MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Exploiting Intra-Conversation Variability for Speaker Diarization

Stephen Shum*, Najim Dehak*, Ekapol Chuangsuwanich*, Douglas Reynolds^, Jim Glass*

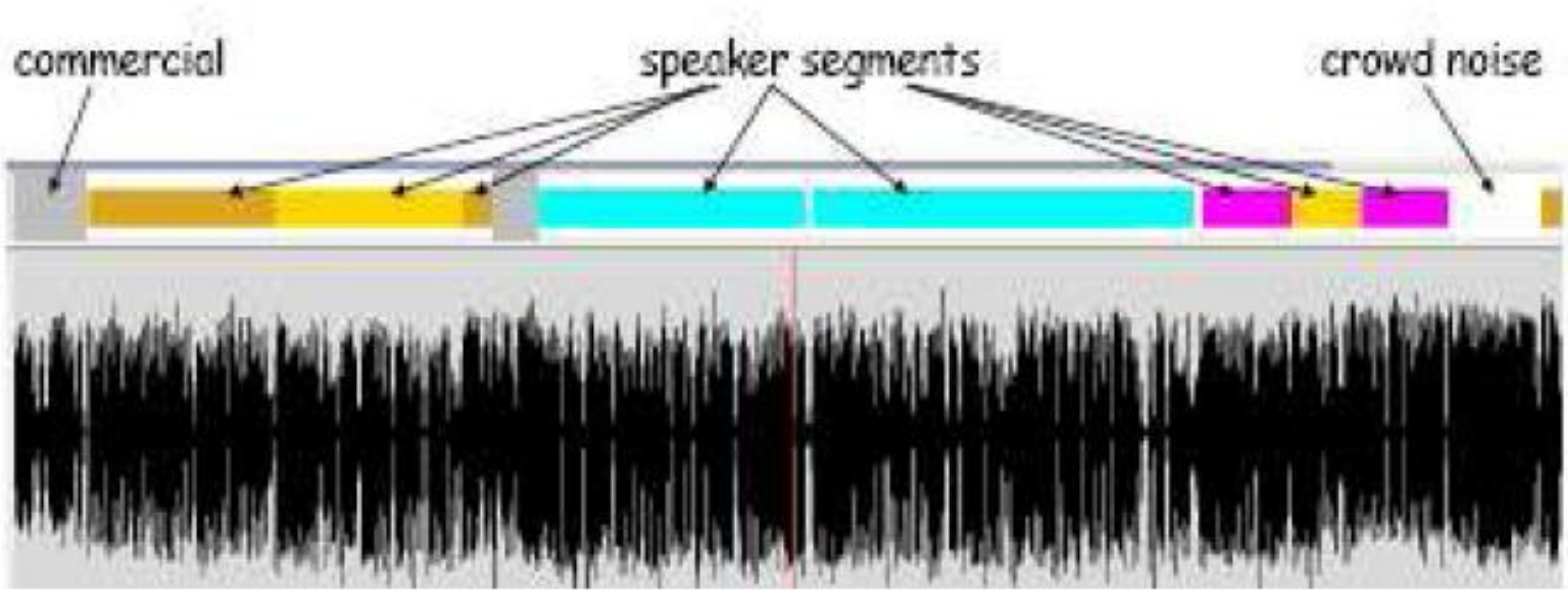
**MIT Computer Science and Artificial Intelligence Laboratory*

^MIT Lincoln Laboratory

August 31, 2011

Audio Diarization

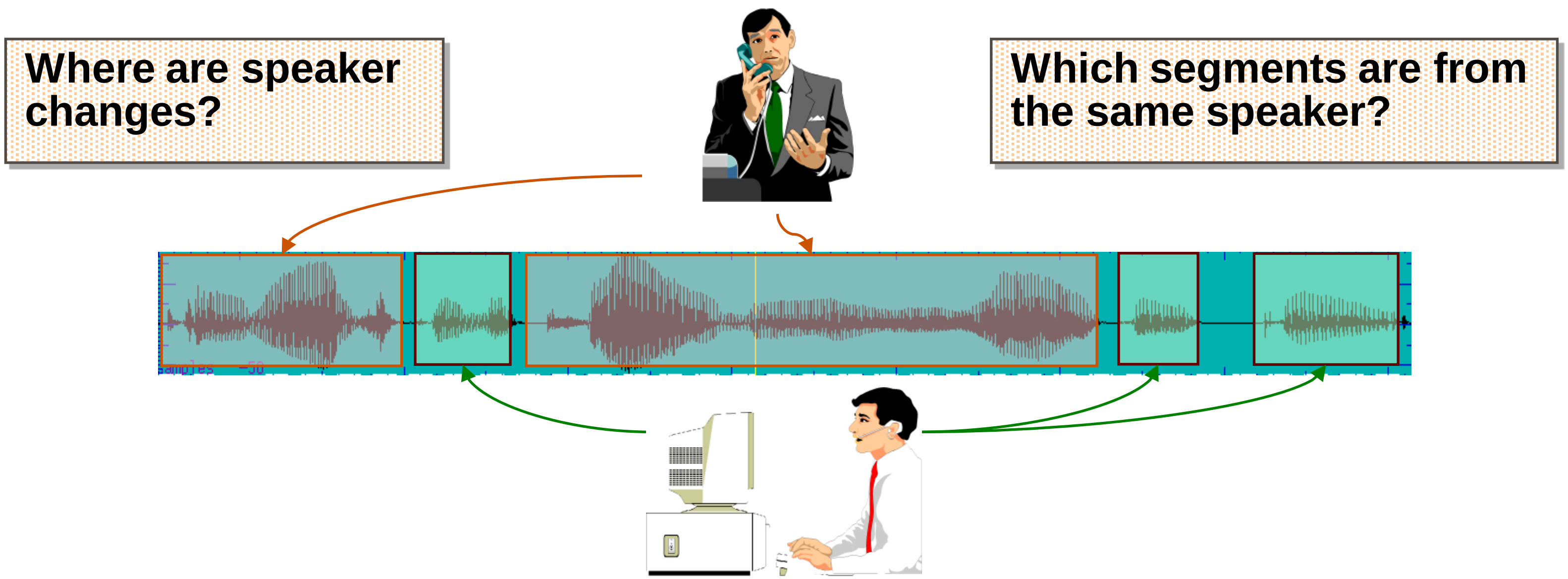
The task of marking and categorizing the different audio sources within an unmarked audio sequence



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



Towards Factor Analysis

- **At the heart of the speaker diarization problem is the problem of speaker modeling**
  - Factor analysis-based methods have recently achieved success in the speaker recognition community.
- **Previous work in FA-based diarization**
  - Stream-based, on-line system (Castaldo, 2008)
  - Variational Bayesian system (Kenny, 2010)
- **Difficulties**
  - Decisions made on very short (~1 second) speech segments
  - Poor speaker change detection can corrupt speaker models

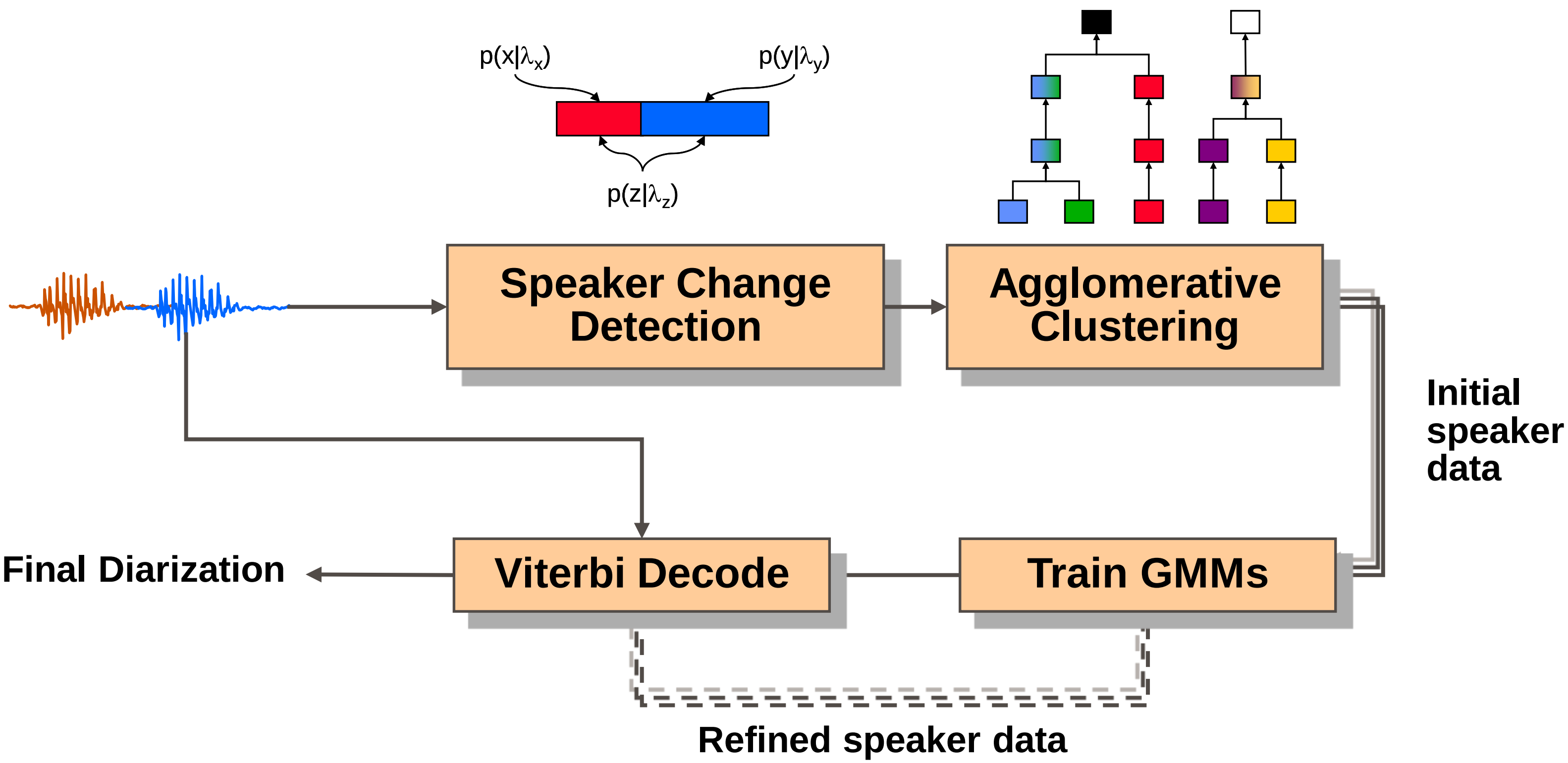
Roadmap

- **Introduction**
- **A BIC-based Baseline System**
- **A Total Variability-based Approach**
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- **System Evaluation**
- **Discussion and Conclusion**

Roadmap

- Introduction
- A BIC-based Baseline System
- A Total Variability-based Approach
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- System Evaluation
- Discussion and Conclusion

BIC-based Baseline System



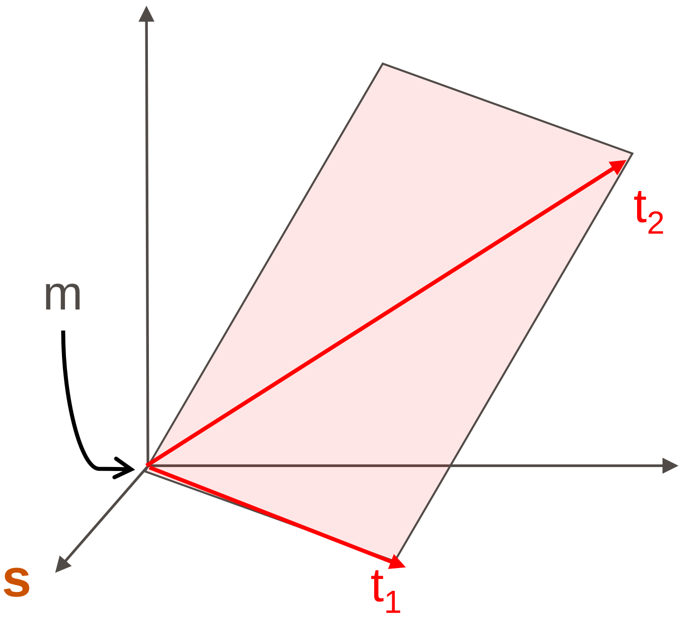
- **Bayesian Information Criterion (BIC)**
  - BIC-based speaker change detection
  - Agglomerative hierarchical clustering with BIC-based stopping criterion
  - Iterative re-segmentation with GMM-Viterbi decoding

A Review of Total Variability

- **Definition**
  - A ***supervector*** is created by concatenating all the mixture mean components in a GMM.

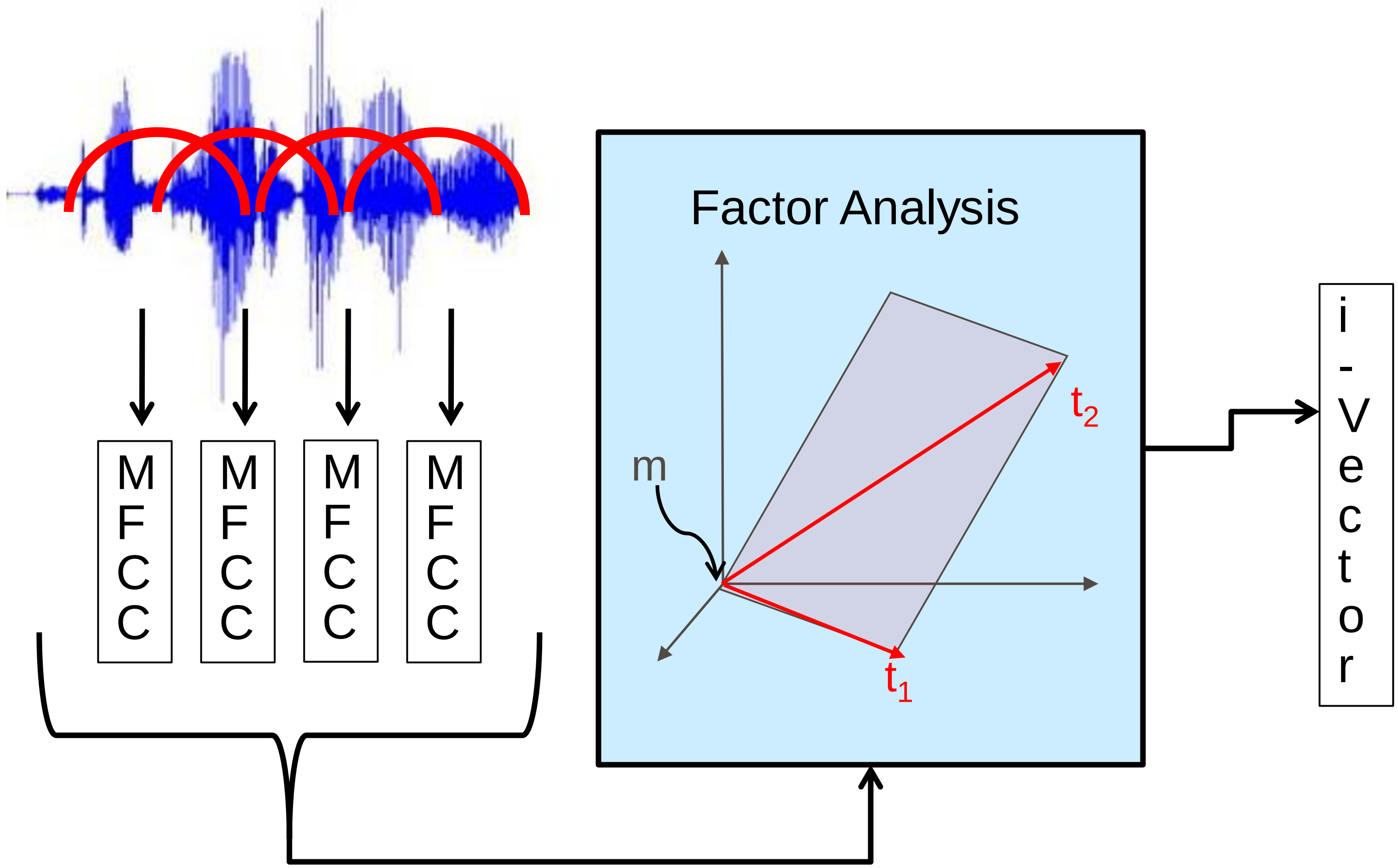
- **Assumption (Dehak, 2009)**
  - All pertinent variabilities lie in some low dimensional subspace T
    - **Call it the Total Variability Space**

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



- $\mathbf{w}$ **is the vector of Total Factors (Identity/Intermediate Vectors or i-vectors)**
- $\mathbf{m}$ **is supervector of un-adapted (UBM) means**
- $\mathbf{M}$ **is supervector of speaker- and channel- dependent means**

i-vector Extraction



Inter-session Compensation and Cosine Scoring

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

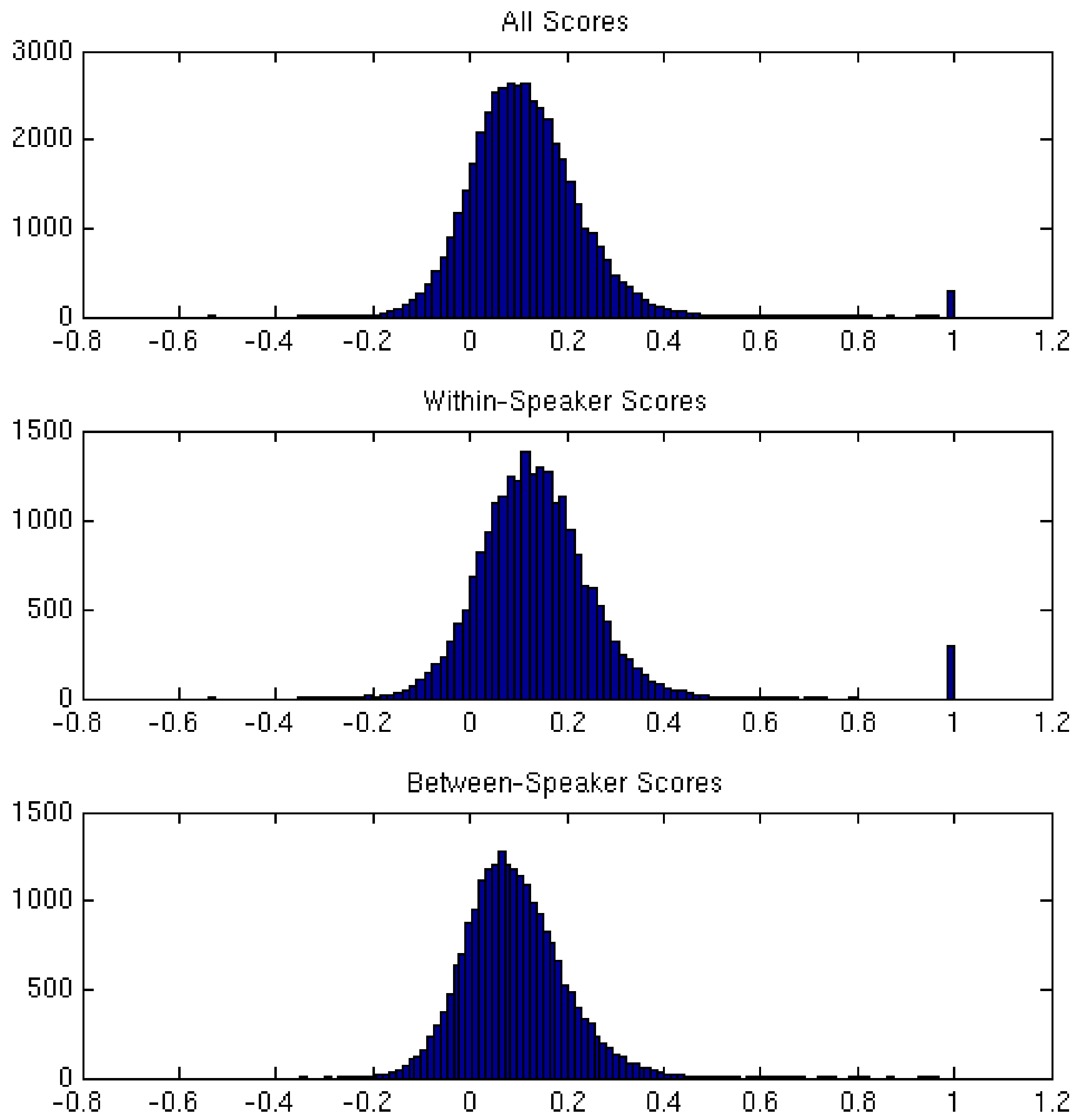
$$score(w_1, w_2) = \frac{(A^t w_1)^t W^{-1} (A^t w_2)}{\sqrt{(A^t w_1)^t W^{-1} (A^t w_1)} \cdot \sqrt{(A^t w_2)^t W^{-1} (A^t w_2)}}$$

A: Linear Discriminant Analysis (LDA) projection matrix

W: Within Class Covariance Normalization (WCCN) matrix

Inter-session Compensation

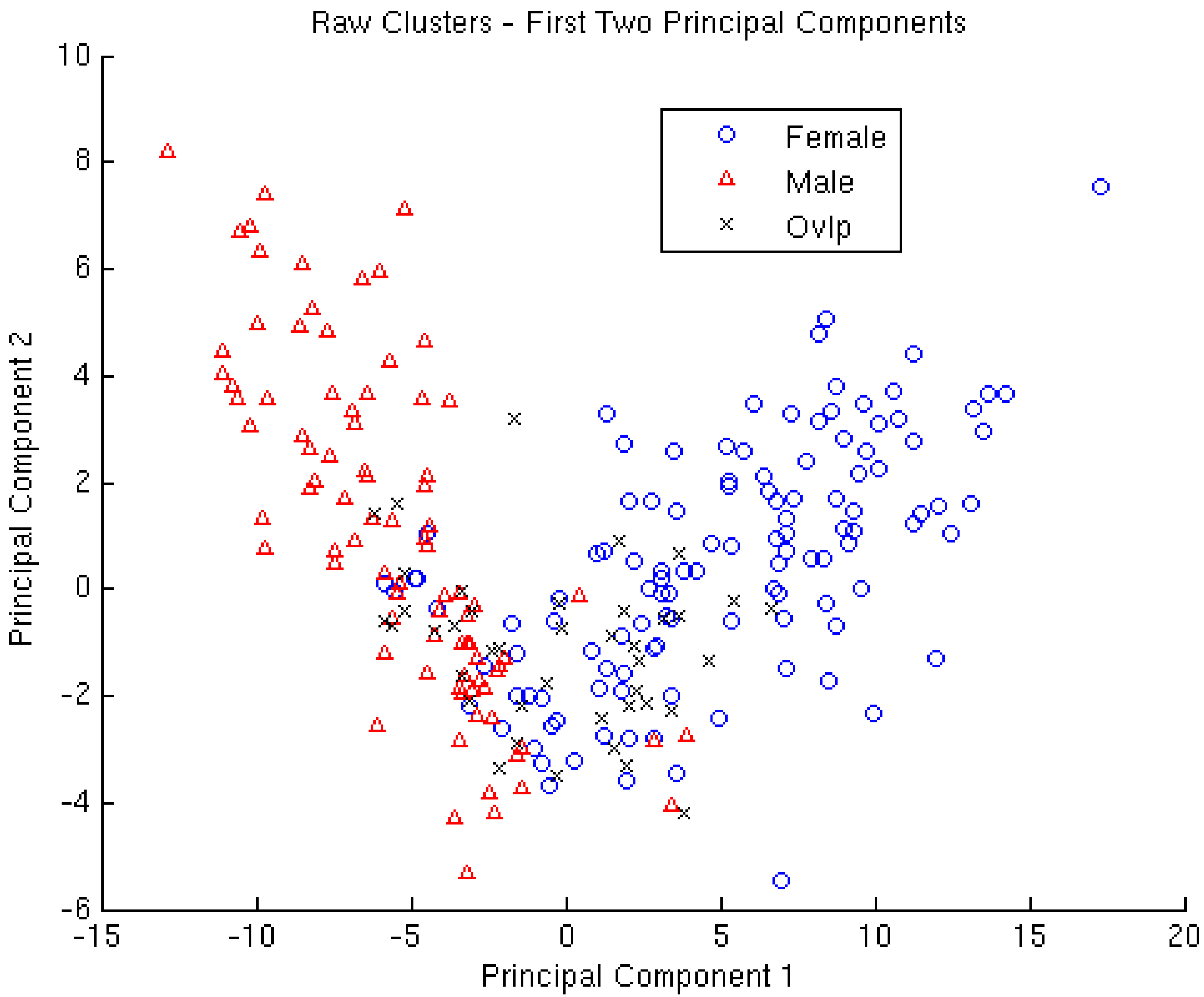
However, we ran into some issues…



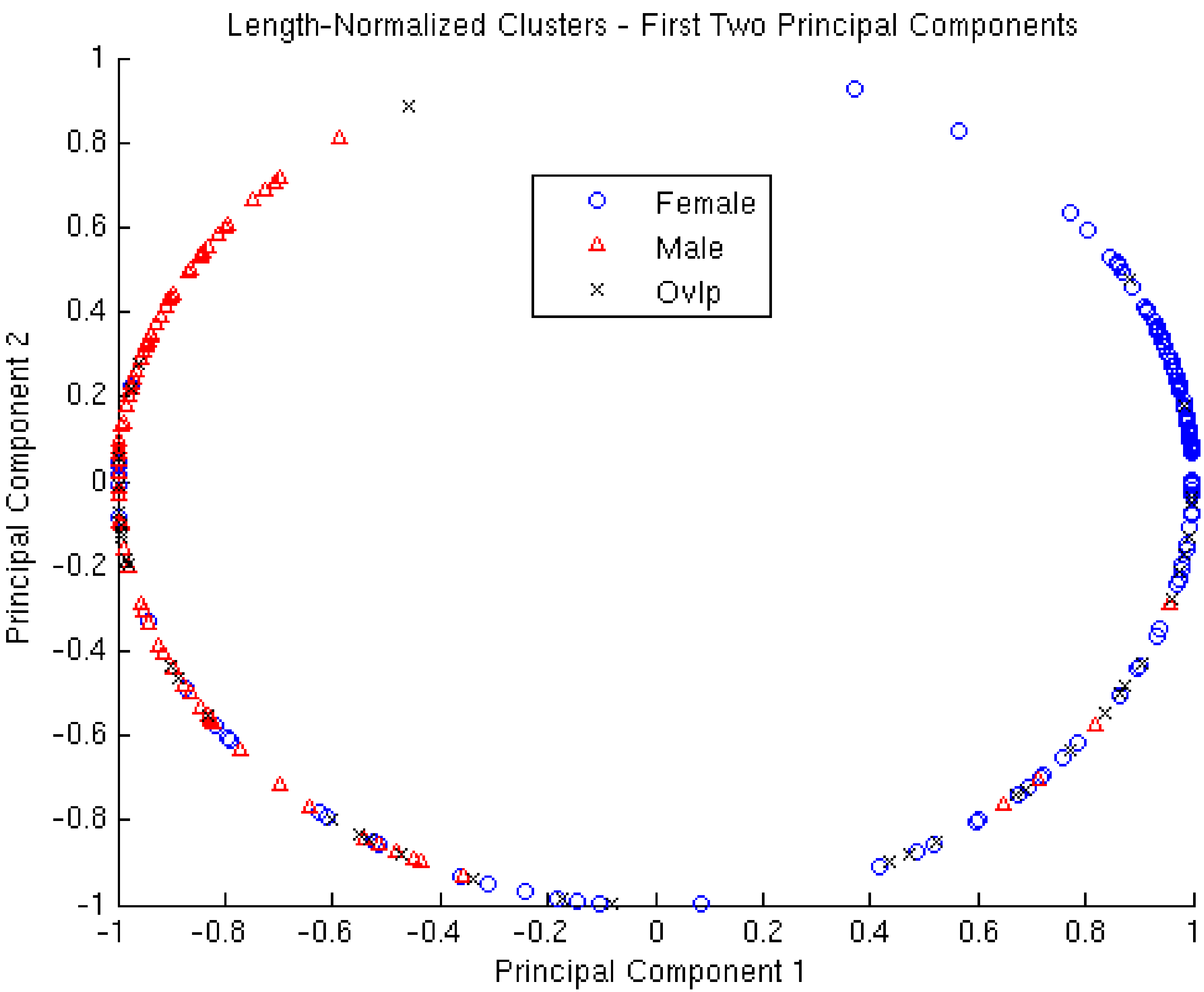
~~Inter-session Compensation~~
Intra-session Exploitation

- **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 <u>different</u> speakers.**

i-vector Visualization



i-vector Visualization



Intra-session Exploitation

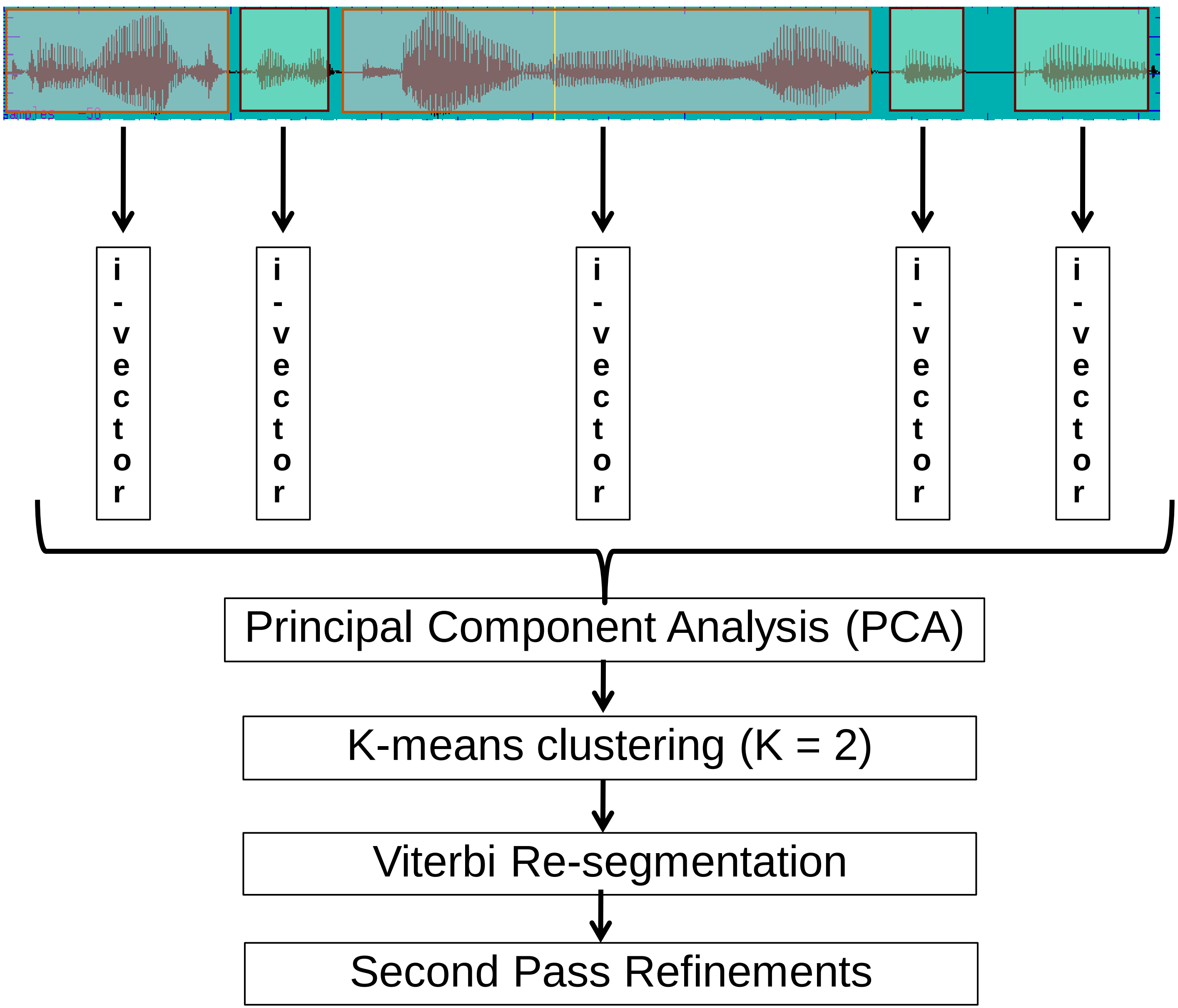
- **Could further emphasize the importance of principal directions with the most variability**
  - i.e. the *most* principal components have the largest eigenvalues

$$score(w_1', w_2') = \frac{(w_1')^t \Lambda (w_2')}{\left\| \Lambda^{1/2} w_1' \right\| \cdot \left\| \Lambda^{1/2} w_2' \right\|}$$

w_i' : PCA-projected i-vector

Λ : Corresponding diagonal matrix of eigenvalues

System Diagram



Viterbi Re-segmentation

– Operate at the acoustic feature level

– Initialize a 32-mixture GMM for each cluster
  * **Speaker A, Speaker B, Non-speech N**

– Obtain a posterior probability for each cluster given each feature vector
  * $P(A|x_t)$, $P(B|x_t)$, $P(N|x_t)$

– Pool these probabilities across the entire conversation ($t = 1, \ldots, T$) and use them to re-estimate each respective speaker's GMM
  * **The Non-speech GMM is never re-trained.**

– The Viterbi algorithm re-assigns each frame to the speaker/non-speech model with highest posterior probability.

Second Pass Refinements

– Extract a single i-vector for each respective speaker
  * **Using the newly defined re-segmentation assignments**

– Re-assign each newly-extracted segment i-vector w_i to the speaker i-vector $\{w_A, w_B\}$ that is closer in cosine similarity

– Iterate until convergence
  * **i.e. when segment-speaker assignments no longer change**

– Similar to Re-segmentation algorithm
  * **But makes hard decisions at the i-vector level instead of soft (posterior-based) decisions at the cepstral level**

– Also similar to K-means
  * **Except we determine the "means" $\{w_A, w_B\}$ via i-vector extraction**

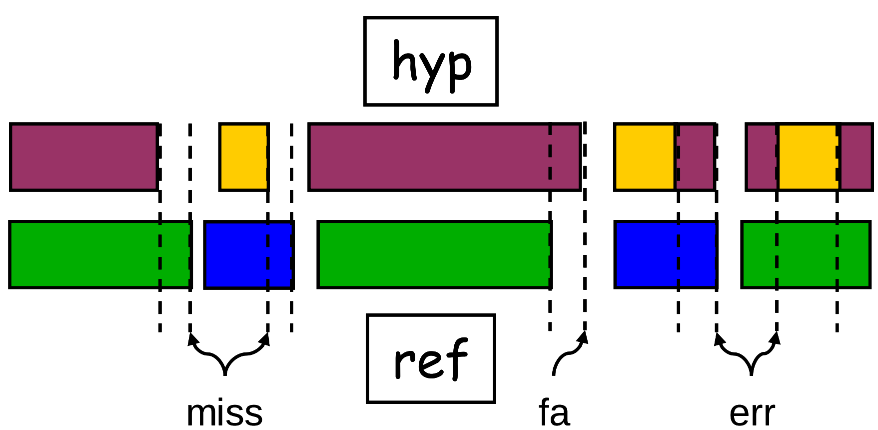
Roadmap

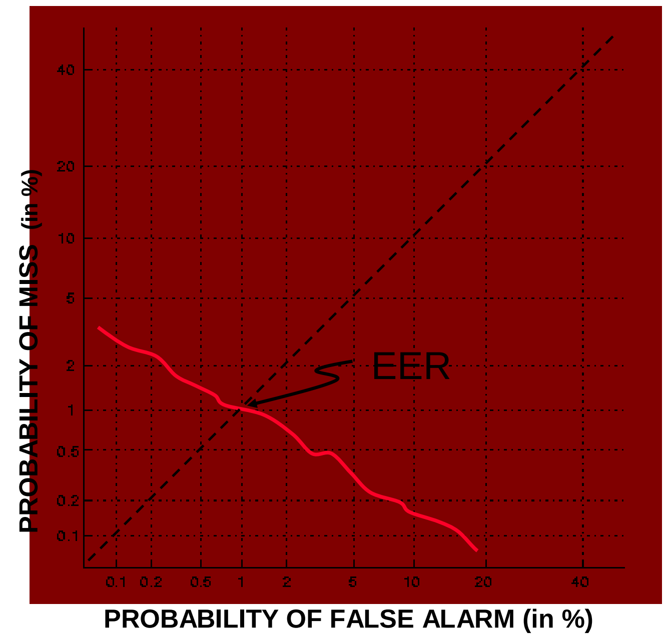
- Introduction
- A BIC-based Baseline System
- A Total Variability-based Approach
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- **System Evaluation**
- Discussion and Conclusion

Measuring Diarization Error

- **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)

- **Note**
  - Scoring protocol ignores overlapped speech segments





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
    - **Thanks to Brno University of Technology for providing these reference transcripts.**

Experiment Results

- **Initial Approach – TV400**

	Error Breakdown				
	Miss	False Alarm	Confusion	**DER (%)**	**σ (%)**
First Pass	**7.7**	2.0	4.0	**13.8**	**9.6**
Re-segmentation	**0.3**	2.3	**2.9**	**5.2**	**8.6**
Second Pass	0.3	2.3	**1.5**	**4.2**	**7.0**

Experiment Results

- **Initial Approach – TV400**

	Error Breakdown				
	Miss	False Alarm	Confusion	**DER (%)**	**σ (%)**
First Pass	**7.7**	2.0	4.0	**13.8**	**9.6**
Re-segmentation	**0.3**	2.3	**2.9**	**5.2**	**8.6**
Second Pass	0.3	2.3	**1.5**	**4.2**	**7.0**

- **After Parameter Optimization – TV100**

	Error Breakdown				
	Miss	False Alarm	Confusion	**DER (%)**	**σ (%)**
First Pass	7.7	2.0	2.8	**12.5**	**8.2**
Re-segmentation	0.3	2.3	2.6	**5.2**	**8.2**
Second Pass	0.3	2.3	**1.1**	**3.7**	**6.4**

Experiment Results

- **Using Non-reference Segmentation (TV100)**

	Error Breakdown				
	Miss	False Alarm	Confusion	**DER (%)**	**σ (%)**
First Pass	7.7	2.0	2.8	**12.5**	**8.2**
Re-segmentation	0.3	2.3	2.6	**5.2**	**8.2**
Second Pass	0.3	2.3	**1.1**	**3.7**	**6.4**

- **Using Reference Segmentation**

	Speaker Confusion (%)	**σ_C (%)**
BIC-based Baseline	3.5	8.0
VB-based FA	1.0	3.5
Ref VAD + TV100	**0.9**	**3.2**
Own VAD + TV100	1.1	3.3

Roadmap

- **Introduction**
- **A BIC-based Baseline System**
- **A Total Variability-based Approach**
  - Factor Analysis Re-visited
  - Exploiting Intra-Conversation Variability
- **System Evaluation**
- **Discussion and Conclusion**

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
  – Potential to corrupt our PCA
    * **Can mislead our system into finding fruitless directions of variabilities that we do not mean to address**
  – Not too much previous work on this… (Boakye, 2008 & 2011)

- **"Bag of i-vectors" approach is limiting**
  – Would be nice to incorporate temporal dynamics (i.e. HMMs)
  – Can draw from plenty of previous work

Conclusions

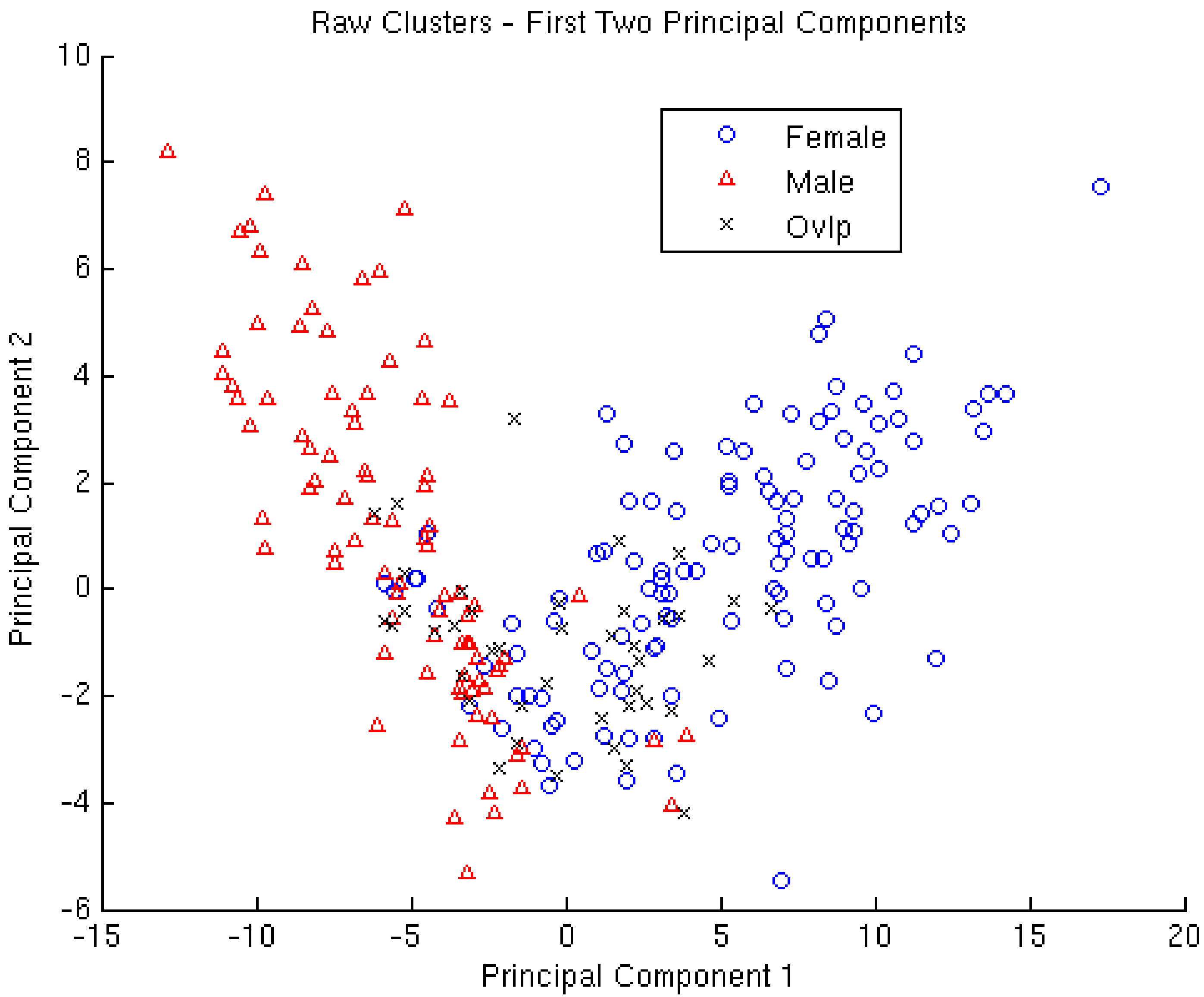
- **Factor analysis-based approach to speaker diarization**
  - Inspired by Total Variability and i-vectors
  - Key Insight
    * **Exploiting Intra-Conversation Variability**
  - Attained state of the art results on a test set of 2-speaker conversations

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

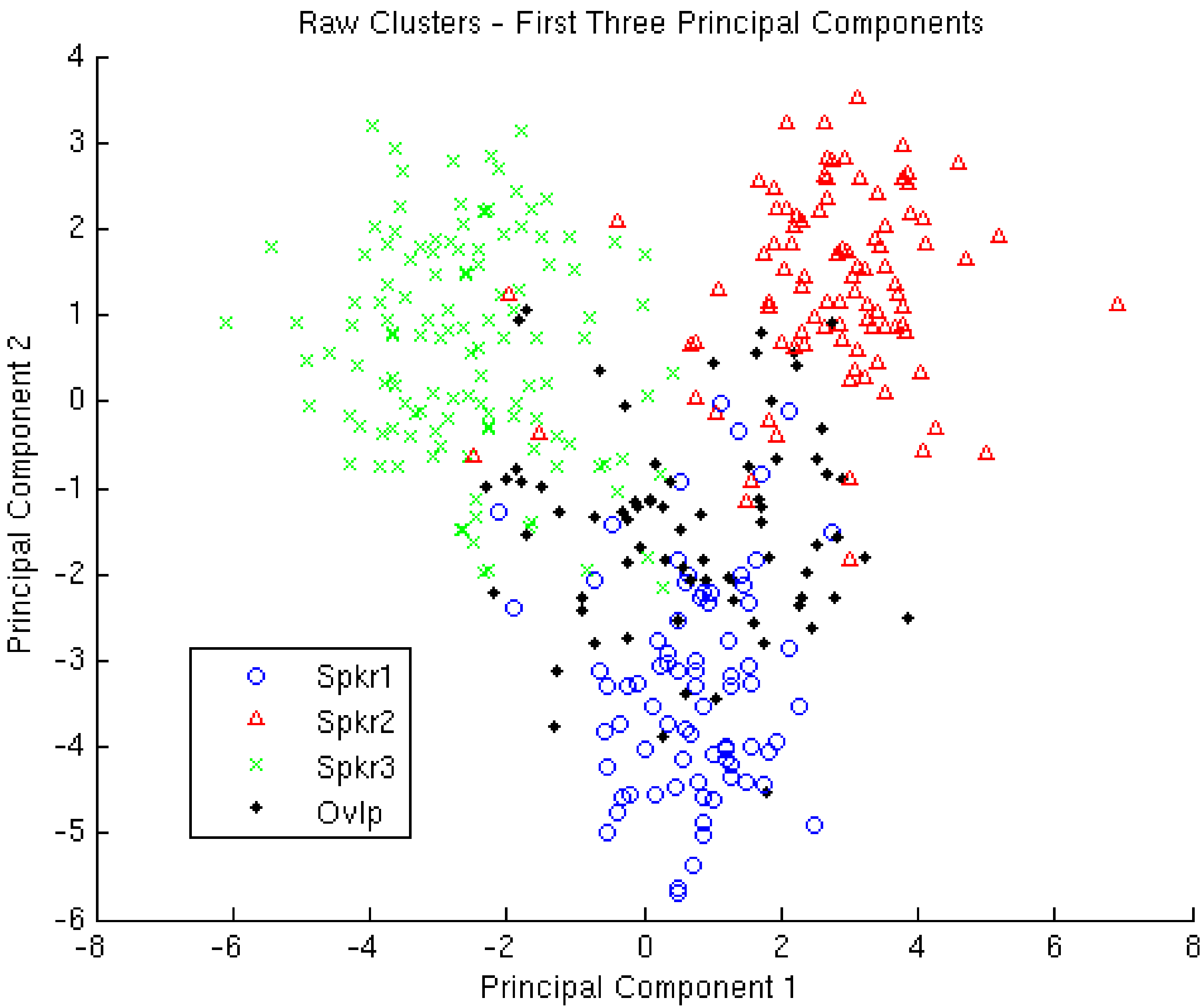
Questions?

Bonus Slides

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, these 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** $P(x, y \mid w) \approx q(x) \cdot q(y)$

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

Other Issues

- **Cosine similarity → data lie on the unit hypersphere**
  - Poorly modeled by a GMM
- **Data sparsity**
  - A speaker may speak very infrequently
  - All i-vectors are weighted equally, but some are more equal than others
    * **Need some way of incorporating information about the duration of speech used to extract a given i-vector**

Conclusions

- **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**
  - Incorporating temporal dynamics
  - Addressing problems of data sparsity

Questions?