A Text-Independent Speaker Recognition System

Catie Schwartz
Advisor: Dr. Ramani Duraiswami

Final Presentation AMSC 664
May 10, 2012

Outline

- Introduction
- Review of Algorithms
  1. Mel Frequency Cepstral Coefficients (MFCCs)
  2. Voice Activity Detector (VAD)
  3. Expectation-Maximization (EM) for Gaussian Mixture Models (GMM)
  4. Maximum A Posteriori (MAP) Adaptation
  5. Block Coordinate Decent Minimization (BCDM) for Factor Analysis (FA)
  6. Linear Discriminant Analysis (LDA)
- Review of Classifiers
  1. Log-Likelihood Test (LLT)
  2. Cosine Distance Scoring (CDC)
- Databases
  1. TIMIT
  2. SRE 2004, 2005, 2006, 2010
- Results
- Summary

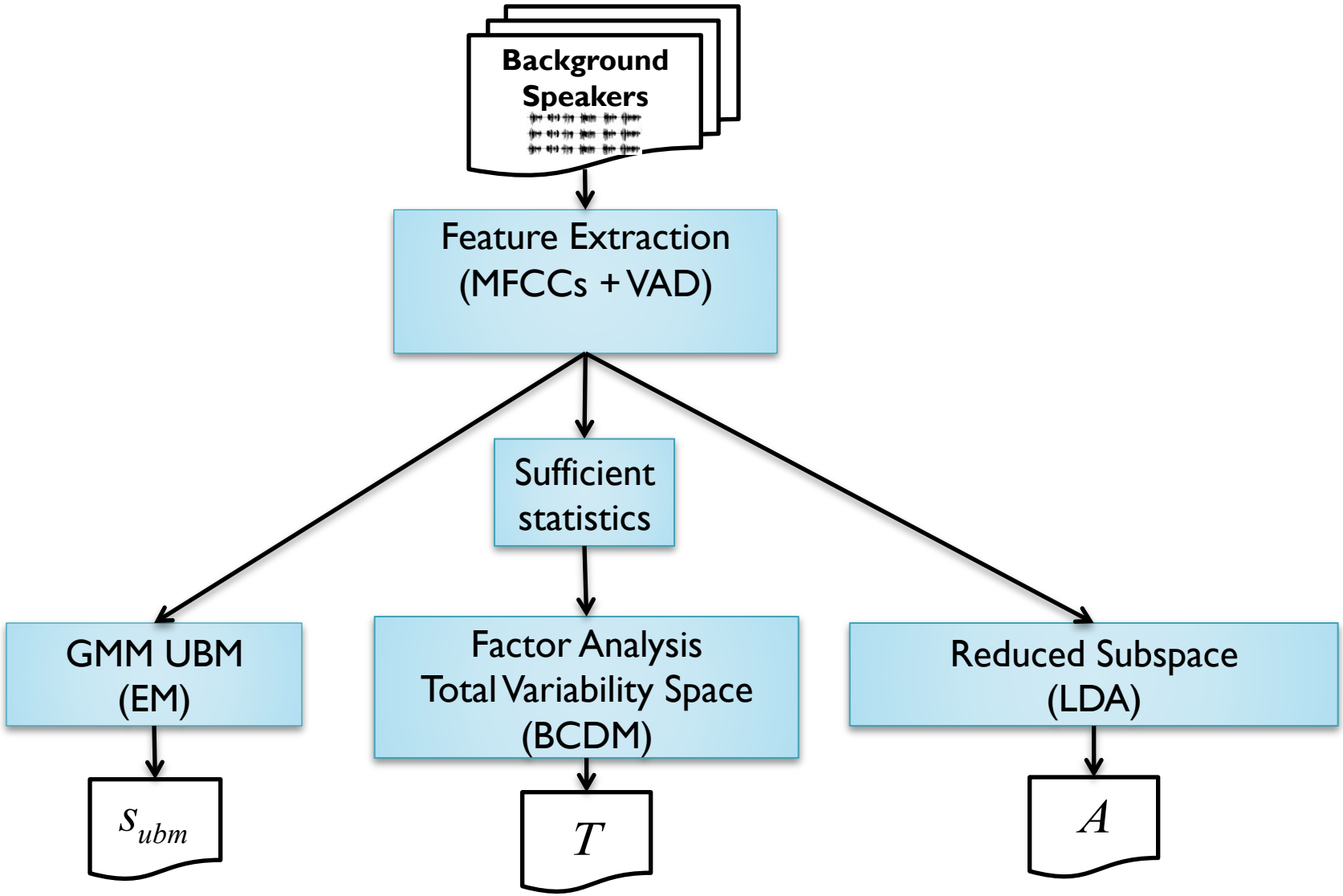
Introduction: Speaker Recognition

- Given two audio samples, do they come from the same speaker?
- Text-independent: no requirement on what the test speaker says in the verification phase
- Needs to be robust against channel variability and speaker dependent variability

Introduction: 663/664 Project

- 3 different speaker recognition systems have been implemented in MATLAB
  - GMM Speaker Models
  - i-vector models using Factor Analysis techniques
  - LDA reduced i-vectors models
- All build off of a "Universal Background Model" (UBM)

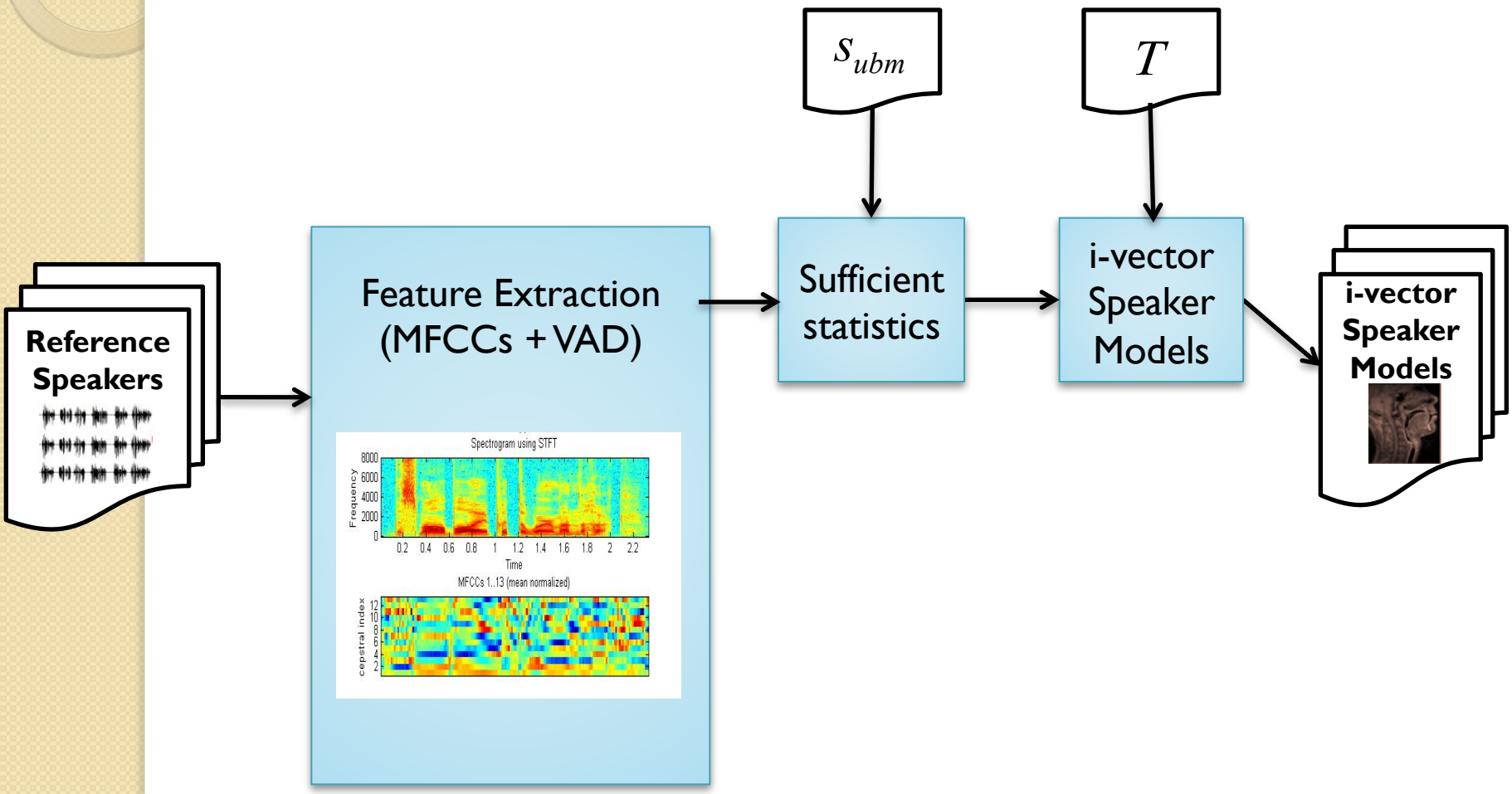
Algorithm Flow Chart Background Training



Algorithm Flow Chart

Factor Analysis/i-vector Speaker Models

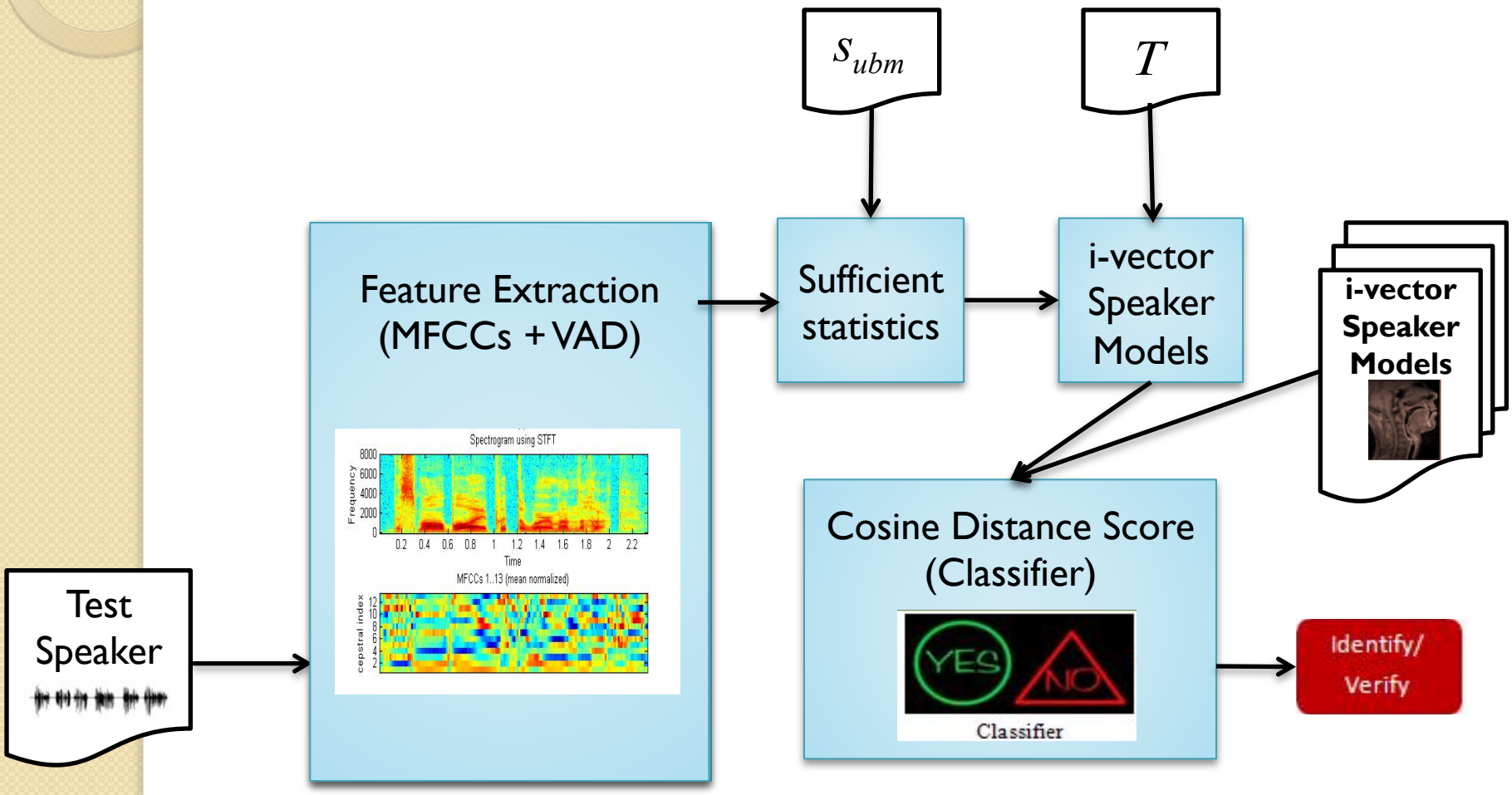
Enrollment Phase



Algorithm Flow Chart

Factor Analysis/i-vector Speaker Models

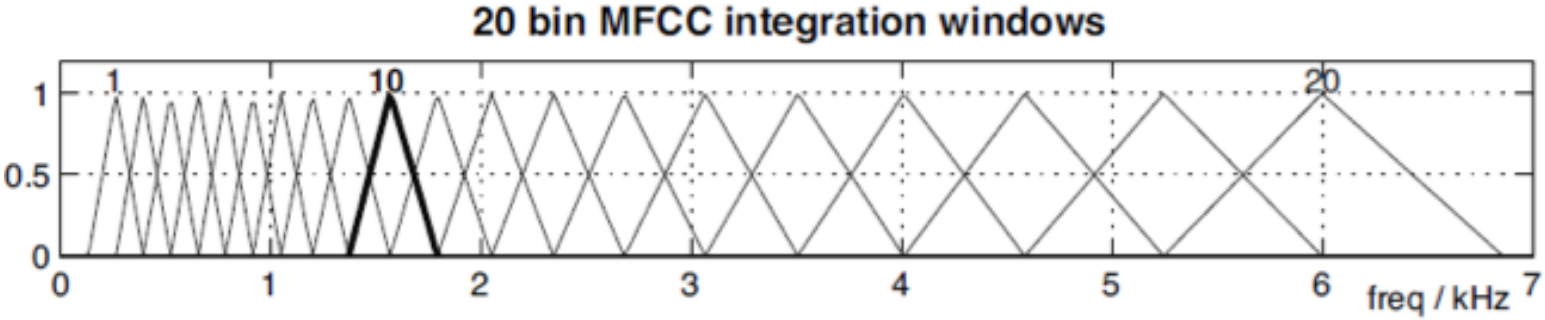
Verification Phase



Review of Algorithms

1. Mel-frequency Cepstral Coefficients (MFCCs)

- Low-level feature (20 ms)
- Bin in frequency domain



- Convert to cepstra

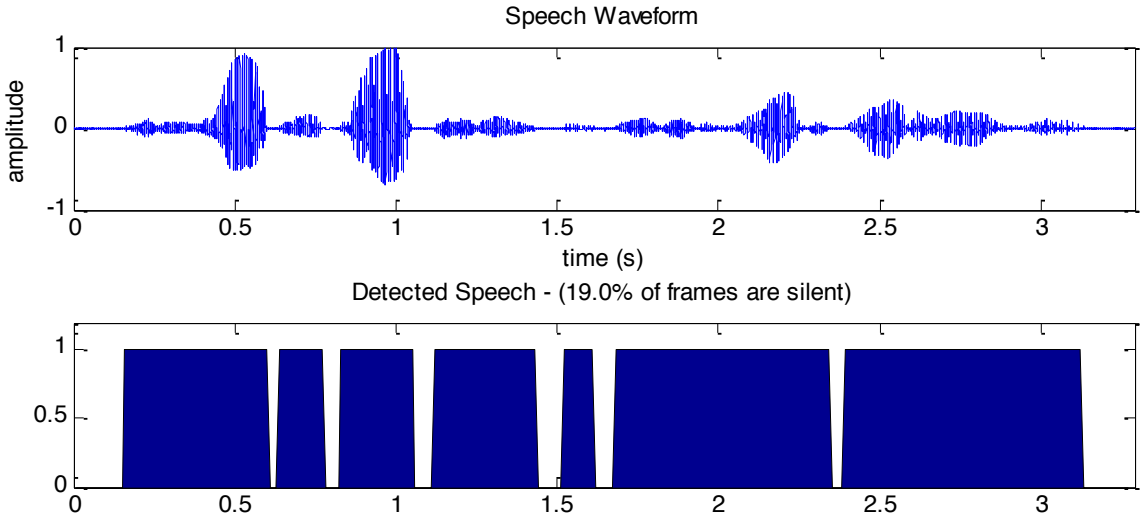
$$c_n = \sum_{m-1}^{M} [\log Y(m)] \cos\left[\frac{\pi n}{m}\left(m - \frac{1}{2}\right)\right]$$

- Derivatives of MFCCs can be used as features as well

Ellis, Daniel P.W. *PLP and RASTA (and MFCC, and Inversion) in Matlab. PLP and RASTA (and MFCC, and Inversion) in Matlab.* Vers. Ellis05-rastamat. 2005. Web. 1 Oct. 2011. <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>.

Review of Algorithms
2. Voice Activity Detector (VAD)

- Energy based
  - Remove any frame with either less than 30 dB of maximum energy or less than -55dB overall



- Can be combined with Automatic Speech Recognition (ASR) if provided

Kinnunen, Tomi, and Haizhou Li. "An Overview of Text-independent Speaker Recognition: From Features to Supervectors." *Speech Communication* 52.1 (2010): 12-40. Print.

Review of Algorithms

3. Expectation Maximization (EM) for Gaussian Mixture Models (GMM)

- The EM Algorithm is used for training the Universal Background Model (UBM)
- UBM assume speech in general is represented by a finite mixture of multivariate Gaussians

$$p(x_t \mid s) = \sum_{k=1}^{K} \pi_k N(x_t \mid \boldsymbol{\mu}_k, \Sigma_k)$$

Reynolds, Douglas A., and Richard C. Rose. "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Transations on Speech and Audio Processing* IEEE 3.1 (1995): 72-83. Print.

Review of Algorithms

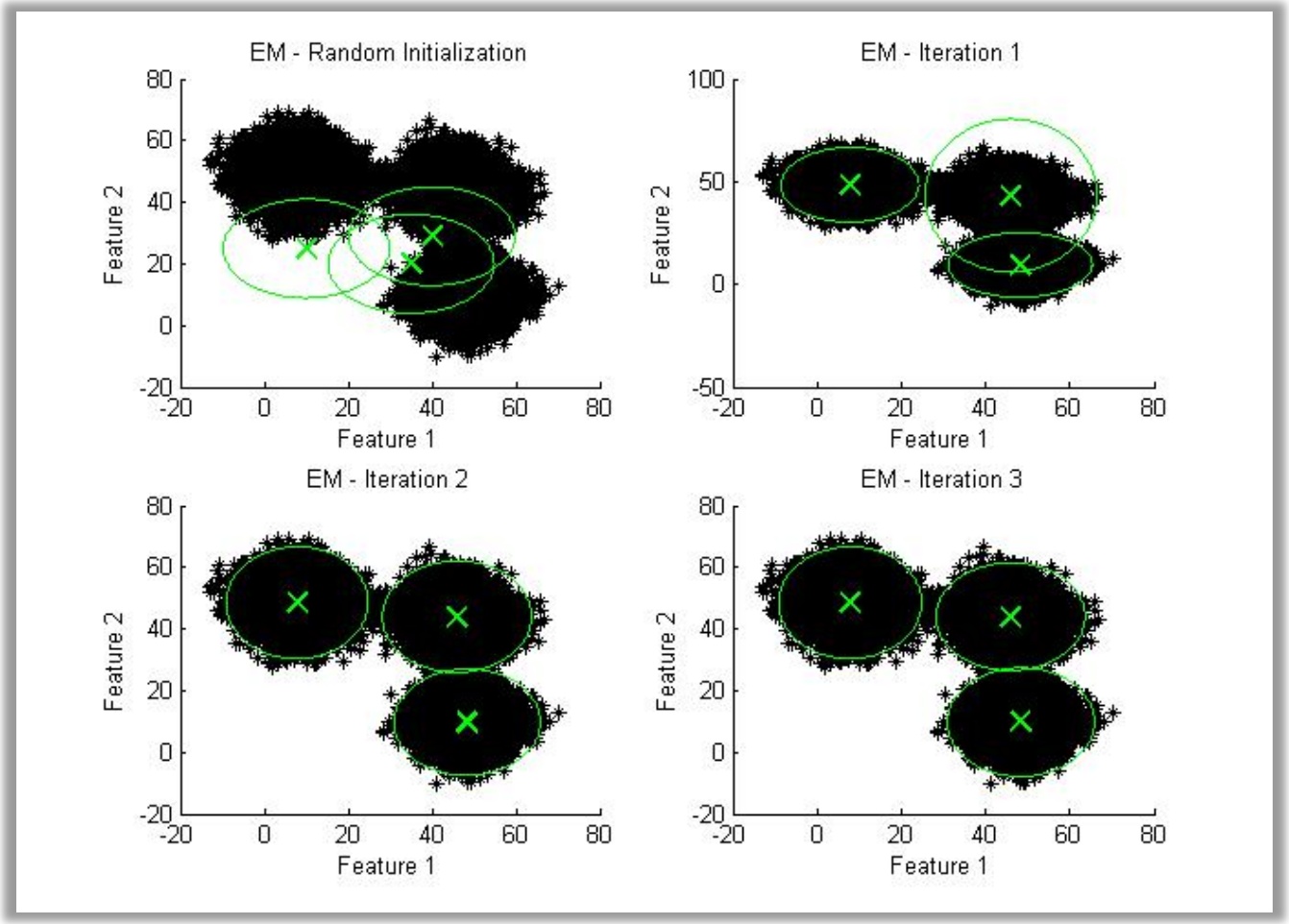
3. Expectation Maximization (EM) for Gaussian Mixture Models (GMM)

1. Create a large matrix containing all features from all background speakers
   - Can down-sample to 8 times the number of variables
2. Randomly initialize parameters (weights, means, covariance matrices) or use k-means
3. Obtain conditional distribution of each component c
4. Maximize mixture weights, means and covariance matrices
5. Repeat steps 3 and 4 iteratively until convergence

Reynolds, Douglas A., and Richard C. Rose. "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Transations on Speech and Audio Processing* IEEE 3.1 (1995): 72-83. Print.

Review of Algorithms

3. Expectation Maximization (EM) for Gaussian Mixture Models (GMM)



Review of Algorithms

4. Maximum a Posteriori (MAP) Adaptation

- Used to create the GMM Speaker Models (SM)

1. Obtain conditional distribution of each component c based on UBM:

$$\gamma_t(c) = p(c \mid \mathbf{x}_t^i, s^{UBM}) = \frac{\pi_c^{UBM} N(\mathbf{x}_t^i \mid \mu_c^{UBM}, \Sigma_c^{UBM})}{\sum_{k=1}^{K} \pi_c^{UBM} N(\mathbf{x}_t^i \mid \mu_c^{UBM}, \Sigma_c^{UBM})}$$

2. Maximize mean:

$$\mu_c = \frac{\sum_{t=1}^{T} \gamma_t(c) \mathbf{x}_t^i}{\sum_{t=1}^{T} \gamma_t(c)}$$

3. Calculate: $\mu_c^{sm_i} = \alpha_c^m \mu_c + (1 - \alpha_c^m) \mu_c^{ubm}$

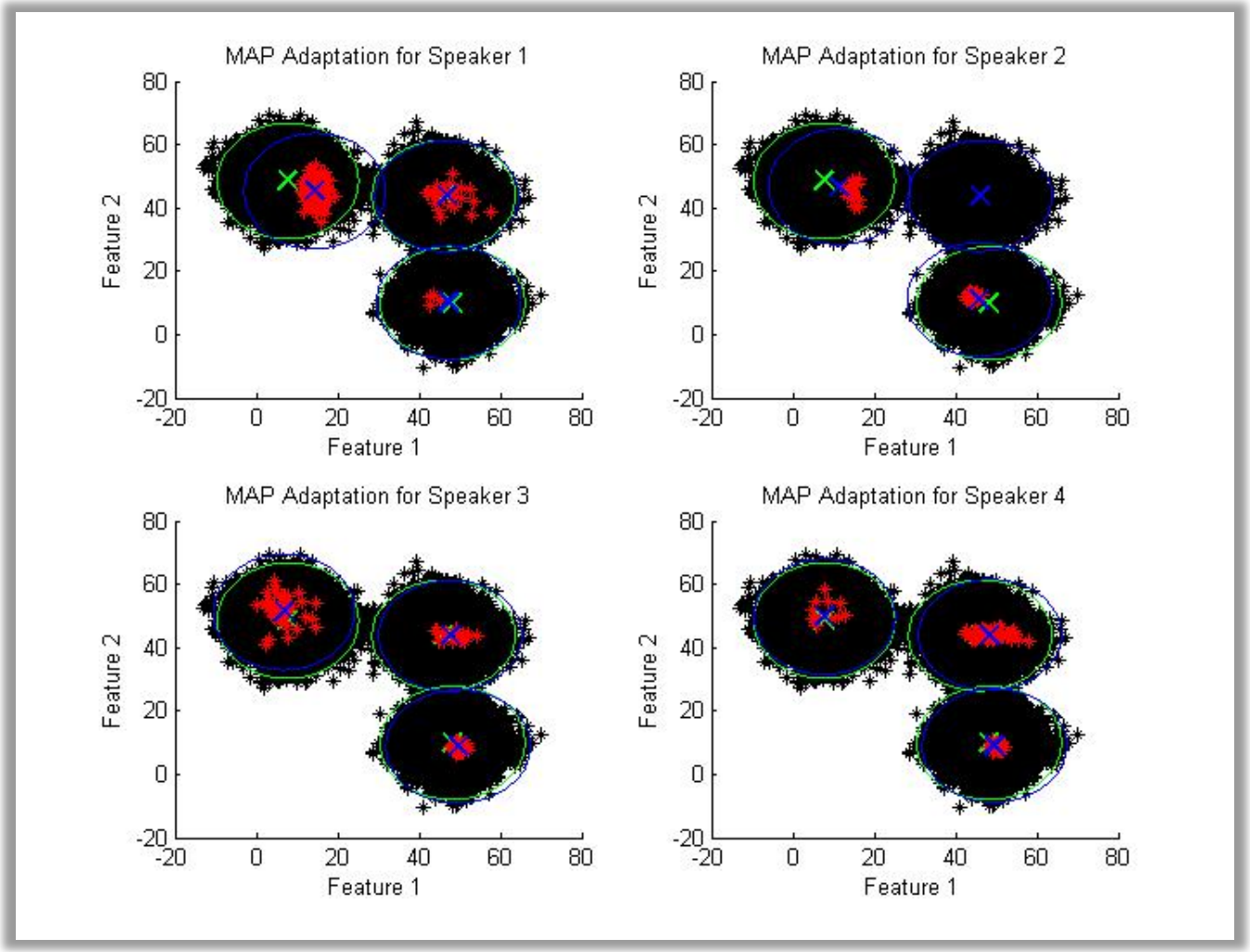
where

$$\alpha_c^m = \frac{\sum_{t=1}^{T} \gamma_t(c)}{\sum_{t=1}^{T} \gamma_t(c) + r}$$

Reynolds, D. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing* 10.1-3 (2000): 19-41. Print.

Review of Algorithms

4. Maximum a Posteriori (MAP) Adaptation



Review of Algorithms

5. Block Coordinate Descent Minimization (BCDM) for Factor Analysis (FA)

- Assume that the MFCCs from each utterance comes from a 2-stage generative model
  1. K-component GMM where the weights and covariance matrices come from a UBM

$$p(x_t \mid s) = \sum_{k=1}^{K} \pi_k N(x_t \mid \mu_k, \Sigma_k)$$

  2. Means of the GMM come from a second stage generative model called a factor analysis model

$$\mu = m + Tw$$

where $\mu, m \in \Re^{KD}$, $T \in \Re^{KD \times pT}$ and $w \in \Re^{pT}$

D. Garcia-Romero and C.Y. Espy-Wilson, "Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries," in *Proceedings of Odyssey 2010:The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, pp. 43-51.

Review of Algorithms

5. Block Coordinate Descent Minimization (BCDM) for Factor Analysis (FA)

i-vector training:

- Given $\mathbf{x} = \{x_t\}_{t=1}^{T}$ and fixed **T**, we want

$$\max_{\mu} p(\mu \mid \mathbf{x}) = \max_{\mu} p(\mathbf{x} \mid \mu) p(\mu)$$
$$= \min_{\mu} \{-\log(p(\mathbf{x} \mid \mu = m + Tw) - \log(p(w))\}$$

where $p(w) = N(0, I)$

- This turns into minimizing:

$$\Psi(w) = \frac{1}{2} \left\| \mathbf{W}^{\frac{1}{2}} (\eta - \mathbf{T}w) \right\|_2^2 + \frac{1}{2} \|w\|_2^2$$

where $\mathbf{W} = \Gamma \Sigma^{-1}, \Gamma = diag(\Gamma_k), \ \Gamma_k = \gamma_k \mathbf{I}$

and $\eta = \mu - m$ are the sufficient statistics

Review of Algorithms

5. Block Coordinate Descent Minimization (BCDM) for Factor Analysis (FA)

i-vector training (cont):

1. Obtain sufficient statistics $(\eta, \mathbf{W})$

2. Given $\mathbf{W}$ is positive semi-definite,

$$\Psi(w) = \frac{1}{2}\left\|\mathbf{W}^{\frac{1}{2}}(\eta - \mathbf{T}w)\right\|_2^2 + \frac{1}{2}\|w\|_2^2$$

is strongly convex and therefore the minimization problem has a closed form solution:

$$w = (\mathbf{I} + \mathbf{T}^T\mathbf{W}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{W}\eta$$

D. Garcia-Romero and C.Y. Espy-Wilson, "Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries," in *Proceedings of Odyssey 2010:The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, pp. 43-51.

Review of Algorithms

5. Block Coordinate Descent Minimization (BCDM) for Factor Analysis (FA)

Total variability space training:

- Given $D = \{\mathbf{X}_r\}_{r=1}^{R}$ with *R* utterances, and each *r* being represented by $(\eta_r, \mathbf{W}_r)$:

$$\min_{\mathbf{T},\{w_r\}} \sum_{r=1}^{R} \left(\left\| \mathbf{W}_r^{\frac{1}{2}} (\eta_r - \mathbf{T} w_r) \right\|_2^2 + \left\| w_r \right\|_2^2 \right)$$

- Solved using block coordinate descent minimization

D. Garcia-Romero and C.Y. Espy-Wilson, "Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries," in *Proceedings of Odyssey 2010:The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, pp. 43-51.

Review of Algorithms

5. Block Coordinate Descent Minimization (BCDM) for Factor Analysis (FA)

Total variability space training (cont):

Alternating optimization:

1. Assume $\mathbf{T}$ is fixed, minimize w_r

$$w_r = (\mathbf{I} + \mathbf{T}^T \mathbf{W}_r \mathbf{T})^{-1} \mathbf{T}^T \mathbf{W}_r \eta_r$$

2. Assume w_r are fixed, minimize $\mathbf{T}$

$$\min_{\mathbf{T}} \sum_{r=1}^{R} \left\| \mathbf{W}_r^{\frac{1}{2}} (\eta_r - \mathbf{T} w_r) \right\|_2^2$$

with the solution $\mathbf{T}_{new}$ where

$$\mathbf{T}_{new}^{(k)} \left(\sum_{r=1}^{R} \gamma_{rk} w_r w_r^T \right) = \sum_{r=1}^{R} \gamma_{rk} \eta_r^{(k)} w_r^T$$

Review of Algorithms

6. Linear Discriminant Analysis (LDA)

Find matrix $\mathbf{A}$ such that for $\omega = \mathbf{A}^T w$ the between speaker covariance of ω:

$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t$$

is maximized while the within-speaker covariance:

$$S_W = \sum_{i=1}^{c} \sum_{x \in D_i} (x - m_i)(x - m_i)^t$$

is minimized, which leads to the eigenvalue problem

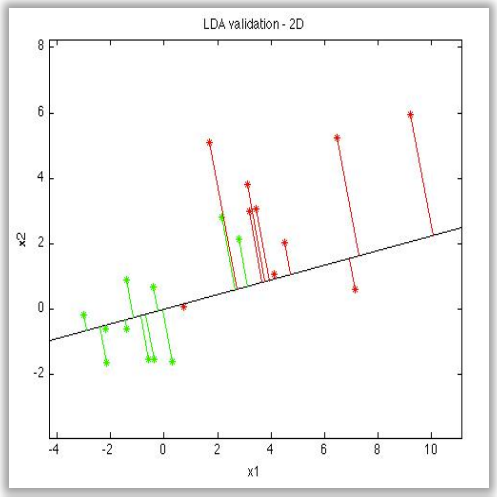
$$S_B a_i = \lambda_i S_W a_i$$

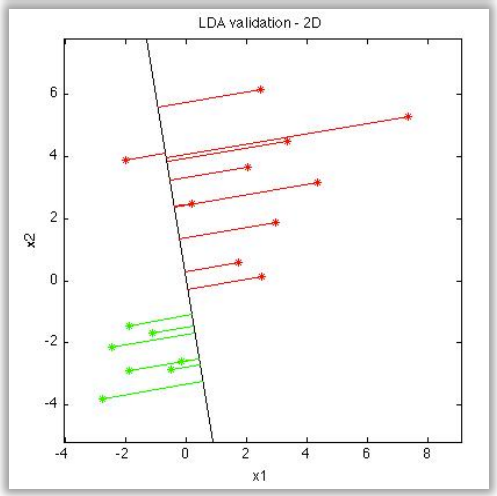
Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*. New York: Wiley, 2001. Print.

Review of Algorithms

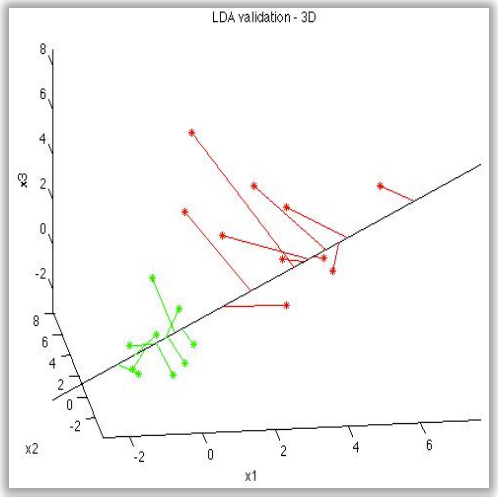
6. Linear Discriminant Analysis (LDA)

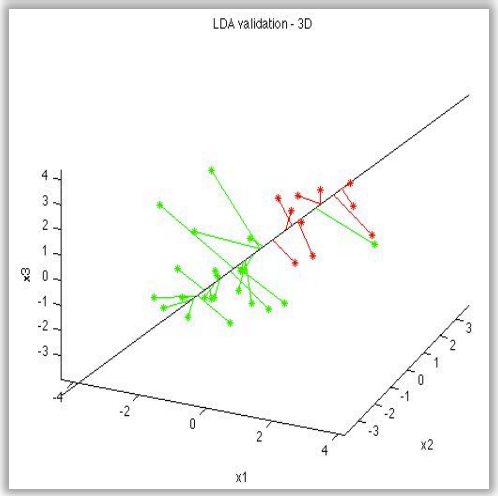
2-D





3-D





Summary of Algorithm Validation

Algorithm	Validation Technique
MFCCs	Compared modified code to original code by Dan Ellis
VAD	Audio evaluation and visual inspection
EM for GMM	Visual inspection for 2D feature space with 3 Gaussian components to check for convergence • Analysis on various k values used in algorithm ($k < 3, k = 3, k > 3$) Compare results with vetted system (Lincoln Labs)
MAP Adaptation	Visual inspection for 2D features space with 3 Gaussian components • Analysis on speakers with varying levels of representation for different components Compare results with vetted system (Lincoln Labs)
BCDM for FA	Attempted to create dataset and compare orthonormal project onto the range of the total variability space • Determined too non-linear for validation method to be valid Compare results with vetted system (BEST Project Code)
LDA	Visual inspection in 2D and 3D space

Review of Classifiers

1. Log-Likelihood Test (LLT)

- Used for GMM speaker models
- Compare a sample speech to a hypothesized speaker

$$\Lambda(x) = \log p(x \mid s_{hyp}) - \log p(x \mid s_{ubm})$$

where $\Lambda(x) \geq \theta$ leads to verification of the hypothesized speaker and $\Lambda(x) < \theta$ leads to rejection

Reynolds, D. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing* 10.1-3 (2000): 19-41. Print.

Review of Classifiers

1. Cosine Distance Scoring (CDC)

- Used for FA i-vector and LDA reduced i-vector models
- $$score(w_1, w_2) = \frac{w_1^* \cdot w_2}{\|w_1\| \cdot \|w_2\|} = \cos\left(\theta_{w_1, w_2}\right)$$

  where $score(w_1, w_2) \geq \varphi$ leads to verification of the hypothesized speaker and $score(w_1, w_2) < \varphi$ leads to rejection

Lei, Howard. "Joint Factor Analysis (JFA) and i-vector Tutorial." *ICSI.Web. 02 Oct. 2011.* http://www.icsi.berkeley.edu/Speech/presentations/AFRL_ICSI_visit2_JFA_tutorial_icsitalk.pdf

Databases

1. TIMIT

- 2/3 of the database used as background speakers (UBM, TVS and LDA training) and 1/3 as reference/test speakers

Dialect Region	#Male	#Female	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
8	438 (70%)	192 (30%)	630 (100%)

- Each speaker had 8 usable utterances
- Each utterance short sentence

Databases
2. SRE 2004, 2005, 2006, 2010

- SRE 2004, 2005, 2006 used as background speakers (UBM, TVS and LDA training)
  - 1364 different speakers with over 16,000 wav files
- Reference/test speakers from SRE 2010
  - Over 20,000 wav files compared against each other in 9 different conditions
- wav files are long, typically consist of conversations with automatic speech recognition (ASR) files provided

Databases

SRE 2010 Conditions

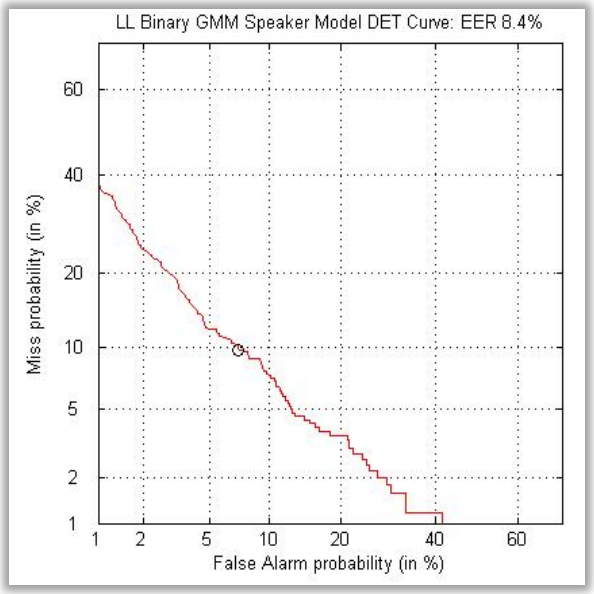
1. All Interview speech from the same microphone in training and test
2. All Interview speech from different microphones in training and test
3. All Interview training speech and normal vocal effort in telephone test speech
4. All Interview training speech and normal vocal effort telephone test speech recorded over a room microphone channel
5. All trials involving normal vocal effort conversational telephone speech in training and test
6. Telephone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test
7. All room microphone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test
8. All telephone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test
9. All room microphone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test

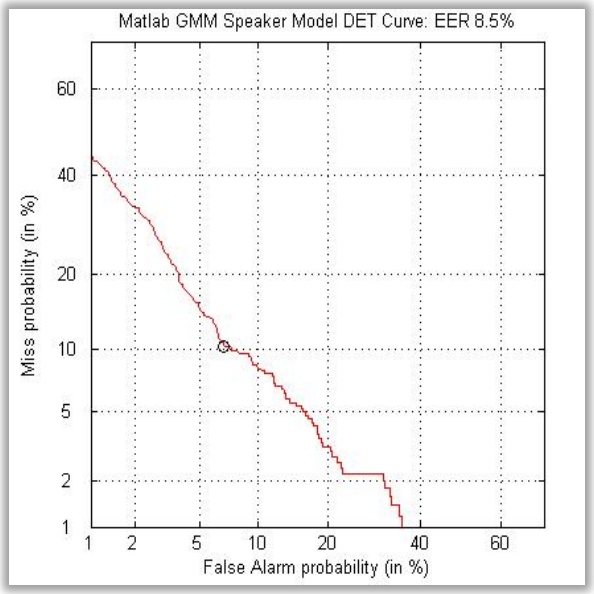
Results

TIMIT

12 normalized MFCCs generated by Lincoln Lab executable, 512 GMM components

- GMM Speaker Model Results





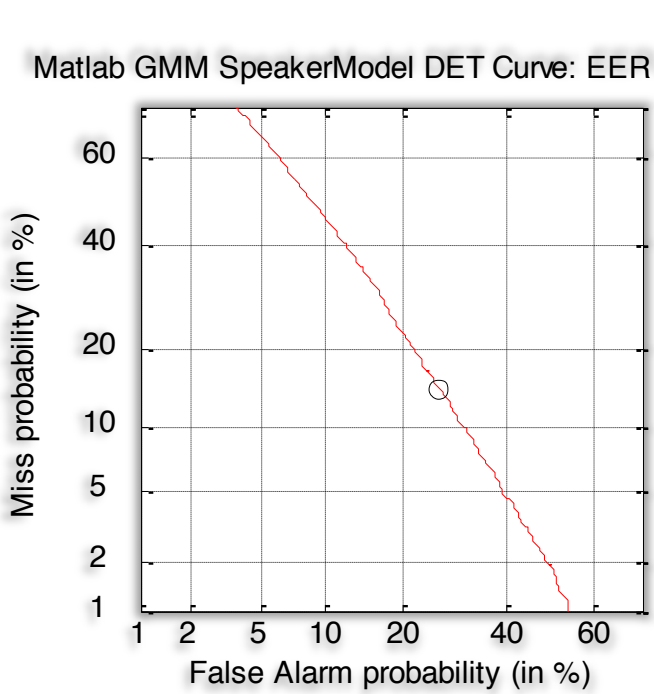
- i-vector and LDA reduced i-vector models resulted in unexpected behaviors
  - After analysis, root cause discovered to be insufficient amount of data in database for higher level methods

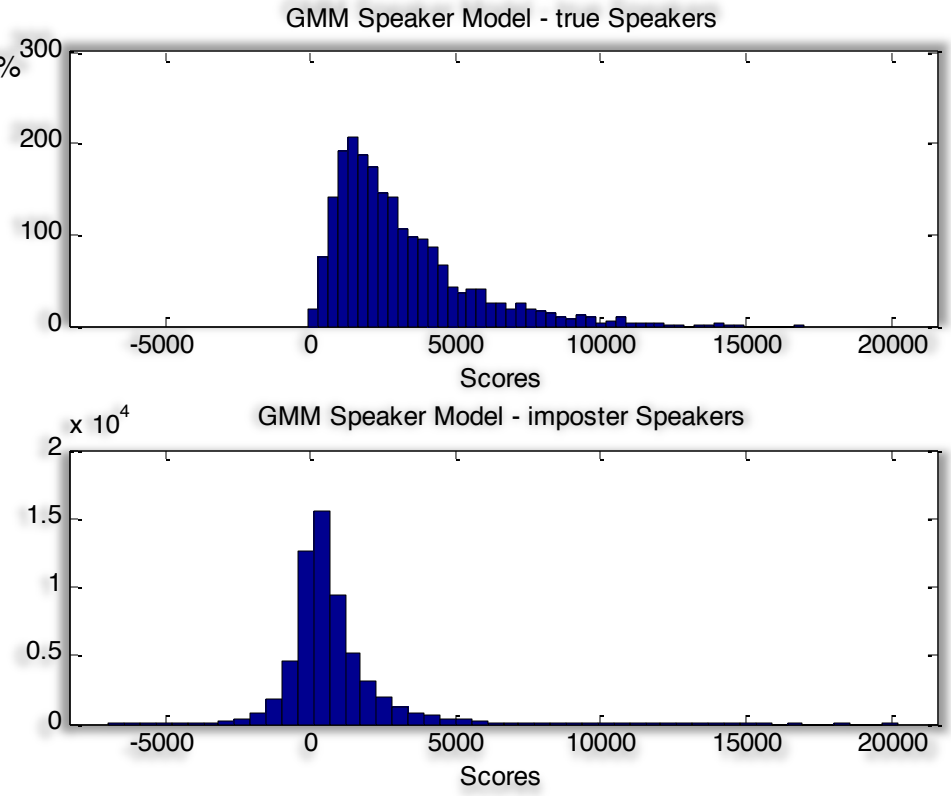
Results

SRE 2010 Condition 1

GMM scored using LLT

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components



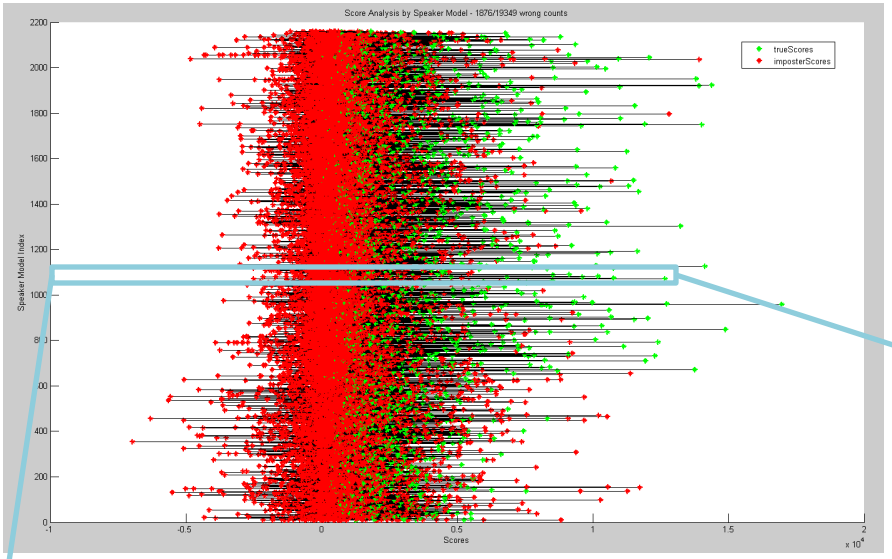


Results

SRE 2010 Condition 1

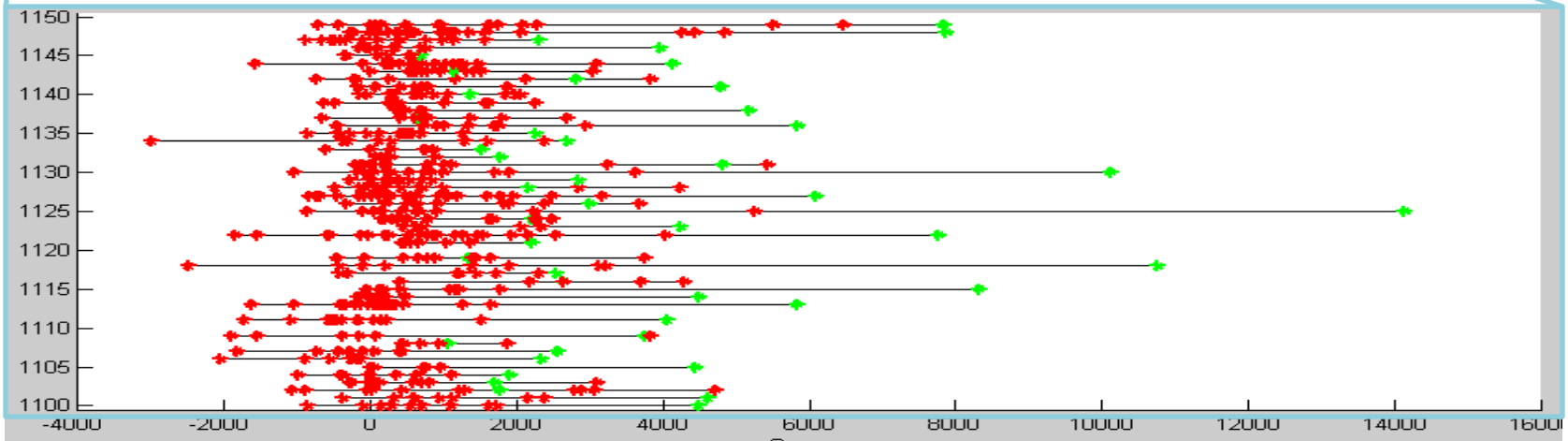
GMM scored using LLT

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components



4628/46923 (9.9%)

"wrong counts"

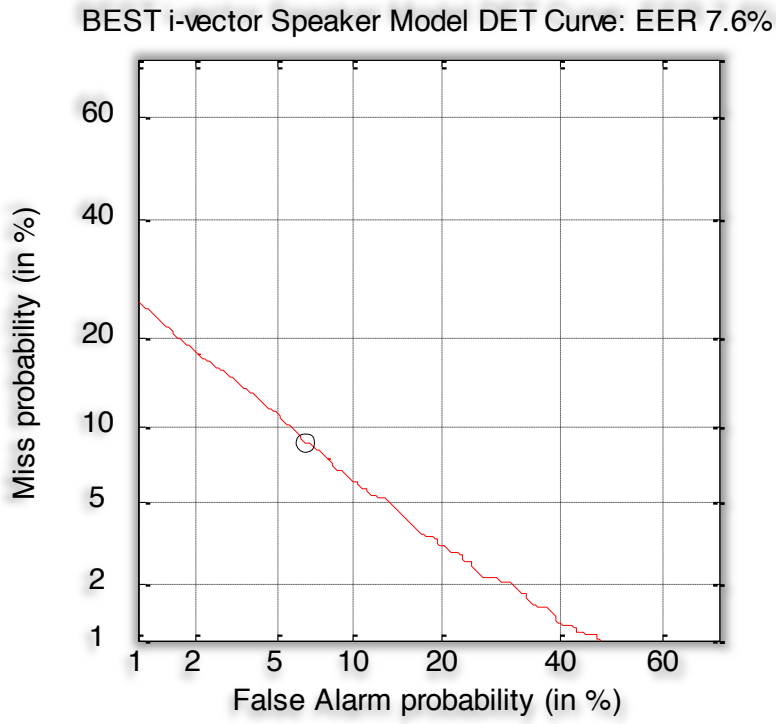


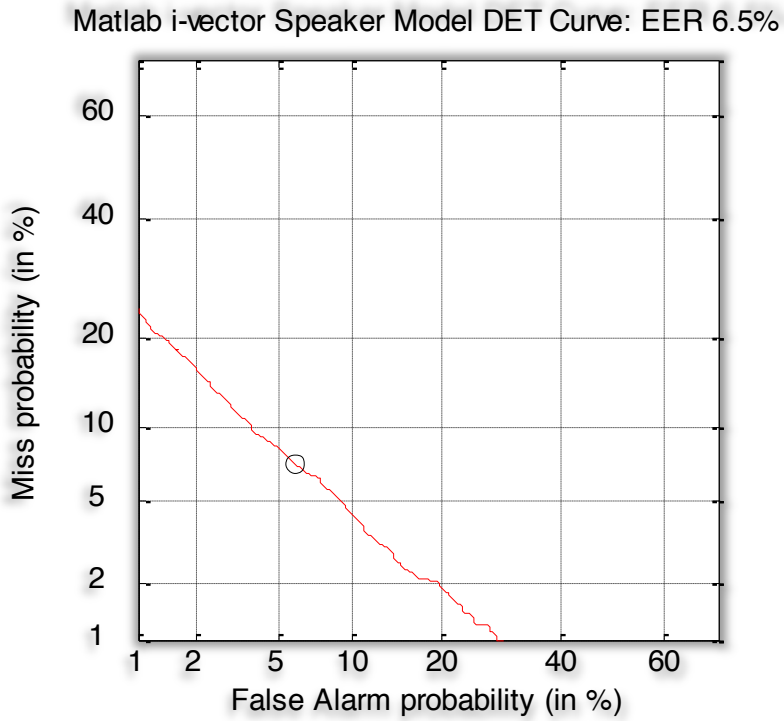
Results

SRE 2010 Condition 1

FA i-vectors scored using CDC

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components



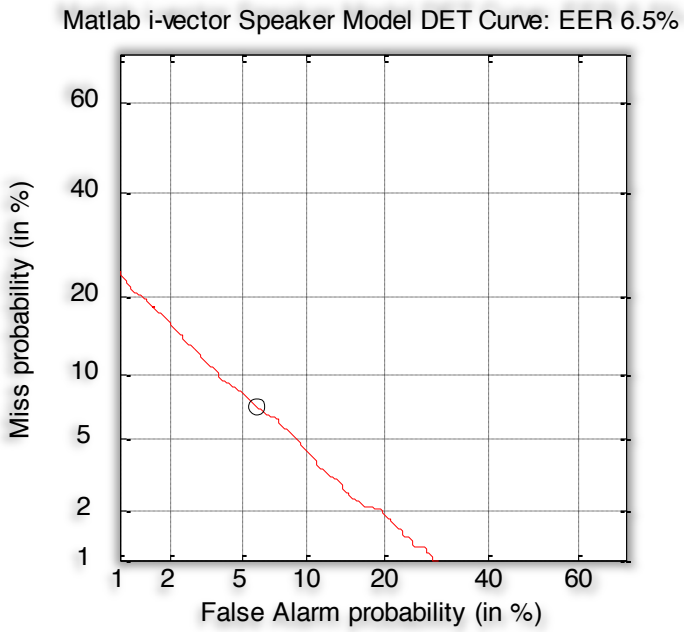


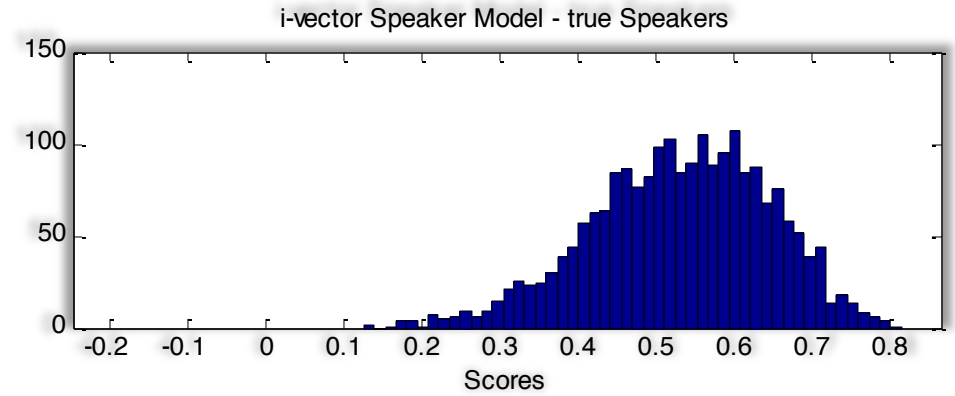
Results

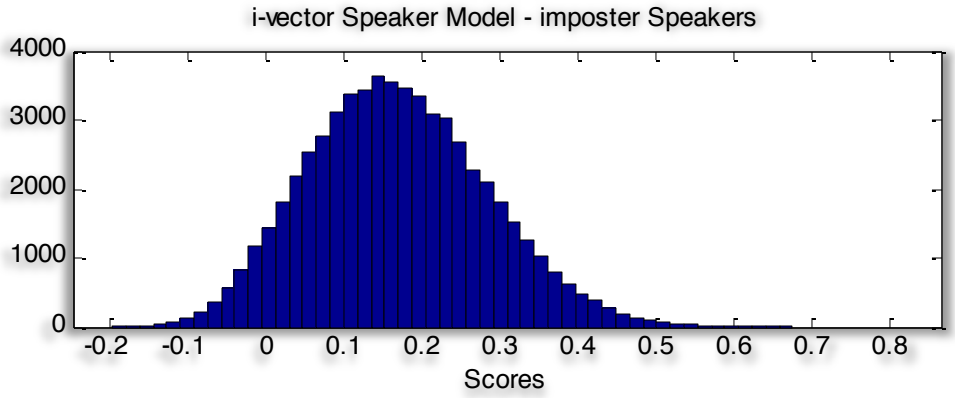
SRE 2010 Condition 1

FA i-vectors scored using CDC

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components





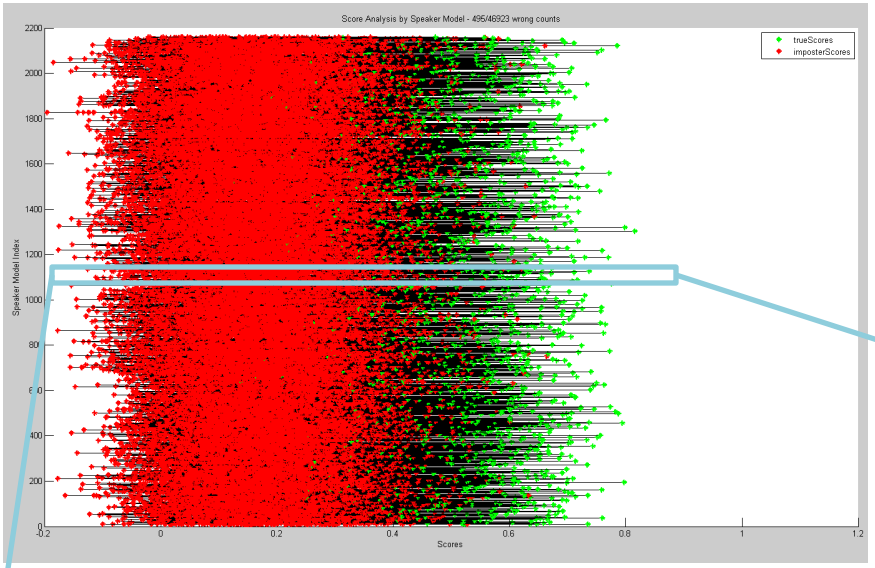


Results

SRE 2010 Condition 1

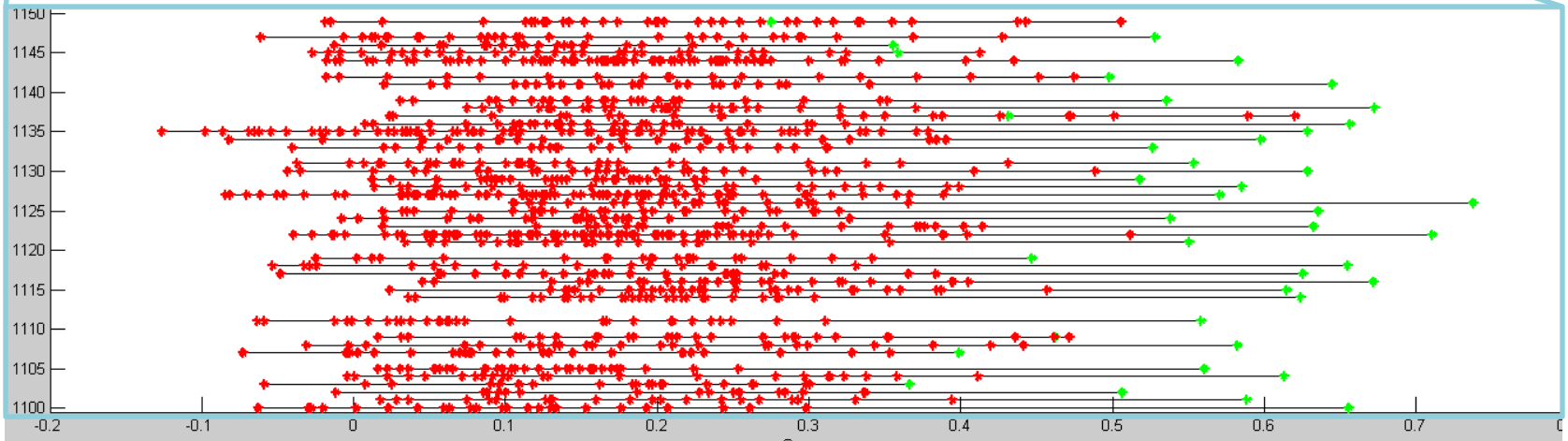
FA i-vectors scored using CDC

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components



495/46923 (1.1%)

"wrong counts"

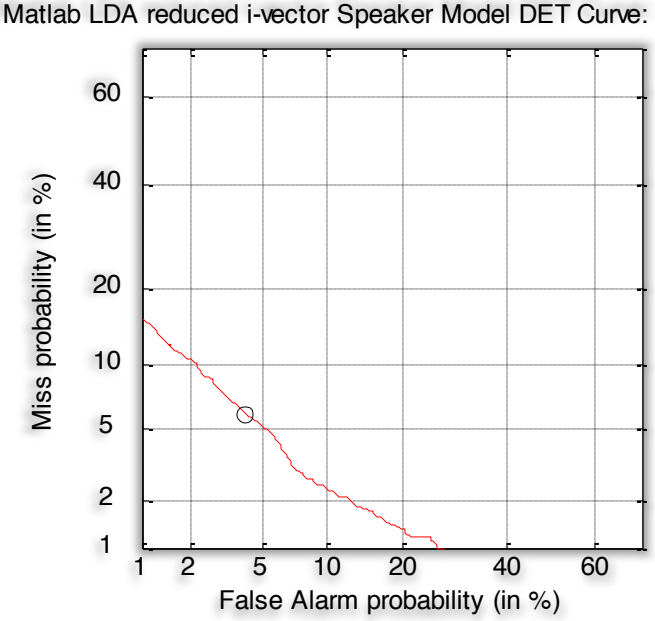


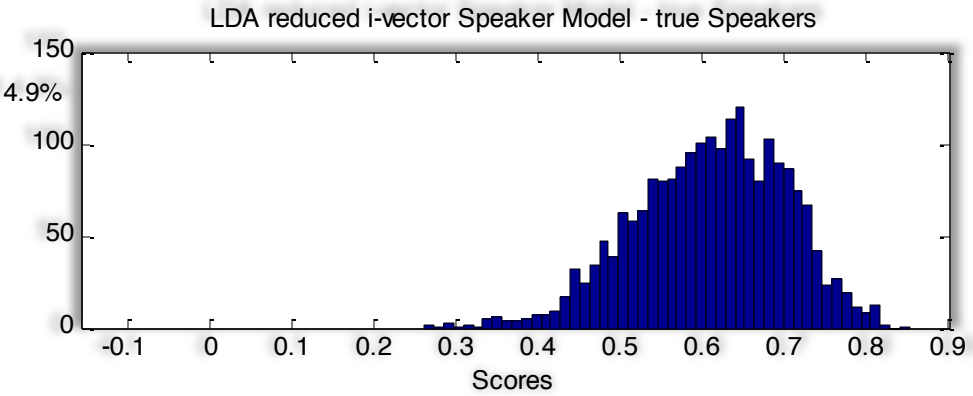
Results

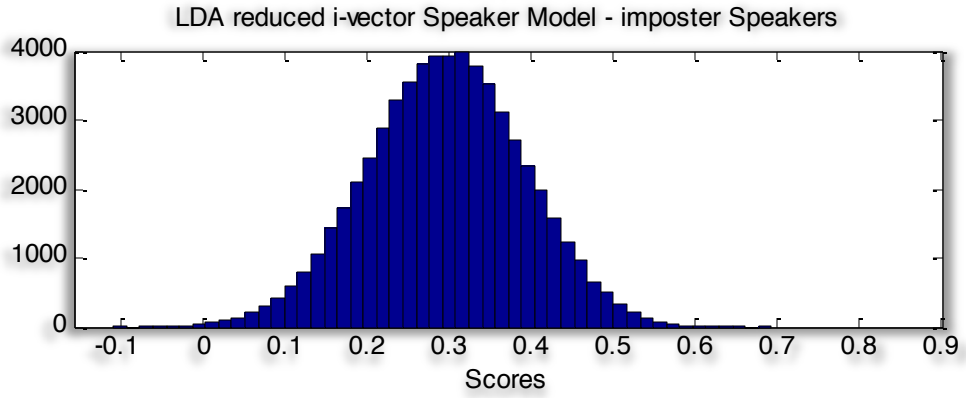
SRE 2010 Condition 1

LDA reduced i-vectors scored using CDC

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components





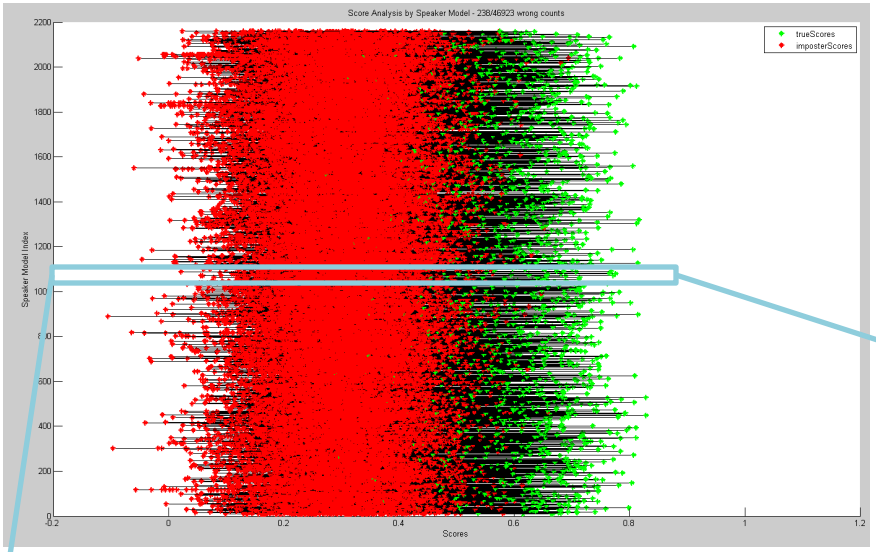


Results

SRE 2010 Condition 1

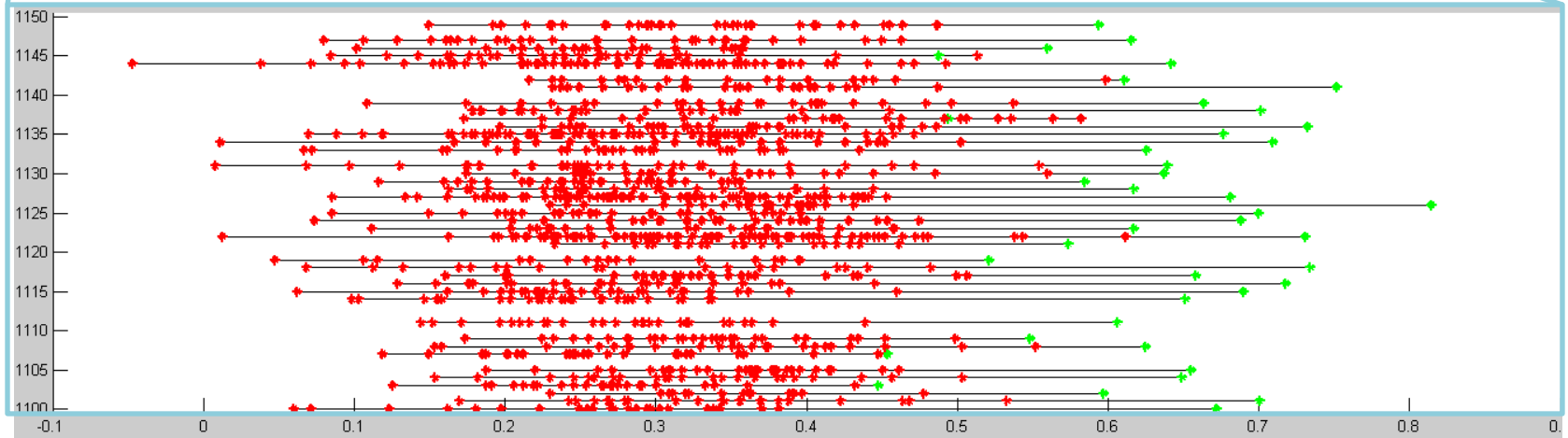
LDA reduced i-vectors scored using CDC

12 normalized MFCCs generated with Dan Ellis code, 512 GMM components



238/46923 (0.5%)

"wrong counts"



Results

SRE 2010 - Conditions 1-9

EERs of i-vectors and LDA reduced i-vectors using CDC

	Cond1 (46,923)	Cond2 (219,842)	Cond3 (58,043)	Cond4 (85,902)	Cond5 (30,373)	Cond6 (28,672)	Cond7 (28,356)	Cond8 (28,604)	Cond9 (27,520)
i-vectors (12 MFCCs, 512 GC)	6.5%	14.5%	10.9%	11.5%	12.2%	16.1%	16.6%	5.3%	8.2%
LDA reduced i-vectors (12 MFCCs, 512 GC)	4.9%	9.9%	10.1%	7.7%	10.7%	16.2%	14.4%	5.4%	4.9%
i-vectors (57 MFCCs, 1024 GC)	8.1%	17.2%	10.7%	14.4%	10.5%	16.7%	17.0%	3.6%	8.5%
LDA reduced i-vectors (57 MFCCs, 1024 GC)	4.6%	8.5%	8.5%	7.8%	8.8%	14.5%	15.1%	3.4%	3.5%
i-vectors BEST (12MFCCs, 512 GC)	7.6%	14.8%	14.5%	12.2%	13.8%	17.7%	18.0%	6.0%	7.8%

Summary

Schedule/Milestones

Fall 2011

October 4	✓	Have a good general understanding on the full project and have proposal completed. Present proposal in class by this date.
		✓ *Marks completion of Phase I*
November 4	✓	Validation of system based on supervectors generated by the EM and MAP algorithms
		✓ *Marks completion of Phase II*
December 19	✓	Validation of system based on extracted i-vectors
	✓	Validation of system based on nuisance-compensated i-vectors from LDA
	✓	Mid-Year Project Progress Report completed. Present in class by this date.
		✓ *Marks completion of Phase III*

Spring 2012

Feb. 25	✓	Testing algorithms from Phase II and Phase III will be completed and compared against results of vetted system. Will be familiar with vetted Speaker Recognition System by this time.
		✓ *Marks completion of Phase IV*
March 18	✓	Decision made on next step in project. Schedule updated and present status update in class by this date.
April 20	✓	Completion of all tasks for project.
		Marks completion of Phase V
May 10	➢	Final Report completed. Present in class by this date.
		Marks completion of Phase VI

Summary
Promised Results

- ✓ A fully validated and complete MATLAB implementation of a speaker recognition system will be delivered with at least two classification algorithms.
- ✓ Both a mid-year progress report and a final report will be delivered which will include validation and test results.

References

- [1] Kinnunen, Tomi, and Haizhou Li. "An Overview of Text-independent Speaker Recognition: From Features to Supervectors." *Speech Communication* 52.1 (2010): 12-40. Print.
- [2] Ellis, Daniel. "An introduction to signal processing for speech." *The Handbook of Phonetic Science,* ed. Hardcastle and Laver, 2nd ed., 2009.
- [3] Reynolds, D. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing* 10.1-3 (2000): 19-41. Print.
- [4] Reynolds, Douglas A., and Richard C. Rose. "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Transations on Speech and Audio Processing* IEEE 3.1 (1995): 72-83. Print.
- [5] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*. New York: Wiley, 2001. Print.
- [6] Dehak, Najim, and Dehak, Reda. "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification." *Interspeech 2009 Brighton.* 1559-1562.
- [7] Kenny, Patrick, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel. "A Study of Interspeaker Variability in Speaker Verification." *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008): 980-88. Print.
- [8] Lei, Howard. "Joint Factor Analysis (JFA) and i-vector Tutorial." *ICSI. Web. 02 Oct. 2011.* http://www.icsi.berkeley.edu/Speech/presentations/AFRL_ICSI_visit2_JFA_tutorial_icsitalk.pdf
- [9] Kenny, P., G. Boulianne, and P. Dumouchel. "Eigenvoice Modeling with Sparse Training Data." *IEEE Transactions on Speech and Audio Processing* 13.3 (2005): 345-54. Print.
- [10] Bishop, Christopher M. "4.1.6 Fisher's Discriminant for Multiple Classes." *Pattern Recognition and Machine Learning*. New York: Springer, 2006. Print.
- [11] Ellis, Daniel P.W. *PLP and RASTA (and MFCC, and Inversion) in Matlab. PLP and RASTA (and MFCC, and Inversion) in Matlab*. Vers. Ellis05-rastamat. 2005. Web. 1 Oct. 2011. <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>.
- [12] D. Garcia-Romero and C.Y. Espy-Wilson, "Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries," in *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, pp. 43-51.

Questions?