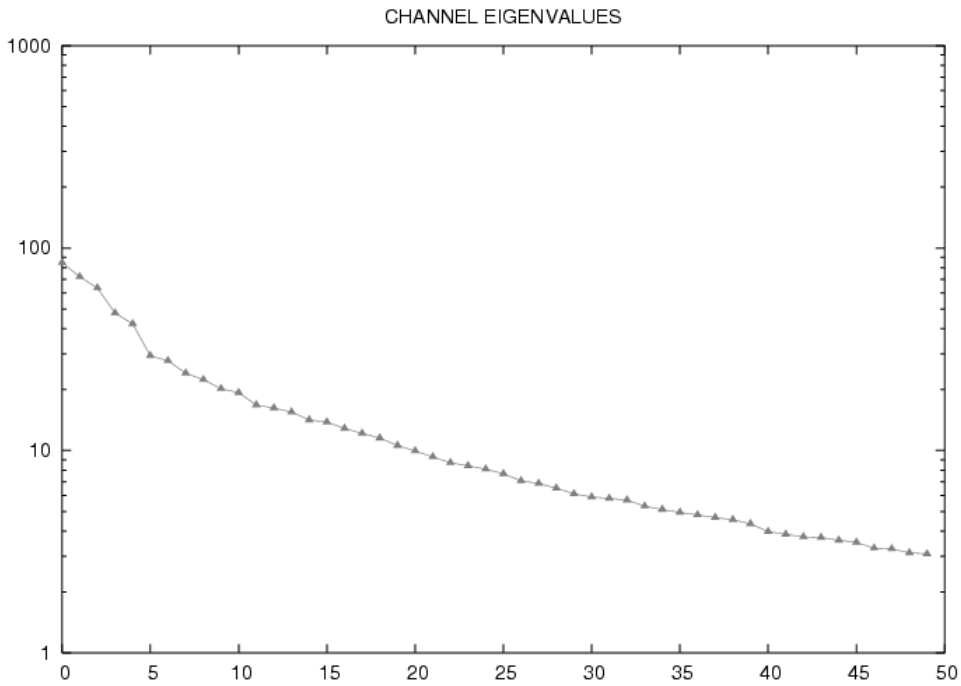
Components of x are *channel factors*

channel space = range (uu^*)

eigenchannels = eigenvectors of uu^*

Channel Variablity (50 eigenchannels)

$$\mathrm{tr}(uu^{*}) = 742$$



Likelihood Function for Speaker Recognition

Use the enrollment data to estimate supervector M for hypohtesized speaker s (factor analysis MAP)

$M' = \text{supervector for test recording}$

$$M' = M + ux$$

$$P(\text{test data} \mid s) = \int P(\text{test data} \mid M' = M + ux) N(x \mid 0, I) dx$$

Model Configuration (official submission)

- Gender-dependent
- 2,048 Gaussians per GMM
- 25 speaker factors, 50 channel factors
- 12 cepstral coefficients + log energy
- Feature warping
- Silence detection with .ctm files
- Only attempted core condition
- T-norm

Various Model Configurations (2048 Gaussians, T-Norm)

Eigen Voices	Eigen Channels	EER	DCF
25	50	11.7%	0.042
5	50	11.7%	0.036
5	**25**	**10.2%**	**0.036**
0	25	11.7%	0.038

T-Norm, Z-Norm and ZT-norm
5 eigenvoices, 25 eigenchannels

	EER	**DCF**
T- NORM	10.2%	0.036
Z-NORM	13.8%	0.047
ZT-NORM	**9.5%**	**0.034**

T-Norm, Z-Norm and ZT-norm
0 eigenvoices, 25 eigenchannels

	EER	**DCF**
T- NORM	11.7%	0.038
Z-NORM	9.9%	0.034
ZT-NORM	**8.7%**	**0.029**

Factor analsysis beats eigenchannels (zt norm)

	EER	DCF
Forward	6.9%	0.022
Reverse	6.4%	0.023
Combined	**6.6%**	**0.021**

Conclusions

- It seems to have been a mistake to use eigenchannel modeling
  - Factor analysis harder but more powerful
- Could we please have some more data?
  - Eg. Impossible to use these methods on auxiliary microphone data
  - How about a **training set** as well as development and test sets**?**