CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Low-dimensional speech representation based on Factor Analysis and its applications

Najim Dehak and Stephen Shum

Spoken Language System Group

MIT Computer Science and Artificial Intelligence Laboratory

(With a lot of help from Douglas Reynolds, Reda Dehak, Zahi Karam, Pedro Torres-Carrasquillo)

Goals

- **Aim**
  - To provide an overview of theory and operation of modern low-dimensional speech representations and their application to automatic speaker, language, emotion recognition and diarization

- **Participants should gain an introduction to and understanding of:**
  - *Subspace Representation of Speech Signals*
  - *Algorithms for Joint-Factor Analysis and Total-Variability Modeling*
  - *Application of subspace representations to automatic speaker and language recognition systems*

Roadmap

- **Introduction**
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- **Applications**
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

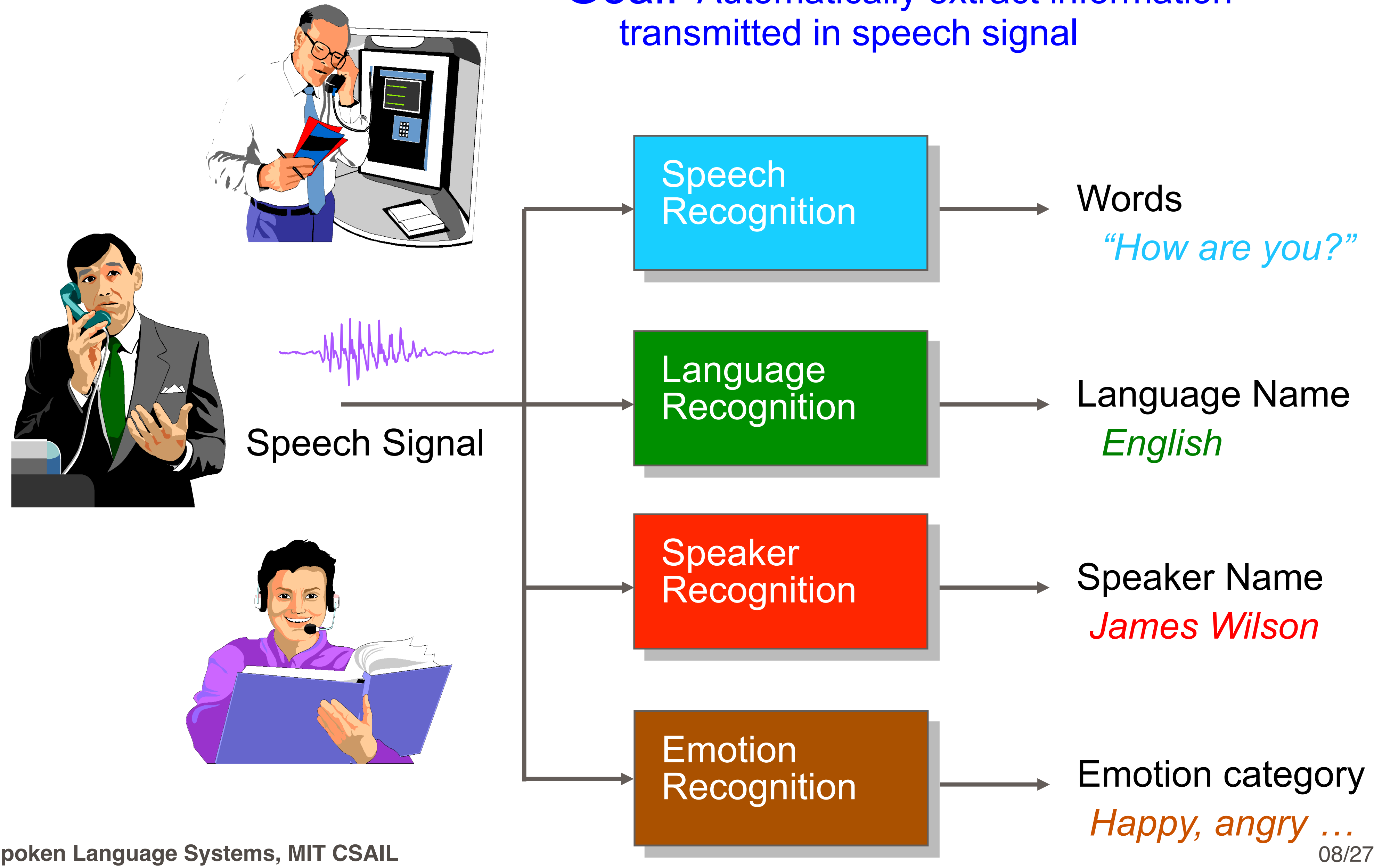
Roadmap

- **Introduction**
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- **Applications**
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

Extracting Information from Speech

CSAIL

Goal: Automatically extract information transmitted in speech signal



Speech Signal

Speech Recognition

Words

"How are you?"

Language Recognition

Language Name

English

Speaker Recognition

Speaker Name

James Wilson

Emotion Recognition

Emotion category

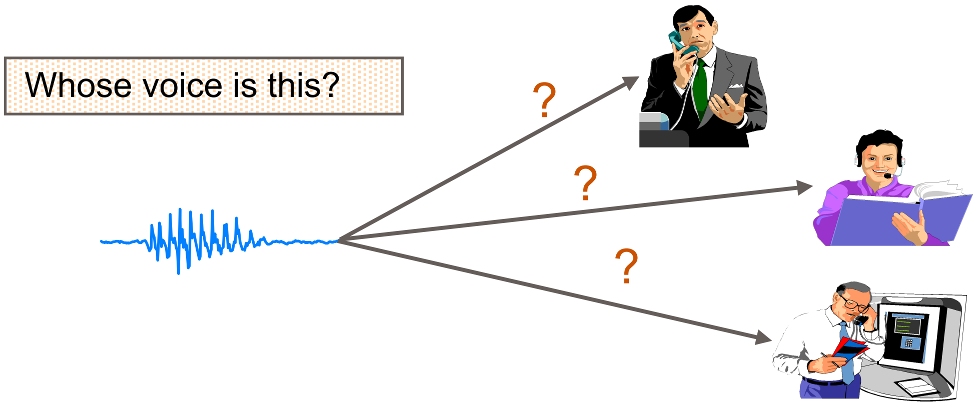
Happy, angry …

Identification

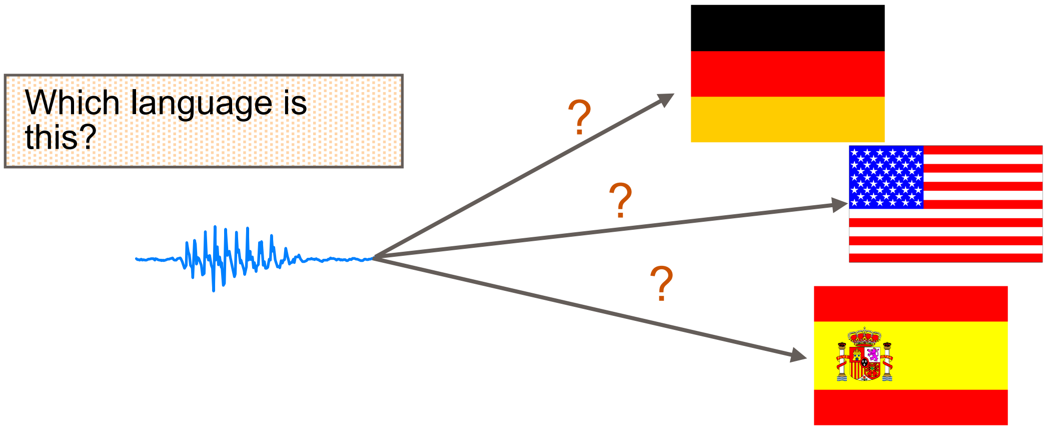
- Determine whether unknown speaker (language) matches one of a set of known speakers (languages)

- One-to-many mapping

- Often assumed that unknown voice must come from a set of known speakers – referred to as **closed-set** identification



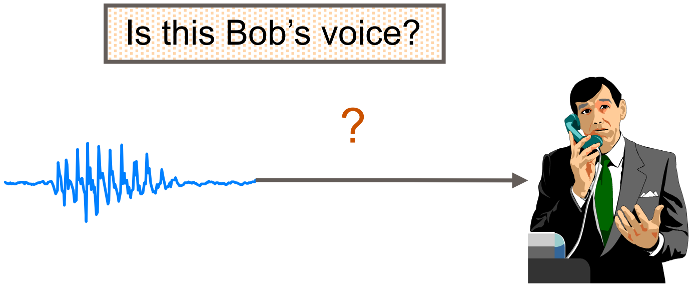
Whose voice is this?

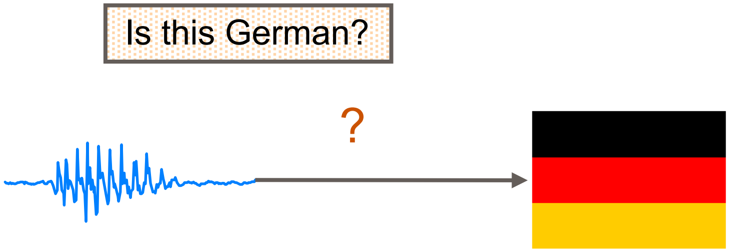


Which language is this?

Verification/Authentication/Detection

- Determine whether unknown speaker (language) matches a specific speaker (language)
- One-to-one mapping
- Unknown speech could come from a large set of unknown speakers (languages) – referred to as open-set verification
- Adding "none of the above" option to closed-set identification gives open-set identification

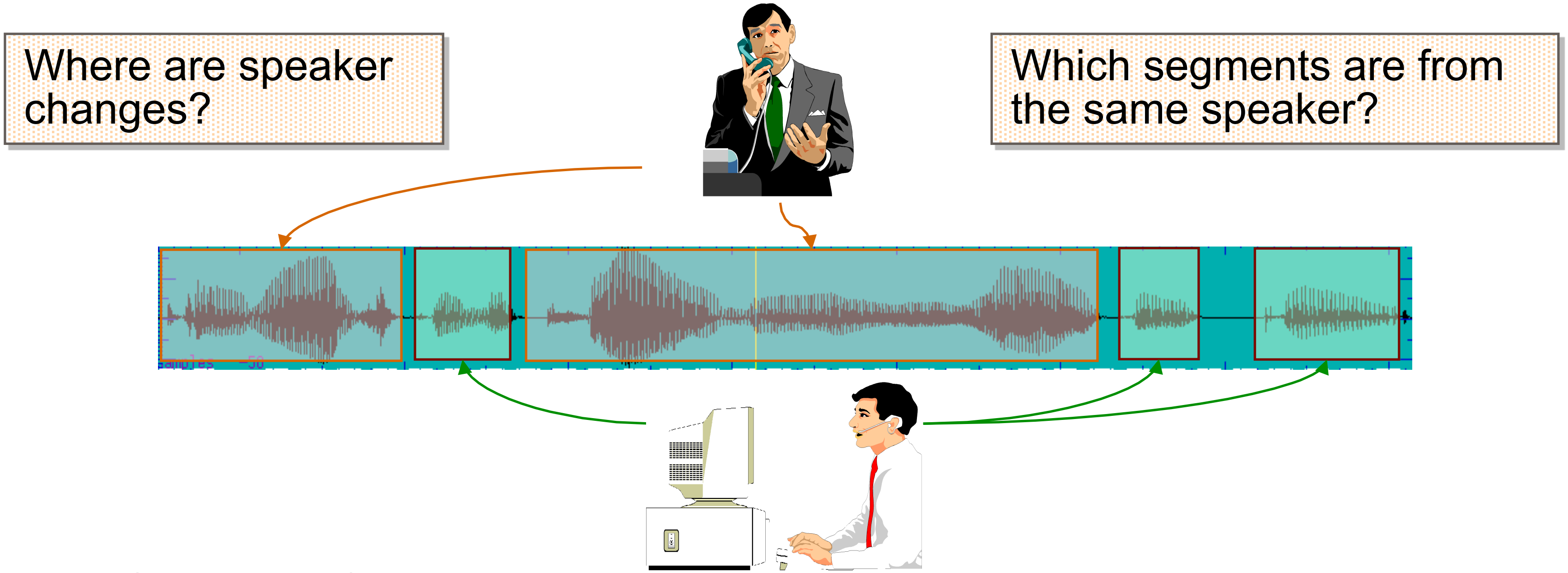

Is this Bob's voice?


Is this German?

Diarization

Segmentation and Clustering

- Determine when a speaker change has occurred in the speech signal (segmentation)
- Group together speech segments corresponding to the same speaker (clustering)
- Prior speaker information may or may not be available



Where are speaker changes?

Which segments are from the same speaker?

Speech Modalities

Application dictates different speech modalities:

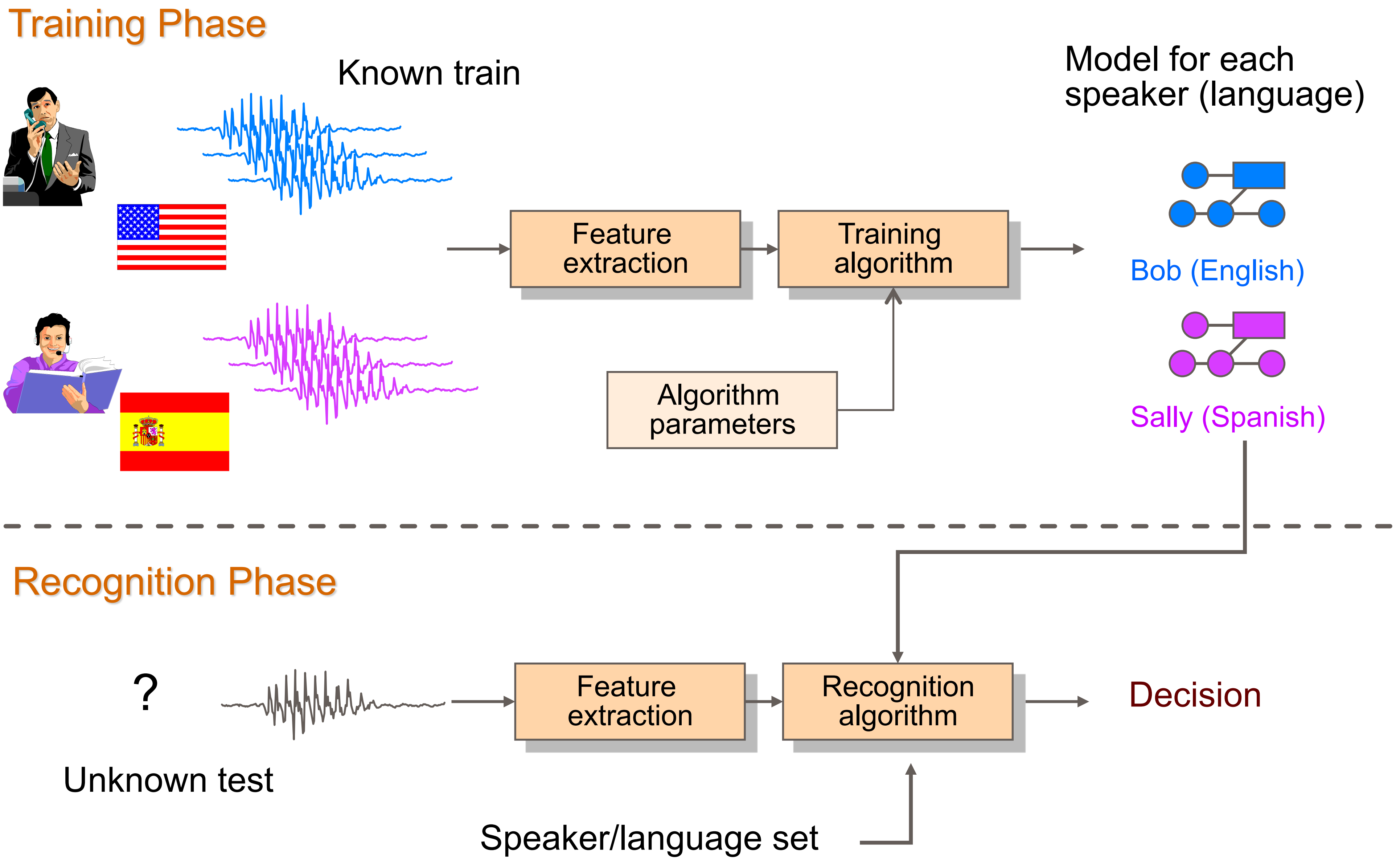
Text-dependent

Text-independent

- Recognition system knows text spoken by person
- Examples: fixed phrase, prompted phrase
- Used for applications with strong control over user input
- Knowledge of spoken text can improve system performance

- Recognition system does not know text spoken by person
- Examples: User selected phrase, conversational speech
- Used for applications with less control over user input
- More flexible system but also more difficult problem
- Speech recognition can provide knowledge of spoken text

Framework for Speaker/Language Recognition Systems



Training Phase

Known train

Model for each speaker (language)

Feature extraction

Training algorithm

Bob (English)

Algorithm parameters

Sally (Spanish)

Recognition Phase

?

Feature extraction

Recognition algorithm

Decision

Unknown test

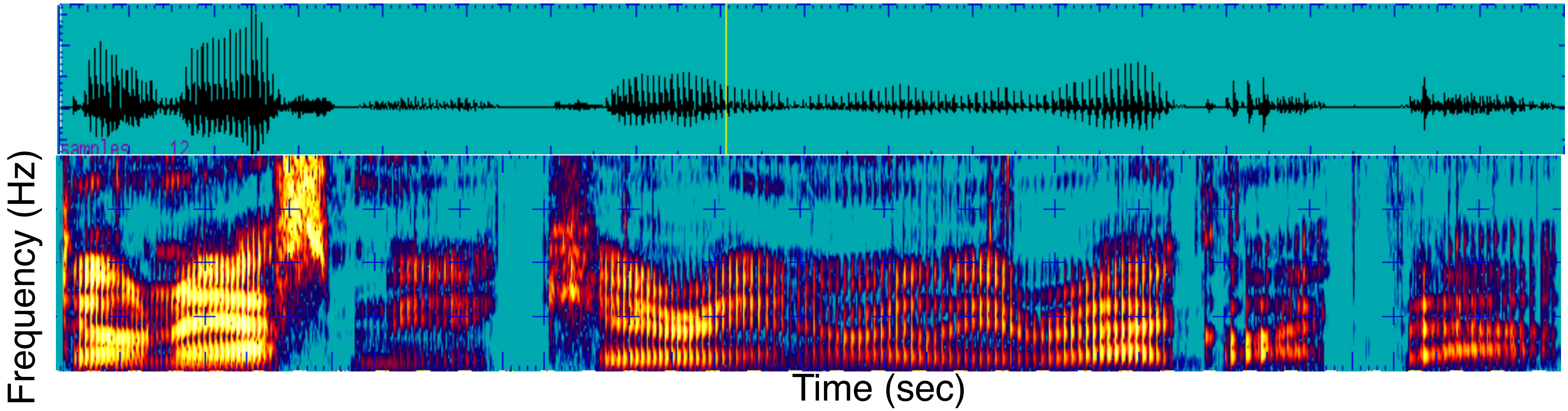
Speaker/language set

Roadmap

- Introduction
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

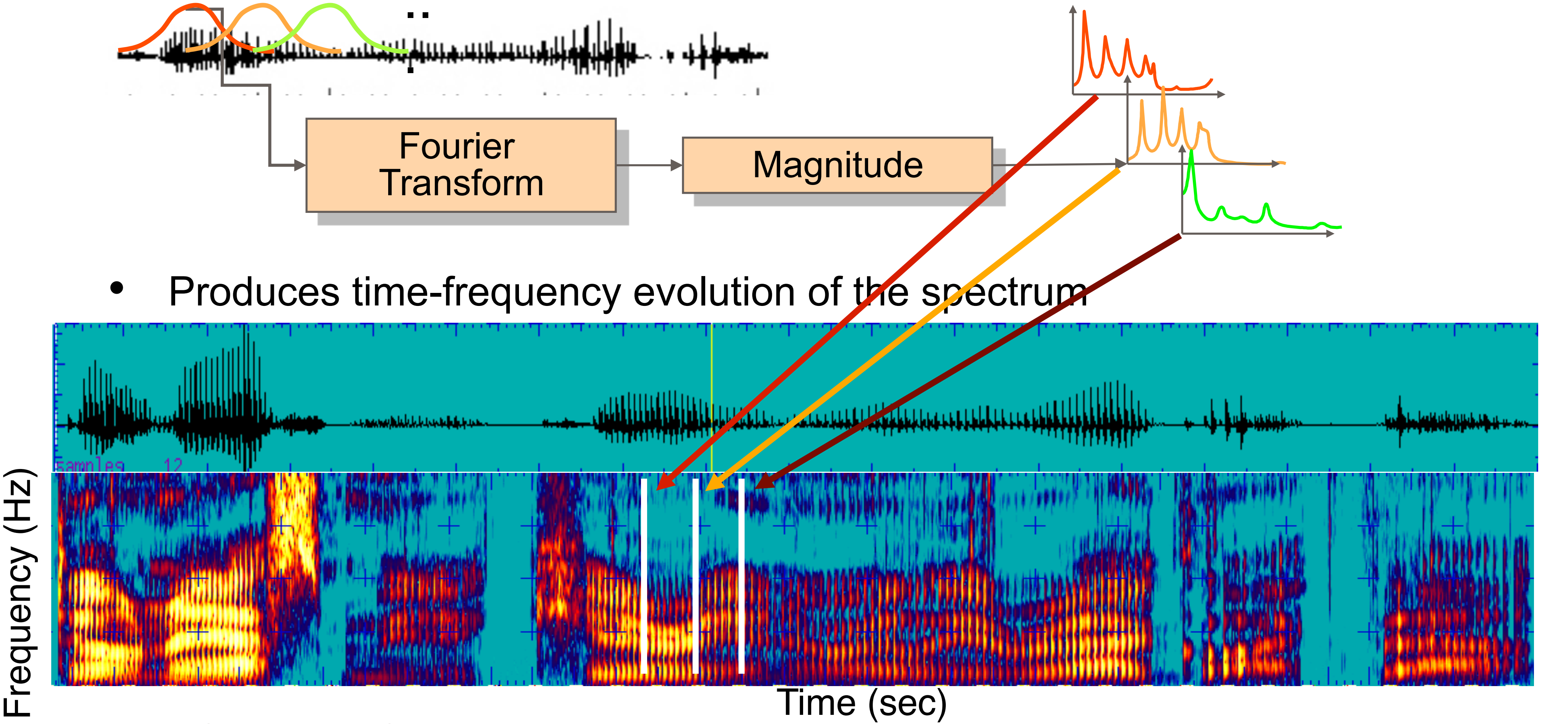
Information in Speech

- **Speech is a time-varying signal conveying multiple layers of information**
  - Words
  - Speaker
  - Language
  - Emotion
- **Information in speech is observed in the time and frequency domains**



Feature Extraction from Speech

- **A time sequence of features is needed to capture speech information**
  - Typically some spectra based features are extracted using sliding window - 20 ms window, 10 ms shift

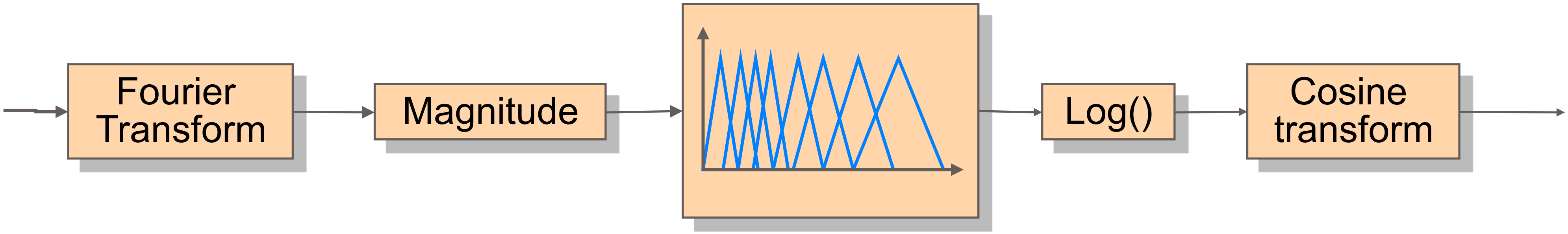


Fourier Transform

Magnitude

- Produces time-frequency evolution of the spectrum

Cepstral Features



Fourier Transform

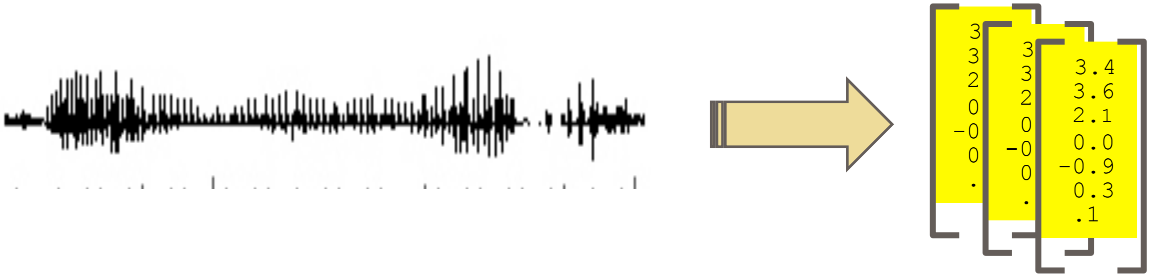
Magnitude

Log()

Cosine transform

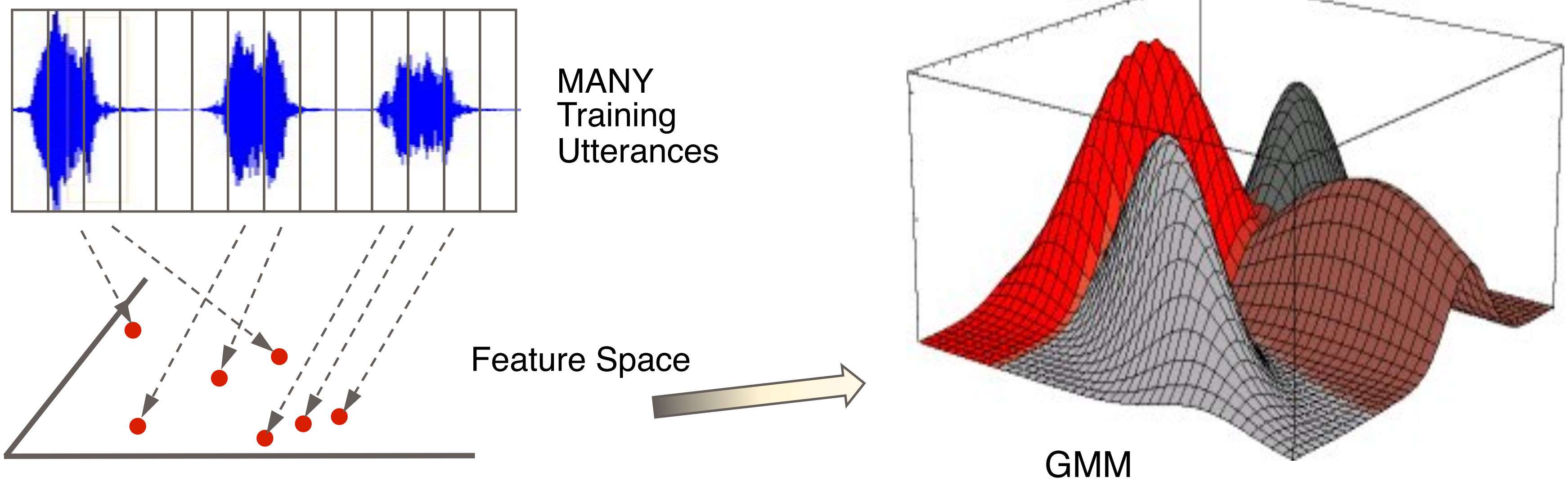
Modeling Sequence of Features
Gaussian Mixture Models

- **For most recognition tasks, we need to model the distribution of feature vector sequences**



100 vec/sec

- In practice, we often use Gaussian Mixture Models (GMMs).



Signal Space

MANY Training Utterances

Feature Space

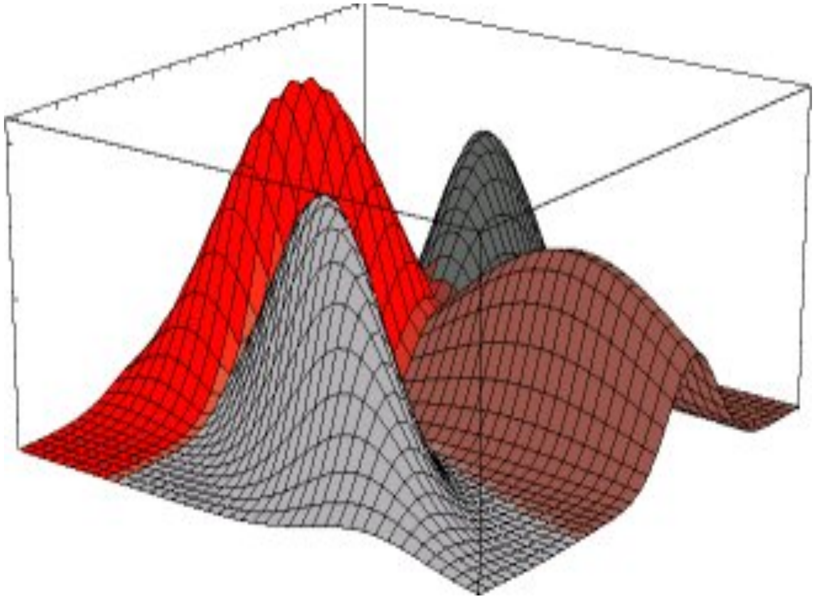
GMM

Gaussian Mixture Models

- A GMM is a weighted sum of Gaussian distributions

$$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

$$\lambda_s = (p_i, \vec{\mu}_i, \Sigma_i)$$



p_i = mixture weight (Gaussian prior proability)

$\vec{\mu}_i$ = mixture mean vector

Σ_i = mixture covariance matrix

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i\right|^{1/2}} \exp(-\tfrac{1}{2}(\vec{x} - \vec{\mu}_i)'\Sigma_i^{-1}(\vec{x} - \vec{\mu}_i))$$

Gaussian Mixture Models

Log Likelihood

- **To use a GMM, we need to do two things**

  1 – Compute the likelihood of a sequence of features given a GMM

  2 – Estimate the parameters of a GMM given a set of feature vectors

- **If we assume independence between feature vectors in a sequence, then we can compute the likelihood as**

$$p(\vec{x}_1,\ldots,\vec{x}_N \mid \lambda) = \prod_{n=1}^{N} p(\vec{x}_n \mid \lambda)$$

- **Usually written as log likelihood**

$$\log p(\vec{x}_1,\ldots,\vec{x}_N \mid \lambda) = \sum_{n=1}^{N} \log p(\vec{x}_n \mid \lambda)$$

$$= \sum_{n=1}^{N} \log\left(\sum_{i=1}^{\mathrm{M}} p_i b_i(\vec{x}_n) \right)$$

Gaussian Mixture Models

Parameter Estimation

- GMM parameters are estimated by maximizing the likelihood of on a set of training vectors

$$\lambda^{*} = \arg\max_{\lambda} \sum_{n=1}^{N} \log p(\vec{x}_n \mid \lambda)$$

- Setting the derivatives with respect to model parameters to zero and solving

$$\Pr(i \mid \vec{x}) = \frac{p_i b_i(\vec{x})}{\sum_{j=1}^{\mathrm{M}} p_j b_j(\vec{x})}$$

$$n_i = \sum\nolimits_{n=1}^{N} \Pr(i \mid \vec{x}_t)$$

$$p_i = \frac{1}{N} \sum\nolimits_{n=1}^{N} \Pr(i \mid \vec{x}_n)$$

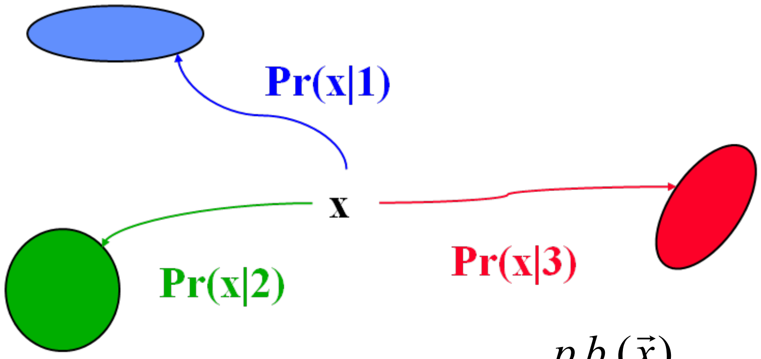
$$\vec{\mu}_i = \frac{1}{n_i} \sum\nolimits_{n=1}^{N} \Pr(i \mid \vec{x}_n)\, \vec{x}_n$$

$$\Sigma_i = \frac{1}{n_i} \sum\nolimits_{n=1}^{N} \Pr(i \mid \vec{x}_n)\, \vec{x}_i \vec{x}_i\,' - \vec{\mu}_i \vec{\mu}_i\,'$$

Gaussian Mixture Models
Expectation Maximization (EM)

E-Step

Probabilistically align vectors to model



$$\Pr(i \mid \vec{x}) = \frac{p_i b_i(\vec{x})}{\sum_{j=1}^{M} p_j b_j(\vec{x})}$$

Accumulate sufficient statistics

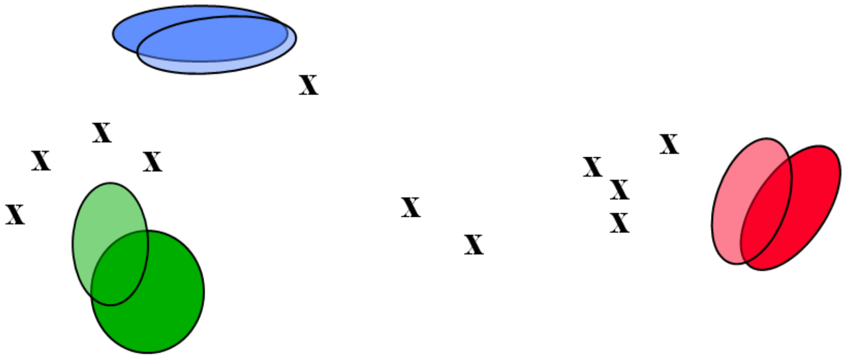
$$n_i = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n)$$

$$E_i(\vec{x}) = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n)\, \vec{x}_n$$

$$E_i(\vec{x}\vec{x}') = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n)\, \vec{x}_n \vec{x}_n'$$

M-Step

Update model parameters



$$p_i = \frac{1}{N} n_i$$

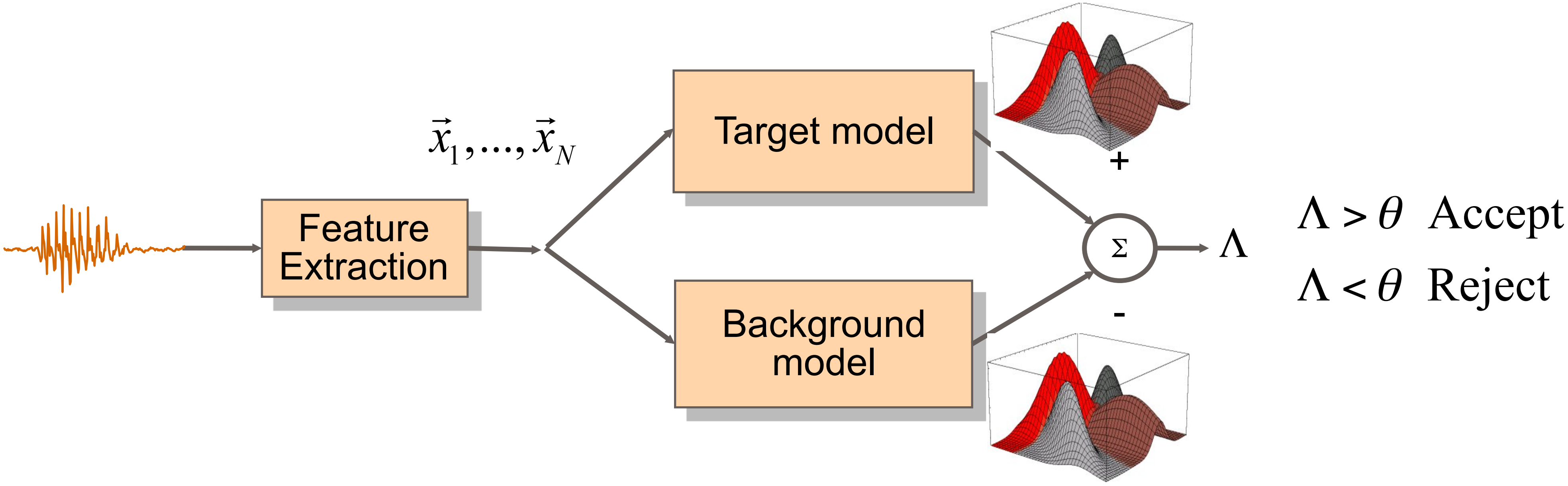
$$\vec{\mu}_i = \frac{1}{n_i} E_i(\vec{x})$$

$$\Sigma_i = \frac{1}{n_i} E_i(\vec{x}\vec{x}') - \vec{\mu}_i \vec{\mu}_i'$$

Detection System
GMM-UBM

- Realization of log-likelihood ratio test from signal detection theory

$$LLR = \Lambda = \log p(X \mid \text{target}) - \log p(X \mid \overline{\text{target}})$$



$\vec{x}_1, \ldots, \vec{x}_N$
Feature Extraction
Target model
Background model
+
-
Σ
Λ
$\Lambda > \theta$ Accept
$\Lambda < \theta$ Reject

- GMMs used for both target and background model
  - Target model trained using enrollment speech
  - Background model trained using speech from many speakers (often referred to as Universal Background Model – UBM)

MAP Adaptation

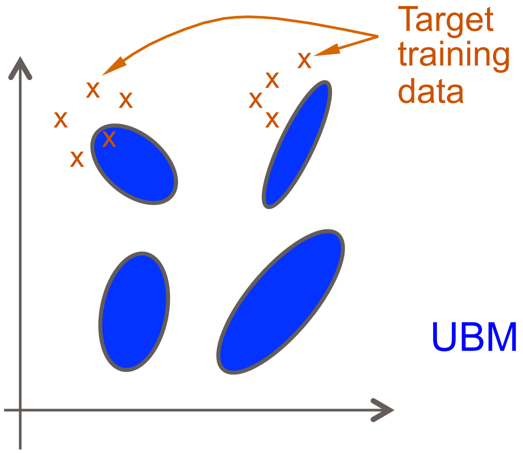
- **Target model is often trained by adapting from background model**
  - Couples models together and helps with limited target training data
- **Maximum A-Posteriori (MAP) training (similar to EM)**
  - Align target training vectors to UBM
  - Accumulate sufficient statistics
  - Update target model parameters with smoothing to UBM parameters
- **Adaptation only updates parameters representing acoustic events seen in target training data**
  - Sparse regions of feature space filled in by UBM parameters
- **Side benefits**
  - Keeps correspondence between target and UBM mixtures (important later)
  - Allows for fast scoring when using many target models (top-M scoring)

Adapted GMMs

Mean-only adaptation

- Probabilistically align target training data into UBM mixture states

$$\Pr(i \mid \vec{x}) = \frac{p_i b_i(\vec{x})}{\sum_{j=1}^{M} p_j b_j(\vec{x})}$$



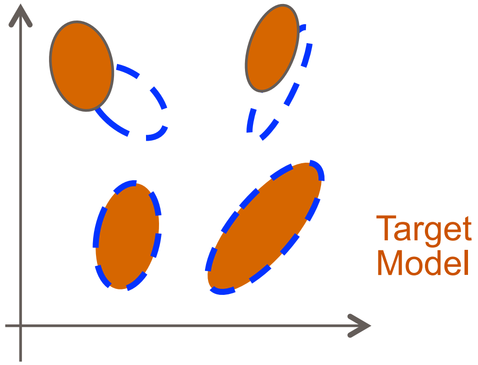
- Accumulate sufficient statistics from probabilistic alignment
  - Mean-only adaptation empirically found to be better

$$n_i = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n)$$

$$E_i(\vec{x}) = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n)\, \vec{x}_n$$

- Update target model parameters using sufficient statistics and adapt parameter (α)
  - *Relevance factor r controls rate of adaptation*
  - *r → 0, MAP → EM*
  - *r → ∞. No adaptation*

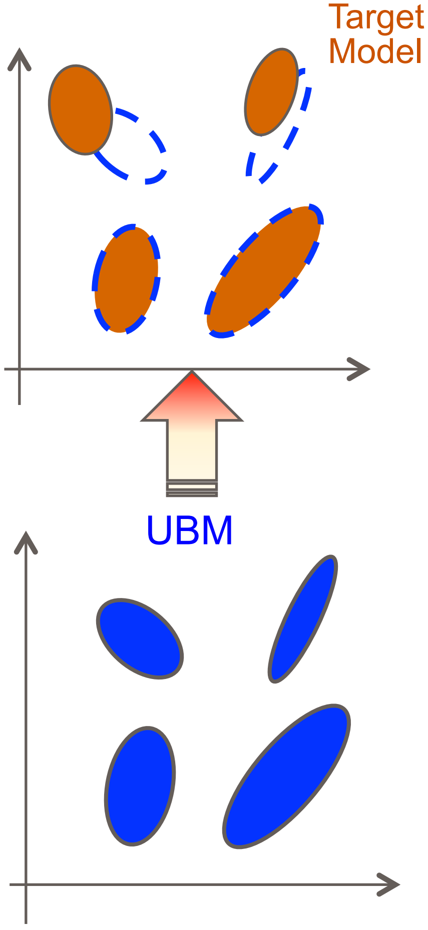
$$\alpha_i = \frac{n_i}{n_i + r}$$



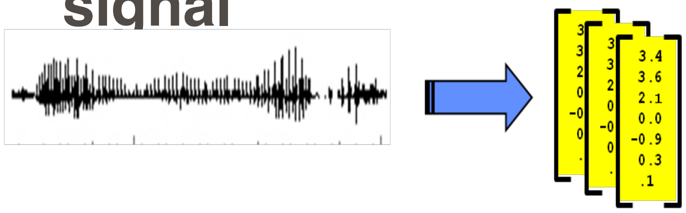
$$\vec{\mu}_i = \alpha_I E_i(\vec{x}) + (1 - \alpha_I)\vec{\mu}_i^{ubm}$$

GMM-UBM Recap

(3) Adapt target model from UBM



(1) Extract feature vector sequence from speech signal



(4) Compute likelihood ratio of test data

$$LLR(X) = \log p(X \mid \lambda_{target}) - \log p(X \mid \lambda_{ubm})$$

(2) Train UBM with speech from many speakers using EM

Another View of log-likelihoods

- We can use sufficient statistics to score a GMM…

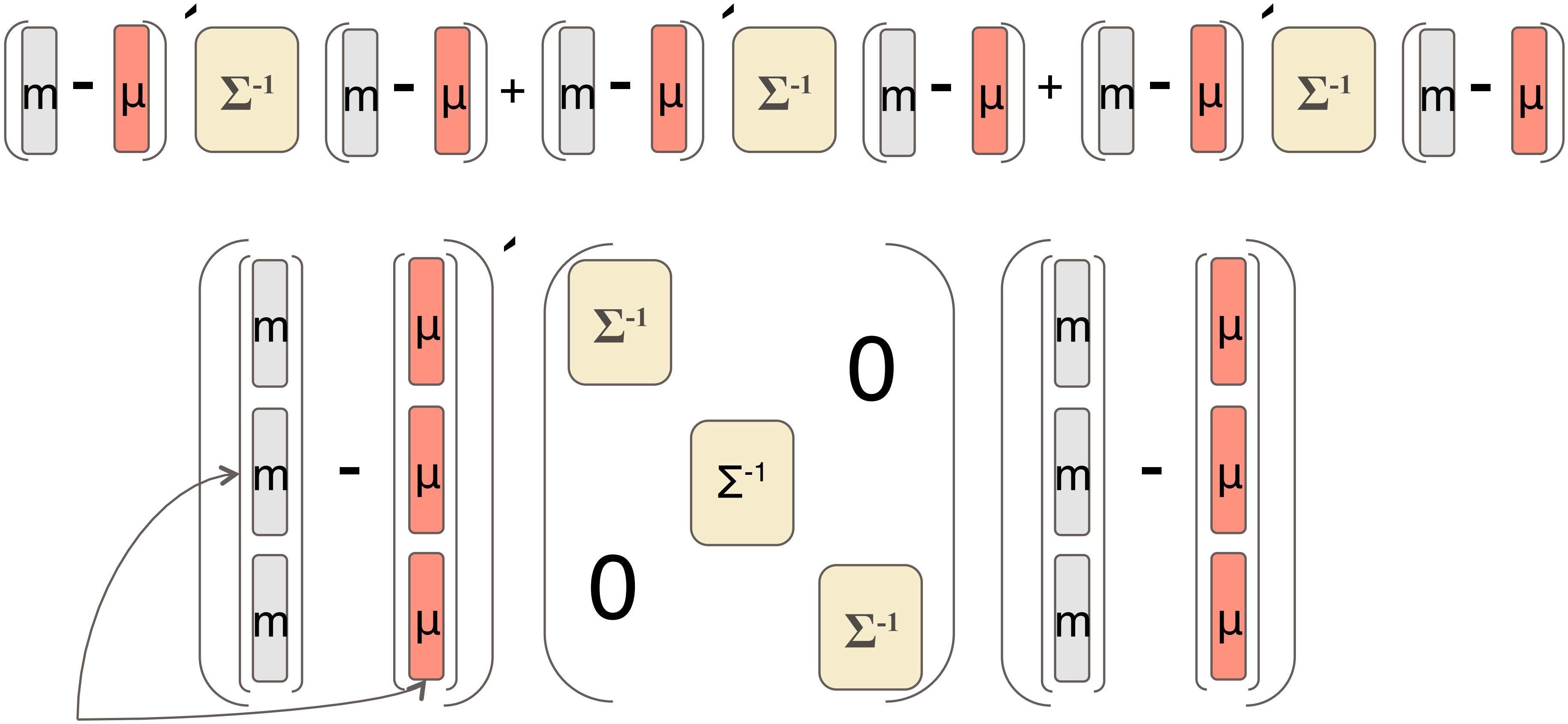
$$n_i = \sum_{n=1}^{N} \Pr(i \mid \vec{x}_n) \qquad \vec{m}_i = E_i\left[\vec{x}\right]/n_i \qquad S_i = E_i(\vec{x}\vec{x}')/n_i$$

- Since we adapt only the UBM means

$$E\left[\log p(X, I \mid \lambda)\right] = -\frac{1}{2}\sum_{i=1}^{M}(\vec{m}_i - \vec{\mu}_i)'\left(\frac{\Sigma_i^{ubm}}{n_i}\right)^{-1}(\vec{m}_i - \vec{\mu}_i) + C_i$$

Supervectors

- By stacking vectors and matrices, we can work directly with vector-matrix manipulations



$$(m-\mu)'\,\Sigma^{-1}\,(m-\mu) + (m-\mu)'\,\Sigma^{-1}\,(m-\mu) + (m-\mu)'\,\Sigma^{-1}\,(m-\mu)$$

$$\left(\begin{pmatrix} m \\ m \\ m \end{pmatrix} - \begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix}\right)' \begin{pmatrix} \Sigma^{-1} & & 0 \\ & \Sigma^{-1} & \\ 0 & & \Sigma^{-1} \end{pmatrix} \left(\begin{pmatrix} m \\ m \\ m \end{pmatrix} - \begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix}\right)$$

Super-vectors

MAP Reformulated

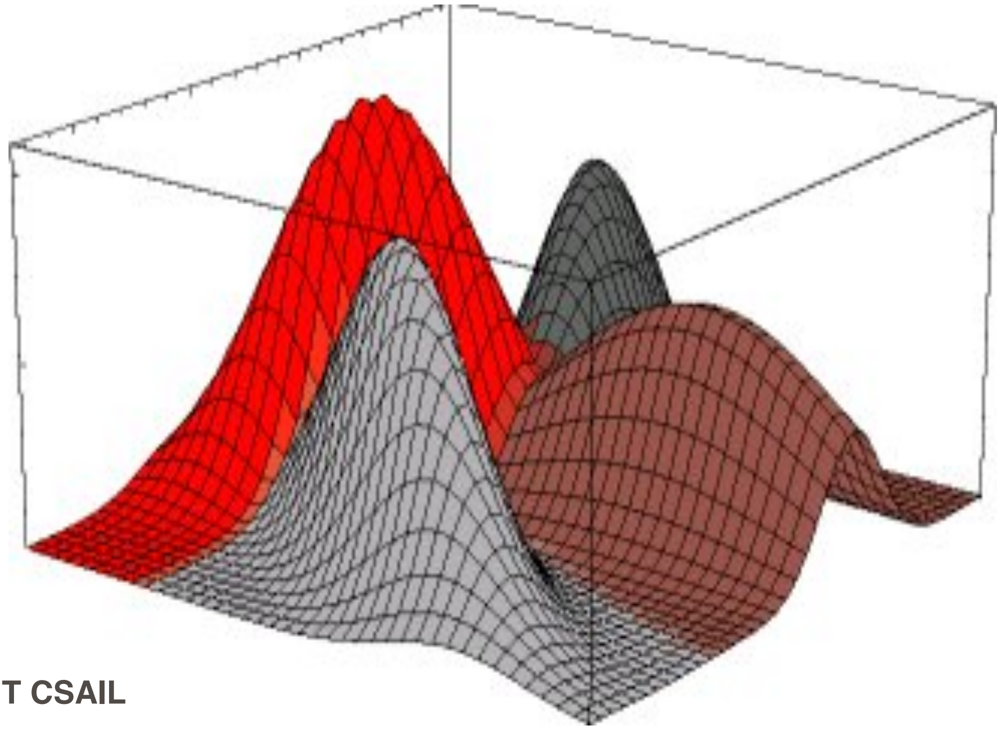
- **New design for MAP adaptation based on Factor Analysis**
- $\boldsymbol{M = m + Dz}$
  - $\boldsymbol{M}$: speaker and channel dependent supervector
  - $\boldsymbol{m}$: speaker and channel independent supervector (UBM)
  - $\boldsymbol{d}$: diagonal matrix
  - $\boldsymbol{z}$ **:** random vectors with a standard normal prior

- $\boldsymbol{M}$ **is normally distributed with mean** $\boldsymbol{m}$ **and covariance** $\boldsymbol{D^2}$**.**
- **Matrix** $\boldsymbol{D}$ **can be trained via maximum likelihood.**

- **If we let** $D^2 = {}^{1}\!/_{r} \cdot \Sigma$
  - then we have the equivalent of Relevance MAP adaptation.

Intuition

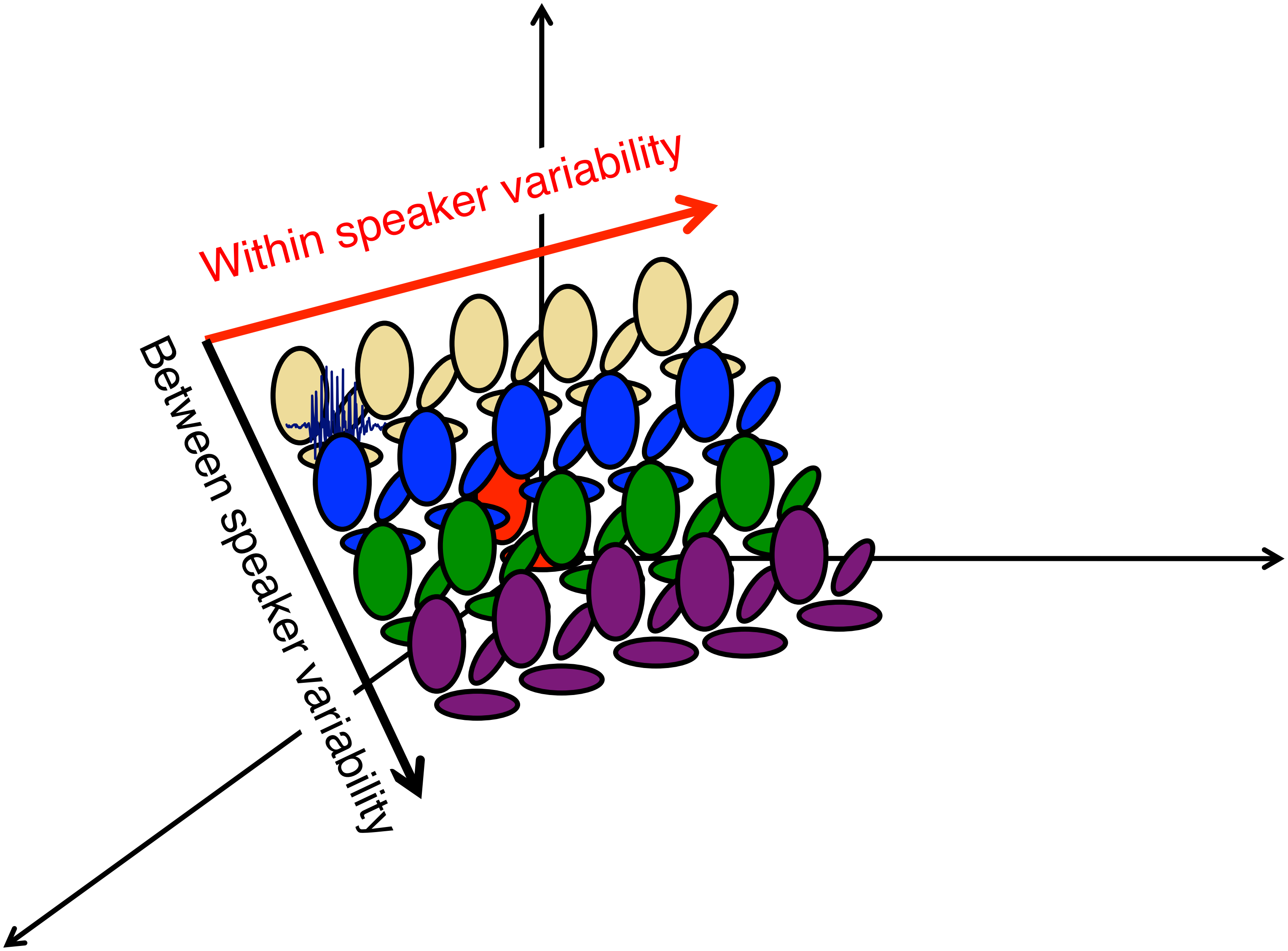
- **The way the UBM adapts to a given speaker ought to be somewhat constrained**
  - There should exist some relationship in the way the mean parameters move relative to speaker to another
  - The Joint Factor Analysis [Kenny 2008] explored this relationship
    - **Jointly model between- and within-speaker variabilities**
  - Support Vector Machine GMM supervector [Campbell 2006]



Roadmap

- Introduction
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

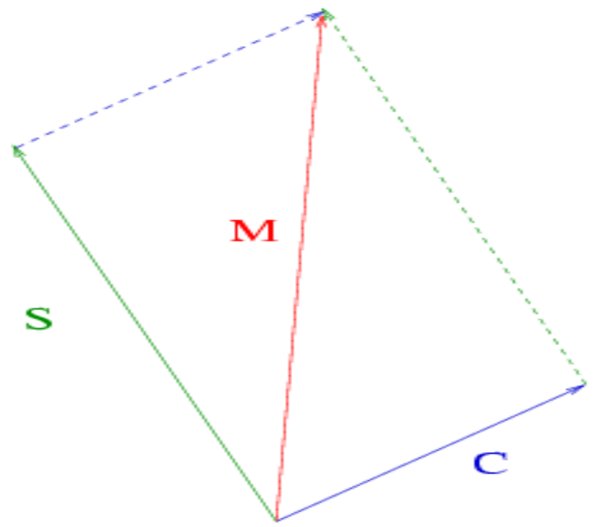
Joint Factor Analysis



Joint Factor Analysis

- Proposed for the GMM frameworks
- Assumption [Kenny2008]

$$M=s+c$$



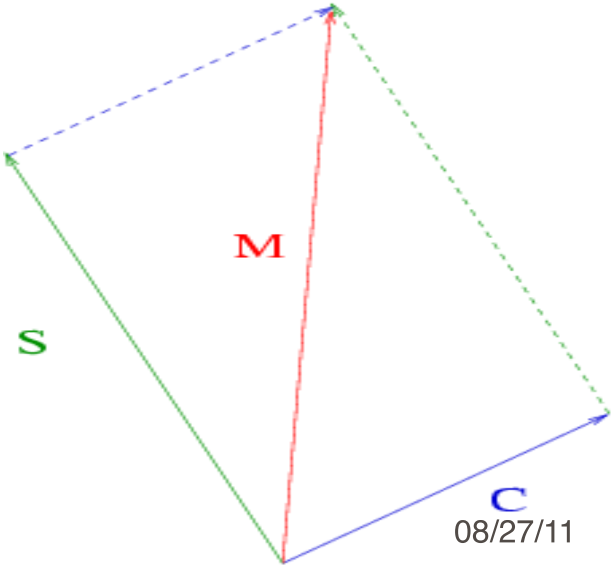
- M : speaker and channel dependent supervector
- s : speaker dependent supervector
- c : channel dependent supervector
  - GMM supervector : concatenation of the means components

Joint Factor Analysis – Details

- $s = m + Vy + Dz$
  - m : speaker- and channel-independent supervector
  - V : rectangular matrix of low-rank (eigenvoices- speaker space)
  - D : diagonal matrix
  - y, z : random vectors with a standard normal prior

- $c = Ux$
  - U : rectangular matrix of low rank (eigenchannels- channel space)
  - x : random vector with standard normal prior

$$M = m + Vy + Dz + Ux$$



Joint Factor Analysis: scoring

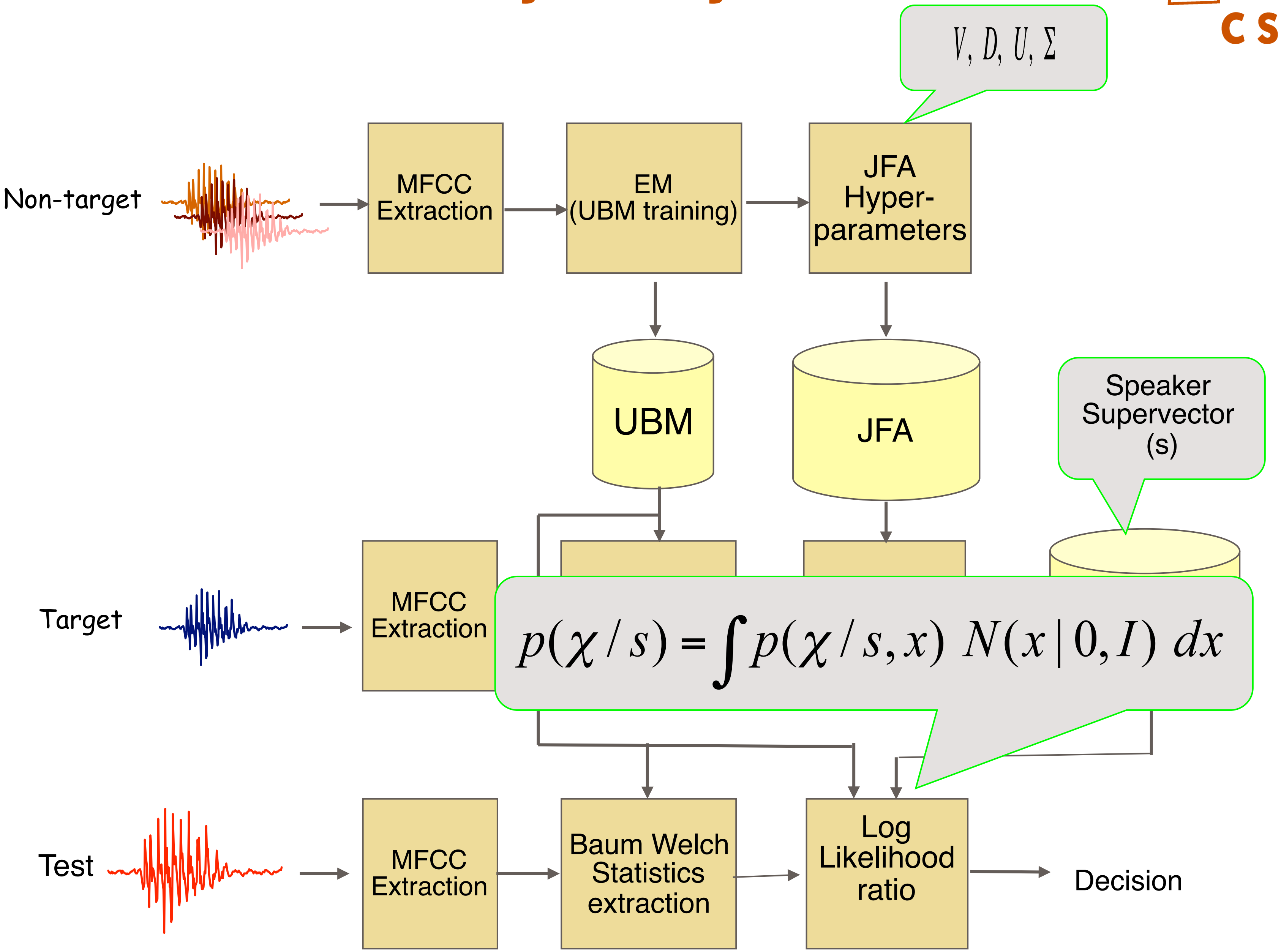
- Likelihood [Kenny2008]

$$p(\chi / s) = \int p(\chi / s, x)\ N(x \mid 0, I)\ dx$$

- Log likelihood ratio

$$score = \ln \frac{p(\chi / s)}{p(\chi / \Omega)} \quad \begin{matrix} \geq \\ < \end{matrix} \quad \theta$$

Joint Factor Analysis System



V, D, U, Σ

Speaker Supervector (s)

$$p(\chi / s) = \int p(\chi / s, x)\ N(x \mid 0, I)\ dx$$

Roadmap

- Introduction
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

The story begins…

Johns Hopkins University
The Center for Language and Speech Processing
2008

Channel factors
Contain speaker information

The importance of each factors v JFA

$$M= m + Vy + Dz + Ux$$

Total variability

- Factor analysis as feature extractor
- Joint factor analysis

$$\mathbf{M}= \boldsymbol{m} + \boldsymbol{Vy} + \boldsymbol{Dz} + \boldsymbol{Ux}$$

- Speaker and channel dependent supervector

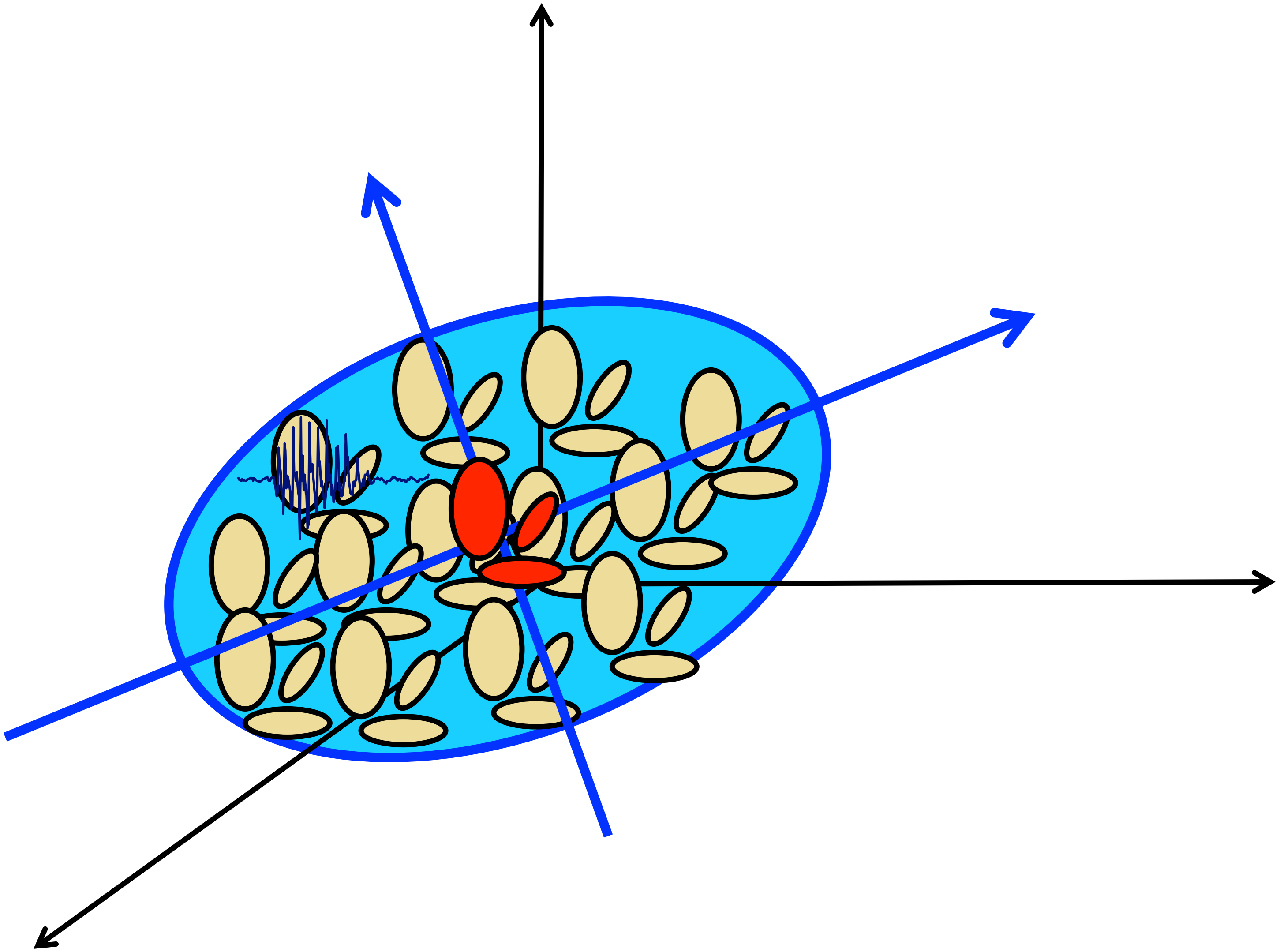
$$\mathbf{M} = \boldsymbol{m} + \boldsymbol{Tw}$$

  – T is rectangular, low rank (total variability matrix)
  – w standard Normal random (total factors – intermediate vector or i-vector)
- Estimate the i-vector: contain both variabilities
- **No distinction between speaker V and channel U variabilities**
- We will apply channel compensation later in the i-vector space.

Najim Dehak, Patrick Kenny, Pierre Dumouchel, Reda Dehak, Pierre Ouellet, «Front-end factor analysis for speaker verification » in IEEE Transactions on Audio, speech and Language Processing 2011.

Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet and Pierre Dumouchel, Support Vector Machine versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In Proc INTERSPEECH 2009, Brighton, UK, September 2009.

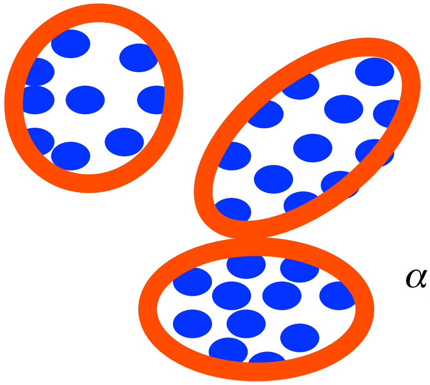
Total variability space



Why call it an i-vector?

$\alpha_1, \mu_1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix}, \Sigma_1$

$\alpha_2, \mu_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix}, \Sigma_2$



$\alpha_3, \mu_3 = \begin{bmatrix} \mu_{31} \\ \mu_{32} \end{bmatrix}, \Sigma_3$

GMM components: 2048
Feature dimension: 60

GMM-SV :
60*2048=122880

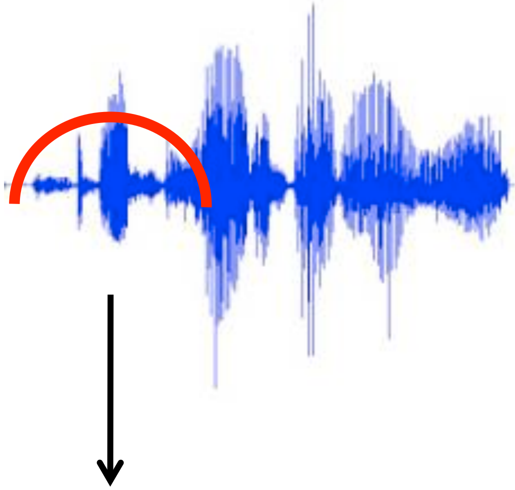
$$\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{bmatrix}$$

It is definitely not an Apple product

I- for Intermediate representation

IVECTOR

Actually between 100 to 1000



MFCC

Feature dimension 60

Extracting the Hyperparameters

- Speaker and channel dependent supervector

$$\mathbf{M} = \boldsymbol{m} + \boldsymbol{Tw}$$

- The i-vector extractor is characterized by
  - m : A supervector mean (can be the UBM)
  - T : low rank Total variability matrix
  - Σ : diagonal covariance matrix
- **Preliminaries**
  - Acoustic observations $u = \left\{\vec{x}_1,\ldots,\vec{x}_L\right\}$
    - **Each y_t has dimension F**
  - Universal Background Model θ_{UBM}
    - **Number of Gaussian components C, indexed by c**
    - **→ Supervector dimension = CF**

Baum-Welch (Sufficient) Statistics

- Zeroth Order

$$N_c(u) = \sum_{t=1}^{L} P(c \mid \vec{x}_t, \theta_{\text{UBM}}) = \sum_t \gamma_t(c)$$

- First Order

$$F_c(u) = \sum_{t=1}^{L} P(c \mid \vec{x}_t, \theta_{\text{UBM}}) \cdot \vec{x}_t = \sum_t \gamma_t(c) \cdot \vec{x}_t$$

- Second Order

$$S_c(u) = \text{diag}\left(\sum_t \gamma_t(c) \cdot \vec{x}_t \vec{x}_t^{\,t} \right)$$

where $c = 1, \ldots, C$ for each UBM component

Simplified Notation I

- Recall

$$\gamma_t(c) = P(c \mid \vec{x}_t, \theta_{\mathrm{UBM}}) = \frac{\pi_c P_c(\vec{x}_t \mid \mu_c, \Sigma_c)}{\sum_{i=1}^{C} \pi_i P_i(\vec{x}_t \mid \mu_i, \Sigma_i)}$$

- Centralized First- / Second-Order Statistics

$$\tilde{F}_c(u) = \sum_t \gamma_t(c) \cdot (\vec{x}_t - m_c)$$

$$\tilde{S}_c(u) = \mathrm{diag}\left(\sum_t \gamma_t(c) \cdot (\vec{x}_t - m_c)(\vec{x}_t - m_c)^t \right)$$

$$m = \begin{bmatrix} m_1 & m_2 & \cdots & m_C \end{bmatrix}^t$$

Simplified Notation II

$$N(u) = \begin{bmatrix} N_1(u) \cdot I_{F \times F} & 0 & \cdots & 0 \\ 0 & N_2(u) \cdot I_{F \times F} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & N_C(u) \cdot I_{F \times F} \end{bmatrix}$$

$$\widetilde{F}(u) = \begin{bmatrix} \widetilde{F}_1(u) \\ \widetilde{F}_2(u) \\ \vdots \\ \widetilde{F}_C(u) \end{bmatrix} \qquad \widetilde{S}(u) = \begin{bmatrix} \widetilde{S}_1(u) & 0 & \cdots & 0 \\ 0 & \widetilde{S}_2(u) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \widetilde{S}_C(u) \end{bmatrix}$$

The EM Algorithm

- **Initialize m and Σ as defined by our UBM covariance matrices**
- **Pick a desired rank R for the Total Variability Matrix T and initialize this CF x R matrix randomly.**

- **E-step:**
  - For each utterance u, calculate the parameters of the posterior distribution of w(u) using the current estimates of m, T, Σ

- **M-step:**
  - Update T and by Σ solving a set of linear equations in which the w (u)'s play the role of explanatory variables

- **Iterate until parameters / data likelihood converges…**

Kenny, P., Boulianne, G. and P. Dumouchel. Eigenvoice Modeling with Sparse Training Data. IEEE Transactions on Speech and Audio Processing, 13 May (3) 2005 : 345-359.

E-step: The Posterior Distribution of w(u)

- For each utterance u, let l(u) be the matrix defined by

$$l(u) = I + T^t \Sigma^{-1} N(u) T$$

- Then the posterior distribution of w(u) conditioned on the acoustic observations of an utterance u is Gaussian with mean

$$E[w(u)] = l^{-1}(u) T^t \Sigma^{-1} \tilde{F}(u)$$

and covariance matrix

$$\mathrm{cov}(w(u), w(u)) = l^{-1}(u)$$

Kenny, P., Boulianne, G. and P. Dumouchel. Eigenvoice Modeling with Sparse Training Data. IEEE Transactions on Speech and Audio Processing, 13 May (3) 2005 : 345-359.

Proof Sketch

- To show this, let $E(u) = l^{-1}(u)T^t\Sigma^{-1}\tilde{F}(u)$

- Then it suffices to show that

$$P_{T,\Sigma}(w \mid u) \propto \exp\left(-\frac{1}{2}(w - E(u))^t l(u)(w - E(u))\right)$$

- First, just apply Bayes' Rule

$$P_{T,\Sigma}(w \mid u) \propto P_{T,\Sigma}(u \mid w) \cdot N(w \mid 0, I)$$

$$= P_{T,\Sigma}(\{\vec{x}_1, \ldots, \vec{x}_L\} \mid w) \cdot N(w \mid 0, I)$$

M-step: Maximum Likelihood Re-estimation 1/2

$$N_c = \sum_u N_c(u)$$

$$A_c = \sum_u N_c(u) E[w(u)w^t(u)]$$

$$C = \sum_u \tilde{F}(u) E[w^t(u)]$$

$$N = \sum_u N(u)$$

$u = 1, \ldots,$ number of utterances

c = 1,..., number of GMM components

$$E[w(u)w^t(u)] = Cov(w(u),w(u)) + E[w(u)].E[w(u)]^t$$

M-step: Maximum Likelihood Re-estimation 2/2

- Update matrix T

$$T(i,:)A_c = C_i$$
$$\text{where } i = (c-1) * D\_F + f$$

- Update the diagonal covariance matrix

$$\Sigma = N^{-1}(\sum_u \tilde{S}(u) - diag(CT^t))$$

$$f = 1,\ldots,\ \text{D\_F}$$
$$\text{D\_F} = \text{dimensionality of features vector}$$
$$\text{c} = 1,\ldots,\ \text{number of GMM components}$$

Roadmap

- Introduction
  - Terminology, tasks, and framework
- **Low-Dimensional Representation**
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

Scoring and channel Compensation

- **Cosine scoring**

$$score = \frac{< w_{t\arg et}, w_{test} >}{\left\| w_{t\arg et} \right\| . \left\| w_{test} \right\|}$$

- **Channel Compensation techniques**
  - – Linear Discriminant Analysis
  - – Within Class Covariance Normalization
  - – Nuisance Attribute projection

Scoring and channel Compensation

- Cosine scoring

$$score = \frac{< w_{t\arg et}, w_{test} >}{\left\| w_{t\arg et} \right\| . \left\| w_{test} \right\|}$$

- Channel Compensation techniques
  - Linear Discriminant Analysis
  - Within Class Covariance Normalization
  - Nuisance Attribute projection

Discussion

- **Low dimensional representation simplify life**
  - Mixture of Gaussians
  - Supervector of Gaussian mean components
  - Low-dimensional i-vector
- **Easy way to compare between sequences of features with different duration**
  - Less frames produce high uncertainty in the estimation of the i-vector
- **Classical pattern recognition approaches like LDA can be easily applied to maximize the discrimination between the different classes**

Roadmap

- Introduction
  - Terminology, tasks, and framework
- Low-Dimensional Representation
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

CSAIL

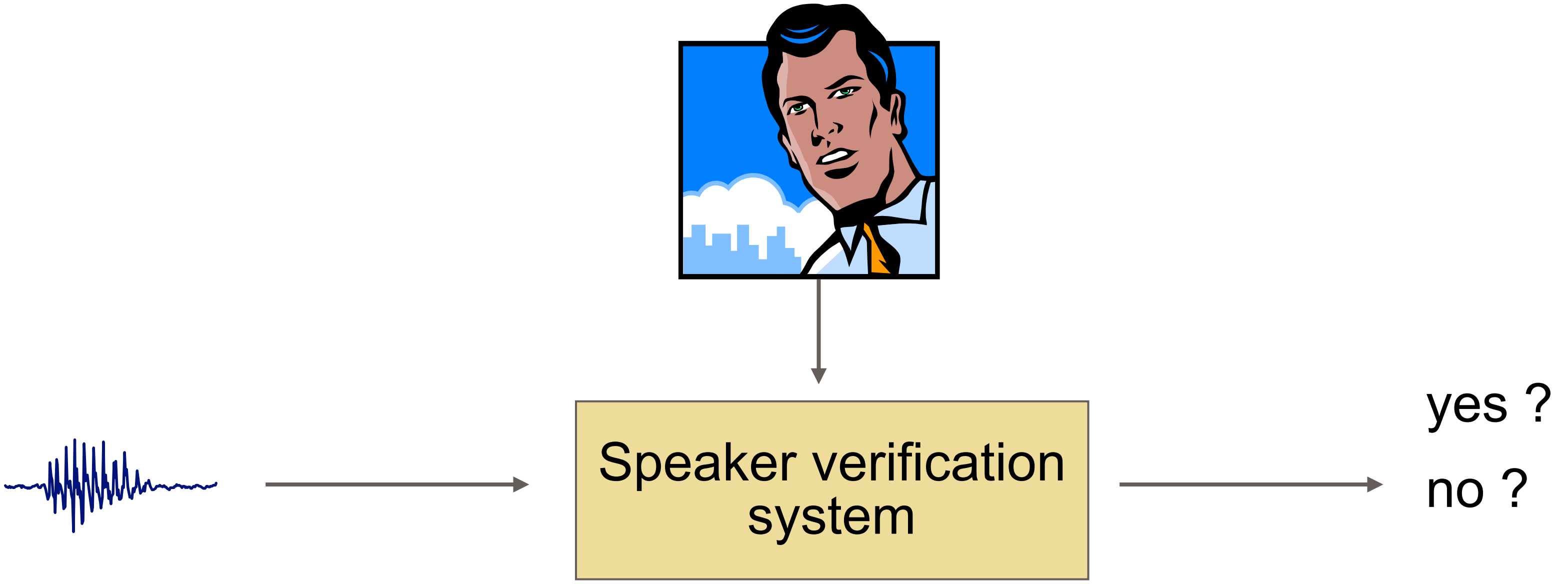
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Speaker Verification

Speaker Verifiation outline

- Speaker Recognition problem
- NIST Speaker Recognition evaluation
  - Performance evaluation metrics
- Feature extraction.
- I-vector for speaker verification
- Intersession compensation
- Experiments and results
- Data mismatch
- Data visualization
- Conclusion

Speaker Verification Problem



Speaker verification system

yes ?

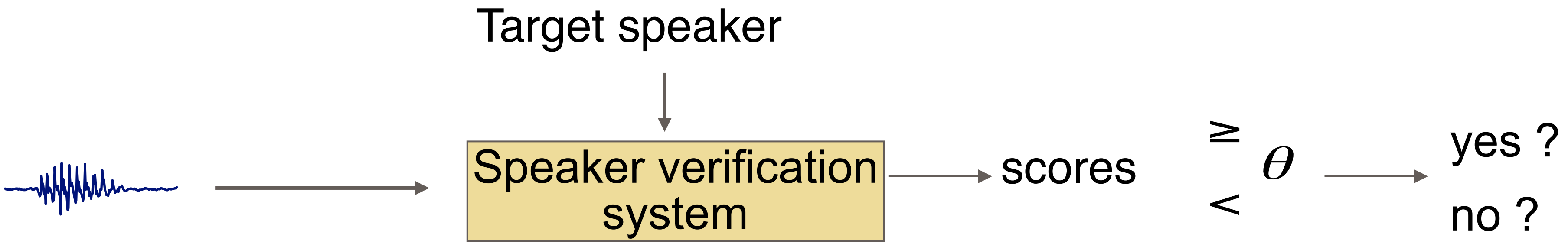
no ?

NIST
Speaker Recognition evaluation

- Several training and testing conditions (dependent on speech duration: 10sec, 1conv , 3conv,…)
- Telephone conversation and microphone data

		Test			
		10sec	1conv	1conv sum	1conv aux mic
Training	10sec	**Opt**			
	1conv	Opt	**Core**	Opt	Opt
	3conv	Opt	Opt	Opt	Opt
	8conv		Opt	Opt	Opt
	3conv sum		Opt	Opt	

Speaker verification system performances



Target speaker

Speaker verification system

scores

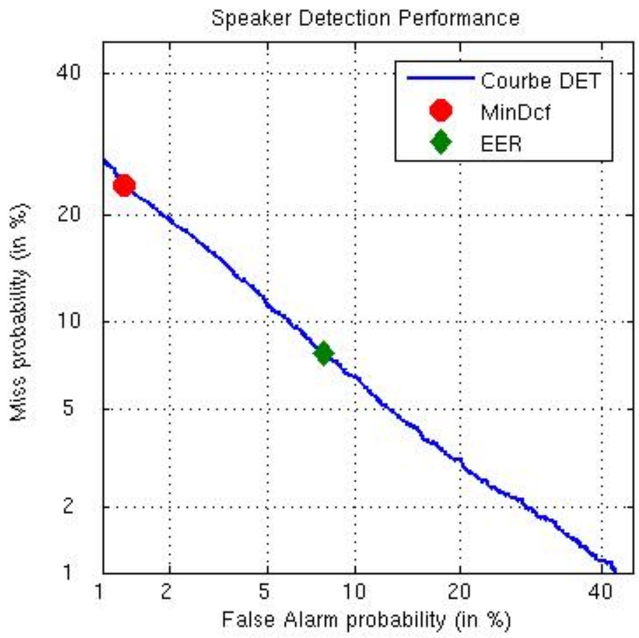
$\geq \theta$

$<$

yes ?

no ?

- **Detcurve**



- **False acceptance and rejection Rates**

$$R_{FA} = \frac{\text{Number of False Acceptance}}{\text{Number of impostors accesses}}$$

$$R_{FR} = \frac{\text{Number of False Rejection}}{\text{Number of target accesses}}$$

- **EER**

$$R_{FA} = R_{FR}$$

- **MinDCF**

$$DCF = C_{FR}.P_{target}.R_{FR} + C_{FA}.P_{imposteur}.R_{FA}$$

Total Variability – I-vector [Dehak 09,11]

- Factor analysis as feature extractor

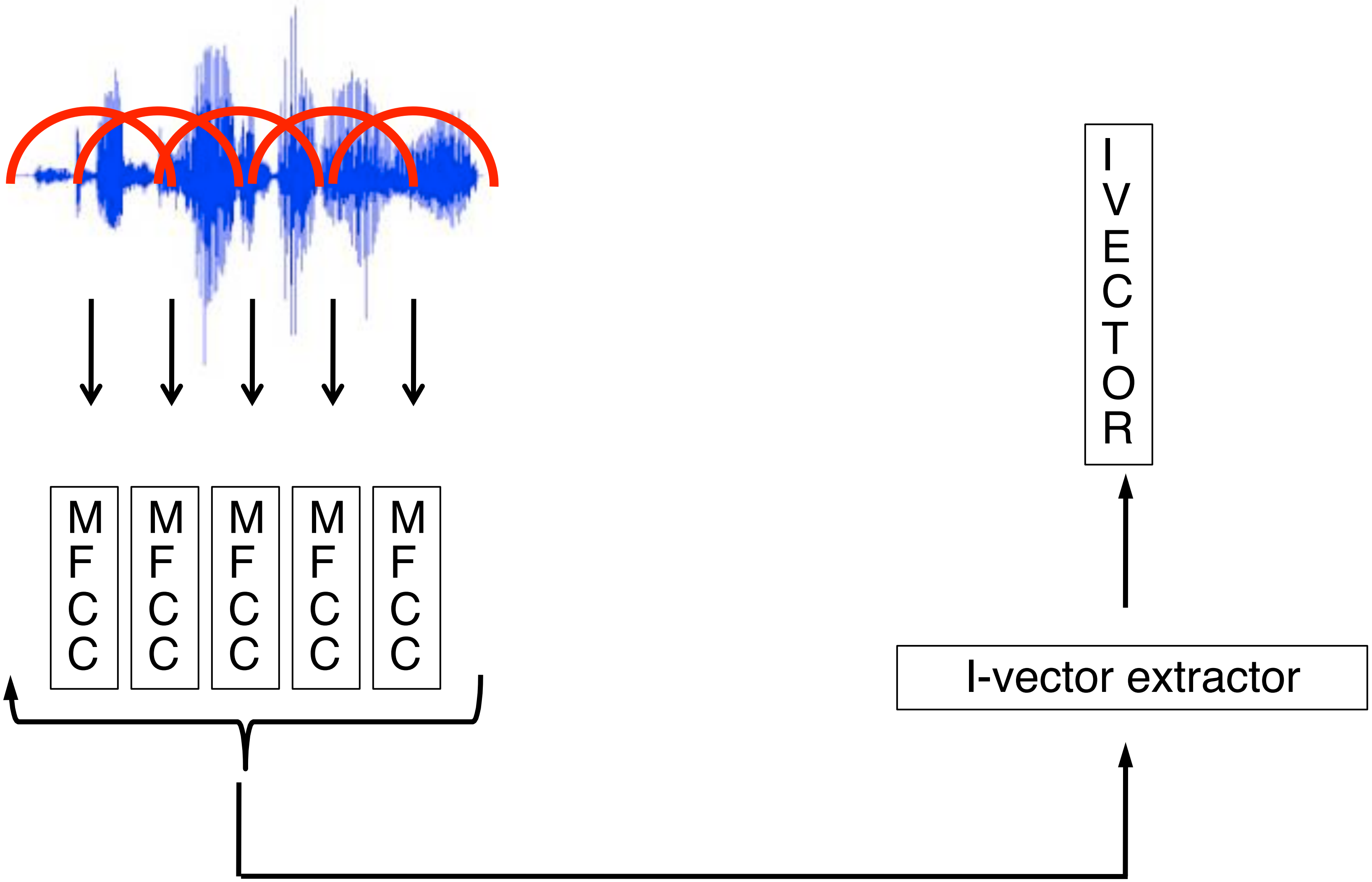
- $\mathbf{M} = \boldsymbol{m} + \boldsymbol{Tw}$
  - T is rectangular, low rank (total variability matrix)
  - w variable with standard Normal prior (i-vectors)

$$w(u) = E[w(u)] = l^{-1}(u)T^{*}\Sigma^{-1}\tilde{F}(u)$$

- Cosine scoring

$$score = \frac{< w_{t\arg et}, w_{test} >}{\left\| w_{t\arg et} \right\| . \left\| w_{test} \right\|}$$

I-vector Extraction



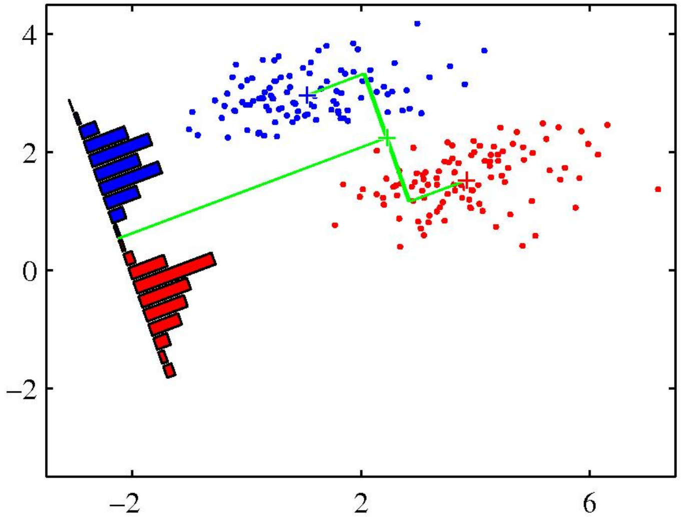
Intersession compensation

- **LDA [Dehak 2009,2011]**

A is matrix of eigenvectors from $S_b.v = \lambda.S_w.v$

$$S_b = \sum_{j=1}^{S}(w_j - \overline{w})(w_j - \overline{w})^t$$

$$S_w = \sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_i^s - w_s)(w_i^s - w_s)^t$$



- **LDA+ WCCN [Hatch2006] , [Dehak 2009,2011]**

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(A^t w_i^s - w_s)(A^t w_i^s - w_s)^t$$

$w_s = \frac{1}{n_s}\sum_{i=1}^{n_s} A^t w_i^s$ mean of utterances of each speaker

S number of speakers
n_s number of utterances for each speaker (s)
$\overline{w}$ the mean of the entire population

Modified Cosine Scoring

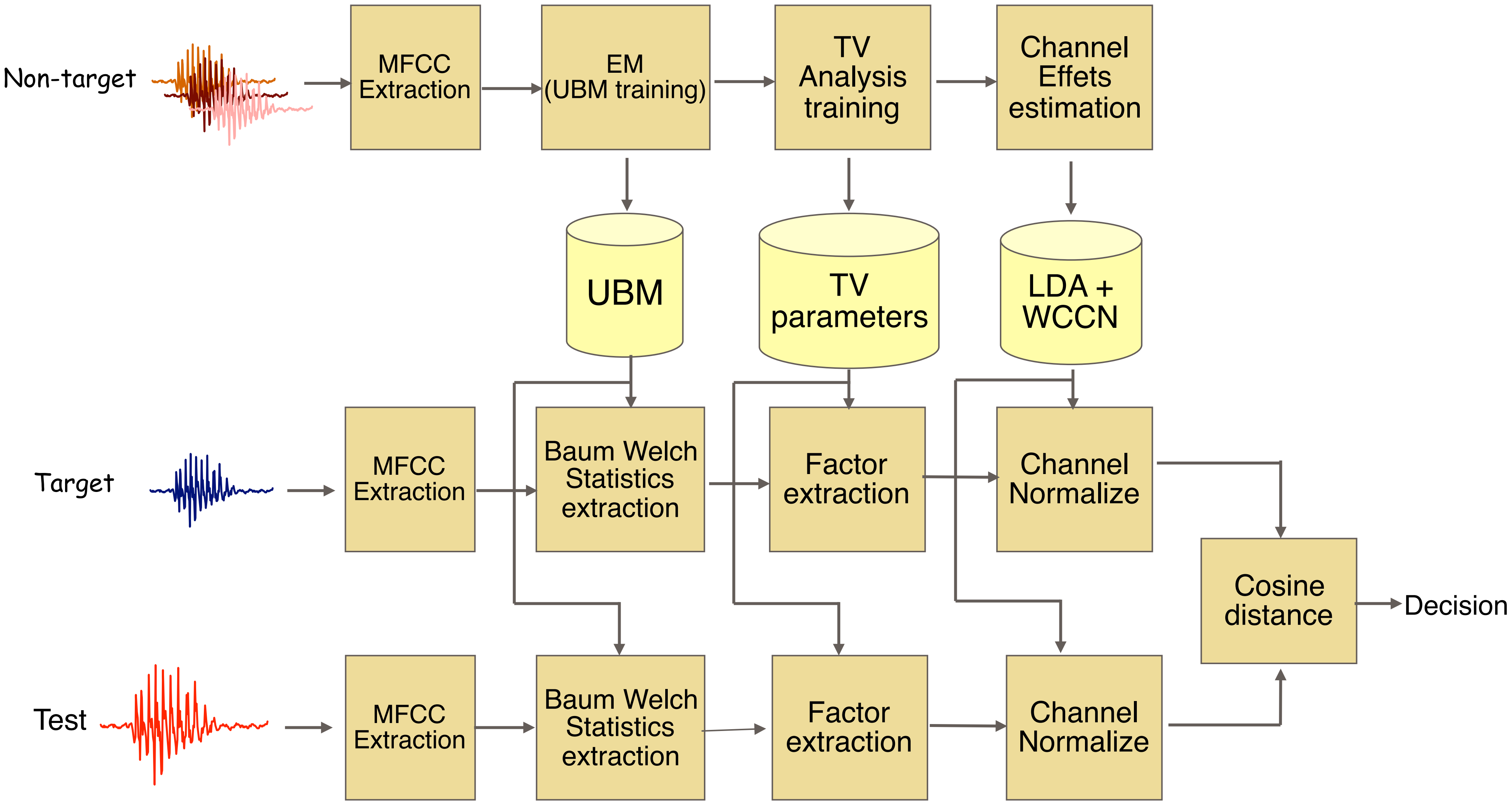
- **LDA and WCCN combination [Dehak 09,11]**

$$score(w_{t\arg et}, w_{test}) = \frac{(A^t w_{t\arg et})^t W^{-1} (A^t w_{test})}{\sqrt{(A^t w_{t\arg et})^t W^{-1} (A^t w_{t\arg et})} \cdot \sqrt{(A^t w_{test})^t W^{-1} (A^t w_{test})}} \quad \begin{matrix} \geq \\ < \end{matrix} \quad \theta$$

A : Linear Discriminant Analysis

W : Within Class Covariance Normalization

Telephone System



JFA/TV Comparison experiments

- **Gender dependent UBM**
  - 2048 Gaussians
  - 60 dimensional features : 19 Gaussianized MFCC's + energy + delta + double delta
- **JFA**
  - 300 speaker factors, 100 channel factors, common factors.
  - 1000 z-norm utterances and around 200 t-norm impostor models
- **Cosine distance scoring**
  - i-vector dim=400
  - LDA (dim=200) +WCCN
  - 1000 z-norm utterances and around 200 t-norm impostor models

Results on core condition
NIST 2008 SRE- JFA/TV comparison

- NIST 2008 SRE : female trials

	English trials		All trials	
	EER	M	[illegible]	MinDCF
JFA scoring	3.17%	0.0[illegible]	6.15%	0.032
Cosine distance scoring	**2.90%**	**0.012**	**5.76%**	**0.032**

9.5% relative improvement

- NIST 2008 SRE : male trials

	English trials		All trials	
	EER	M	[illegible]	MinDCF
JFA scoring	2.64%	0.0[illegible]	5.15%	0.027
Cosine distance scoring	**1.12%**	**0.009**	**4.48%**	**0.024**

57% relative improvement

Results on 10sec-10sec
NIST 2008 SRE- JFA/TV comparison

- NIST 2008 SRE : female trials

	English trials		All trials	
	EER	M		inDCF
JFA scoring	16.01%	0.0	17.99%	0.075
Cosine distance scoring	12.19%	0.057	16.59%	0.072

25% relative improvement

- NIST 2008 SRE : male trials

	English trials		All trials	
	EER	M		MinDCF
JFA scoring	15.20%	0.0	15.45%	0.068
Cosine distance scoring	11.09%	0.047	14.44%	0.063

26% relative improvement

Telephone and Microphone mismatch

- **NIST 2010 Speaker Recognition Evaluation core condition**
  - Telephone data
  - Interview data (microphone)
- **All proposed systems are channel dependent (conditioned system)**
- **Contributions**
  - Single system for all conditions (independent to Data Type)
  - Applying a cascade of channel compensation techniques to remove the mismatches between Telephone and interview data
  - The score calibration is still an issue
  - Data visualization for speaker recognition
    - * **Channel effects**

Channel-Blind System

- Single system applied for all different sub-condition - no conditioning
- Total variability space in composed of 800 dimensions [Senoussaoui 10] from
  - 600 telephone i-vector
  - 200 interview and microphone i-vector

$$M = m + Sw \qquad S=[T^{tel}, T^{mic}]$$

- Probabilistic LDA [Prince 07] is similar to JFA but in the i-vector space
  - Extension of the total variability framework for the interview data
- Using Probabilistic LDA to project interview and telephone i-vectors in the same speaker space

$$w = \mu+Vy+Ux+\varepsilon$$

μ means of the entire ivectors (telephone data, (800,1))

V speaker space (telephone data, (800,600))

U channel space (trained on microphone and interview data, (800,200))

ε noise modeled by full covariance matrix trained in telephone data of dimension (800,800**)**

Channel-Blind System

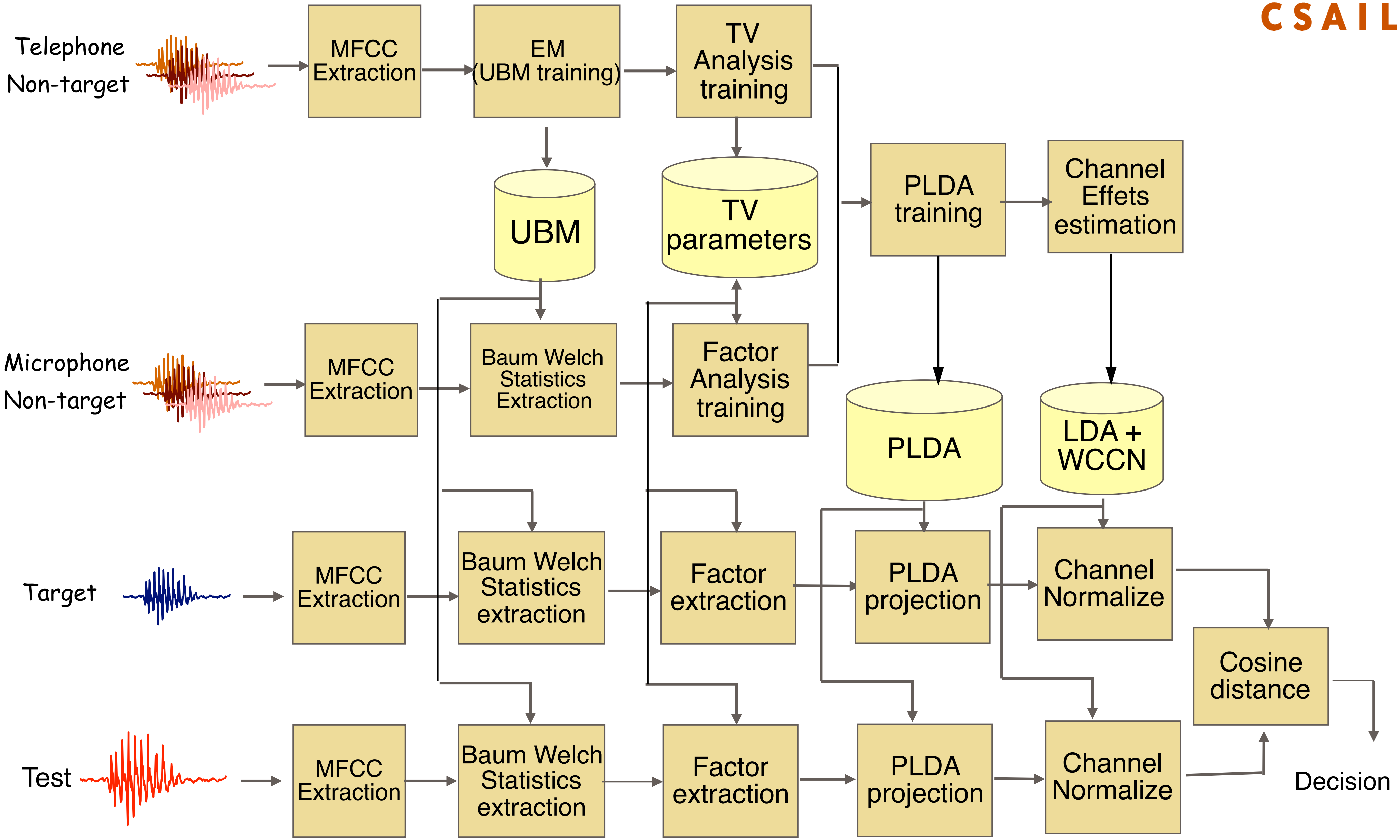
Cosine Scoring

- **LDA and WCCN combination**

$$score(y_{target}, y_{test}) = \frac{(A^t y_{target})^t W^{-1} (A^t y_{test})}{\sqrt{(A^t y_{target})^t W^{-1} (A^t y_{target})} \cdot \sqrt{(A^t y_{test})^t W^{-1} (A^t y_{test})}} \underset{<}{\geq} \theta$$

- **LDA matrix trained in the speaker space of the PLDA**
  - Trained on both microphone and telephone data (Switchboard and MIXER)
  - Dimension reduction from 600 to 250
- **Within Class Covariance**
  - Trained on both microphone and telephone data (NIST 2005,2006,2008 SRE)
  - In the interview data every speaker session is considered as new class

Channel-Blind System

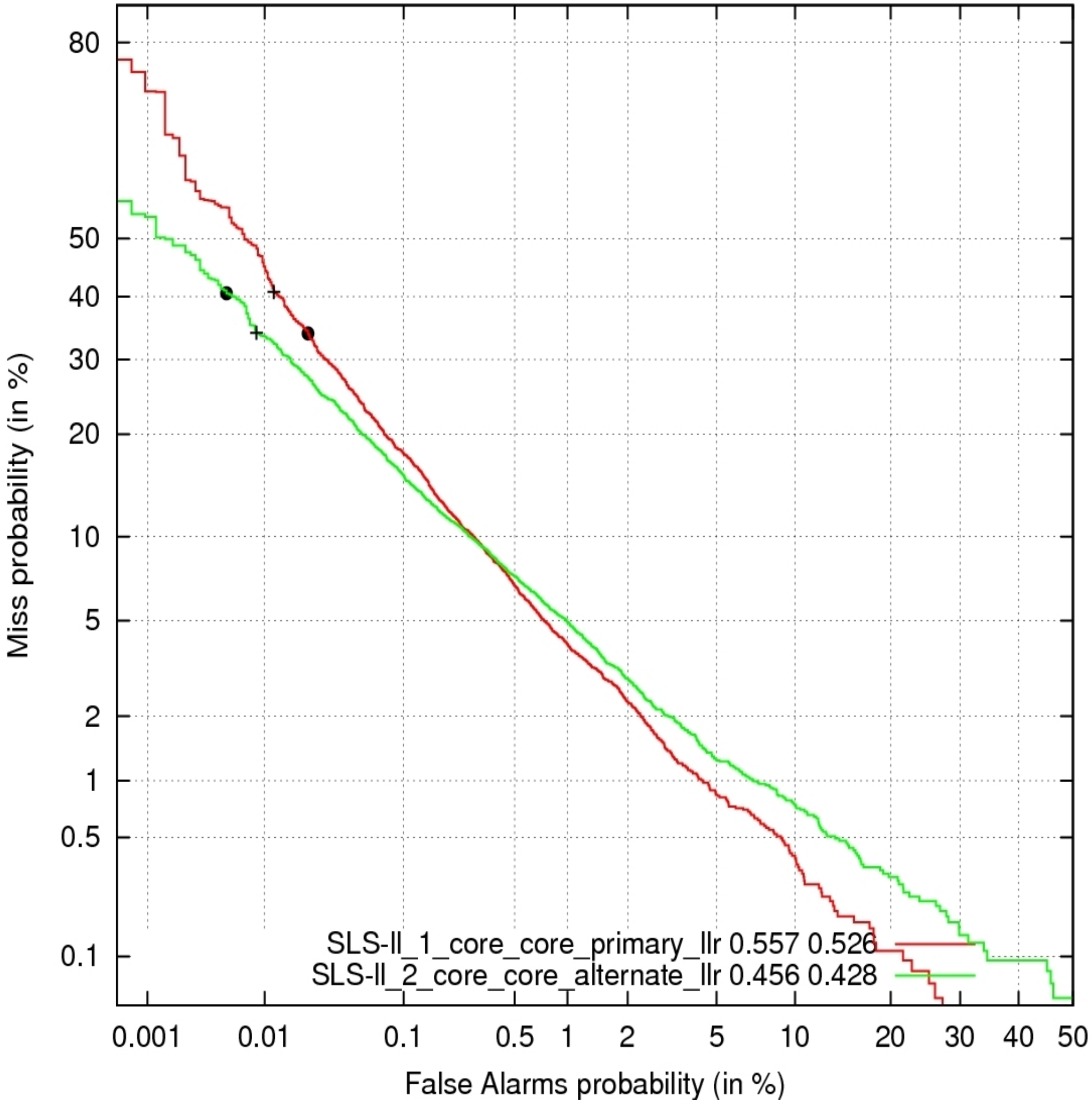


System Details

- **General**
  - Gender dependent UBMs
  - 2048 Gaussians
  - 60 dimensional features : 19 MFCC's + energy + delta + double delta
  - Feature warping
  - SAD
    - **Thanks to BUT for providing us the telephone SAD**
    - **Thanks to CRIM for providing us the interview and microphone SAD**
- **Telephone system**
  - I-vectors dim=600
  - LDA (dim=250) +WCCN
  - zt-norm : 1000 z-norm utterances and 300 t-norm impostor models
- **Blind system**
  - I-vectors dim=800
  - PLDA is used to project in 600 speaker factors
  - LDA (dim=250) +WCCN
  - s-norm : 1000 s-norm

Telephone vs Blind Systems

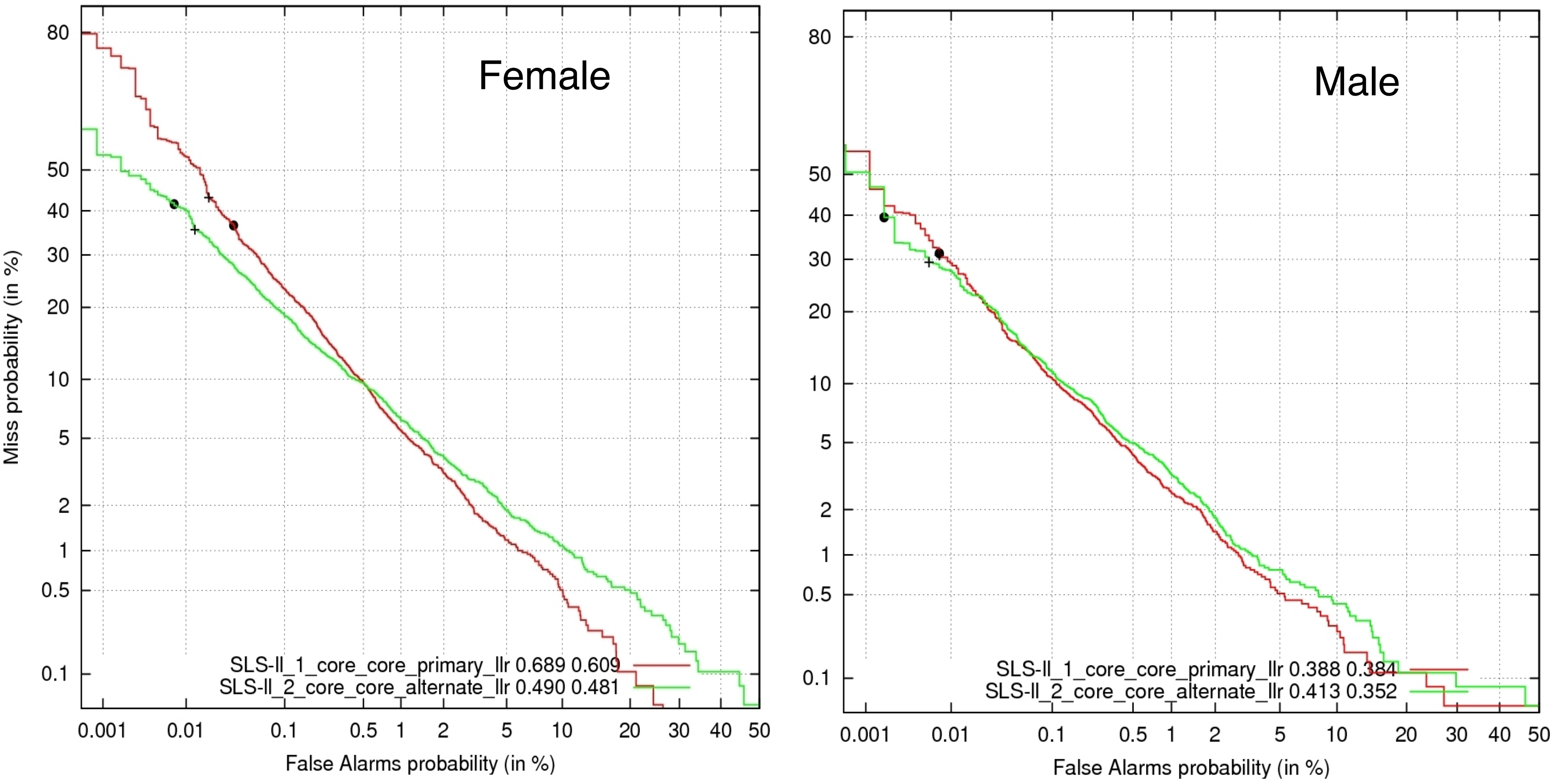
Common Condition 5 (tel-tel) Extended



Telephone vs Blind Systems

Common Condition 5 (tel-tel)

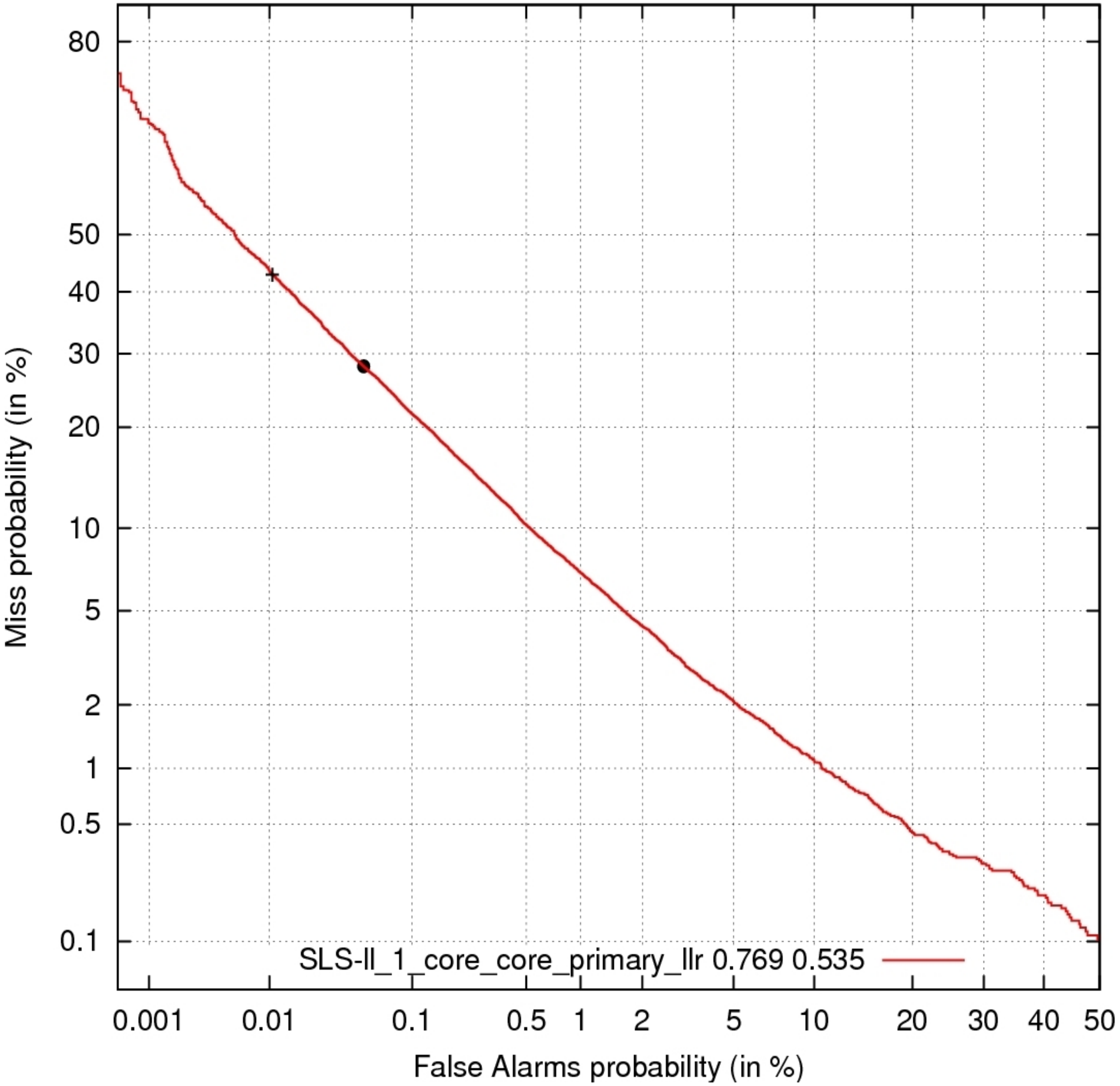
Extended



Interview Microphone

Common Condition 2 (diff mic)

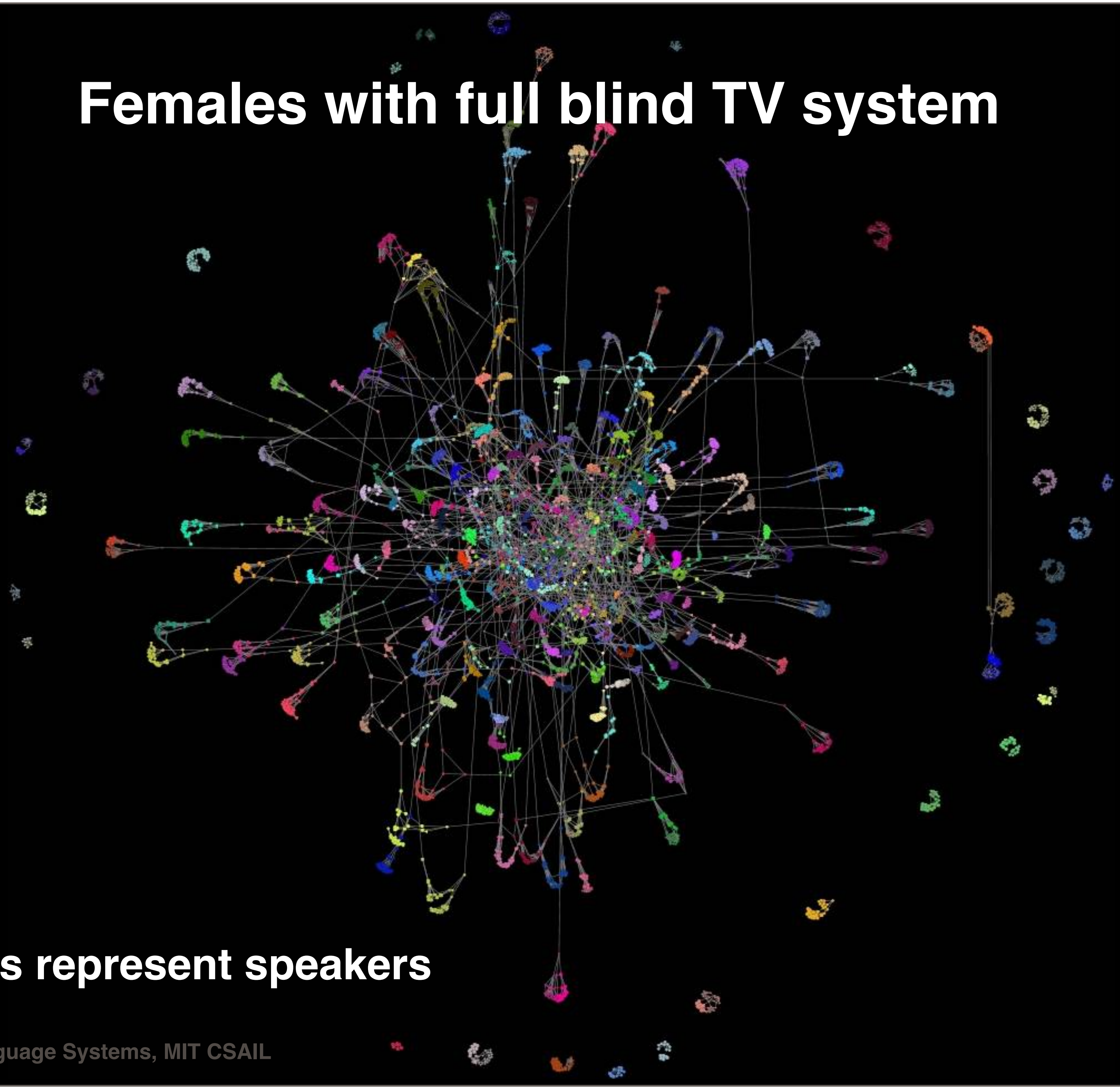
Extended



Graph Visualization

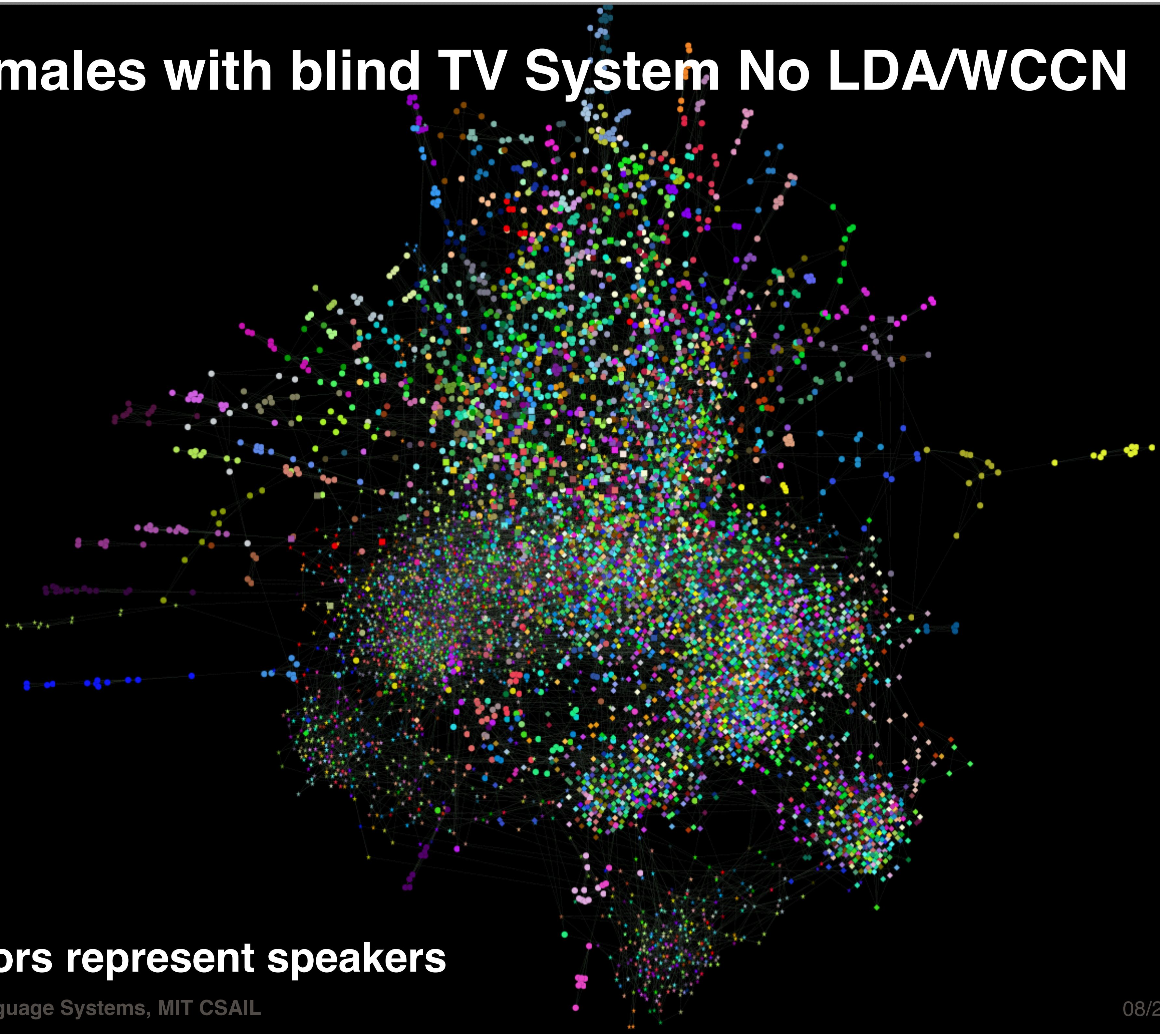
- **Work at exploring behavior of speaker matching for large data set mining (Zahi Karam)**
  - Visualization using the Graph Exploration System (GUESS) [Eytan 06]
- **Represent segment as a node with connections (edges) to nearest neighbors (3 NN used)**
  - NN computed using blind TV system (with and without channel normalization)
- **Applied to 5438 utterances from the NIST SRE10 core**
  - Multiple telephone and microphone channels
- **Absolute locations of nodes not important**
- **Relative locations of nodes to one another is important:**
  - The visualization clusters nodes that are highly connected together
- **Meta data (speaker ID, channel info) not used in layout**
- **Colors and shapes of nodes used to highlight interesting phenomena**

Females with full blind TV system



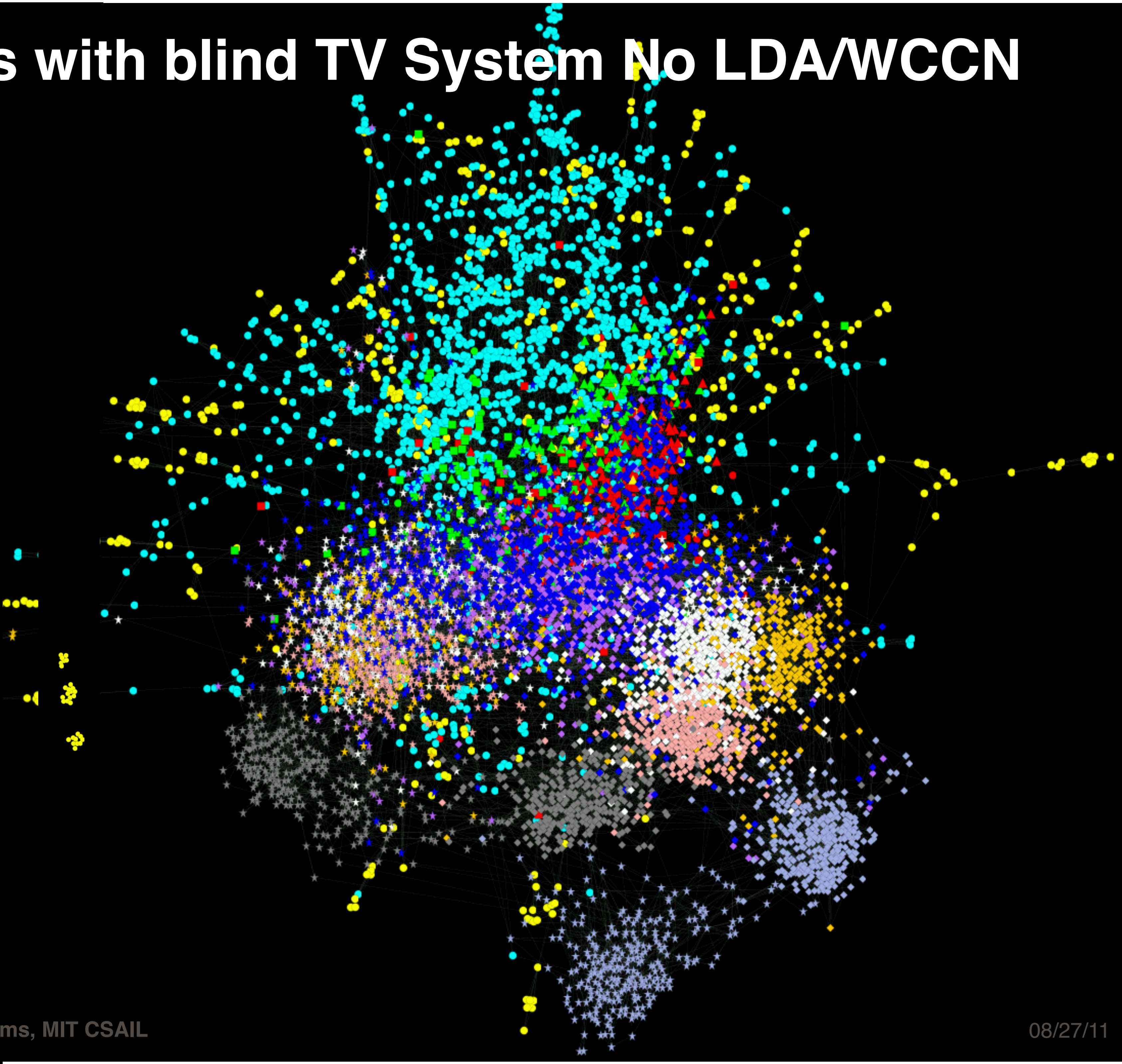
Colors represent speakers

Females with blind TV System No LDA/WCCN



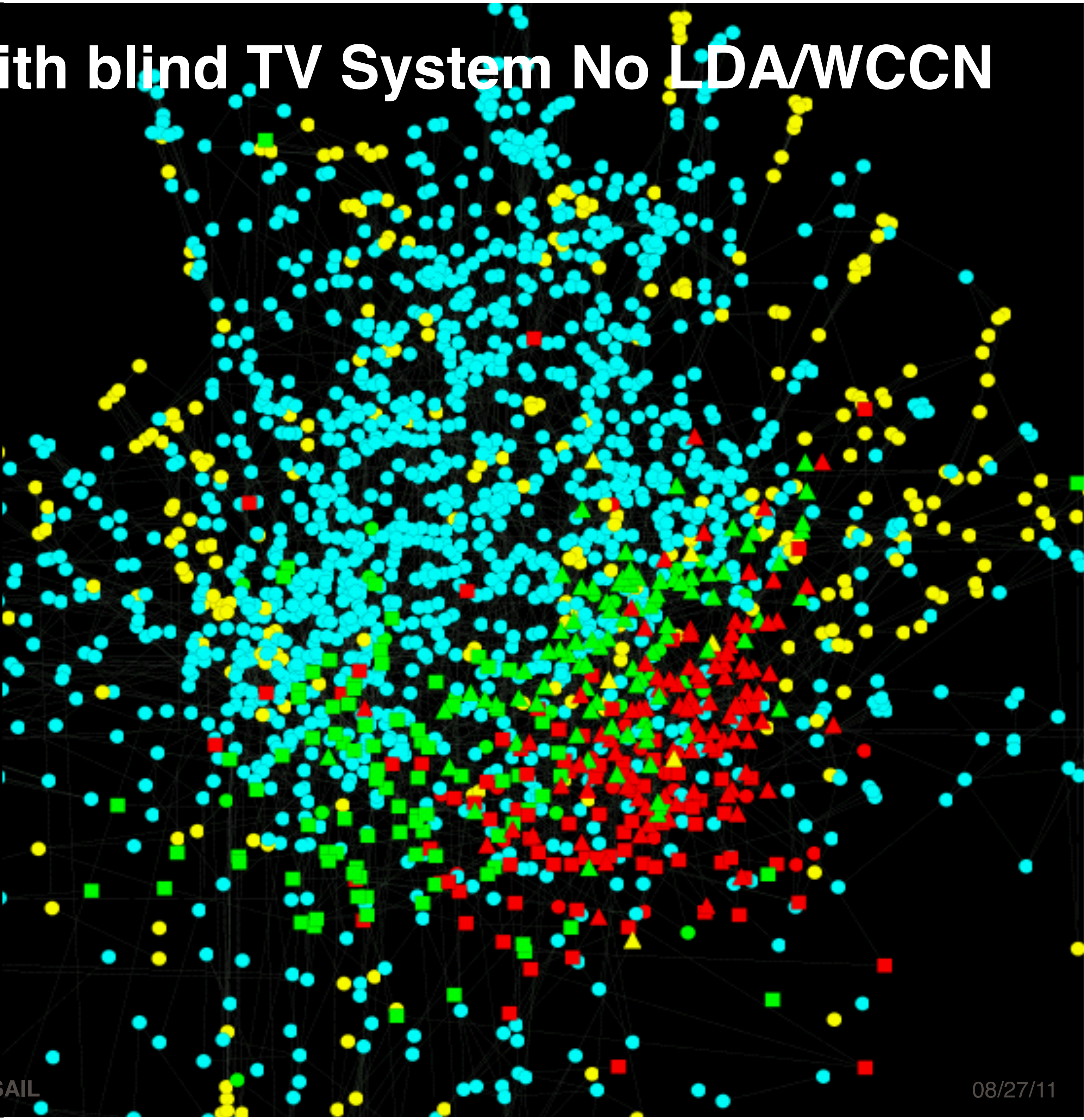
Colors represent speakers

Females with blind TV System No LDA/WCCN



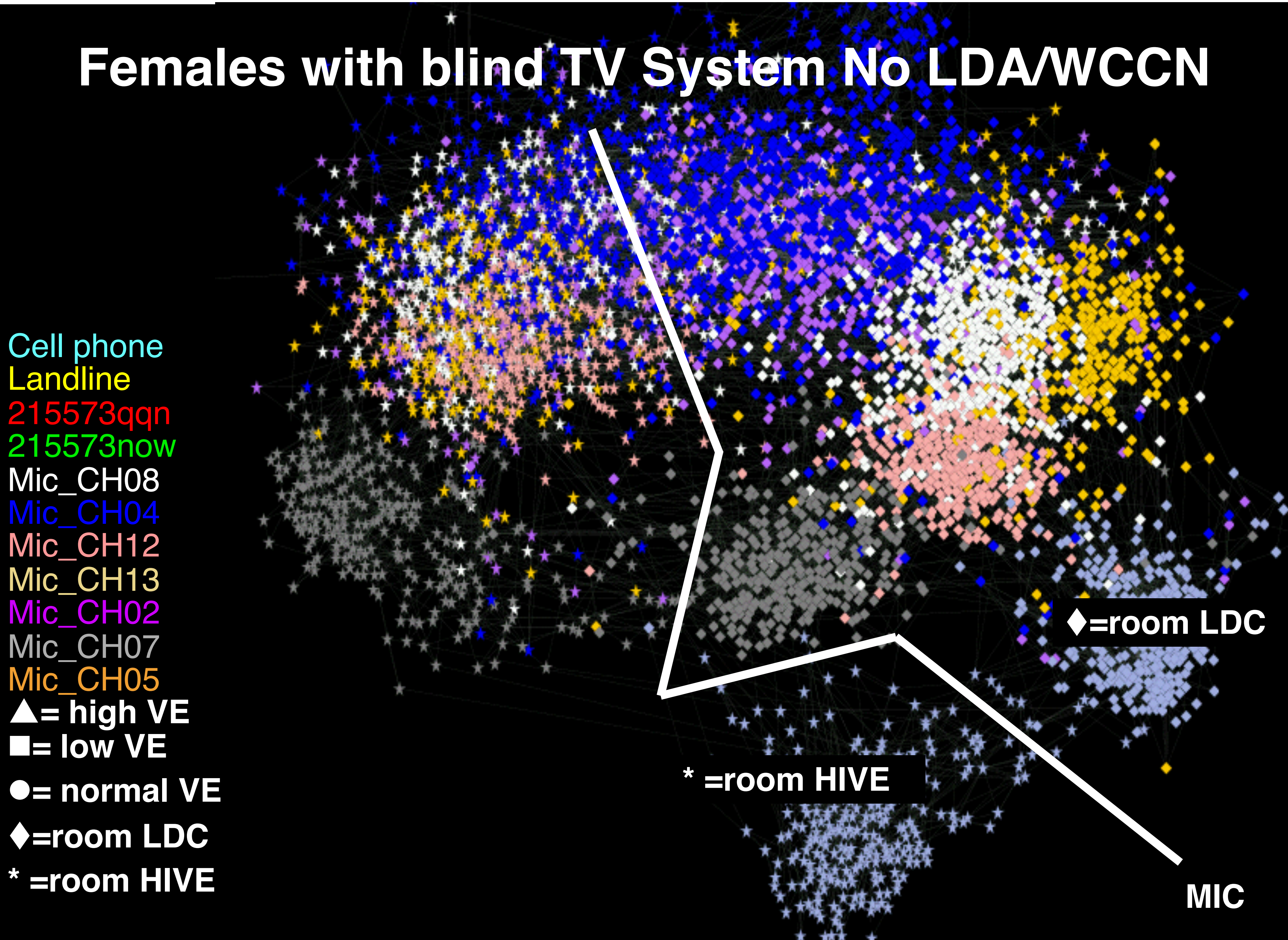
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

Females with blind TV System No LDA/WCCN

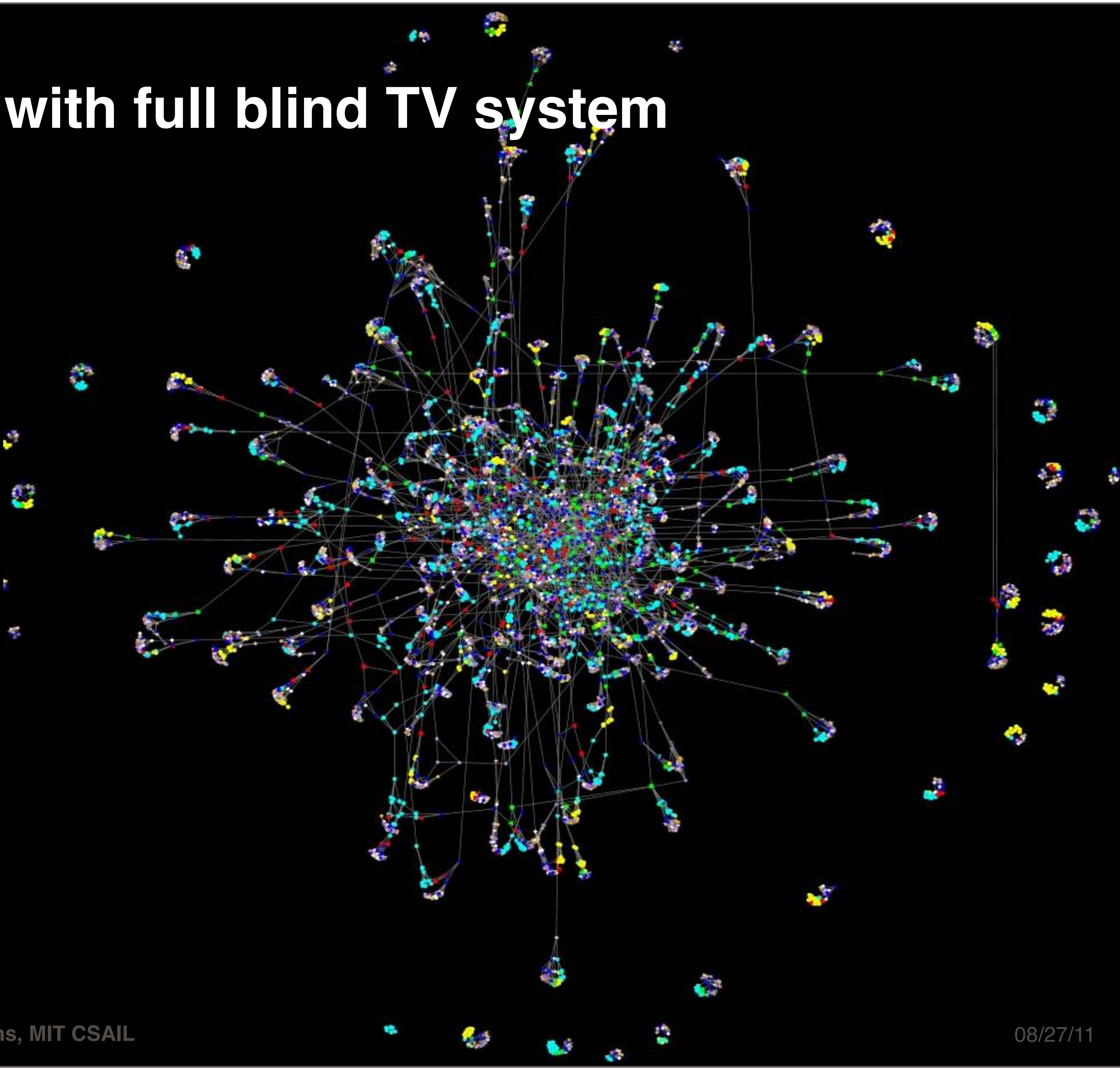


Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
*** =room HIVE**

Females with blind TV System No LDA/WCCN

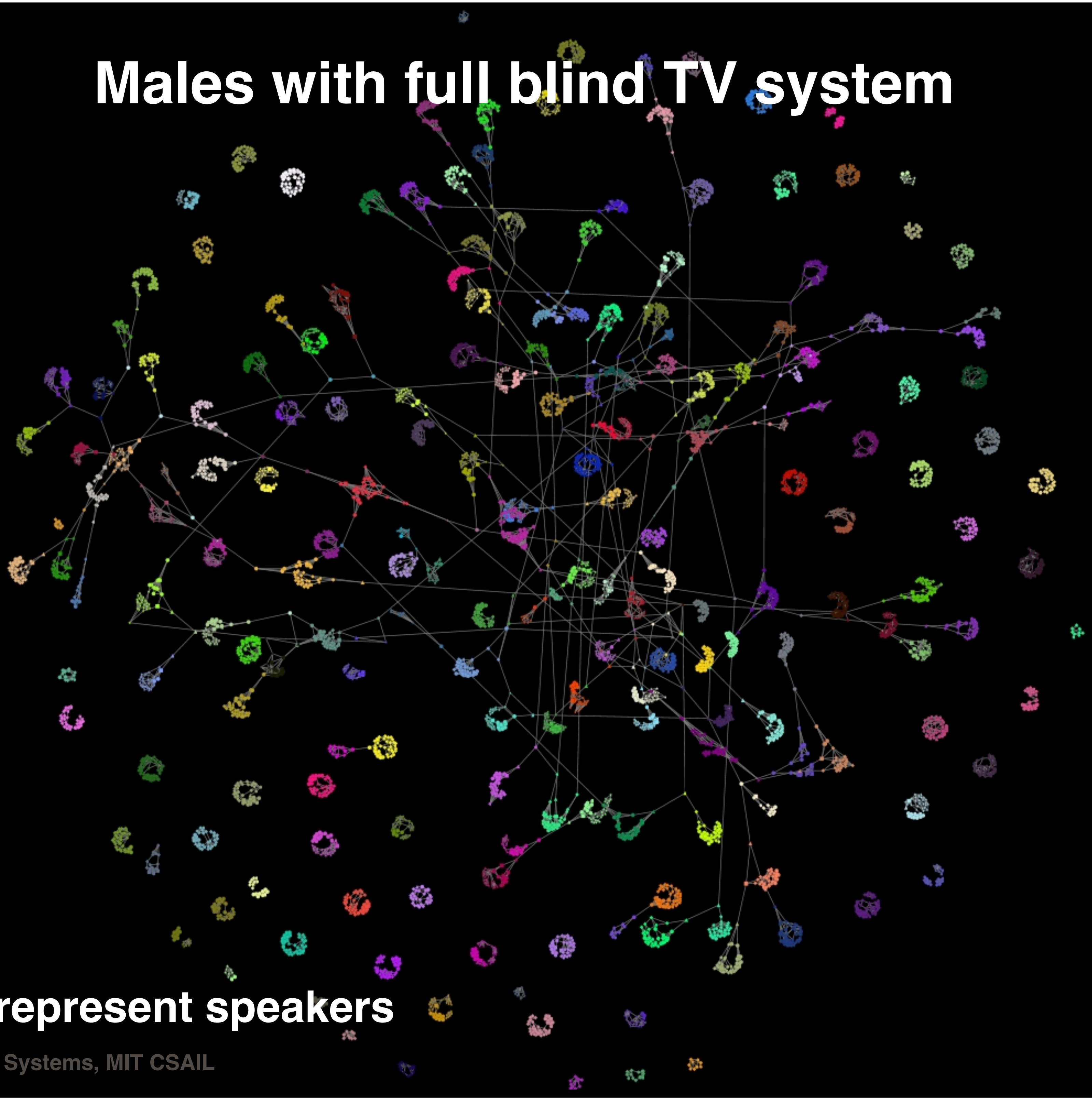


Females with full blind TV system



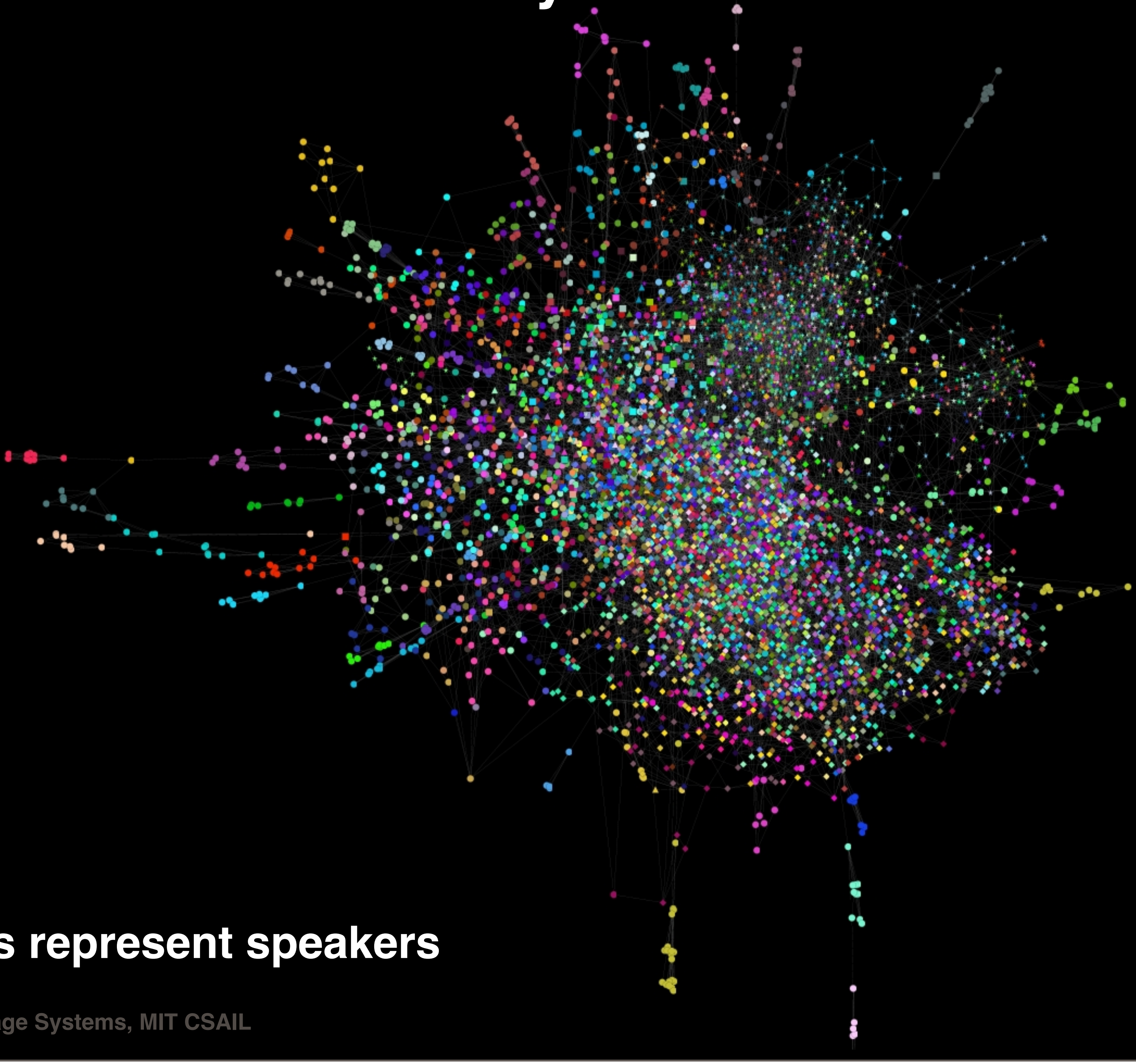
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

Males with full blind TV system



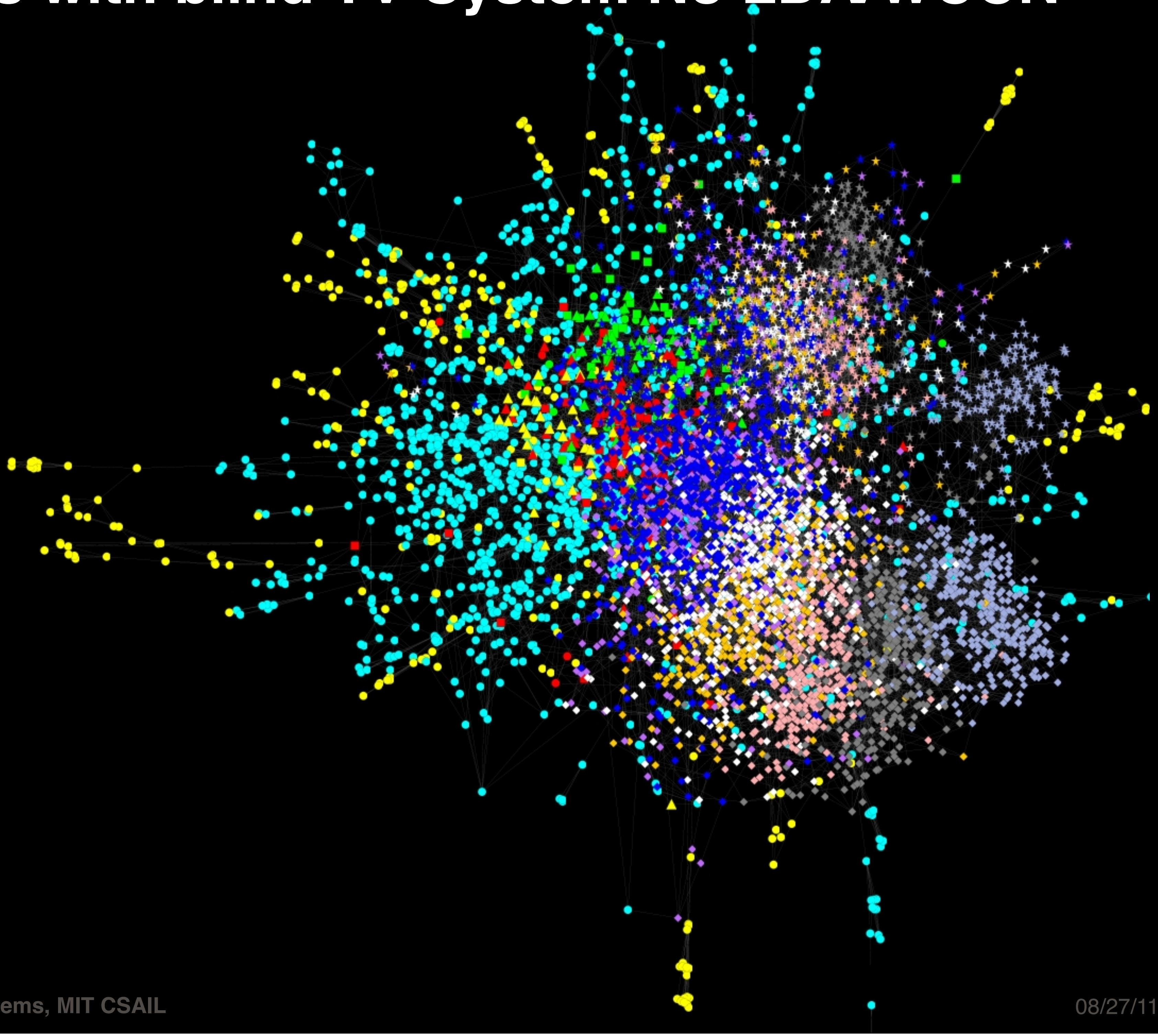
Colors represent speakers

Males with blind TV System No LDA/WCCN



Colors represent speakers

Males with blind TV System No LDA/WCCN



Cell phone

Landline

215573qqn

215573now

Mic_CH08

Mic_CH04

Mic_CH12

Mic_CH13

Mic_CH02

Mic_CH07

Mic_CH05

▲= high VE

■= low VE

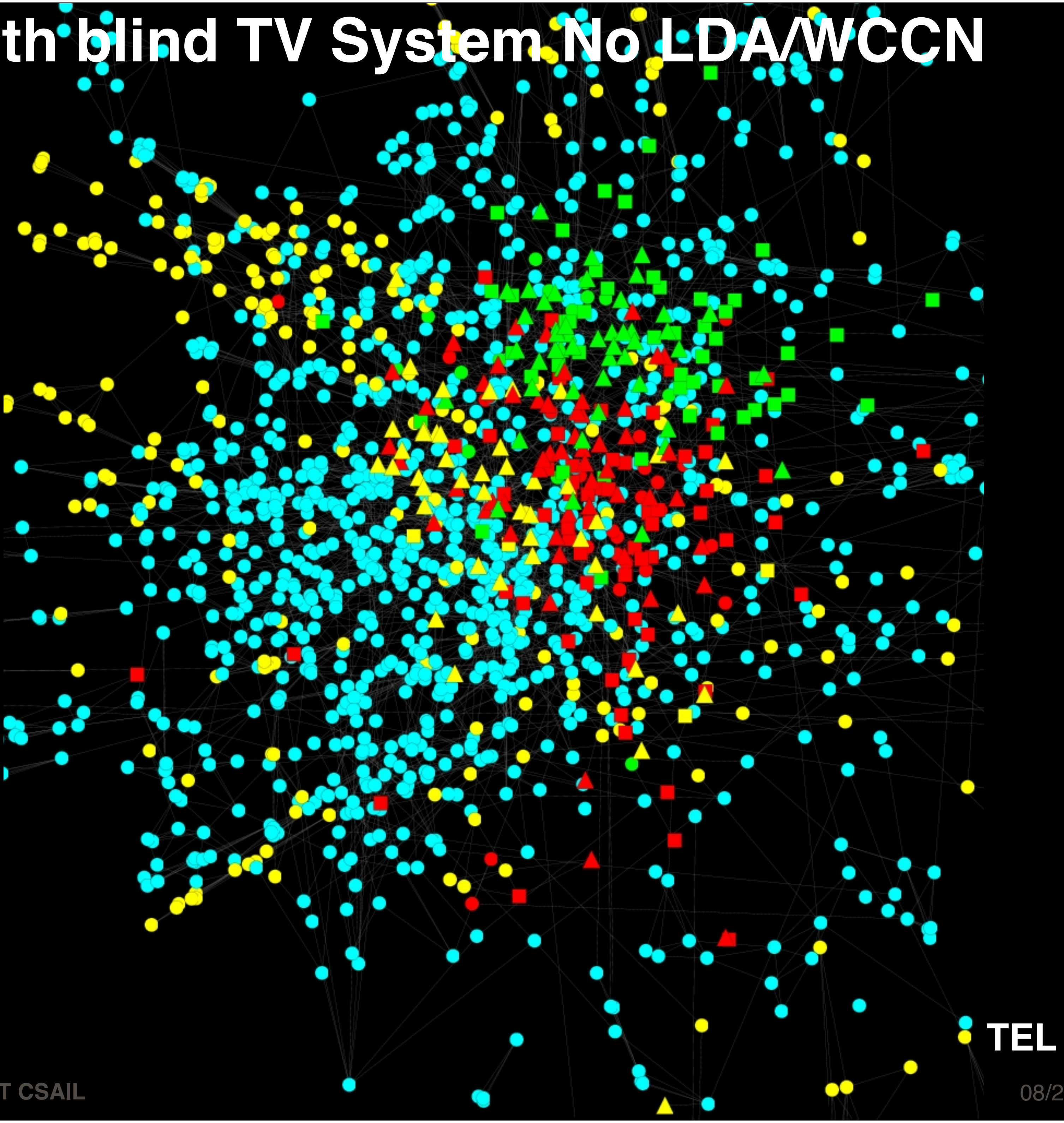
●= normal VE

♦=room LDC

* =room HIVE

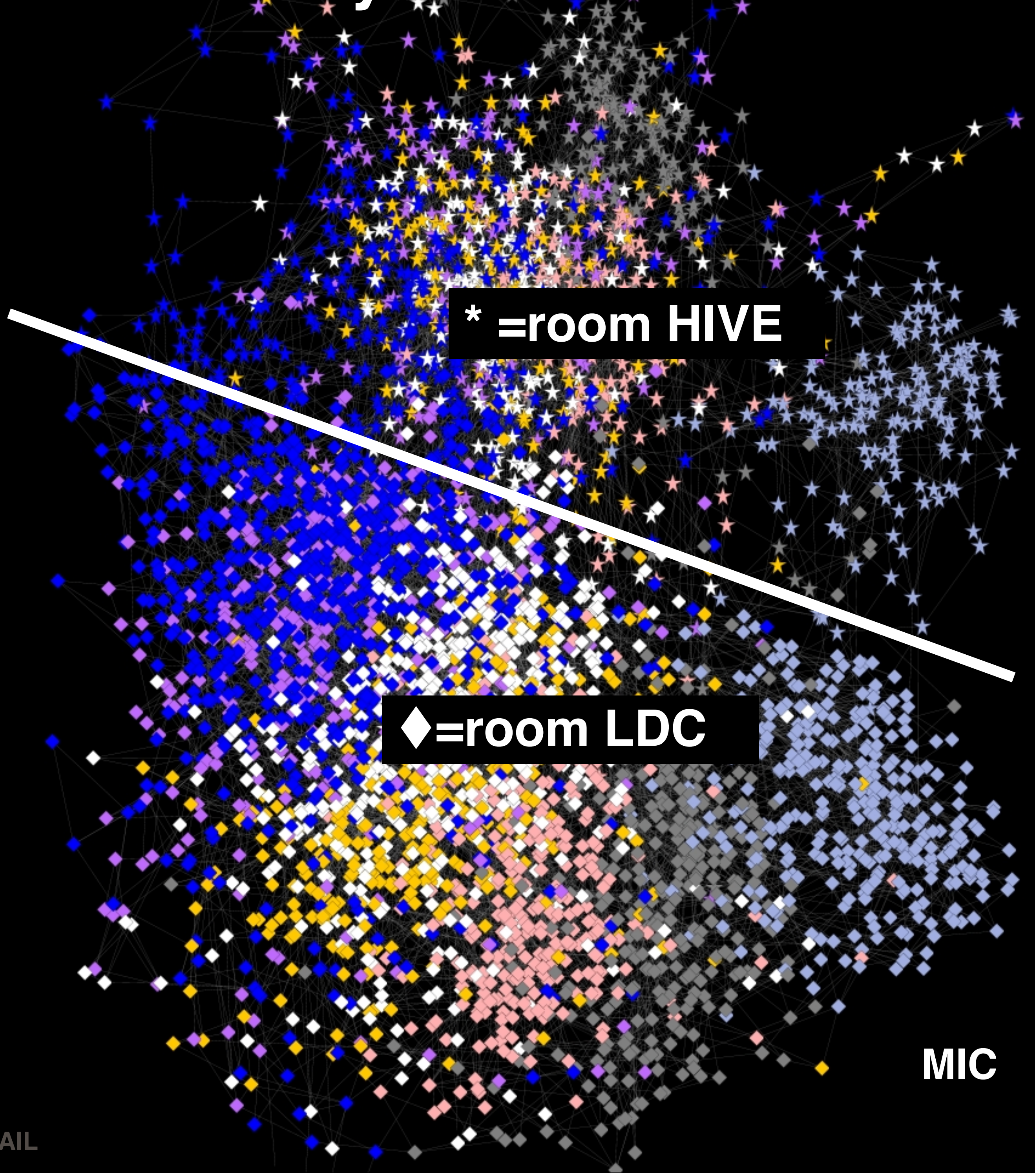
Males with blind TV System No LDA/WCCN

Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

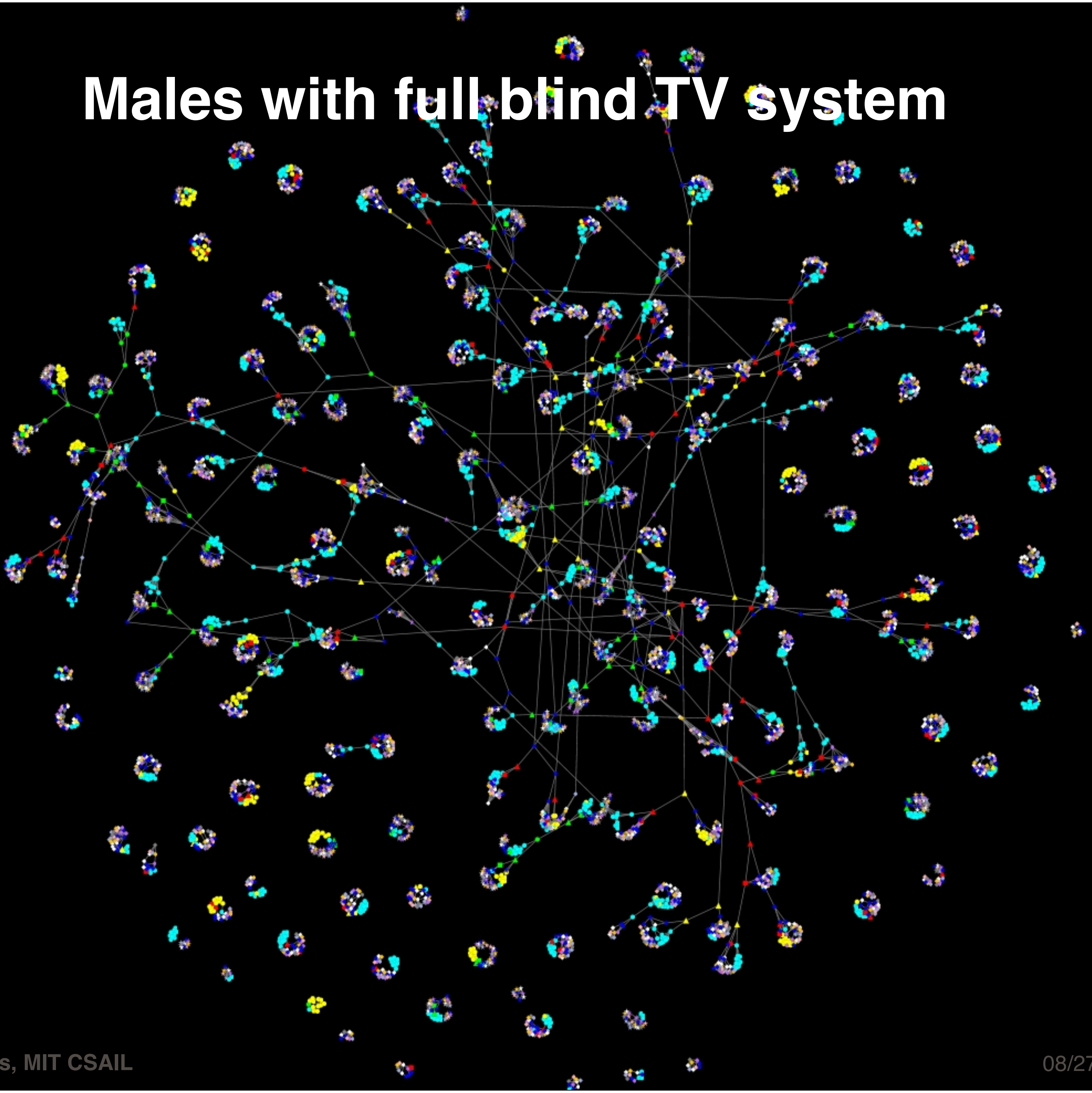


Males with blind TV System No LDA/WCCN

Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE



Males with full blind TV system



Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

Conclusions

- **New powerful speaker representation:**
  - Low dimensional features space (i-vectors)
  - Factor analysis as features extractor
- **The i-vector Blind system demonstrated robustness over conditions**
  - Graph analysis provides new data exploration techniques
- **PLDA can be applied for verification task as well**
  - P. Kenny, Bayesian Speaker Verification with Heavy-Tailed Priors. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, June 2010.
  - Senoussaoui, M., Kenny, P., Dumouchel, P., and Castaldo, F., Well Calibrated Heavy Tailed Bayesian Speaker Verification for Microphone Speech, *Proc ICASSP, Prague, Czech Republic, May 2011*
- **We proposed a cascade of LDA versions to deal with data mismatch and channel effects**
  - Is there any non-linearity affect?

Outline

- **Problem Statement of this talk**
- **History and theory (as a speaker verification problem)**
  - Gaussian Mixture Models w/ Universal Background Model.
  - Joint Factor Analysis
  - Ivector framework
- **Application**
  - Speaker recognition
  - Speaker diarization
  - Language identification
  - Emotion recognition

Roadmap

- Introduction
  - Terminology, tasks, and framework

- Low-Dimensional Representation
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring

- **Applications**
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

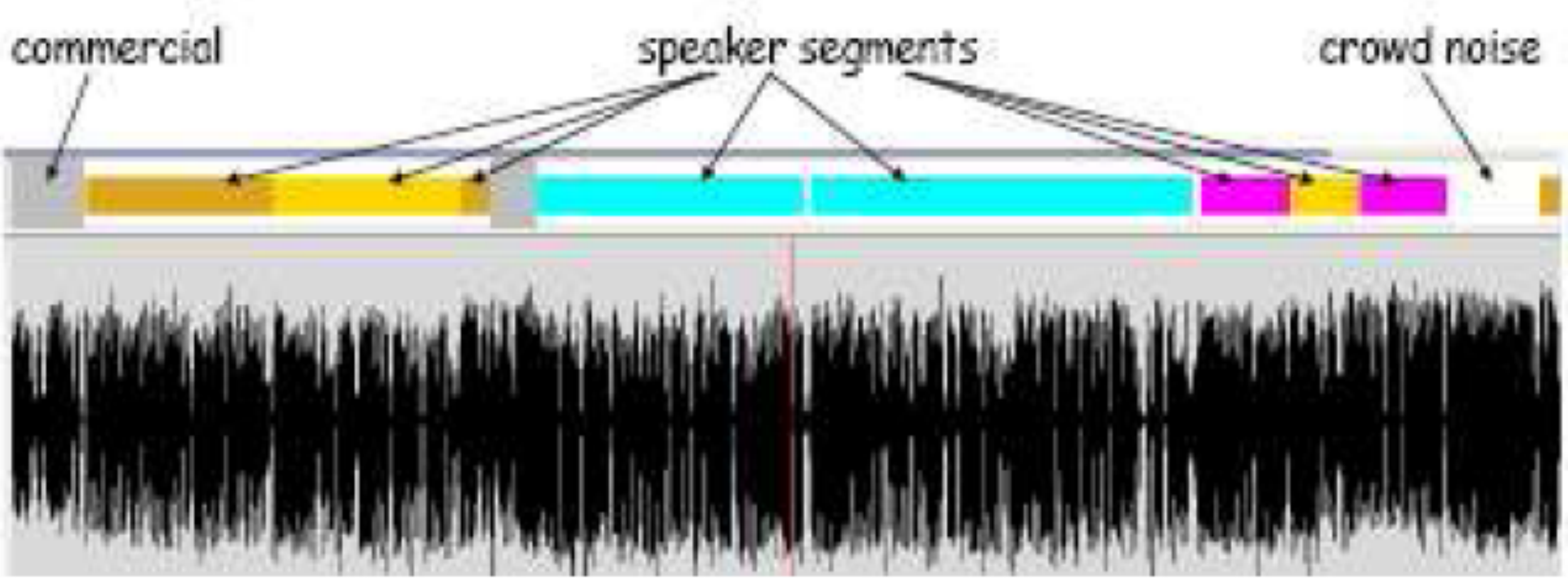
CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Unsupervised Methods for Speaker Diarization

Audio Diarization

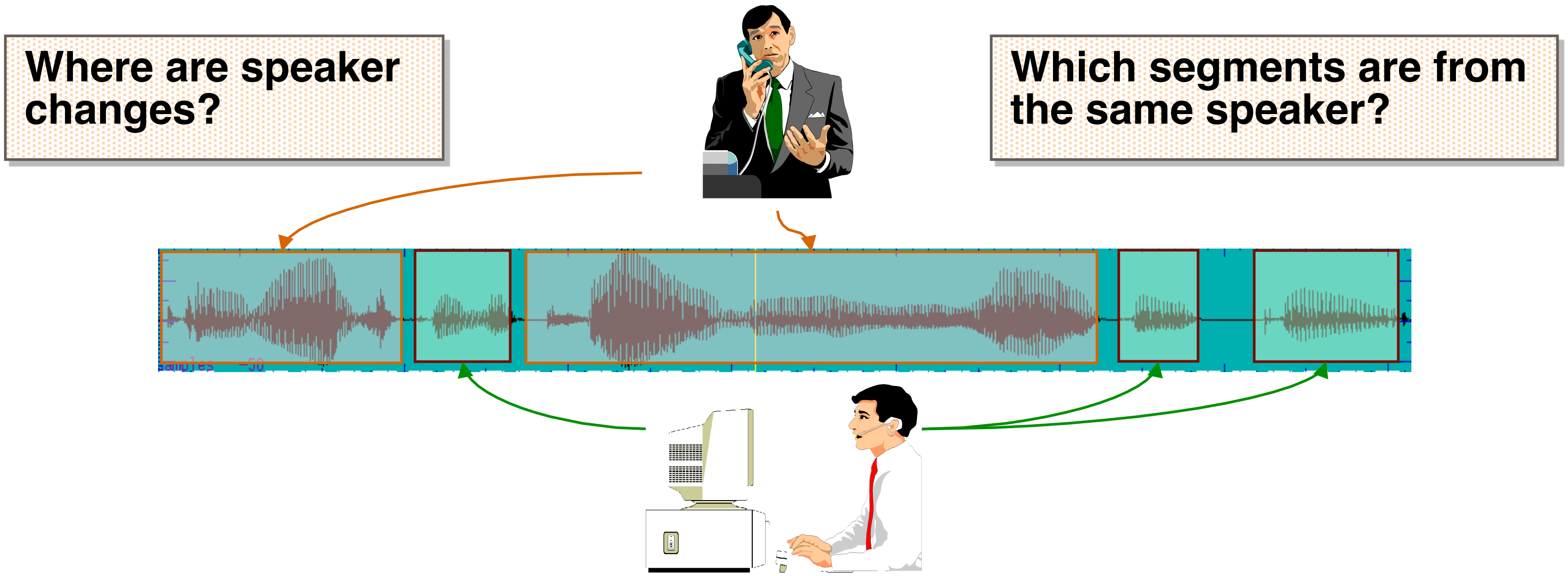
The task of marking and categorizing the different audio sources within an unmarked audio sequence



Speaker Diarization

- “Who is speaking when?”

- **Segmentation**
  - Determine when speaker change has occurred in the speech signal
- **Clustering**
  - Group together speech segments from the same speaker

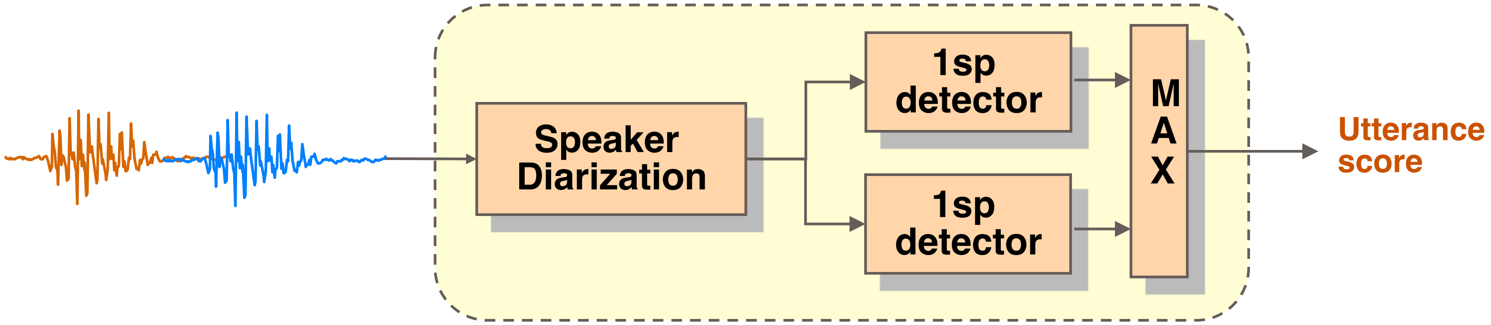


Where are speaker changes?

Which segments are from the same speaker?

Applications

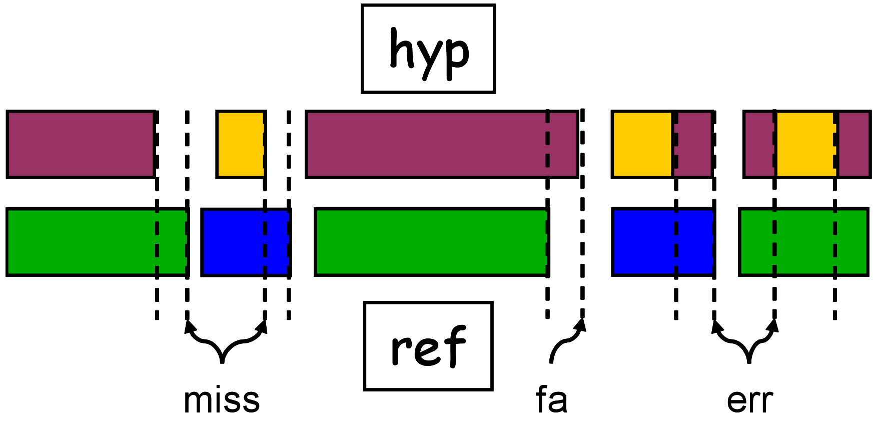
- **As a pre-processing step for other downstream applications**
  - Annotate transcripts with speaker changes and labels
  - Provide an overview of speaker activity
  - Adapt a speech recognition system
  - Do speaker detection on multi-speaker speech

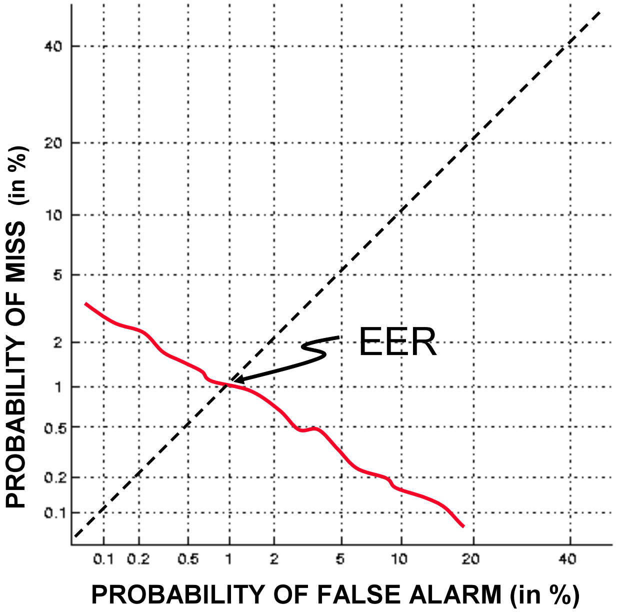


Diarization Error Measures

- **Diarization Error Rate (DER)**
  - Miss (speaker in reference but not in hypothesis)
  - False Alarm (speaker in hypothesis but not in reference)
  - Speaker Confusion (confusing one speaker's speech as from another)

- **Indirect Measures**
  - Effect on the results of a speaker detection system / speech recognizer





Experiment Data

- **Summed-channel telephone speech**
  - 2008 NIST Speaker Recognition Evaluation Test Data
  - 2215 two-speaker telephone conversations (~5min each)
  - Can obtain a reference diarization by applying ASR or Voice Activity Detection on each channel separately
    - * **Scoring ignores overlapped speech**

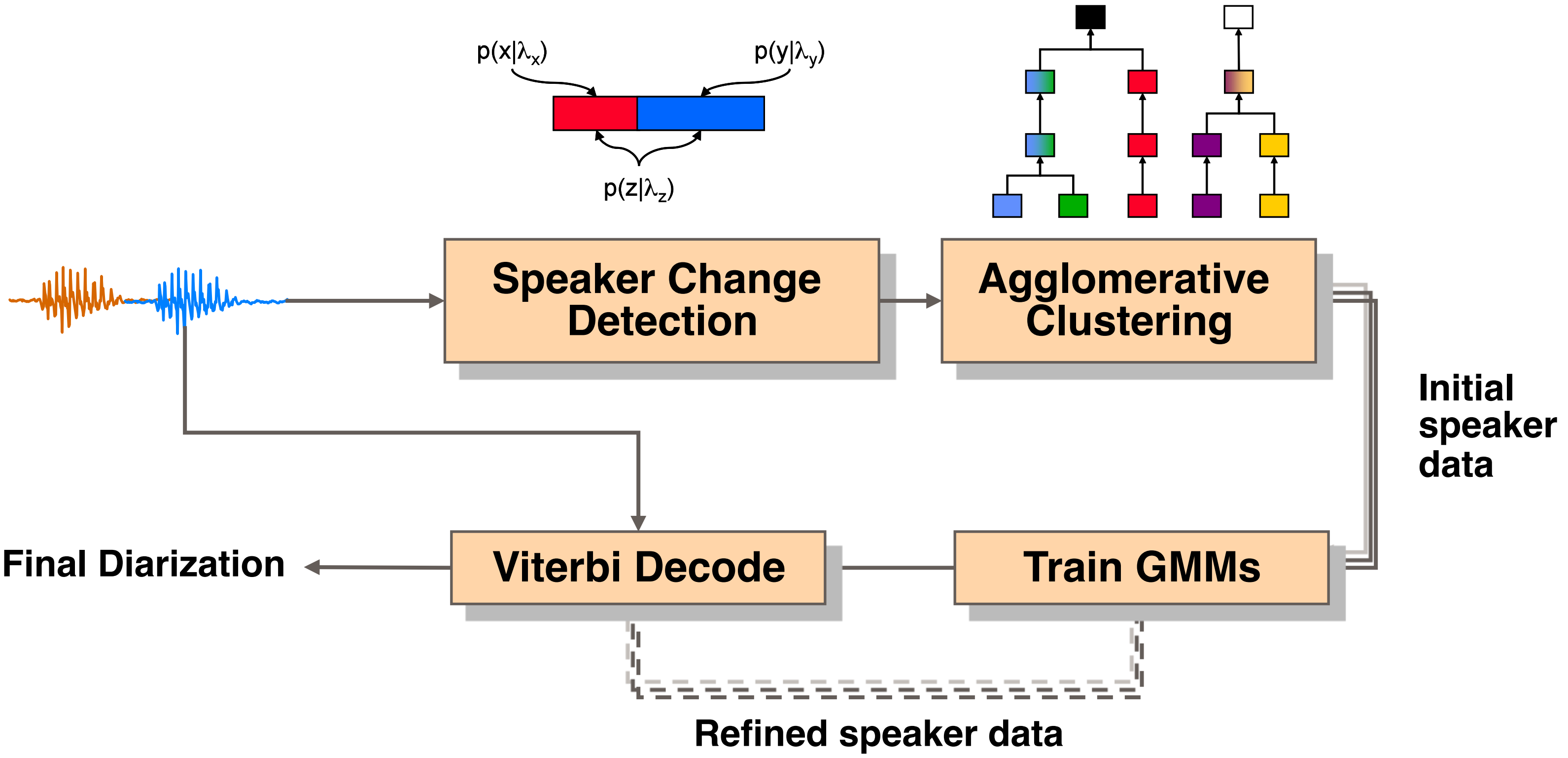
Roadmap

- **Introduction**
- **A BIC-based Baseline System**
- **A Total Variability-based Approach**
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- **Towards Less Supervision**
- **Summary and Outlook**

Roadmap

- Introduction
- A BIC-based Baseline System
- A Total Variability-based Approach
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- Towards Less Supervision
- Summary and Outlook

BIC-based Baseline System



- **Bayesian Information Criterion (BIC)**
  - BIC-based speaker change detection
  - Agglomerative hierarchical clustering with BIC-based stopping criterion
  - Iterative re-segmentation with GMM-Viterbi decoding

Towards Factor Analysis

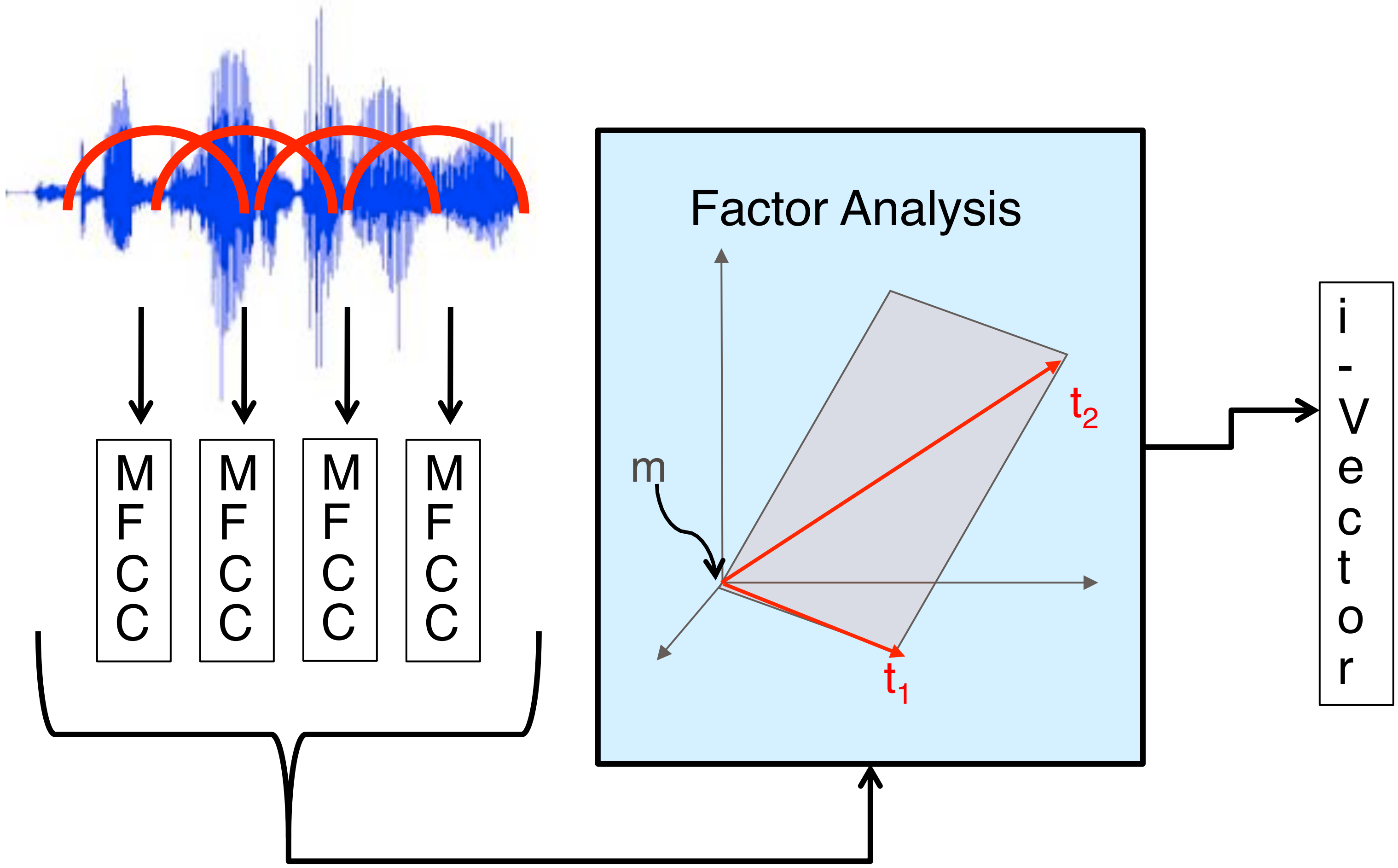
- **At the heart of the speaker diarization problem is the problem of speaker modeling**
  - We have seen how well factor analysis-based methods perform in speaker recognition.

- **Previous work in FA-based diarization**
  - Stream-based, on-line system (Castaldo, 2008)
  - Variational Bayesian system (Kenny, 2010)

Towards Factor Analysis

- **Advantages**
  - We have seen how well factor analysis-based methods perform in speaker recognition.
- **Difficulties**
  - Decisions made on very short (~1 second) speech segments
  - Poor speaker change detection can corrupt speaker models

i-vector Extraction



Inter-session Compensation and Cosine Scoring

IF we were to follow, by rote, the standard recipe, we have …

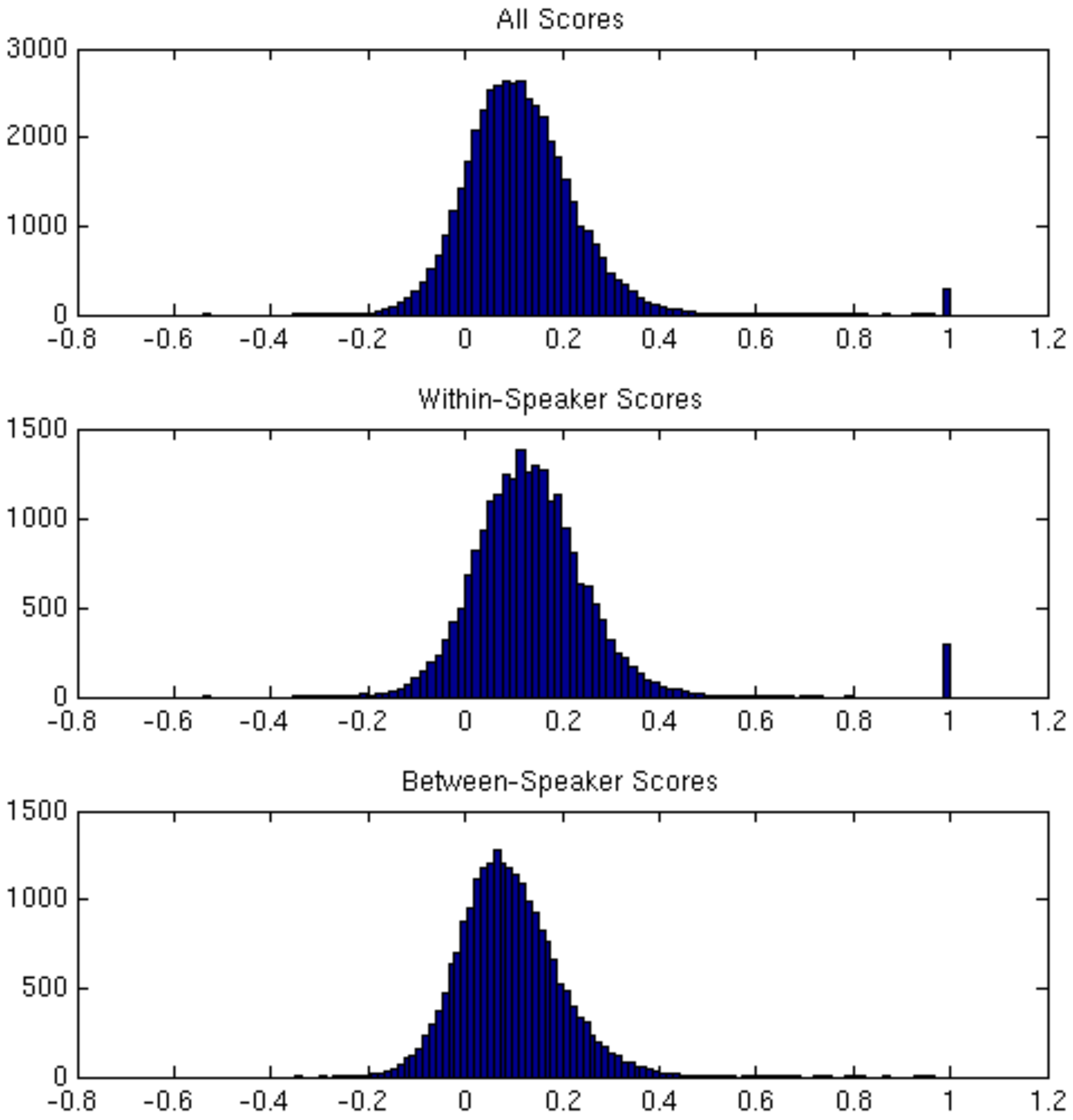
$$score(w_1, w_2) = \frac{(A^t w_1)^t W^{-1} (A^t w_2)}{\sqrt{(A^t w_1)^t W^{-1} (A^t w_1)} \, . \, \sqrt{(A^t w_2)^t W^{-1} (A^t w_2)}}$$

A : Linear Discriminant Analysis (LDA) projection matrix

W : Within Class Covariance Normalization (WCCN) matrix

Inter-session Compensation

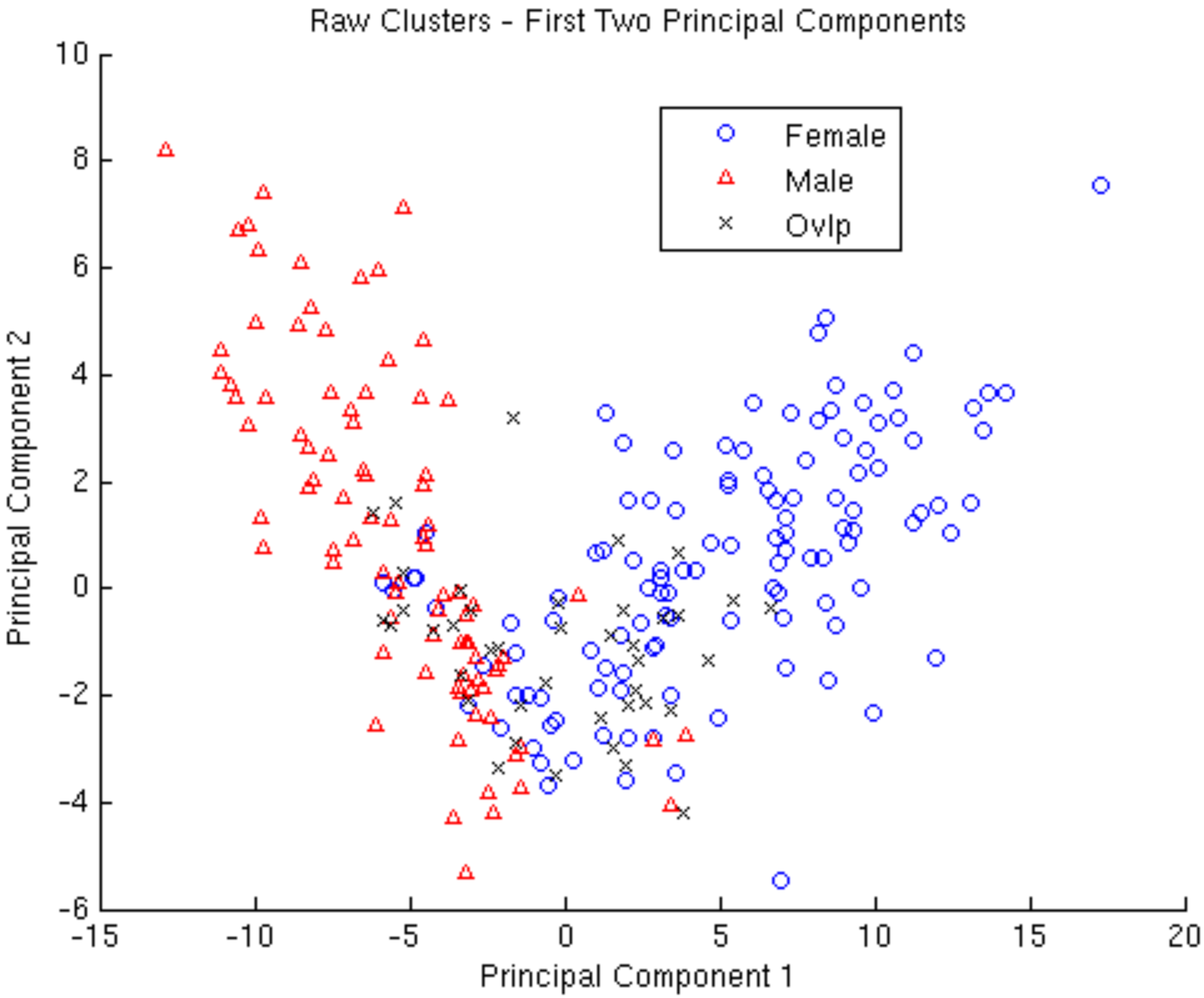
However, we ran into some issues…



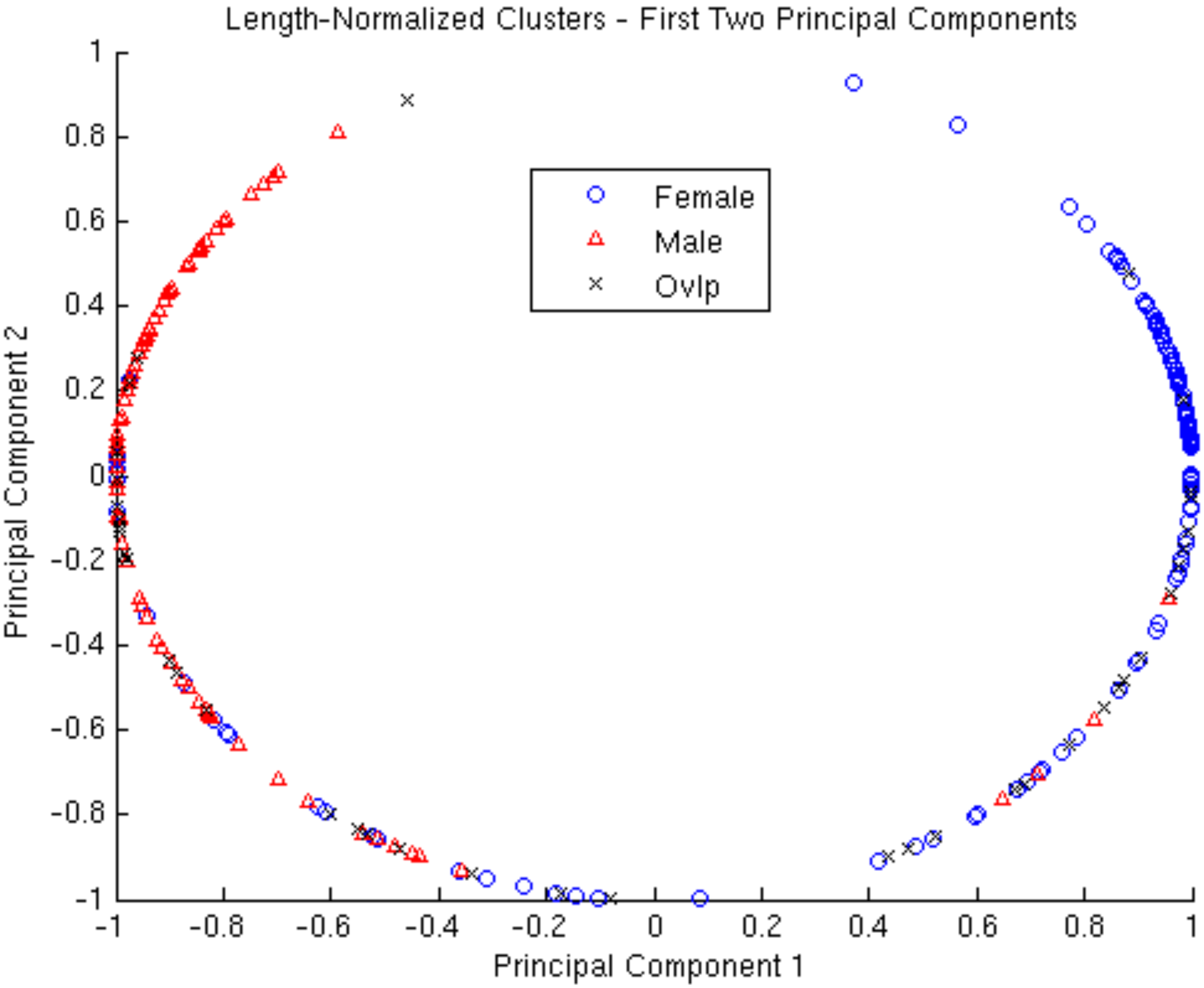
~~Inter-session Compensation~~
Intra-session Exploitation

- **Compensating for inter-session variability is wholly unnecessary in the problem of diarization.**
  - Because we are working on a summed-channel telephone conversation, there is no *inter*-session.
  - What we really care about are the *intra*-session variabilities
    - **And hopefully, the most prominent variabilities correspond to distinctly different speakers.**
- **Interspeech 2011 Paper and Presentation**
  - Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass. "Exploiting Intra-Conversation Variability for Speaker Diarization."

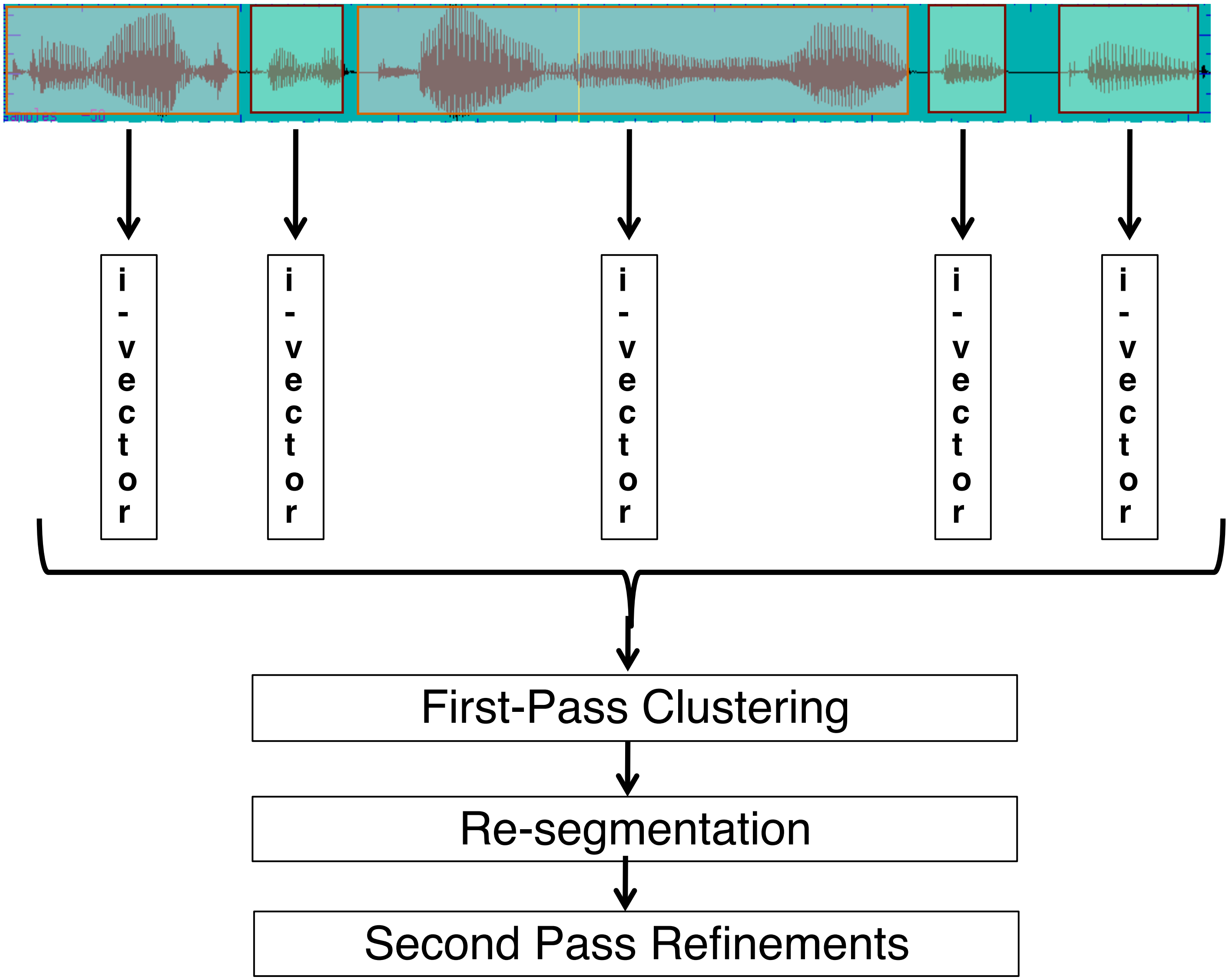
i-vector Visualization



i-vector Visualization



System Diagram



i-vector

i-vector

i-vector

i-vector

i-vector

First-Pass Clustering

Re-segmentation

Second Pass Refinements

Roadmap

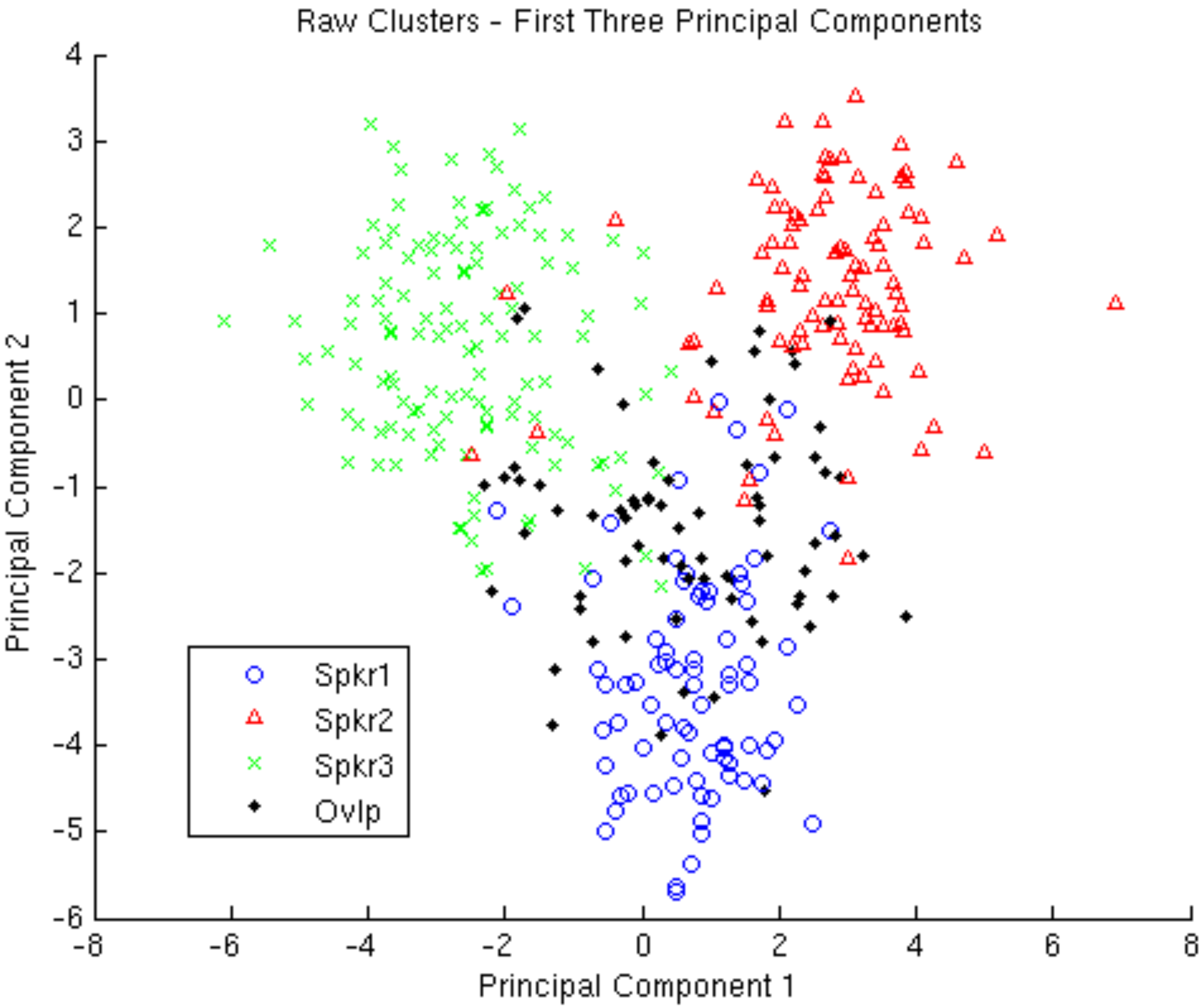
- Introduction
- A BIC-based Baseline System
- A Total Variability-based Approach
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- Towards Less Supervision
- Summary and Outlook

Lingering Issues

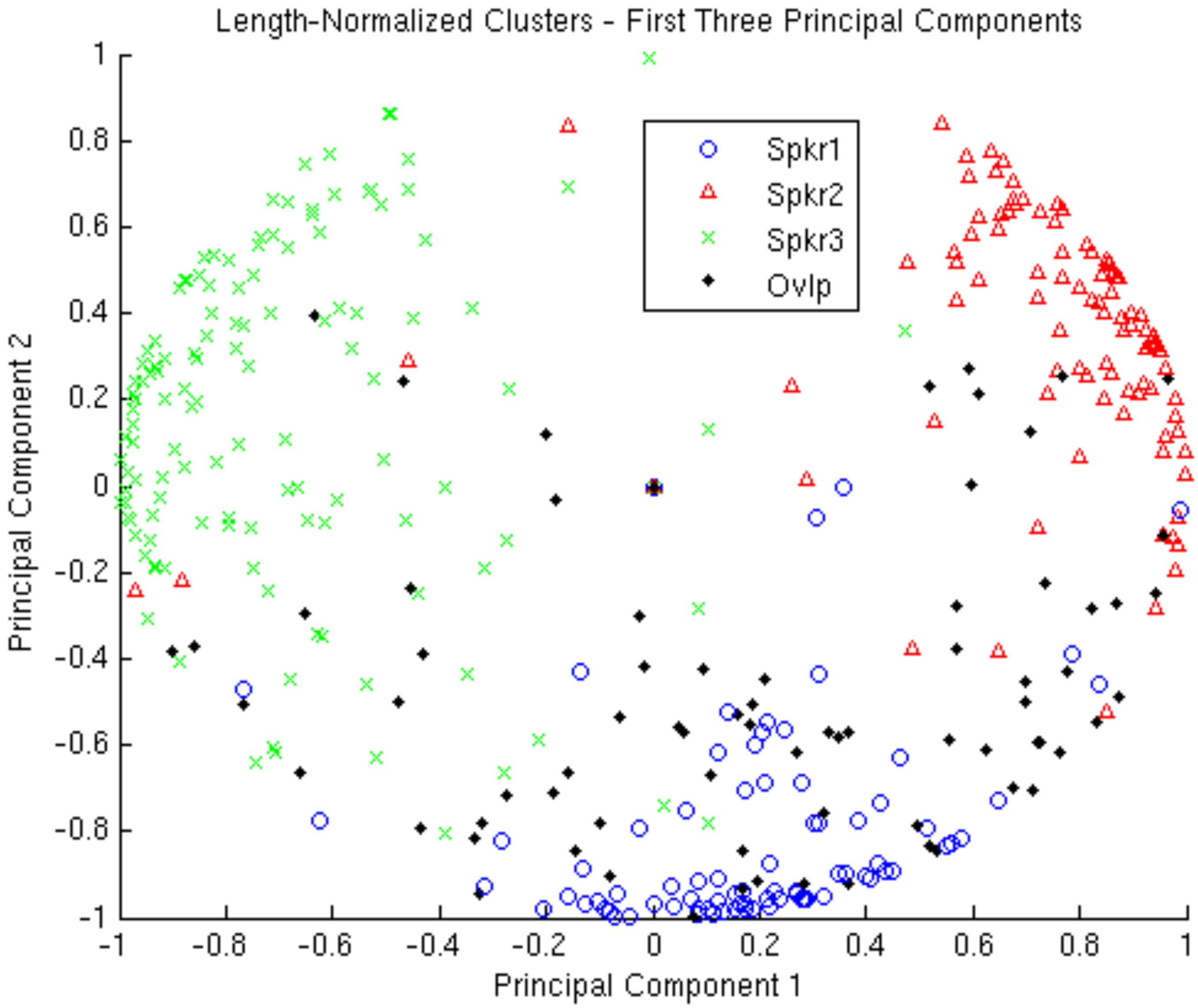
- **Diarization of speech containing more than two speakers**
  - How can we estimate the number of speakers?

- **Overlapped speech segments**
  - Though not scored, we still have to deal with them during diarization
  - Not much previous work on this (Boakye, 2008)

The Problem With Overlap



The Problem With Overlap



Estimating Speaker Number

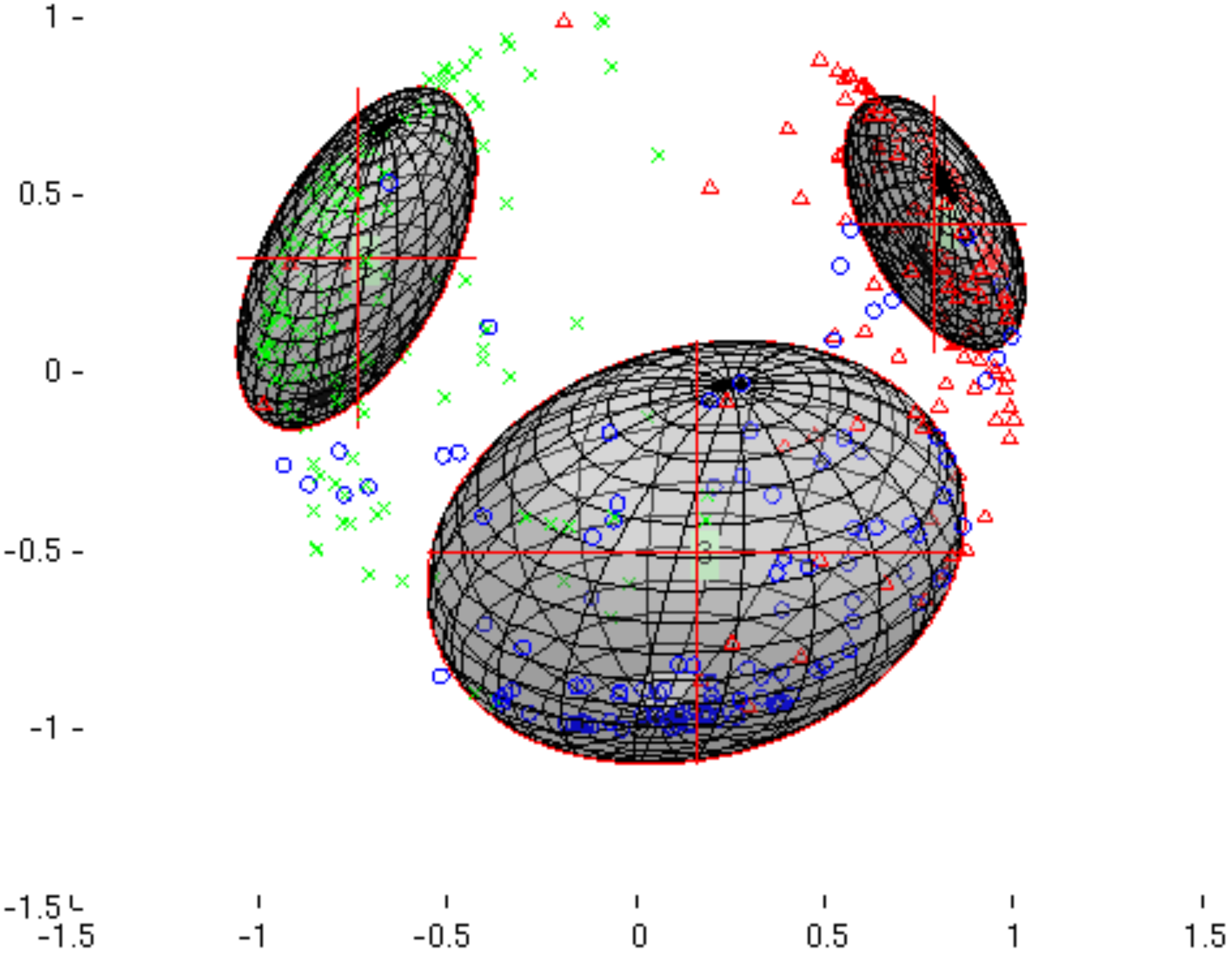
- **Proposed solution: Variational Bayes (VB)**
  - Fabio Valente (2005), Patrick Kenny (2010)

- **Advantages to being Bayesian**
  - In theory, Bayesian methods are not subject to the over-fitting that plagues maximum likelihood methods
    - **Quantitative version of Occam's razor**
    - **Should not need to resort to approximations such as BIC**

- **Variational Approximation** $P(x, y \mid w) \approx q(x) \cdot q(y)$

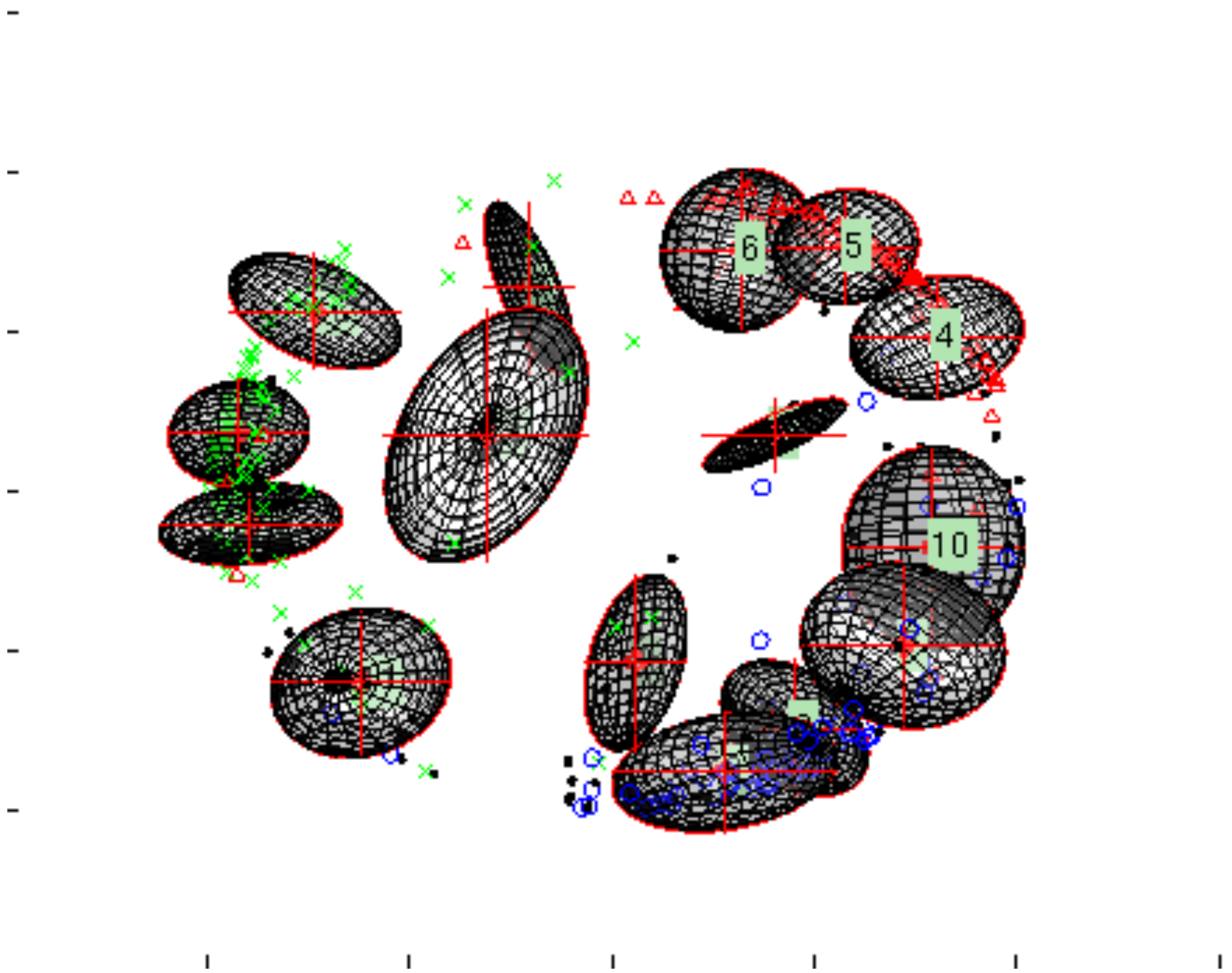
- **Other approaches**
  - *Sticky* HDP-HMM (Fox, 2008) and -HSMM (Johnson, 2010)
    - **Hierarchical Dirichlet Process (HDP)**
    - **Hidden Semi-Markov Model (HSMM)**

A Quick Visualization



A Quick Visualization

Another Visualization



Current Issues

- **All the data lies on the unit hypersphere**
  - Poorly modeled by a GMM
  - One possible direction
    - **Clustering on the Unit Hypersphere using von Mises-Fisher Distributions (Banerjee, 2005)**

- **Data sparsity**
  - A speaker may speak very infrequently
  - All i-vectors are weighted equally, but some are more equal than others
    - **How to incorporate information about the duration of speech used to extract a given i-vector?**

A Sampling Approach

- **So far, we have been using the i-vector w as a point estimate**
  - Recall w = E[w(u)], the expectation of the posterior distribution of w conditioned on the observed acoustic features u.
  - Furthermore, associated with this posterior distribution is a covariance

$$\mathrm{cov}(w,w) = l^{-1}(u) = \left(I + T^{*}\Sigma^{-1}N(u)T\right)^{-1}$$

  - "Size" of covariance is inversely proportional to number of frames N(u)
    - **More frames used to extract i-vector → smaller covariance**

- **Consider sampling this distribution for each i-vector**
  - Let the number of samples drawn be proportional to the number of frames used to extract the i-vector
    - **Shorter segments → larger covariance and fewer samples**
    - **Longer segments → smaller covariance and more samples**

Summary and Outlook

- **Factor analysis-based approach to speaker diarization**
  - Inspired by Total Variability and i-vectors
  - Key Insight
    - **Exploiting Intra-Conversation Variability**
  - Attained state of the art results on a test set of 2-speaker conversations

- **Further Work**
  - Detecting and removing overlapped speech segments
  - Extending to an unknown number of speakers
    - **Variational Bayes**
  - Addressing problems of data sparsity

Roadmap

- Introduction
  - Terminology, tasks, and framework
- Low-Dimensional Representation
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- Applications
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

Language Identification

Language Identification Outline

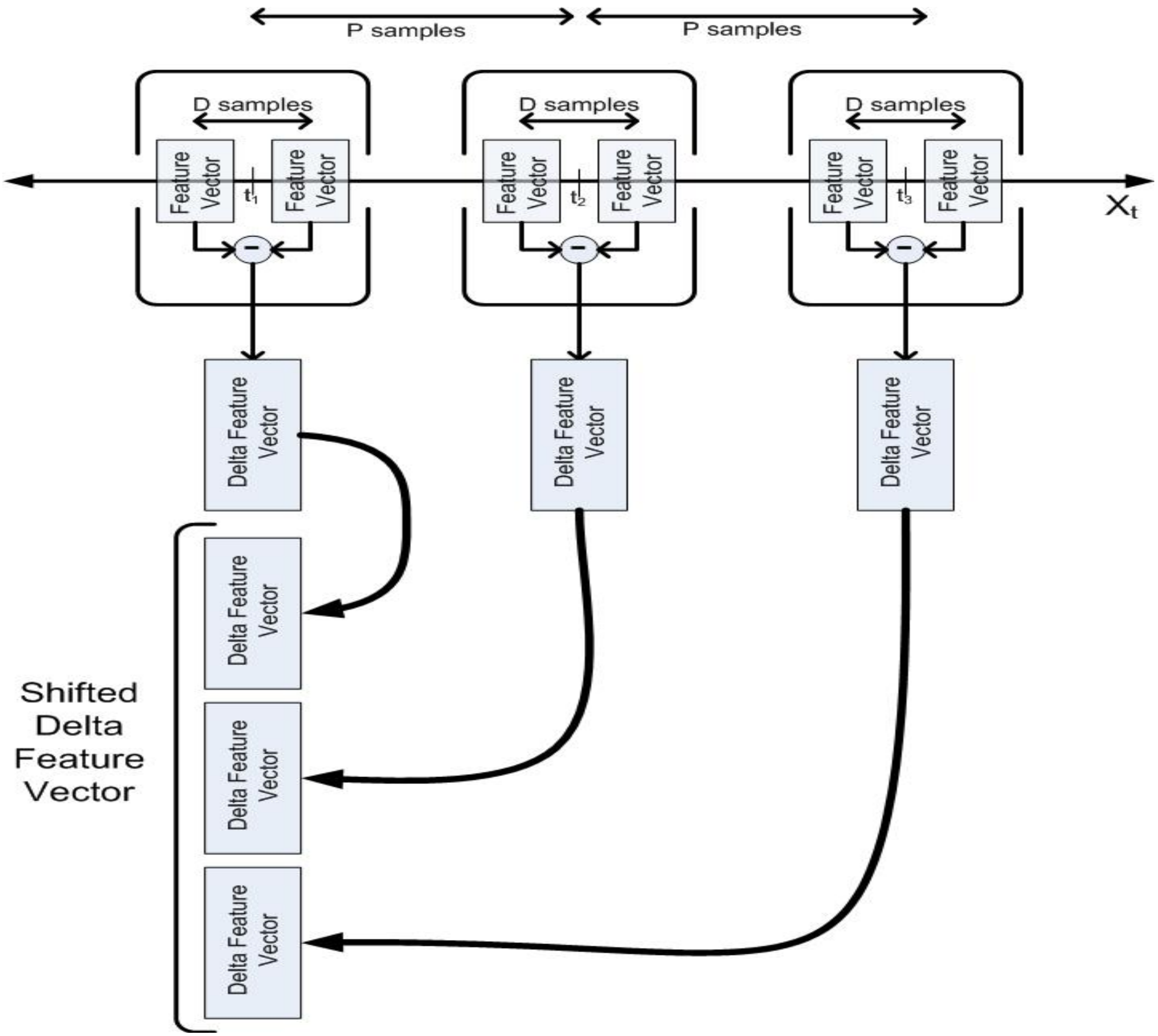
- Motivation
- Features extraction
- Intersession compensation and scoring
- NIST Language Recognition Evaluation
- Experiments and Results
- Interesting data visualization

Motivation

- **Low dimensional speech representation based on the Factor analysis**
  - Each speech recording is mapped on low dimensional vector (400)
- **Factor analysis as feature extractor**
  - Modeling the inter-language variability between different language classes
- **Score decision based on the cosine distance**
  - Simplicity of the system
- **Graph visualization to model connection between different languages**

Feature extraction for language Identification

- Shifted Delta Cepstral



Intersession compensation

- **Linear Discriminant Analysis to maximize the variability between the different language classes [Dehak 2009,2011]**

A is matrix of eigenvectors from $S_b.v = \lambda.S_w.v$

$$S_b = \sum_{i=1}^{L} (w_i - \overline{w})(w_i - \overline{w})^t \qquad \overline{w}: \text{ the mean of the entire population}$$

$$S_w = \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i=1}^{n_l} (w_i^l - w_l)(w_i^l - w_l)^t$$

- **Within Class Covariance Normalization is used to scale the component [Hatch2006] , [Dehak 2009,2011]**

$$W = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i=1}^{n_l} (A^t w_i^l - w_l)(A^t w_i^l - w_l)^t$$

$$w_l = \frac{1}{n_l} \sum_{i=1}^{n_l} A^t w_i^l \quad \text{mean for language class l}$$

$n_l =$ number of files for each language classl

$L =$ total number of language classes

Language Identification Scoring

- **The scoring is based on a dot product**
  - Normalizing the length of the i-vectors
- **Training**
  - Project the i-vectors with LDAA and WCCN$BB^t = W^{-1}$

$$w' = \frac{B^t A^t w}{\left\| B^t A^t w \right\|}$$

  - For each class *i* compute the mean and than normalize the length

$$m_i = \frac{\frac{1}{N}\sum_{j=1}^{N} w'}{\left\| \frac{1}{N}\sum_{j=1}^{N} w' \right\|}$$

- **Test**
  - Project the test i-vector with LDA and WCCN $w'_{test} = \dfrac{B^t A^t w_{test}}{\left\| B^t A^t w_{test} \right\|}$
  - Compute the dot product of the test i-vector with the normalized mean of each class

$$score_i = w'_{test} * m_i$$

NIST 2009 Language Recognition Evaluation

- Current work
  - 23 languages

	languages		languages
1	amharic	13	hindi
2	bosnian	14	korean
3	cantonese	15	mandarin
4	creole	16	pashto
5	croatian	17	portuguese
6	dari	18	russian
7	english_american	19	spanish
8	english_indian	20	turkish
9	farsi	21	ukrainian
10	french	22	urdu
11	georgian	23	vietnamese
12	hausa		

Experimental setup

- **Features**
  - 7-1-3-7 SDC + static ceptral vector
  - Feature normalization to N(0,1)
  - SAD using GMMSAD
- **UBM 2048 Gaussian Components**
- **Ivector of dimension 400 (the best performances)**
- Development set consists of both CTS + VOA Data
- GMM – MMI (2048 mixtures + feature-based FA)
- SVM-GSV (1024 GMM + feature-based NAP)

Results

	30s		10s		3s	
	BB	AB	BB	AB	BB	AB
I-vector	2.2%	-	4.8%	-	13.8%	-
GMM-MMI	7.9%	2.3%	10.8%	4.4%	17.9%	12.9%
SVM-GSV	7.5%	2.3%	11.2%	5.0%	20.4%	15.4%

- **BB : Before Backend**
- **AB : After Backend**
- **Results on Equal Error Rate**

Graph Visualization

- **Work at Exploring the variability between different languages.**
  - Visualization using the Graph Exploration System (GUESS) [Eytan 06]
- **Represent segment as a node with connections (edges) to nearest neighbors (3 NN used)**
  - Euclidean distance after i-vectors length normalization.
  - NN computed using TV system (with and without intersession compensation normalization)
  - Intersession compensation :
    - **Linear Discriminant Analysis + Within Class Covariance Normalization**
- **Applied to 4600 utterances from 30s condition of the NIST LRE09**
  - 200 utterances for Language class
- **Absolute locations of nodes not important**
- **Relative locations of nodes to one another is important:**
  - The visualization clusters nodes that are highly connected together
- **Colors represent Language Classes**

No intersession Compensation

■ georgian

■ hindi

■ french

■ croatian

■ urdu

■ amharic

■ portuguese

■ mandarin

■ korean

■ eng_Indian

■ bosian

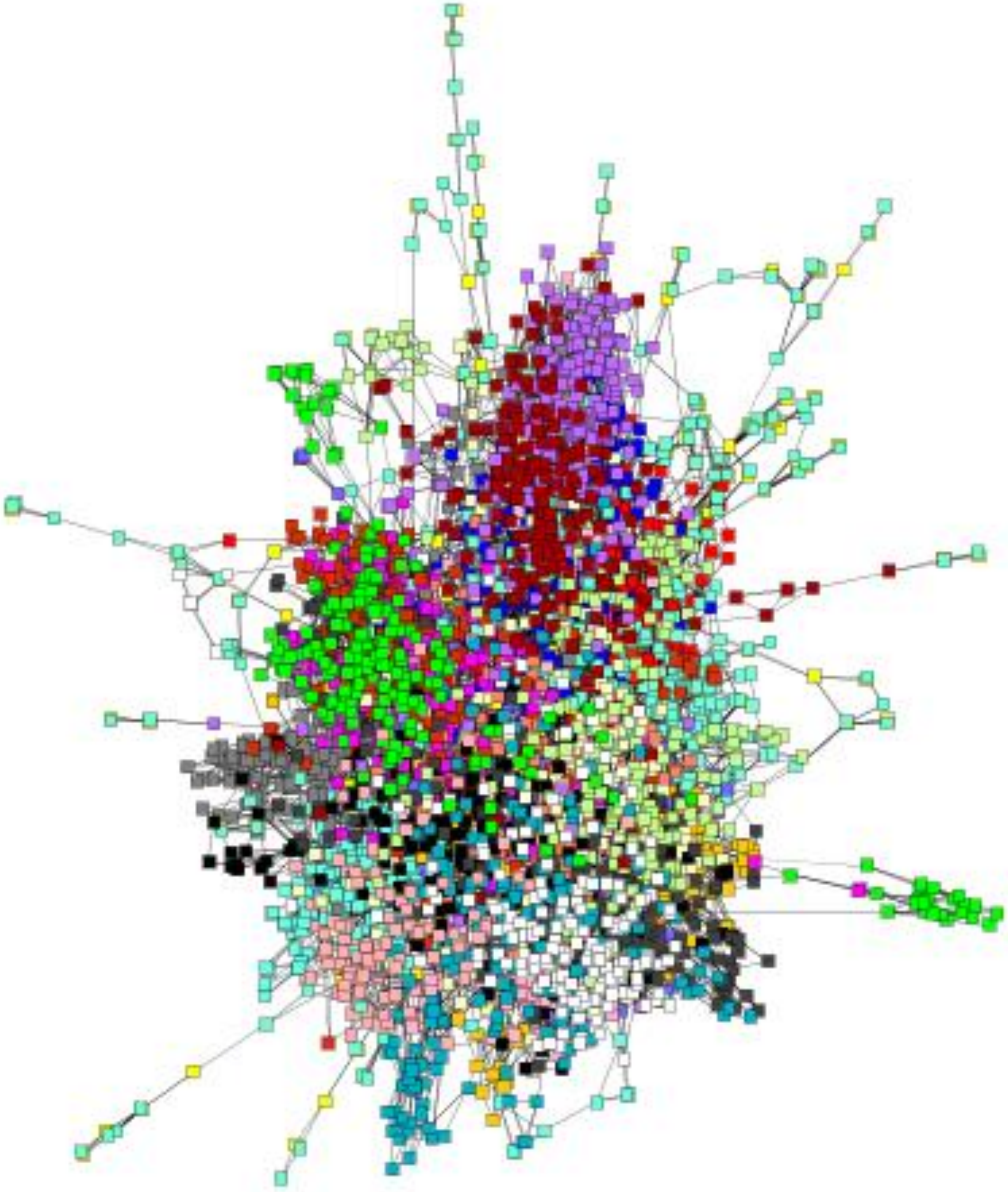
■ hausa

■ russian

■ pashto

■ cantonese

■ ukrainian



turkish ■

spanish ■

dari ■

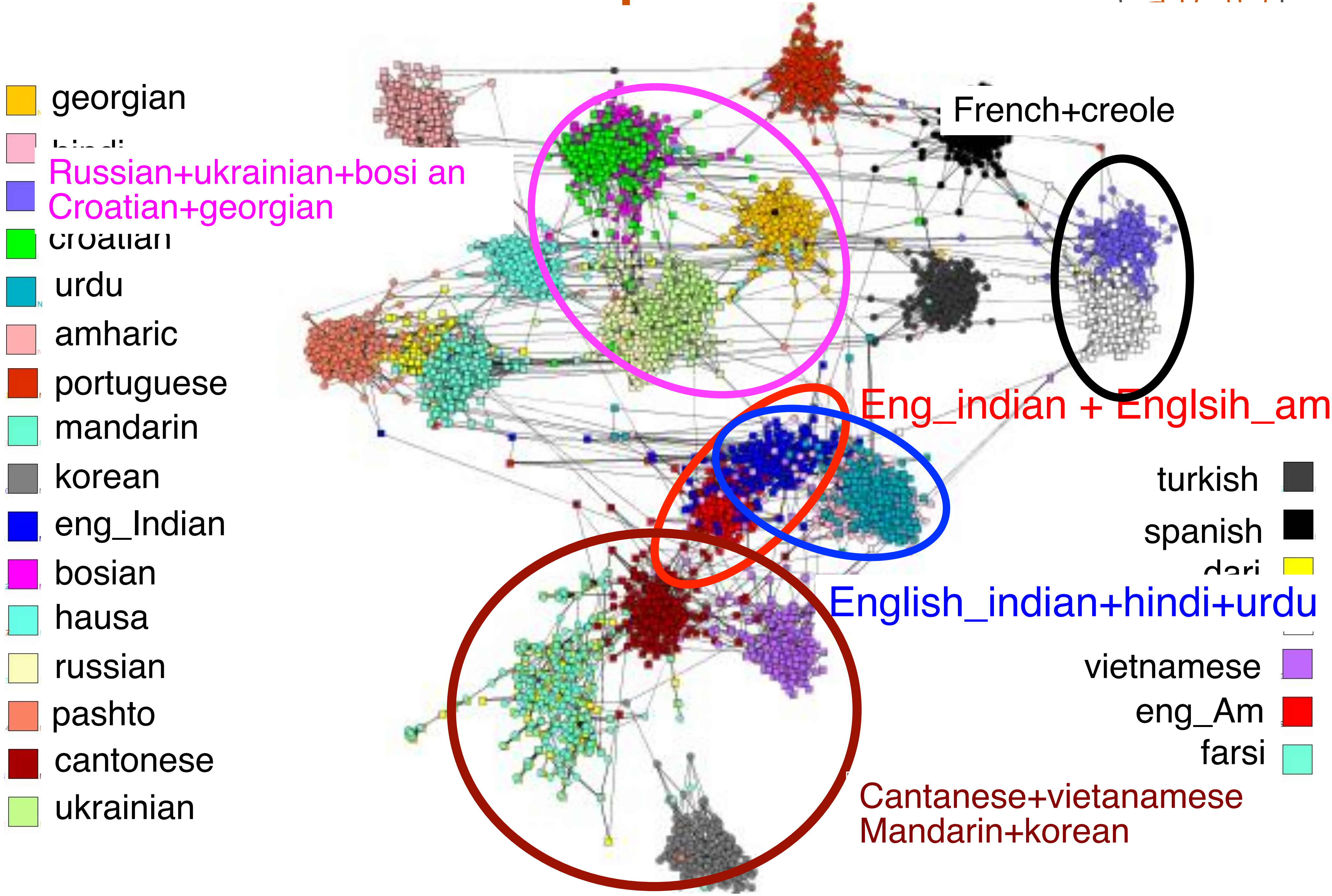
creole □

vietnamese ■

eng_Am ■

farsi ■

With intersession compensation



■ georgian
■ hindi
■ croatian
■ urdu
■ amharic
■ portuguese
■ mandarin
■ korean
■ eng_Indian
■ bosian
■ hausa
■ russian
■ pashto
■ cantonese
■ ukrainian

Russian+ukrainian+bosi an
Croatian+georgian

French+creole

Eng_indian + Englsih_am

English_indian+hindi+urdu

Cantanese+vietanamese
Mandarin+korean

turkish ■
spanish ■
dari ■
vietnamese ■
eng_Am ■
farsi ■

Discussion

- **Introduce the use of the i-vector framework**
- **Dot product (cosine) obtained very comparable results**
- **We show very interesting Data visualization**
  - Show the connections between different languages
- **We have an INTERSPEECH 2011 paper in other Dimensionality reduction techniques for Language Identification using the i-vector space.**
  - Najim Dehak, Pedro A. Torres-Carrasquillo, Douglas Reynolds, Reda Dehak “ Language Recognition via Ivectors and Dimensionality Reduction”

Roadmap

- Introduction
  - Terminology, tasks, and framework
- Low-Dimensional Representation
  - Sequence of features: GMM
  - Super-vectors: JFA
  - Low-dimensional vectors: i-vectors
  - Processing i-vectors: compensation and scoring
- **Applications**
  - Speaker verification
  - Speaker diarization
  - Language recognition
  - Emotion recognition

Emotion recognition

Motivation

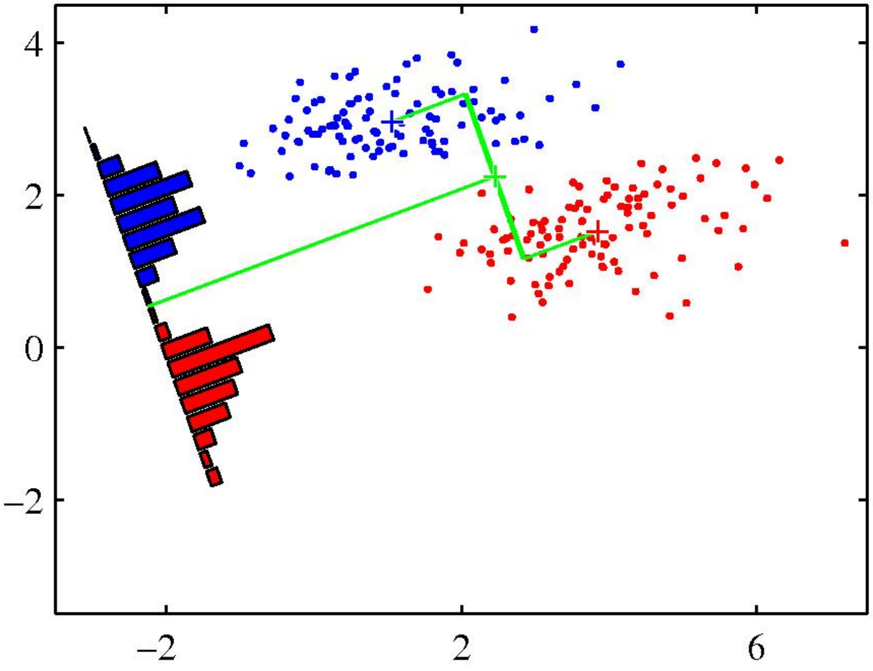
- **Using the ivector framework in two classes emotion recognition problem**
  - Ideal vs Negative
- **INTERSPEECH 2009 Emotion recognition Challenge**
  - FAU, AIBO Emotion corpus
  - Children voice recordings
  - It contains :
    - **9959 train files**
    - **8257 test files**

Fisher Discriminant Analysis

- The mean of each class

$$\overline{w}_1 = \frac{1}{n_1}\sum_{i\in C_1} w_i$$

$$\overline{w}_2 = \frac{1}{n_2}\sum_{i\in C_2} w_i$$



- Within class covariance

$$S_w = \sum_{i\in C_1}(w_i - \overline{w}_1)(w_i - \overline{w}_1)^t + \sum_{i\in C_2}(w_i - \overline{w}_2)(w_i - \overline{w}_2)^t$$

- Direction

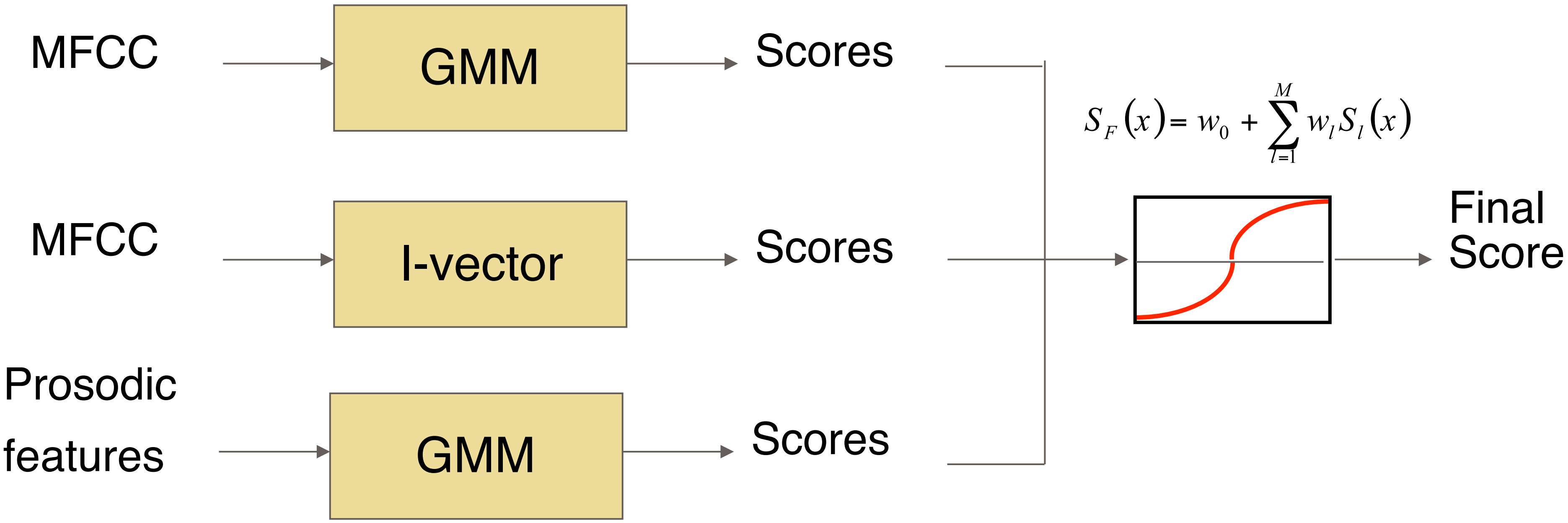
$$v \propto S_w^{-1}(\overline{w}_2 - \overline{w}_1)$$

Experimental Set up

GMM-MFCC system	•MFCC feature (12 MFCC+E+delta+double delta) •512 Gaussian Components
Ivector system	•MFCC feature (12 MFCC+E+delta+double delta) •UBM with 512 Gaussian Components •Ivector dimension of 150 •Fisher Discriminant Analysis
GMM-Prosodic system	•Legendre polynomials coefficients for the pitch and energy contours in the pseudo syllables level. •(6 for pitch+6 for energy + duration of the pseudo-syllables). •UBM with 256 components •MAP adaptation

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R. and Boufaden, N. "Cepstral and Long-Term Features for Emotion Recognition" In *Interspeech 2009, pp. 344-347. Brighton, UK, September 6-10, 2009.*

Score Fusion



MFCC → GMM → Scores

MFCC → I-vector → Scores

Prosodic features → GMM → Scores

$$S_F(x) = w_0 + \sum_{l=1}^{M} w_l S_l(x)$$

Final Score

Results: 2 class problem

System	Unweighted recall
S1 : GMM w/ MFCC	69.72%
S2 : Ivector w/ MFCC	**69.81%**
S3 : GMM-UBM w/ long-term feature	66.61%
S1+S2+S3	**70.54%**

Discussion

- **We introduce the use of the i-vector framework for emotion recognition**
- **A simple Fisher Discriminant Analysis achieved the state of the art results for two emotion classification problem**
- **Future works**
  - Try the i-vector for multiple classes emotion problem
  - Try the i-vector with other classifiers

Final Words

- **Aim**
  - To provide an overview of theory and operation of modern low-dimensional speech representations and their application to automatic speaker, language, emotion recognition and diarization

- **Participants should have gained an introduction to and understanding of:**
  - *Subspace Representation of Speech Signals*
  - *Algorithms for Joint-Factor Analysis and Total-Variability Modeling*
  - *Application of subspace representations to automatic speaker, language, emotion recognition and diarization systems*

References

- Dehak, N " Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristic Modeling : Application to Speaker Verification " PhD thesis, ETS, Montreal 2009.
- Dehak, N., Kenny, P., Dumouchel. P., Dehak, R., Ouellet. P., «Front-end factor analysis for speaker verification » in IEEE Transactions on Audio, speech and Language Processing 2011.
- Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet and Pierre Dumouchel, Support Vector Machine versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In Proc INTERSPEECH 2009, Brighton, UK, September 2009.
- Kenny, P., Ouellet, P., Dehak, N., Gupta, V. and Dumouchel, P. "A Study of Inter-Speaker Variability in Speaker Verification" IEEE Transactions on Audio, Speech and Language Processing, 16 (5) July 2008 : 980-988.
- D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, pp. 19–41, 2000.

Reference

- Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006
- Z. N. Karam, W. M. Campbell “Graph-Embedding for Speaker Recognition”, Submitted to Interspeech 2010

Proof (II)

- Let

$$\bar{Y}_t = \begin{bmatrix} y_t \\ y_t \\ \vdots \\ y_t \end{bmatrix}_{C\times F} \qquad \bar{\gamma}_t = \begin{bmatrix} [\gamma_t(1)]_F \\ [\gamma_t(2)]_F \\ \vdots \\ [\gamma_t(C)]_F \end{bmatrix}_{C\times F}$$

- Then, recalling M = m+Tw,

$$P_{T,\Sigma}(w \mid u) \propto P_{T,\Sigma}(\{y_1,\ldots,y_L\} \mid w) \cdot N(w \mid 0,I)$$

$$= \Pi_{t=1}^{L} P_{T,\Sigma}(y_t \mid w) \cdot N(w \mid 0,I)$$

$$\propto \exp\left(-\frac{1}{2} \sum_t \bar{\gamma}_t (\bar{Y}_t - (m+Tw))^t \Sigma^{-1} (\bar{Y}_t - (m+Tw)) \right)$$

$$\cdot \exp\left(-\frac{1}{2} w^t w \right)$$

Proof (III)

$$
\begin{aligned}
P_{m,T,\Sigma}(w \mid u) &\propto \exp\left(-\frac{1}{2}\sum_t \bar{\gamma}_t(\bar{Y}_t-(m+Tw))^t\Sigma^{-1}(\bar{Y}_t-(m+Tw))\right)\cdot \exp\left(-\frac{1}{2}w^t w\right) \\
&= \exp\left(-\frac{1}{2}\sum_t \bar{\gamma}_t((\bar{Y}_t-m)^t\Sigma^{-1}(\bar{Y}_t-m)\right. \\
&\qquad\qquad \left. -2w^tT^t\Sigma^{-1}(\bar{Y}_t-m)+w^tT^t\Sigma^{-1}Tw)-\frac{1}{2}w^t w\right) \\
&\propto \exp\left(w^tT^t\Sigma^{-1}\sum_t \bar{\gamma}_t(\bar{Y}_t-m)-\frac{1}{2}w^tT^t\Sigma^{-1}Tw\sum_t \bar{\gamma}_t-\frac{1}{2}w^t w\right) \\
&= \exp\left(w^tT^t\Sigma^{-1}\tilde{F}(u)-\frac{1}{2}w^tT^t\Sigma^{-1}N(u)Tw-\frac{1}{2}w^t w\right)
\end{aligned}
$$

Proof (IV)

$$P_{m,T,\Sigma}(w \mid u) \propto \exp\left(w^t T^t \Sigma^{-1} \tilde{F}(u) - \frac{1}{2} w^t T^t \Sigma^{-1} N(u) T w - \frac{1}{2} w^t w \right)$$

$$= \exp\left(w^t T^t \Sigma^{-1} \tilde{F}(u) - \frac{1}{2} w^t (T^t \Sigma^{-1} N(u) T + I) w \right)$$

$$= \exp\left(-\frac{1}{2} (w^t l(u) w - 2 w^t (l(u) \cdot l^{-1}(u)) T^t \Sigma^{-1} \tilde{F}(u)) \right)$$

$$\propto \exp\left(-\frac{1}{2} (w - \underbrace{l^{-1}(u) T^t \Sigma^{-1} \tilde{F}(u)}_{E[w(u)]})^t \, l(u) \, (w - \underbrace{l^{-1}(u) T^t \Sigma^{-1} \tilde{F}(u)}_{E[w(u)]}) \right)$$

$$= \exp\left(-\frac{1}{2} (w - E(u)) \right.$$

$$\mathrm{cov}(w(u), w(u)) = l^{-1}(u)$$