CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach

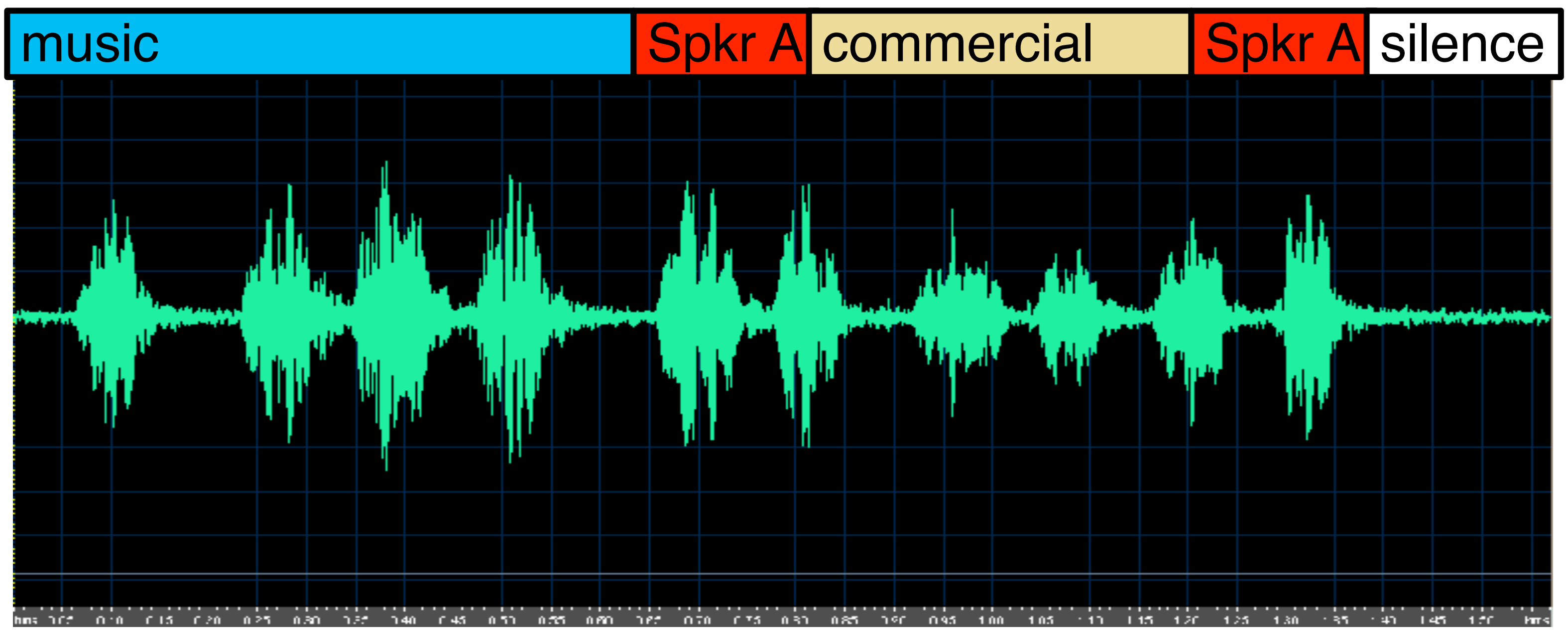
Stephen Shum, Najim Dehak, and Jim Glass

**With help from Reda Dehak, Ekapol Chuangsuwanich, and Douglas Reynolds*

November 29, 2012

Audio Diarization

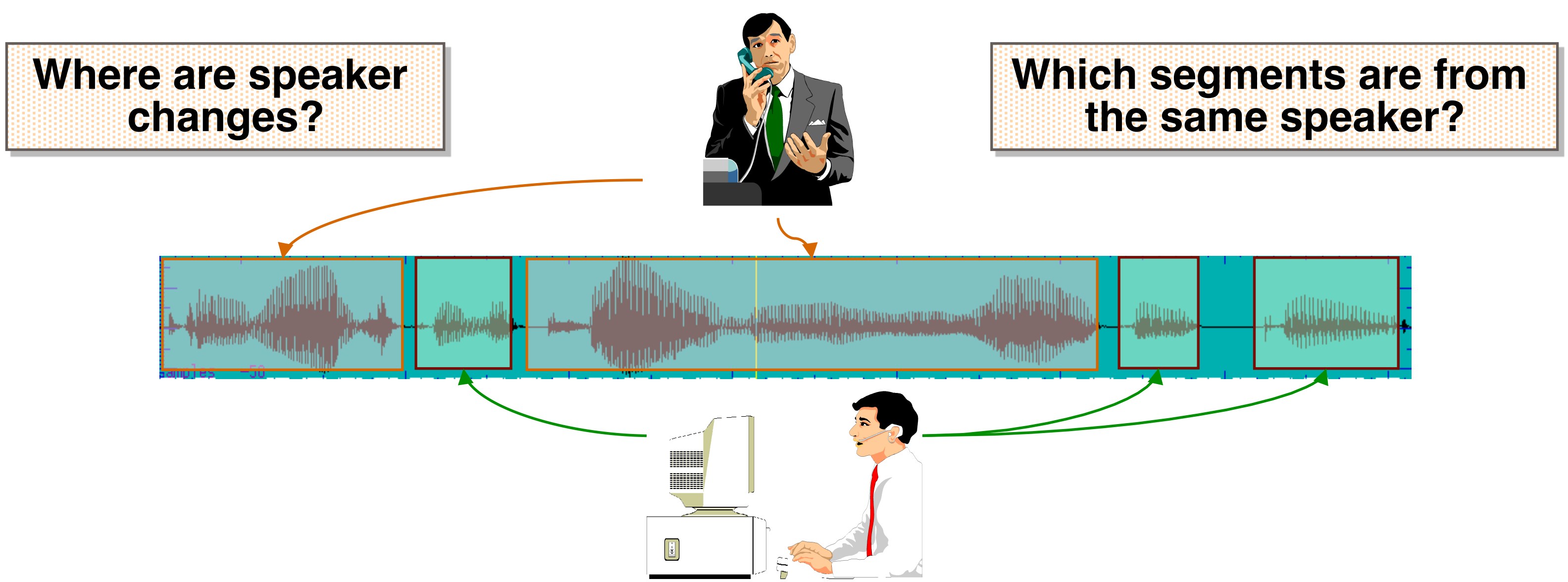
The task of marking and categorizing the different audio sources within an unmarked audio sequence.



Speaker Diarization

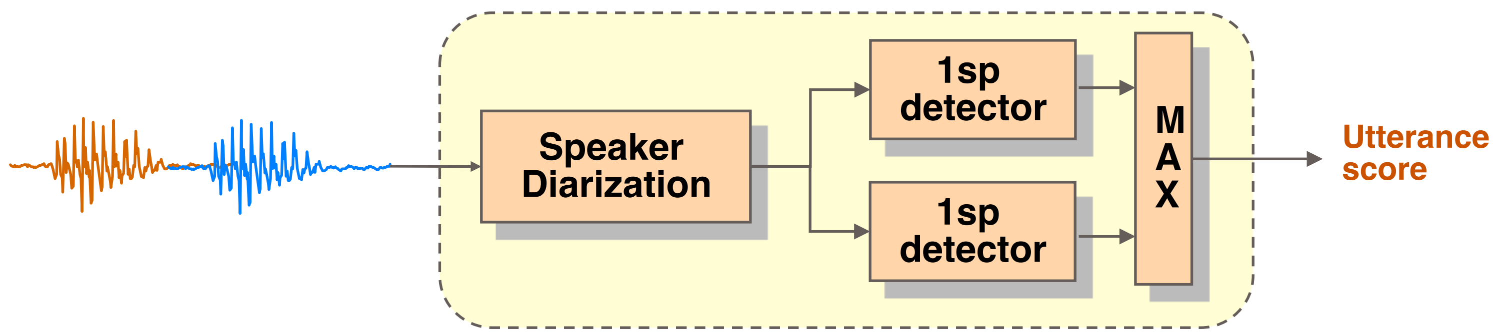
- "Who is speaking when?"

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



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 (i.e., speaker tracking)



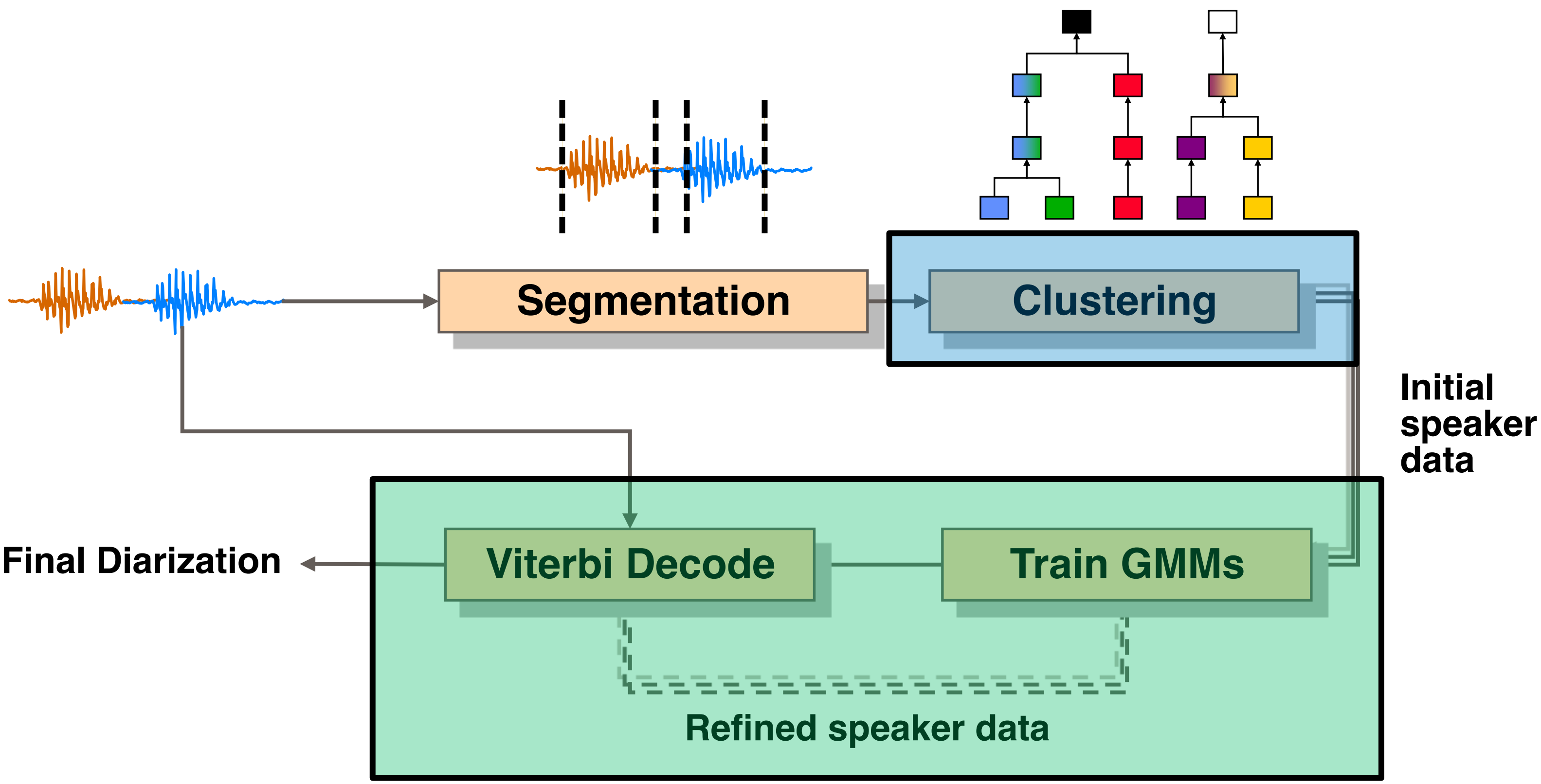
Take-Home Summary

- **Extended previous work in applying factor analysis-based speaker modeling to speaker diarization**
  - Castaldo 2008, Kenny 2010, Interspeech 2011-2012
- **Integrated variational inference into speaker clustering**
  - Valente 2005, Kenny 2010, SM Thesis 2011
- **Validated an iterative optimization procedure to refine clustering and segmentation hypotheses**
  - Interspeech 2012
- **Proposed a duration-proportional sampling scheme to combat issues of i-vector underrepresentation**
  - SM Thesis 2011

Roadmap

- **Introduction**
  - Summary of Contributions
- **Background**
  - Diarization System Overview
  - Speaker Modeling with Factor Analysis
- **Our Incremental Approach**
  - K-means and Spectral Clustering (Interspeech 2011, 2012)
  - Towards Probabilistic Clustering Methods
  - Iterative System Optimization (Re-segmentation/Clustering)
  - Duration-Proportional Sampling
- **Analysis and Discussion**
  - Benchmark Comparison (Castaldo 2008)
- **Conclusion**

Standard Diarization Setup



- **Agglomerative Hierarchical Clustering**
  - Requires methods for model selection
- **Iterative re-segmentation**

Towards Factor Analysis

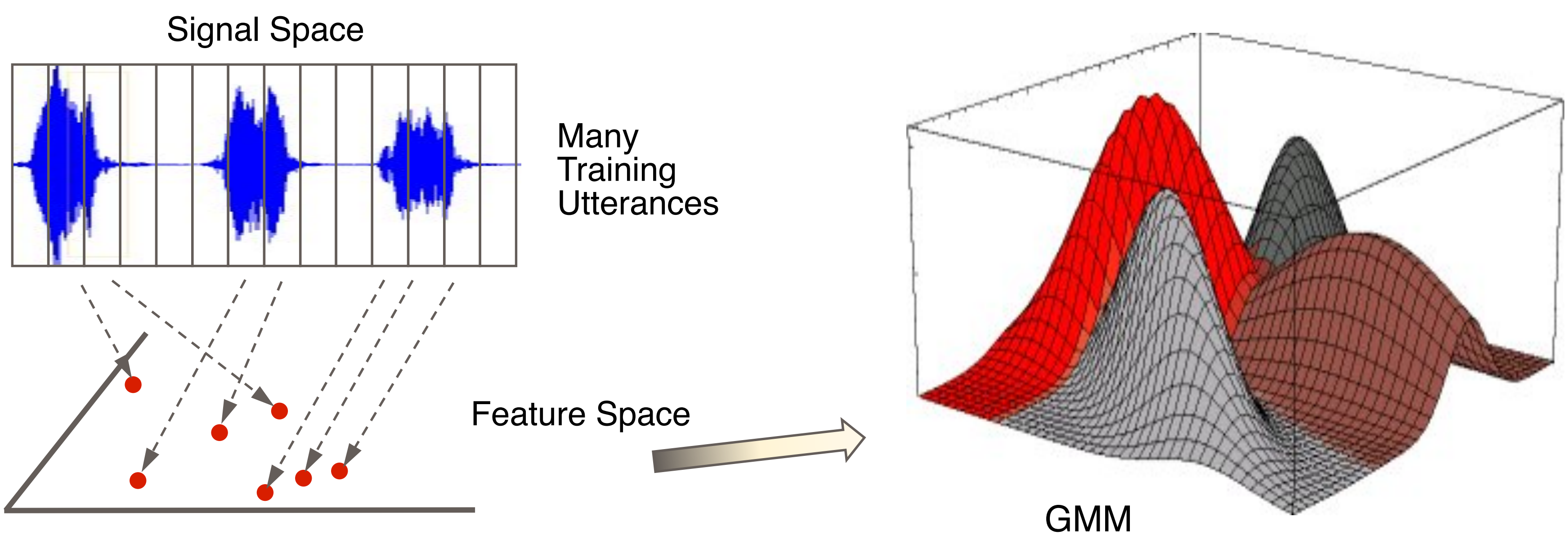
- **At the heart of the speaker diarization problem is the problem of speaker modeling**
  - Factor analysis-based methods have achieved success in the speaker recognition community.
- **Main Idea**
  - Low-dimensional summary of a speaker's distribution of acoustic feature vectors

Modeling Feature Sequences with GMMs

- **We need to model the distribution of feature vector sequences**
  - e.g., Mel Frequency Cepstral Coefficients (MFCCs)

100 vectors/sec

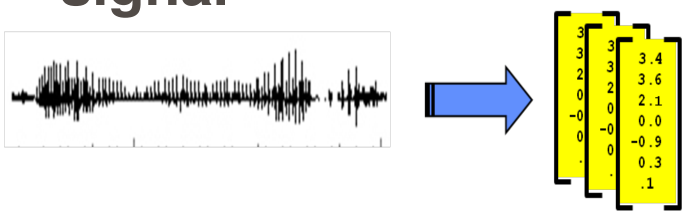
- **Gaussian mixture models (GMMs) are a common representation**

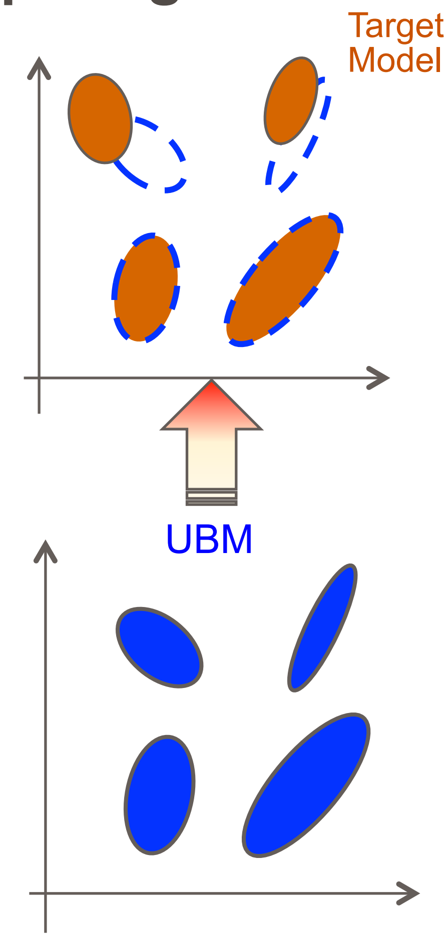


Modeling with Adapted GMM-UBMs

(3) Adapt target model from UBM

(1) Extract feature vector sequence from speech signal





(2) Train UBM with speech from many speakers

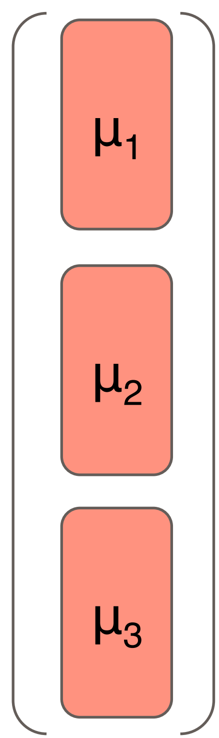
GMM-UBM and MAP Adaptation

- **Target model is trained by adapting from background model**
  - Couples models together and helps with limited target training data
- **Adaptation only updates mean parameters representing acoustic events seen in target training data**
  - Sparse regions of feature space filled in by UBM mean parameters
    - * **Both an advantage and a disadvantage**
- **Disadvantage**
  - Limited target training data can still prevent some UBM components from being adapted.

Intuition

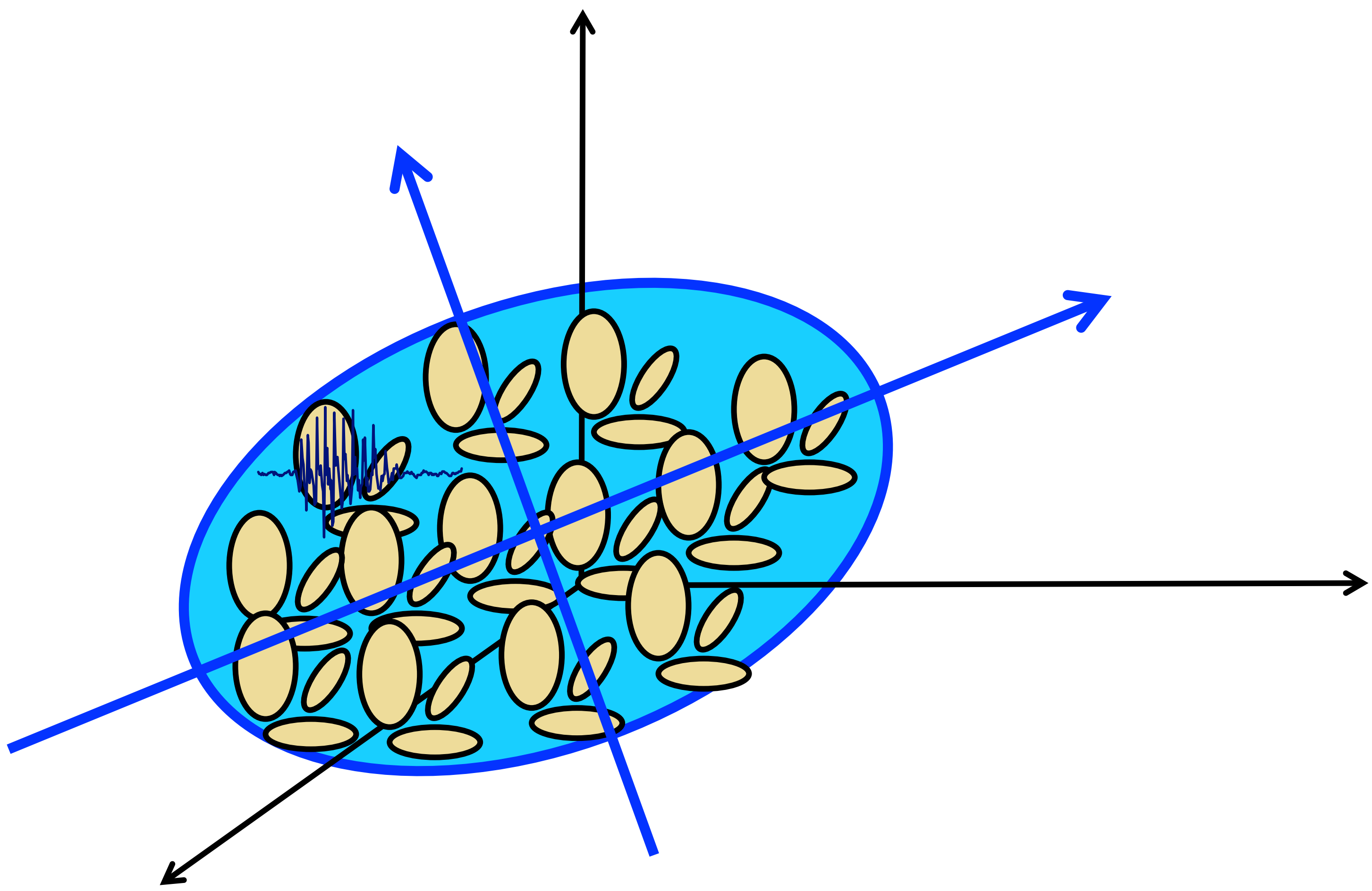
- **The way the UBM adapts to a given speaker ought to be somewhat constrained.**
  - For a particular speaker, there should exist some correspondence in the way the mean parameters move relative to one another.

- **Supervector Re-parameterization**
  - Concatenate all mixture mean components of a GMM.



Total variability space

- A GMM supervector corresponds to a point in space.

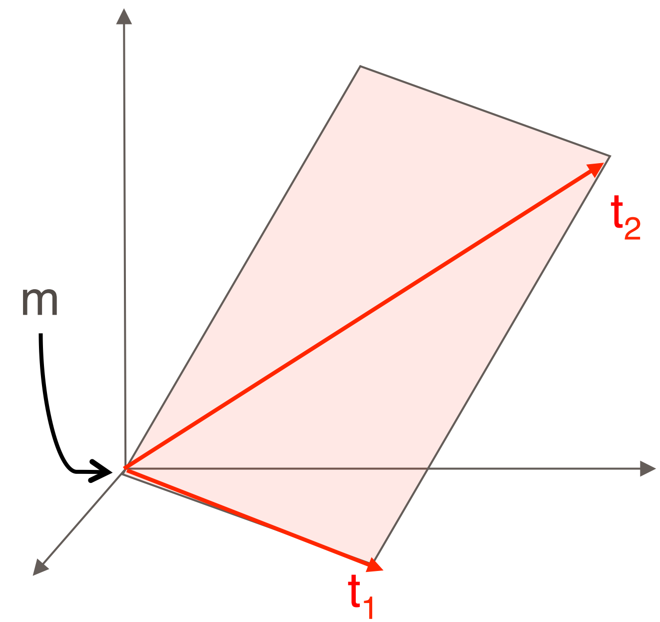


- Factor analysis captures the directions of maximum between-utterance variability.

The Total Variability Approach

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

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}$$



* **$\mathbf{w}$ is the vector of i-vectors (Identity/Intermediate Vectors)**

* **$\mathbf{m}$ is supervector of un-adapted (UBM) means**
* **$\mathbf{M}$ is supervector of speaker- and channel- dependent means**

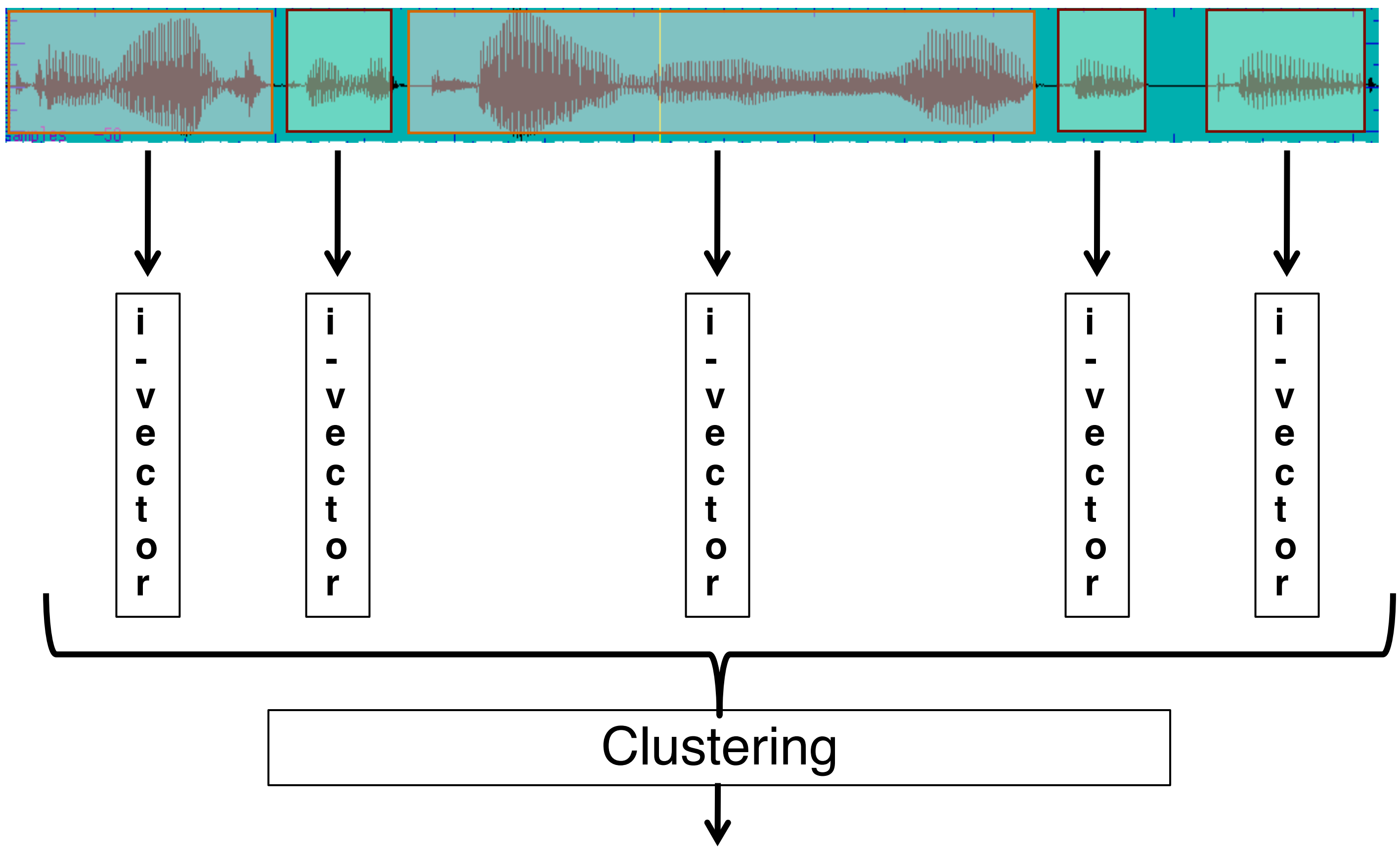
Regarding i-vectors

- **"For some speech segment s, its associated i-vector w_s can be seen as a low-dimensional summary of that segment's distribution of acoustic features with respect to a UBM."**
- **Low-dimensional random vector (100 << 20,000)**
  - Standard normal prior distribution, $\mathcal{N}(0, I)$
- **Given some speech data,**
  - Posterior mean → i-vector
  - Posterior covariance → i-vector covariance
- **Cosine similarity metric**
  - Can also length-normalize i-vectors onto the unit hypersphere

Roadmap

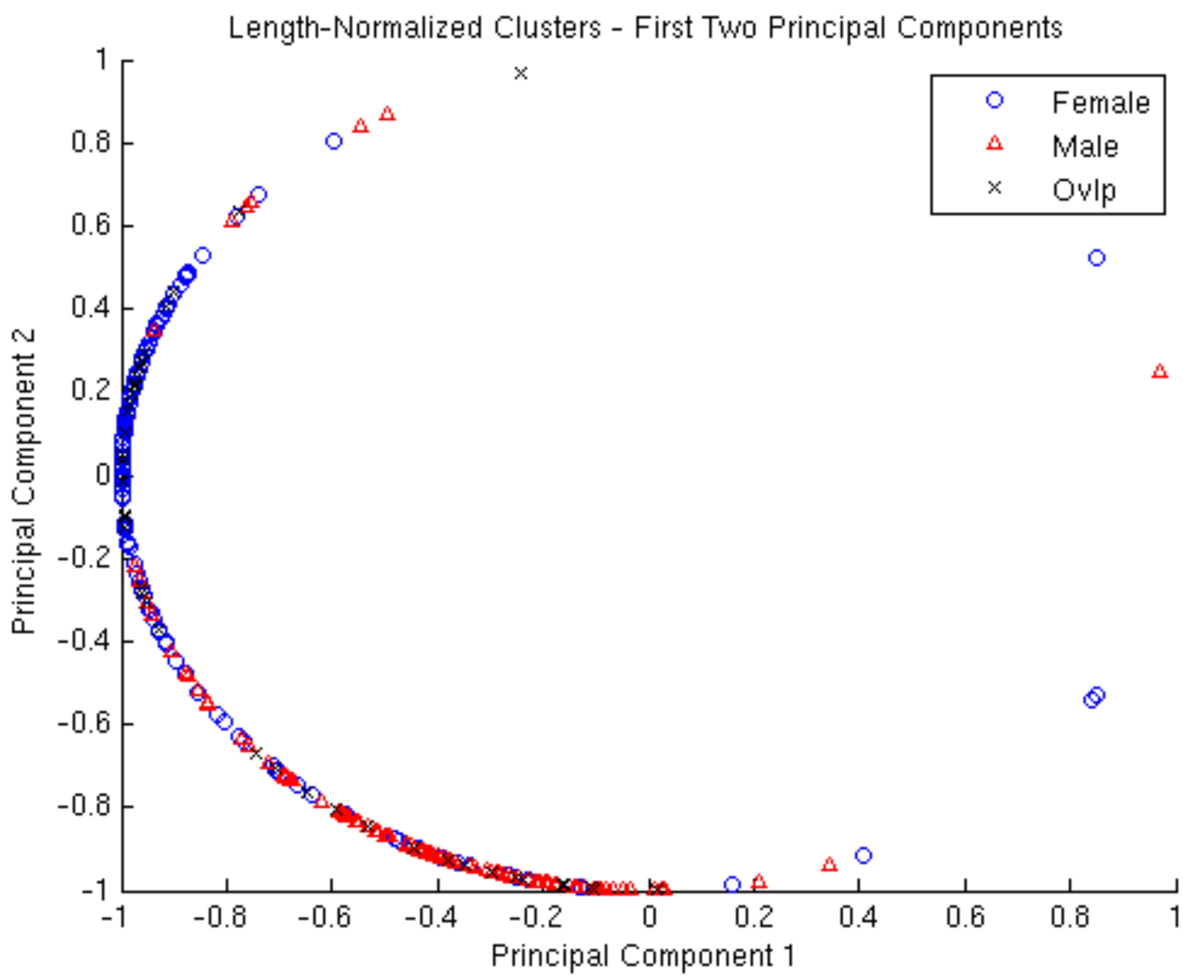
- **Introduction**
  – Summary of Contributions
- **Background**
  – Diarization System Overview
  – Speaker Modeling with Factor Analysis
- **Our Incremental Approach**
  – K-means and Spectral Clustering (Interspeech 2011, 2012)
  – Towards Probabilistic Clustering Methods
  – Iterative System Optimization (Re-segmentation/Clustering)
  – Duration-Proportional Sampling
- **Analysis and Discussion**
  – Benchmark Comparison (Castaldo 2008)
- **Conclusion**

Initialization



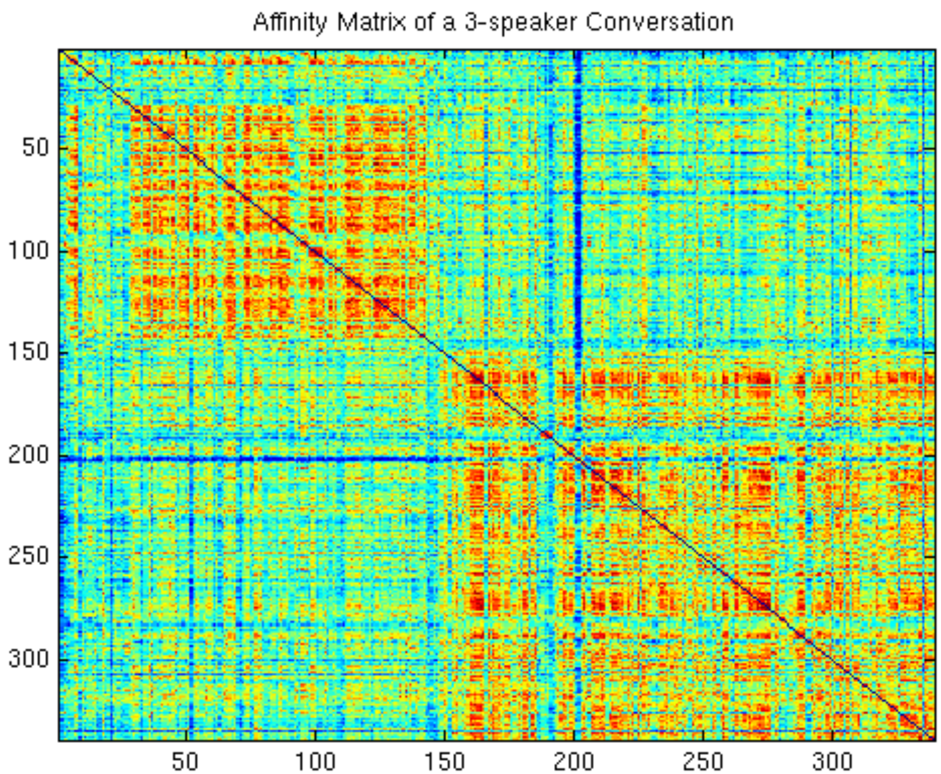
Clustering History

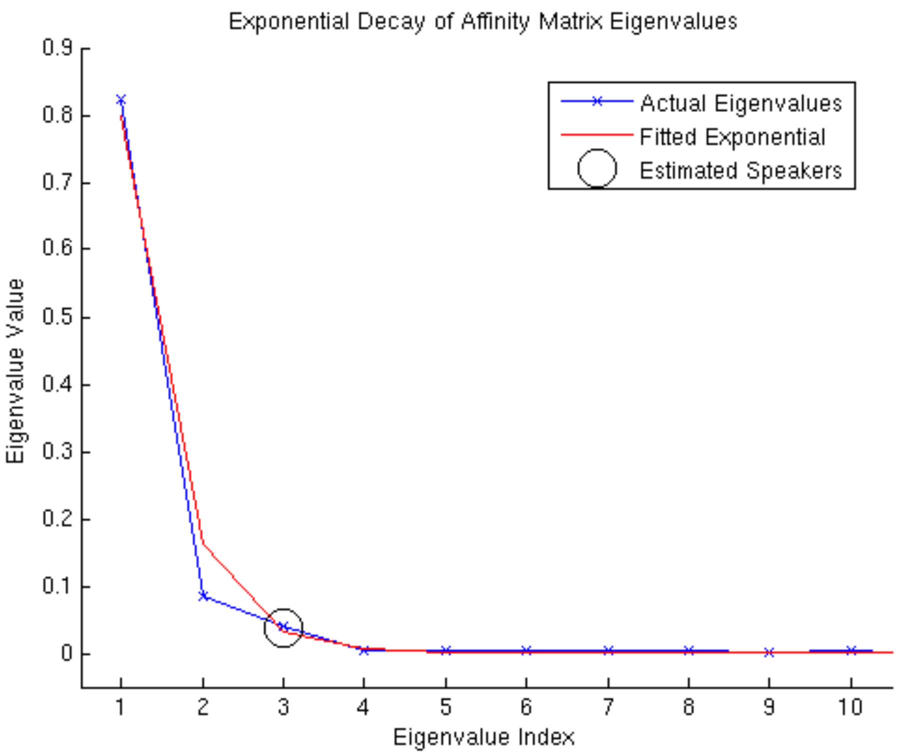
- **K-means on 2-speaker conversations (K = 2 known)**
  - Interspeech 2011



Clustering History

- **K-means on 2-speaker conversations (K = 2 known)**
  – Interspeech 2011
- **K-means and Spectral Clustering on K-speaker telephone conversations (K both known and unknown)**
  – Interspeech 2012





Clustering History

- **K-means on 2-speaker conversations (K = 2 known)**
  - Interspeech 2011
- **K-means and Spectral Clustering on K-speaker telephone conversations (K both known and unknown)**
  - Interspeech 2012

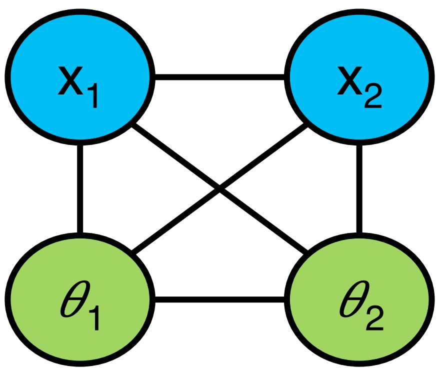
- **Probabilistic Methods (SM Thesis 2011)**
  - K-means → Gaussian Mixture Models
    - **Bayesian model selection via variational inference**

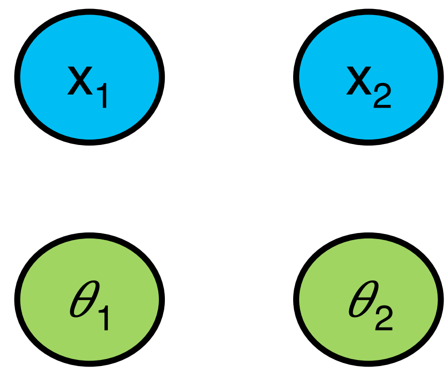
The Need for Approximate Inference

- Consider some observed data *Y*, a hidden variable set *X*, and associated parameters *θ*
- For model selection m, we want to maximize

$$\log P(Y|m) = \log \int P(Y,X,\theta|m)\,dX\,d\theta$$

→ exact computation is intractable in general





- Introduce $q(X,\theta) = q(X)\cdot q(\theta)$ to approximate $P(X,\theta|Y,m)$

$$\log P(Y|m) = F_m(q(X,\theta)) + \mathrm{KL}(q(X,\theta)||P(X,\theta|Y,m))$$

* Maximizing the Free Energy minimizes the KL-divergence between the variational posterior and true posterior distributions

Variational Free Energy

$$F_m(q(X)q(\theta)) = \int \underbrace{q(X)q(\theta) \cdot \log P(Y, X|\theta, m) dX d\theta}_{\text{Expectation, under } q(X,\theta)\text{, of complete data log-likelihood}}$$

$$+\underbrace{\mathrm{H}(q(X))}_{\text{Entropy of X}} - \underbrace{\mathrm{KL}(q(\theta)||P(\theta|m))}_{\text{KL-divergence between variational parameters and actual priors}}$$

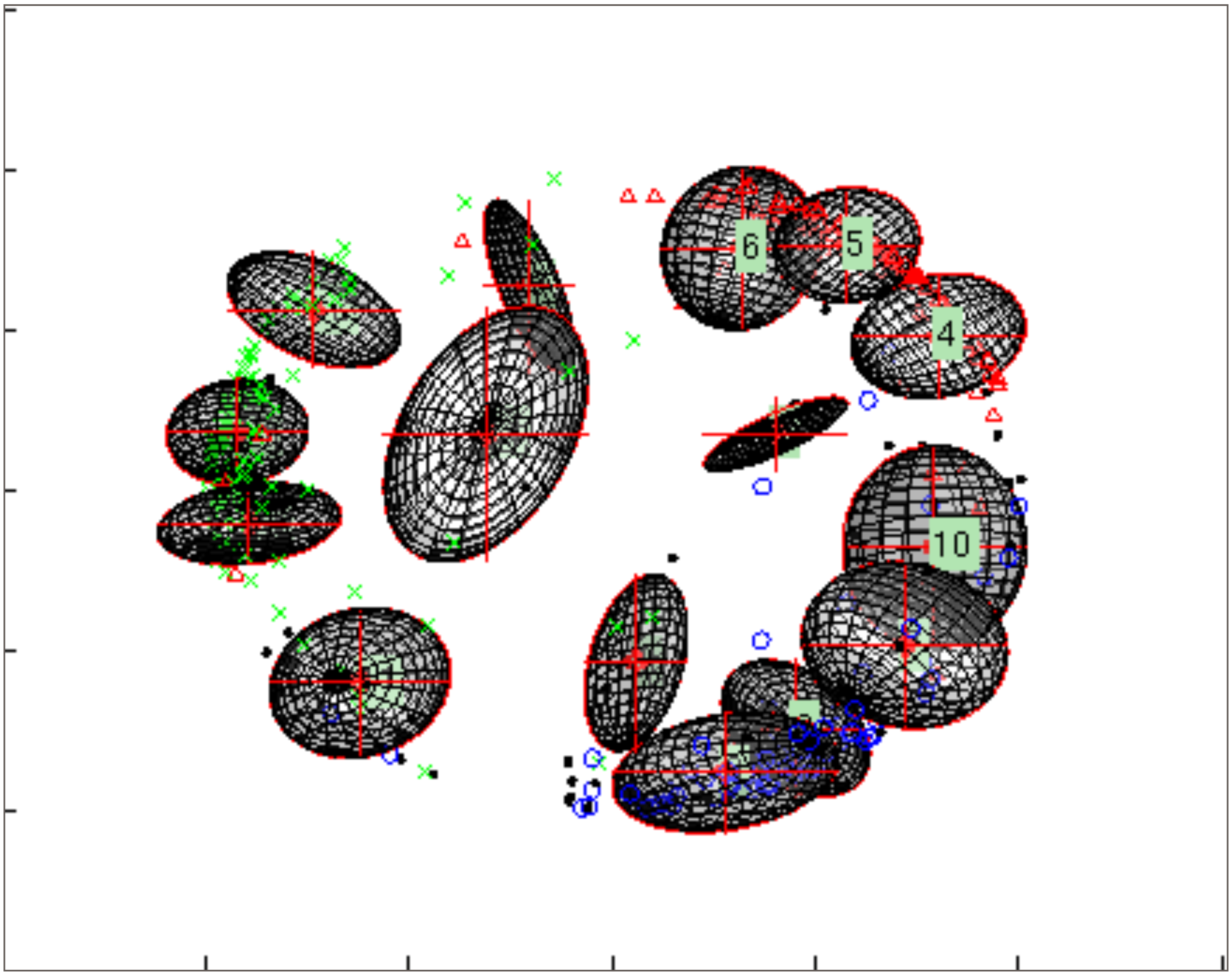
- **The act of maximizing $F_m(q(X)q(\theta))$ yields an EM algorithm**
  - VBEM-GMM

Clustering History

- **K-means on 2-speaker conversations (K = 2 known)**
  - Interspeech 2011
- **K-means and Spectral Clustering on K-speaker telephone conversations (K both known and unknown)**
  - Interspeech 2012

- **Probabilistic Methods (SM Thesis 2011)**
  - K-means → Gaussian Mixture Models
    - **Bayesian model selection via variational inference**
  - Rote application of VBEM-GMM

VBEM-GMM Visualization



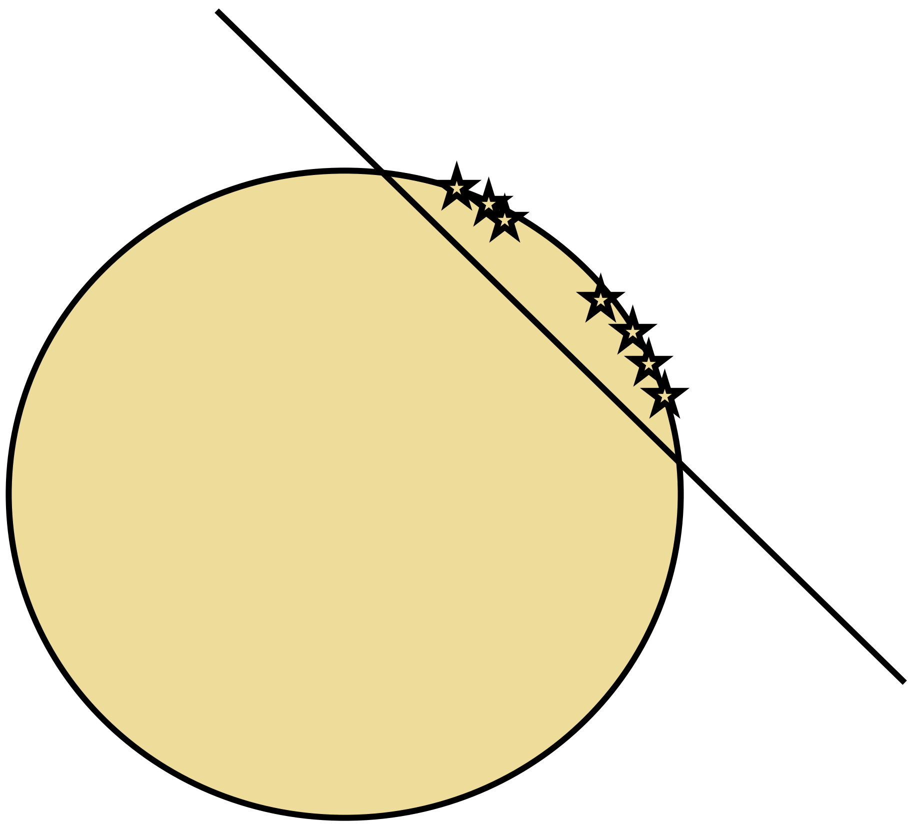
Clustering History

- **K-means on 2-speaker telephone conversations (K known)**
  - Interspeech 2011
- **K-means and Spectral Clustering on K-speaker telephone conversations (K both known and unknown)**
  - Interspeech 2012

- **Probabilistic Methods (SM Thesis 2011)**
  - K-means → Gaussian Mixture Models
    - **Bayesian model selection via the <u>variational approximation</u>**
  - Rote application of VBEM-GMM
    - **GMMs are a poor way to model data living on a unit hypersphere.**

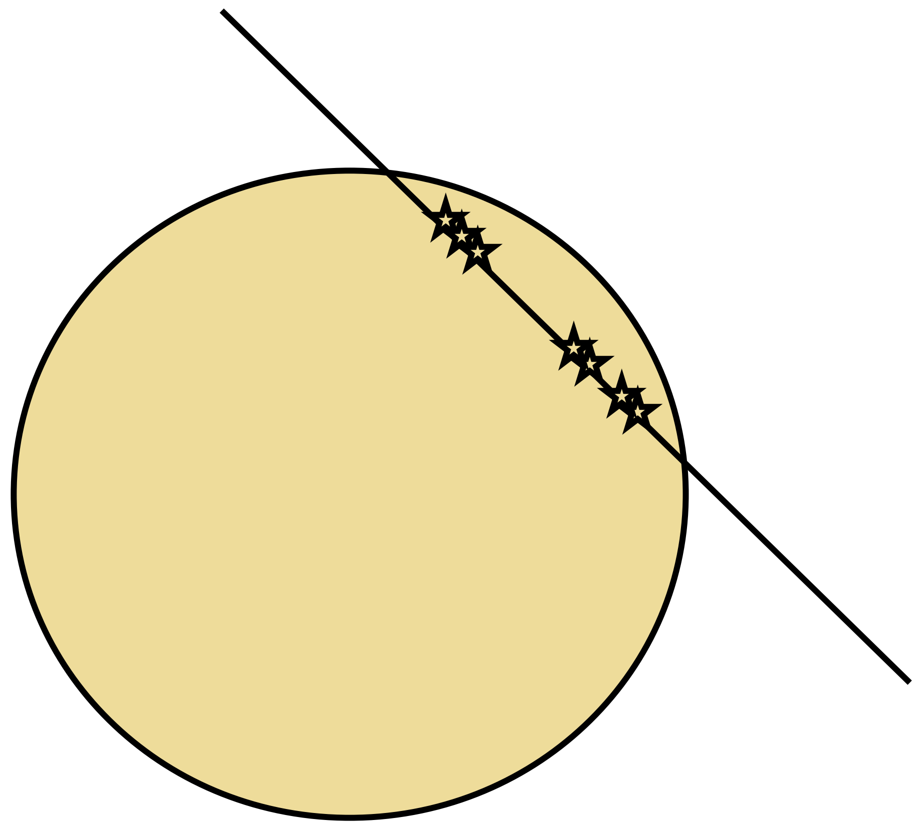
Dimensionality Reduction

- **i-vectors are both speaker- and channel-dependent**
  - Channel effect localizes all i-vectors onto one small region on the unit hypersphere
  - Consider a projection (PCA) onto a lower-dimensional plane

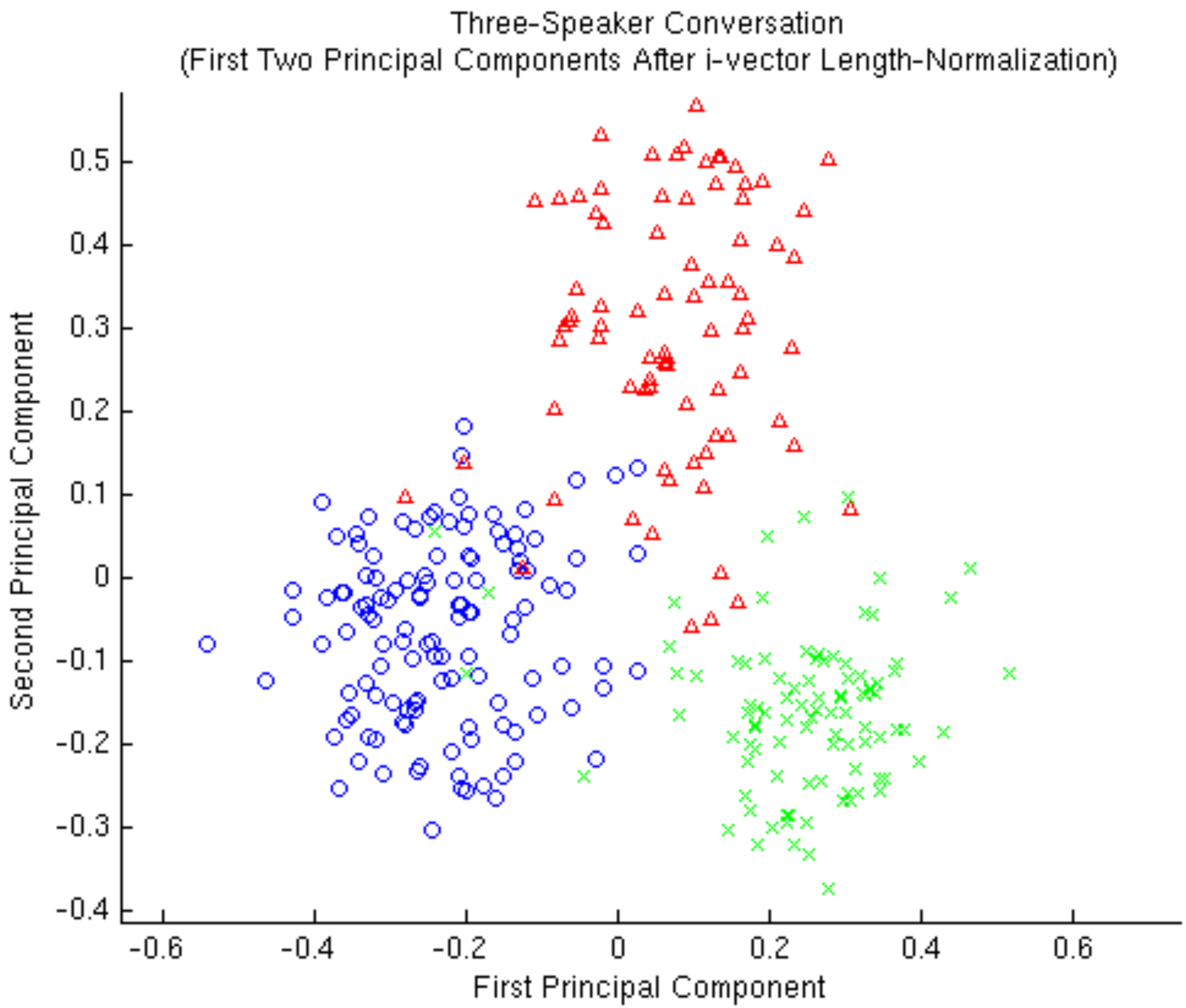


Dimensionality Reduction

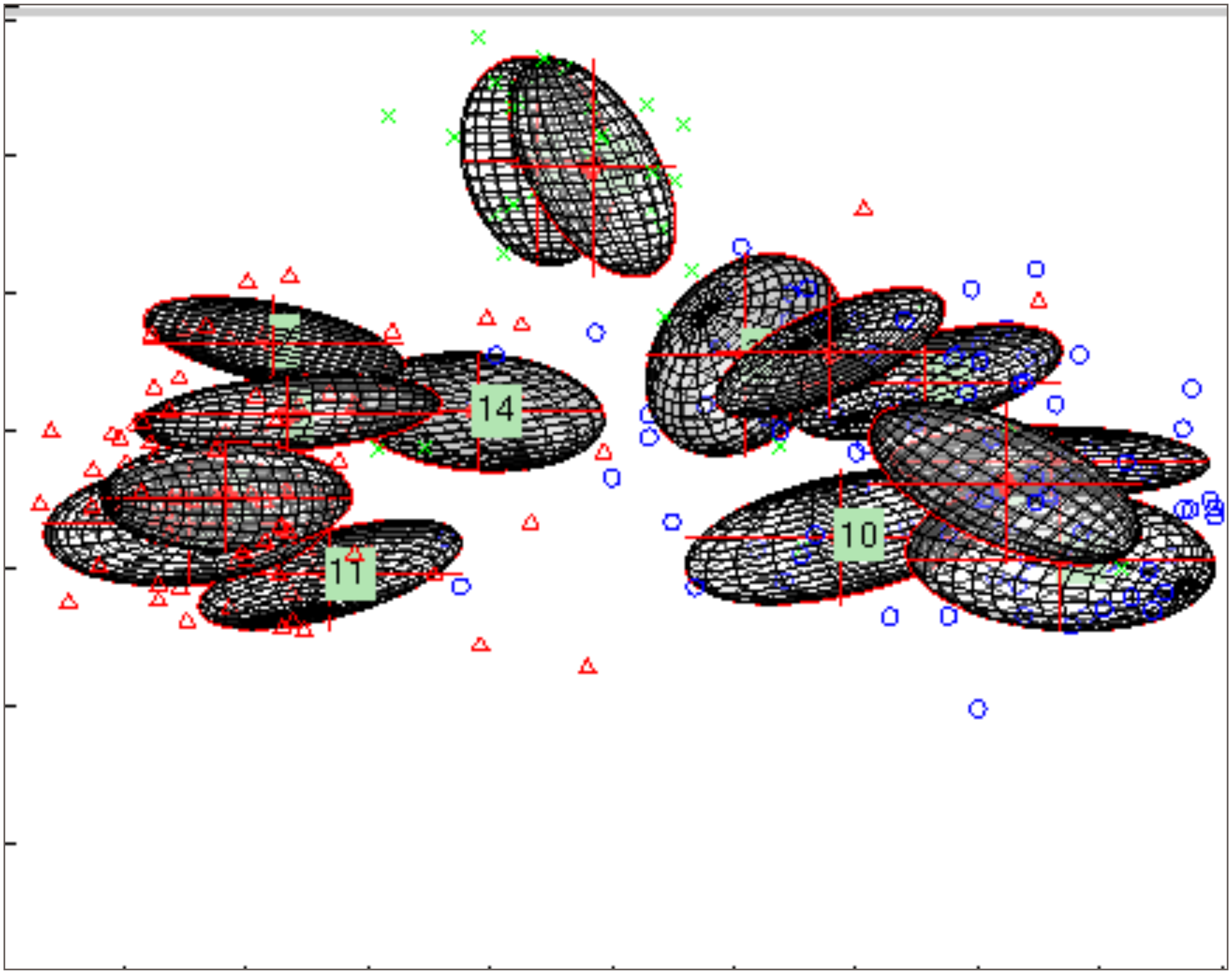
- **i-vectors are both speaker- and channel-dependent**
  - Channel effect localizes all i-vectors onto one small region on the unit hypersphere
  - Consider a projection (PCA) onto a lower-dimensional plane



PCA Visualization



VBEM-GMM Clustering (after PCA)

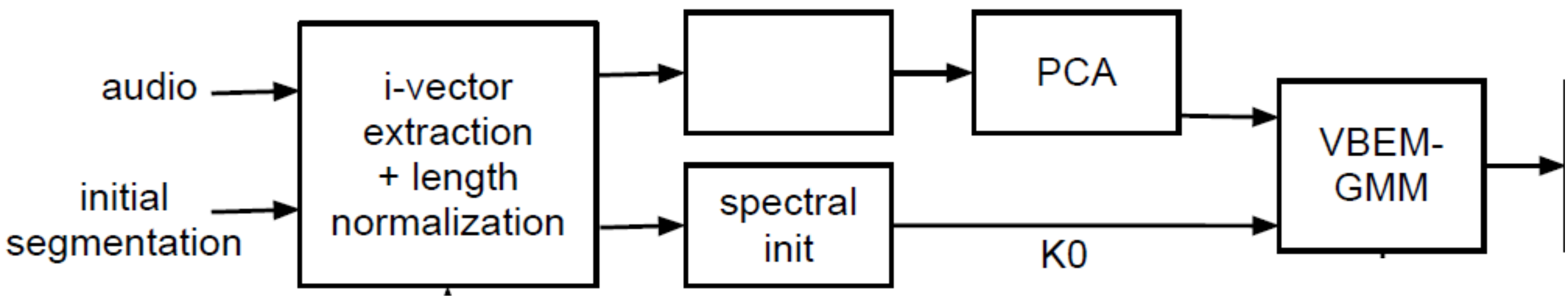


Cluster Initialization

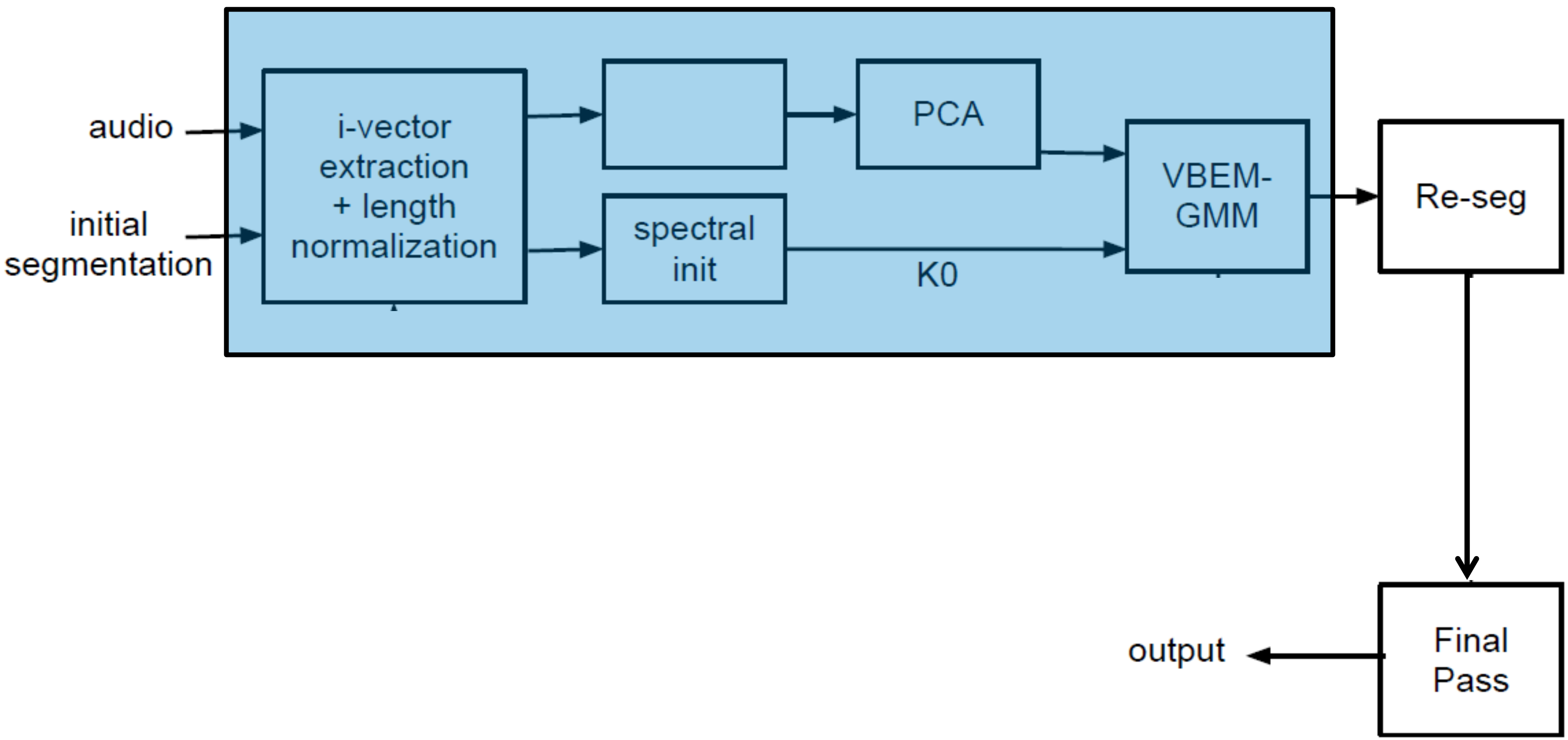
- **Baseline Approach**
  - Over-initialize the number of clusters
    - $K_0 = 15$
  - Remove components iteratively

- **Proposed Refinement**
  - Initialize using eigenvalue roll-off from the affinity matrix generated by the spectral clustering algorithm
    - $K_0 = \hat{K} + \lceil 3 \cdot \sigma_K \rceil$
  - Still want to over-initialize clusters, but in a more informed manner.

System Diagram (Clustering)



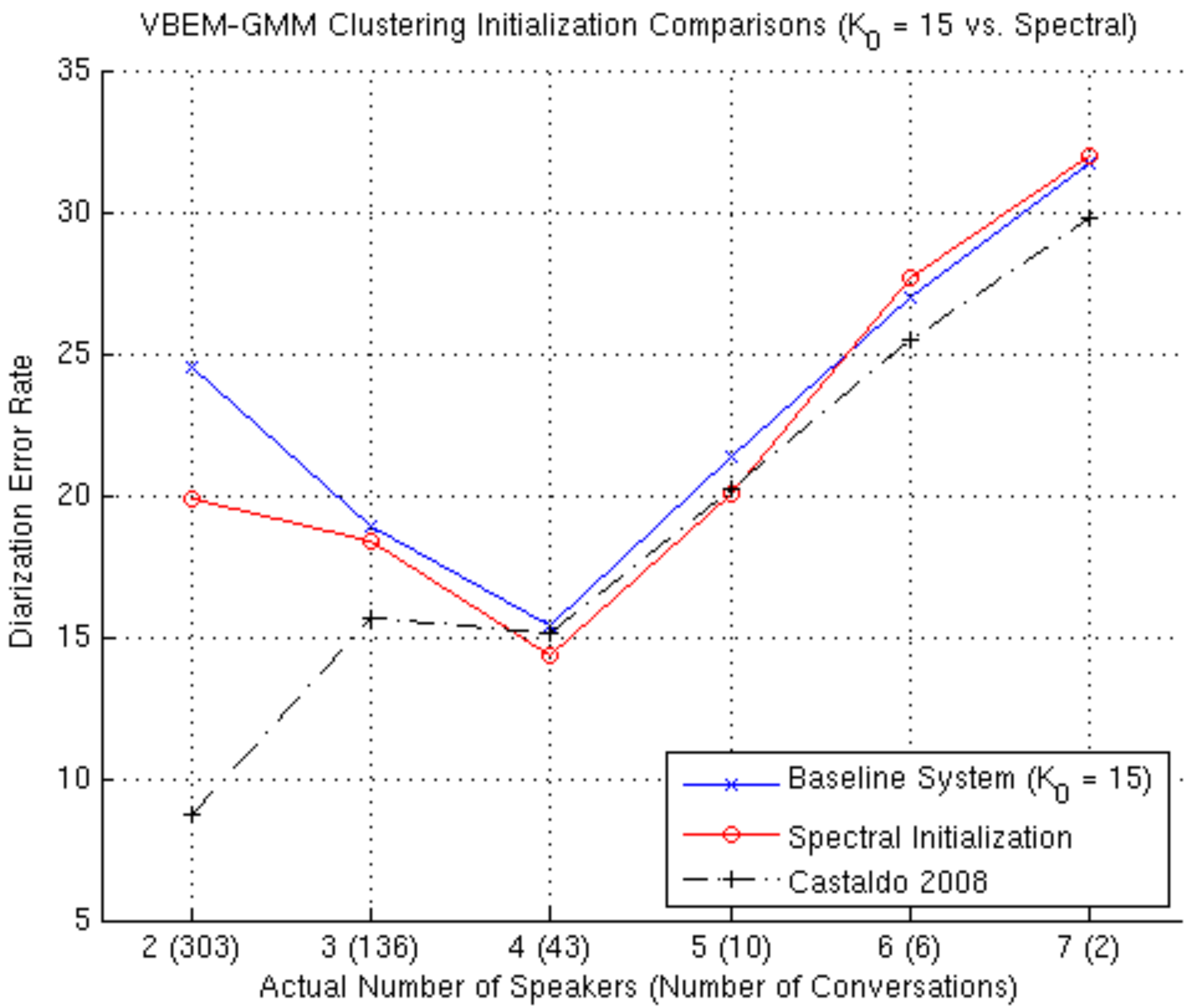
System Diagram (Baseline)



Experiment Details

- **Evaluation Data**
  - Multi-lingual CallHome corpus
    - **500 recordings, 2-5 minutes each, containing 2-7 speakers**
- **Total Variability**
  - 20-dimensional MFCC acoustic feature vectors
  - UBM of 1024 Gaussians
  - Rank of Total Variability matrix = 100
    - **i.e. 100-dimensional i-vectors**
- **Diarization Error Rate (DER)**
  - Amount of time spent confusing one speaker's speech as from another

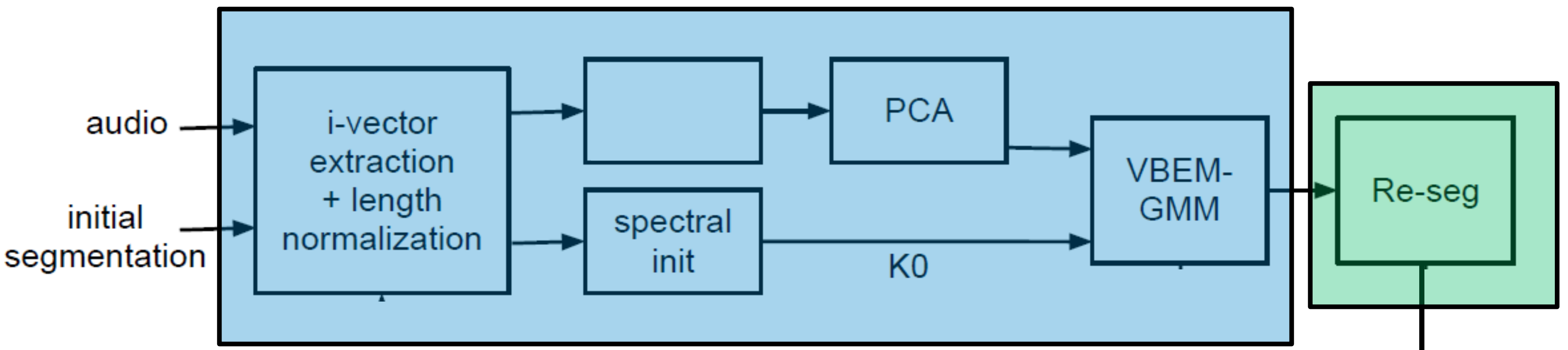
Initial Results

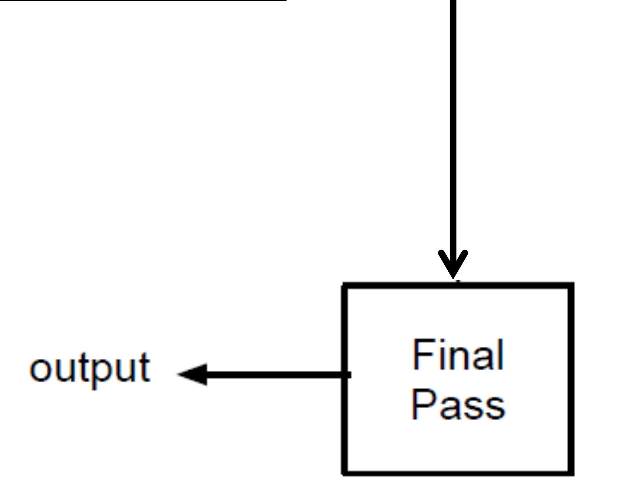


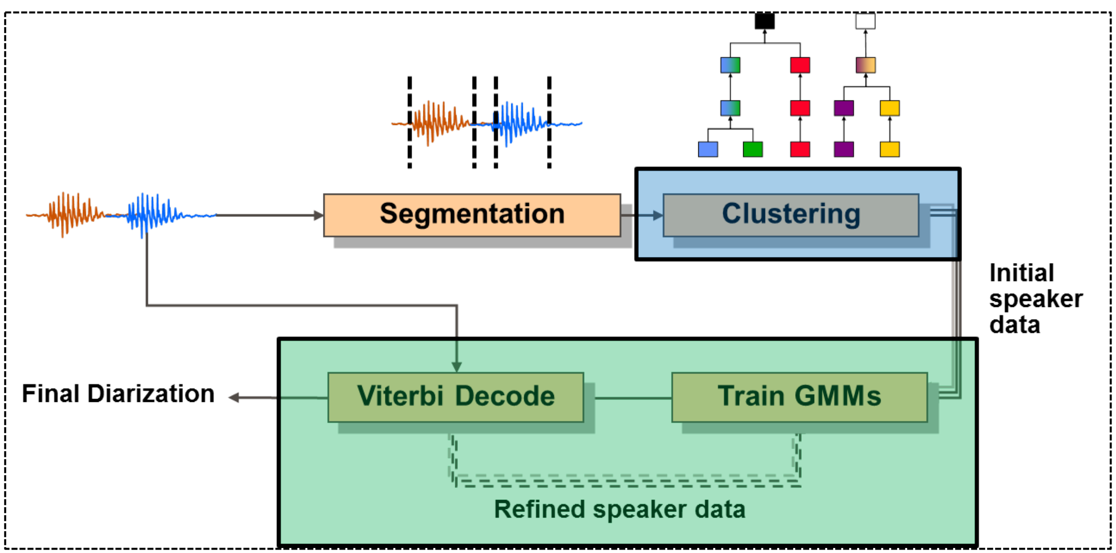
Roadmap

- Introduction
  - Summary of Contributions
- Background
  - Diarization System Overview
  - Speaker Modeling with Factor Analysis
- Our Incremental Approach
  - K-means and Spectral Clustering (Interspeech 2011, 2012)
  - Towards Probabilistic Clustering Methods
  - Iterative System Optimization (Re-segmentation/Clustering)
  - Duration-Proportional Sampling
- Analysis and Discussion
  - Benchmark Comparison (Castaldo 2008)
- Conclusion

System Diagram (Baseline)



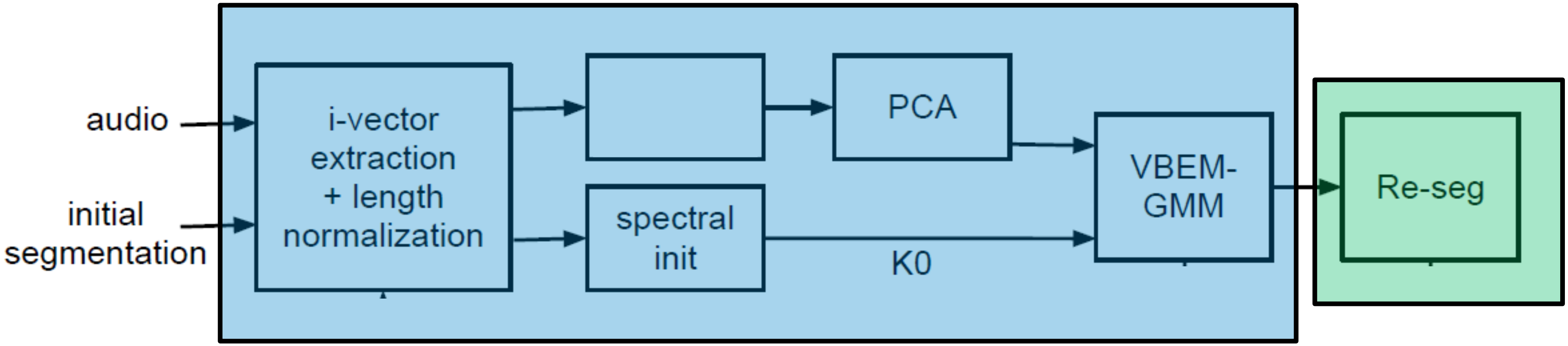




Iterative Re-segmentation

- Initialize a GMM for each cluster.
  - **Speaker 1, Speaker 2, …, Non-speech N**
- Obtain a posterior probability for each cluster given each feature vector.
  - $P(S_1|x_t)$, $P(S_2|x_t)$, …, $P(N|x_t)$
- 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.

A Symbiotic Relationship

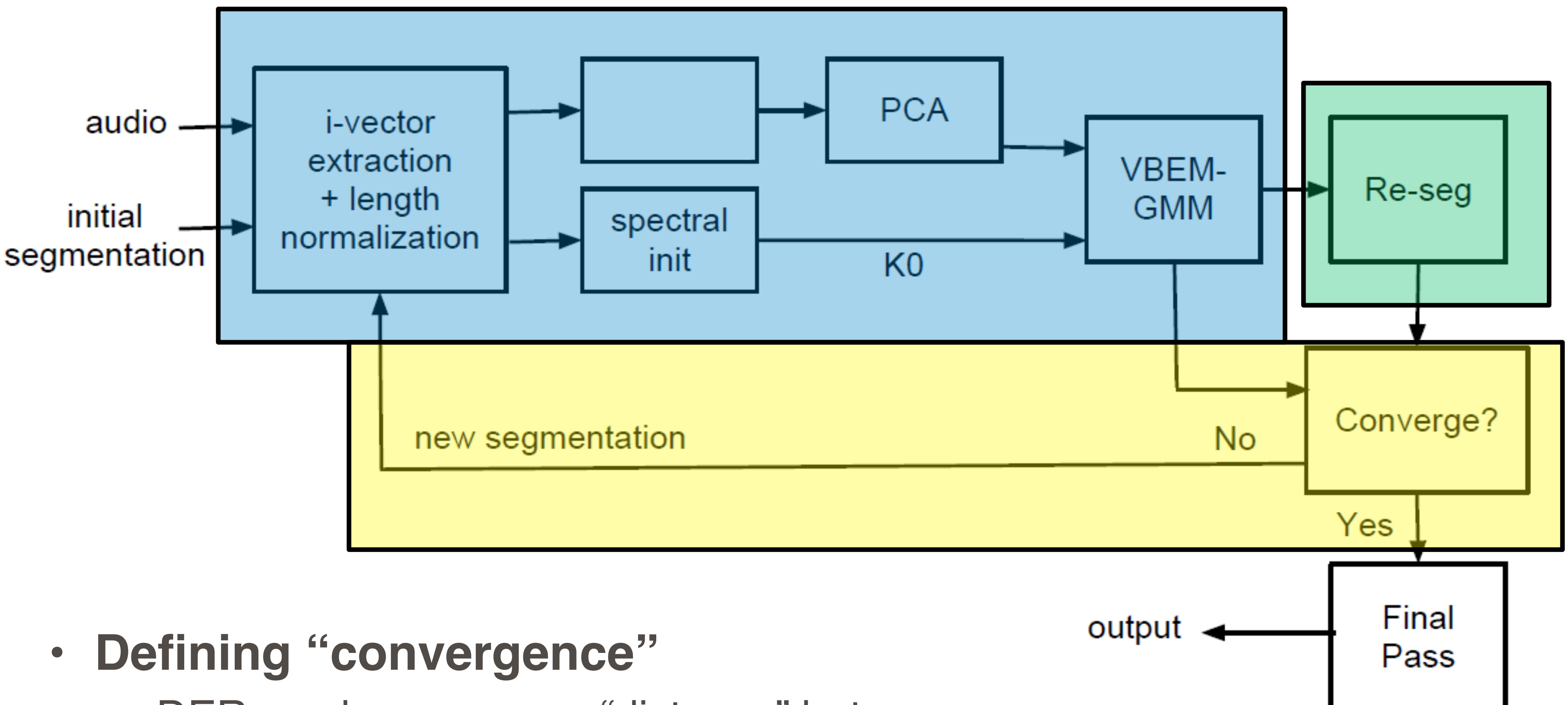


- **Clustering assumes some initial segmentation and clusters at the i-vector level**
  – Better speaker representation

- **Re-segmentation operates at level of acoustic features**
  – Finer temporal resolution

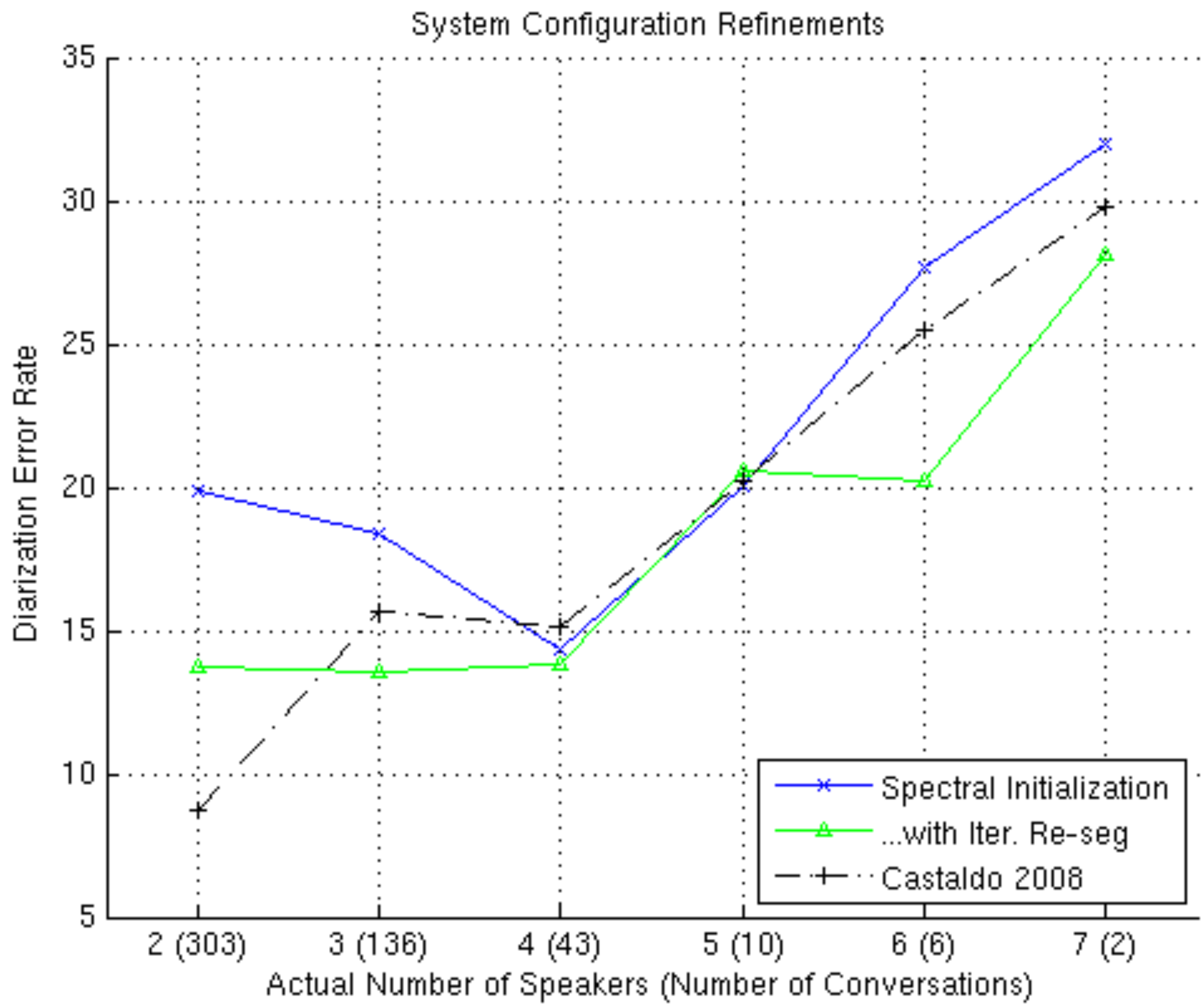
Iterative System Optimization

CSAIL

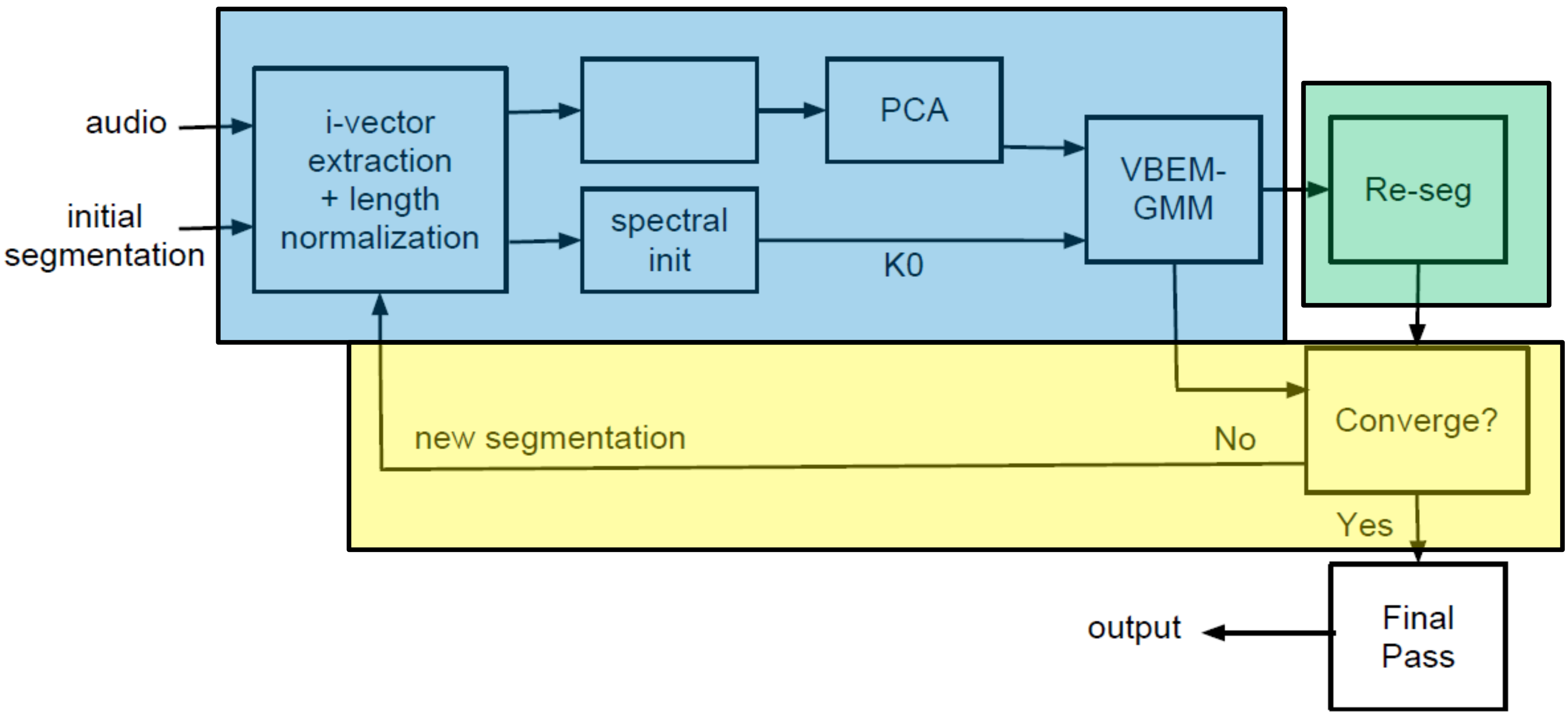


- **Defining "convergence"**
  - DER can be seen as a "distance" between two diarization hypotheses.

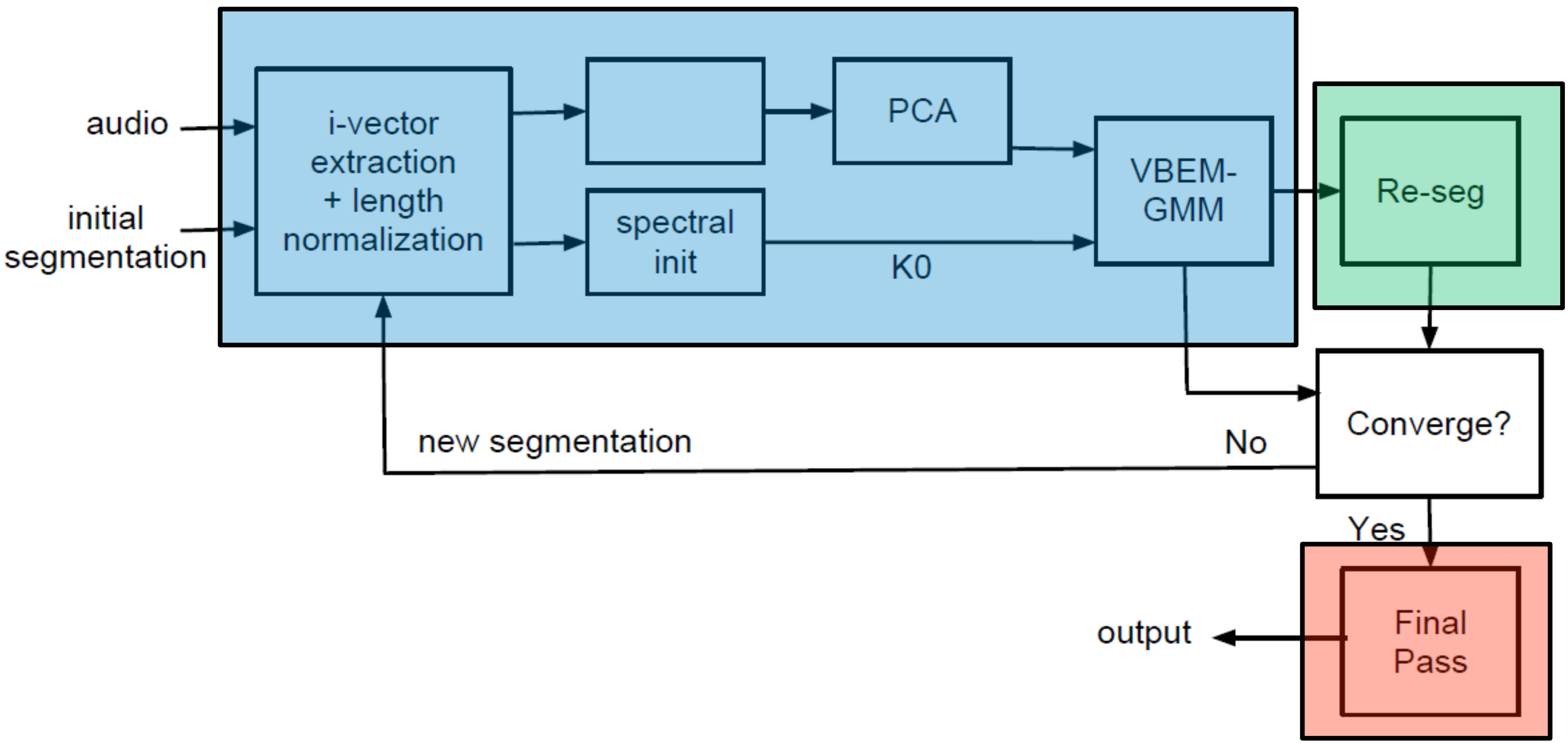
Iterative System Optimization Results



Diarization System So Far



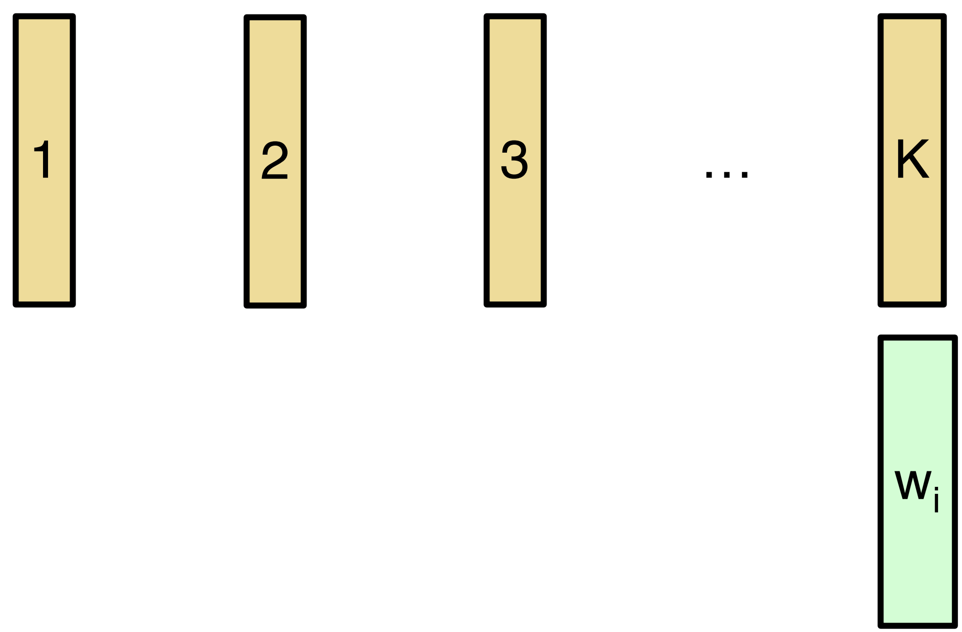
Diarization System So Far



Final Pass Refinements
(Interspeech 2011)

- 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 $\{w_1, w_2, \ldots, w_K\}$ that is closer in cosine similarity.
  - **"Winner Takes All"**



Final Pass Refinements
(Interspeech 2011)

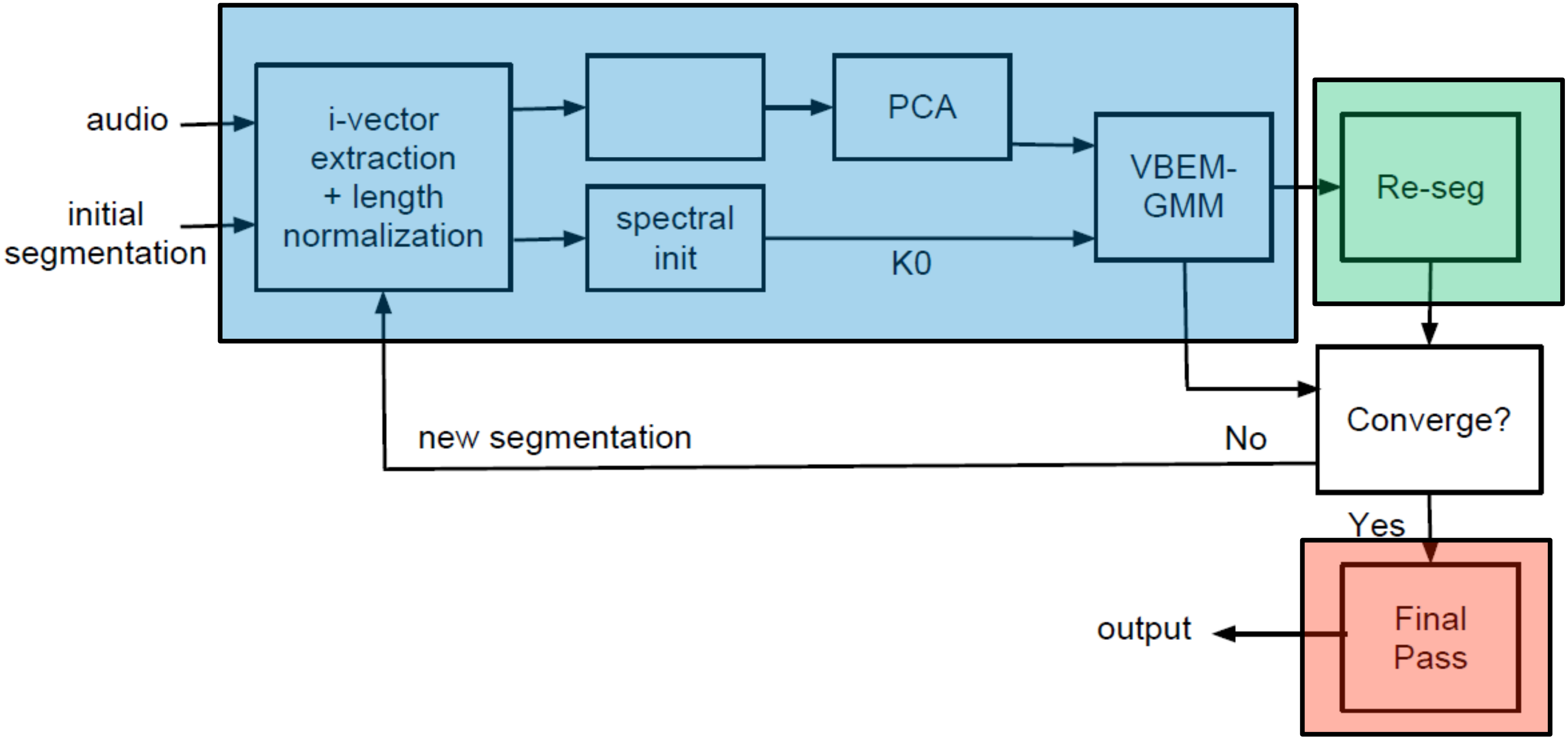
- Extract a single i-vector for each respective speaker.
  - **Using the newly defined re-segmentation assignments**

- Re-assign each newly-extracted segment i-vector w_i to the speaker i-vector $\{w_1, w_2, \ldots, w_K\}$ that is closer in cosine similarity.
  - **"Winner Takes All"**

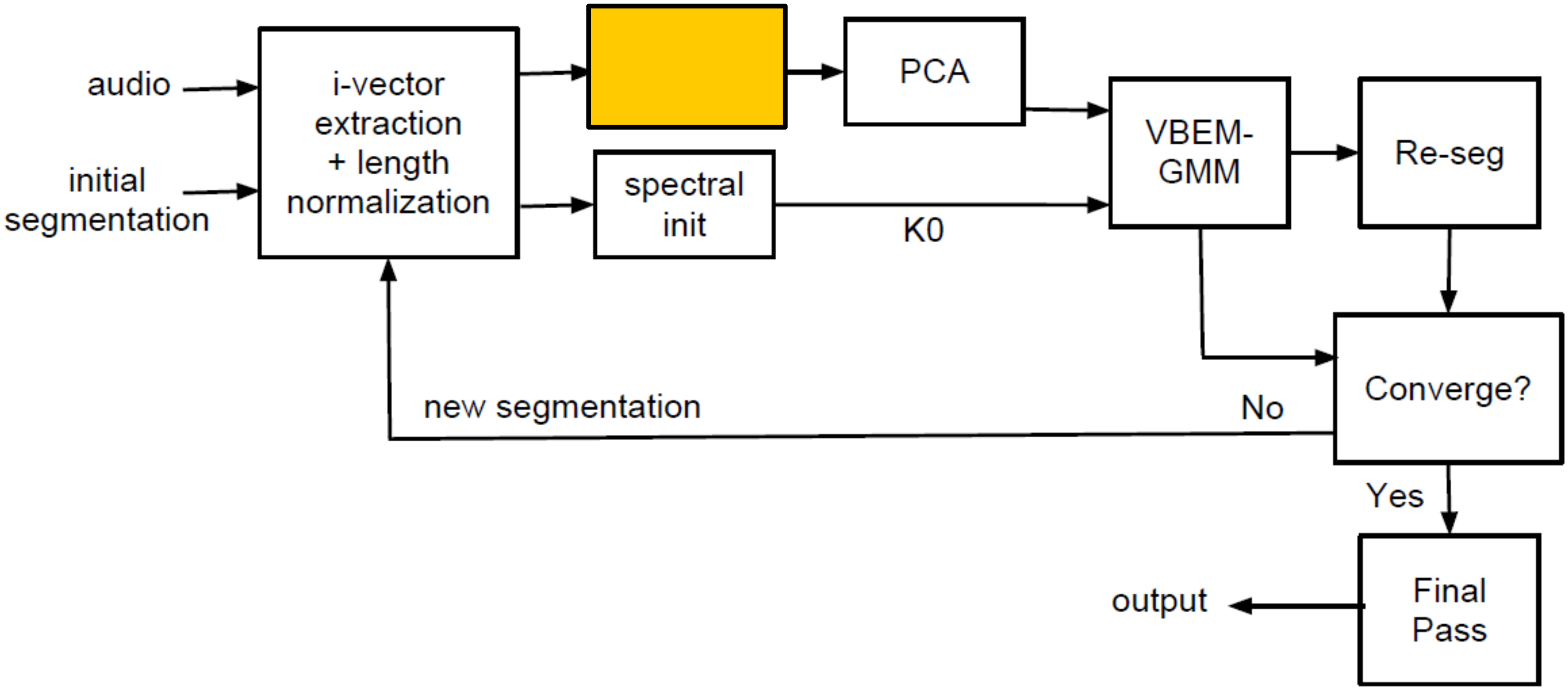
- Iterate until convergence.
  - **i.e. when segment-speaker assignments no longer change**

- Essentially a K-means algorithm
  - **Except determine "means" $\{w_1, w_2, \ldots, w_K\}$ via i-vector extraction**

Diarization System So Far



Diarization System So Far



i-vector Underrepresentation

- **i-vectors have been used as point estimates.**
  - During clustering, we treat them as independent and identically distributed samples from some underlying GMM.

- **However, some i-vectors may be more equal than others.**
  - i-vector from a 5-second speech segment versus 0.5-second segment

- **Recall: Given some speech,**
  - The i-vector is a posterior mean of a Gaussian distribution…
  - With an associated posterior covariance

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

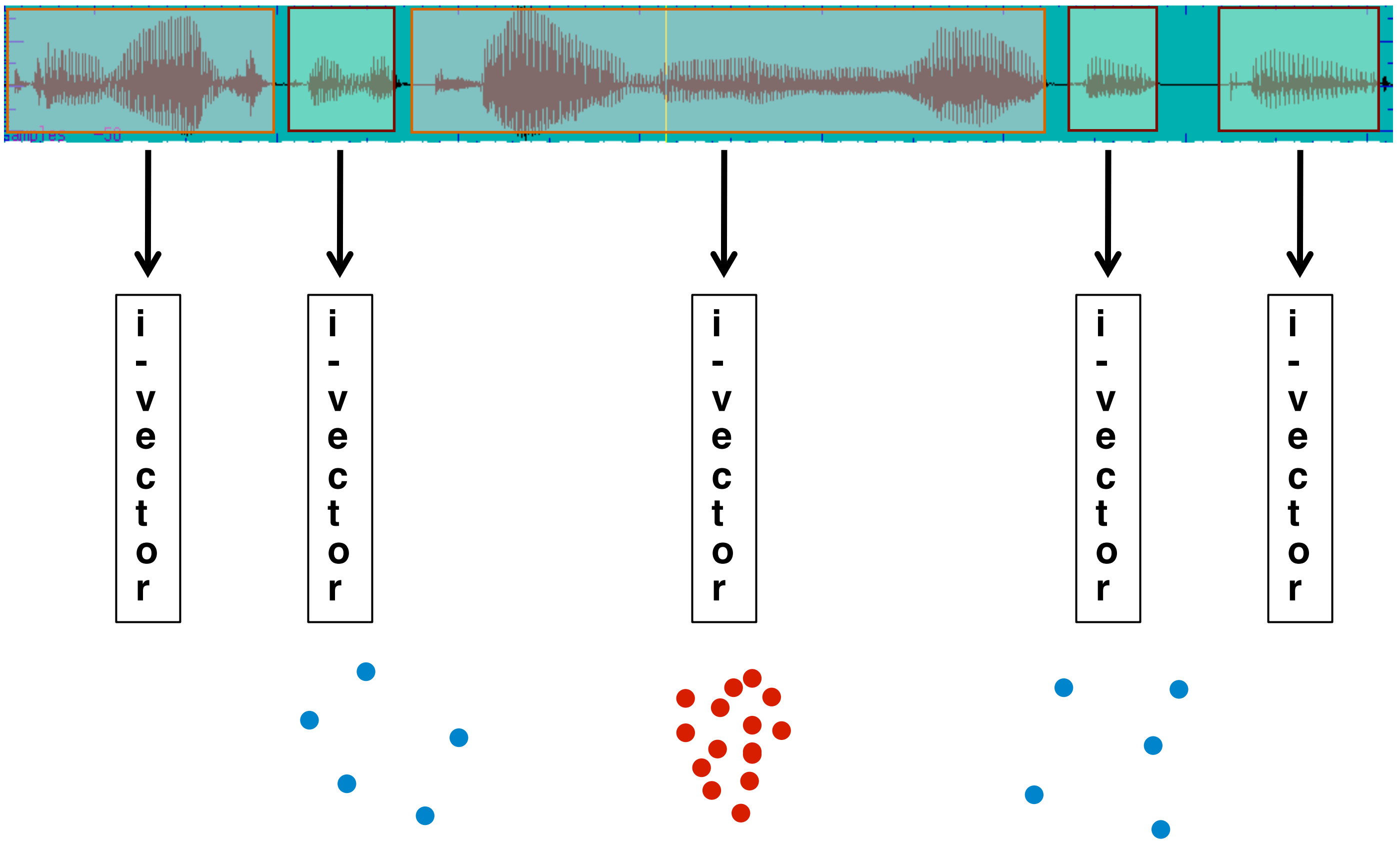
Overcoming Underrepresentation
– A Sampling Approach

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

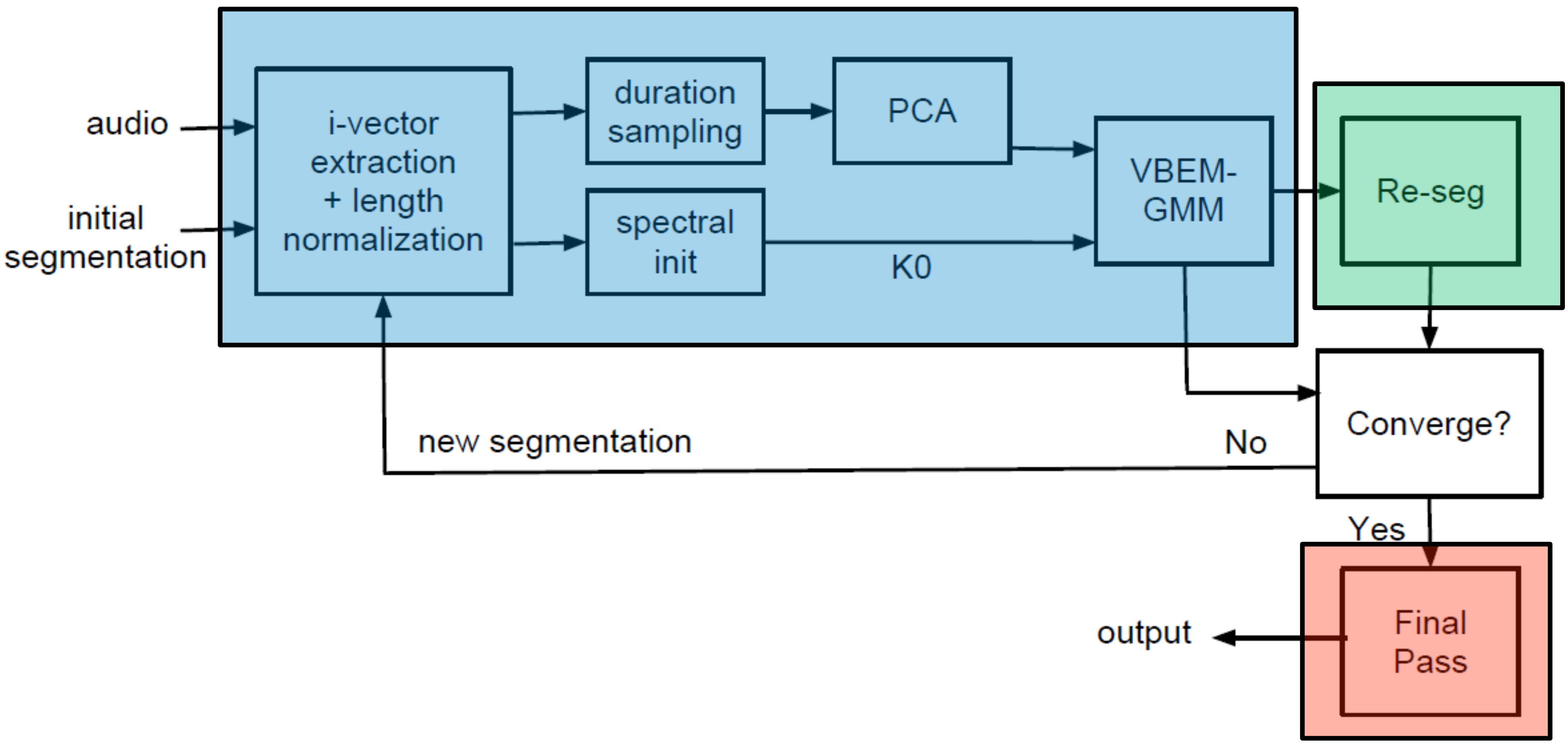
$$\mathrm{cov}(w) = \left(I + T^{*}\Sigma^{-1}\boxed{N(u)}T\right)^{-1}$$

- **Consider sampling the i-vector distribution**
  - Let the number of samples drawn be proportional to the number of frames used to extract the i-vector.
    - **Shorter segments → <u>larger</u> covariance and <u>fewer</u> samples**
    - **Longer segments → <u>smaller</u> covariance and <u>more</u> samples**

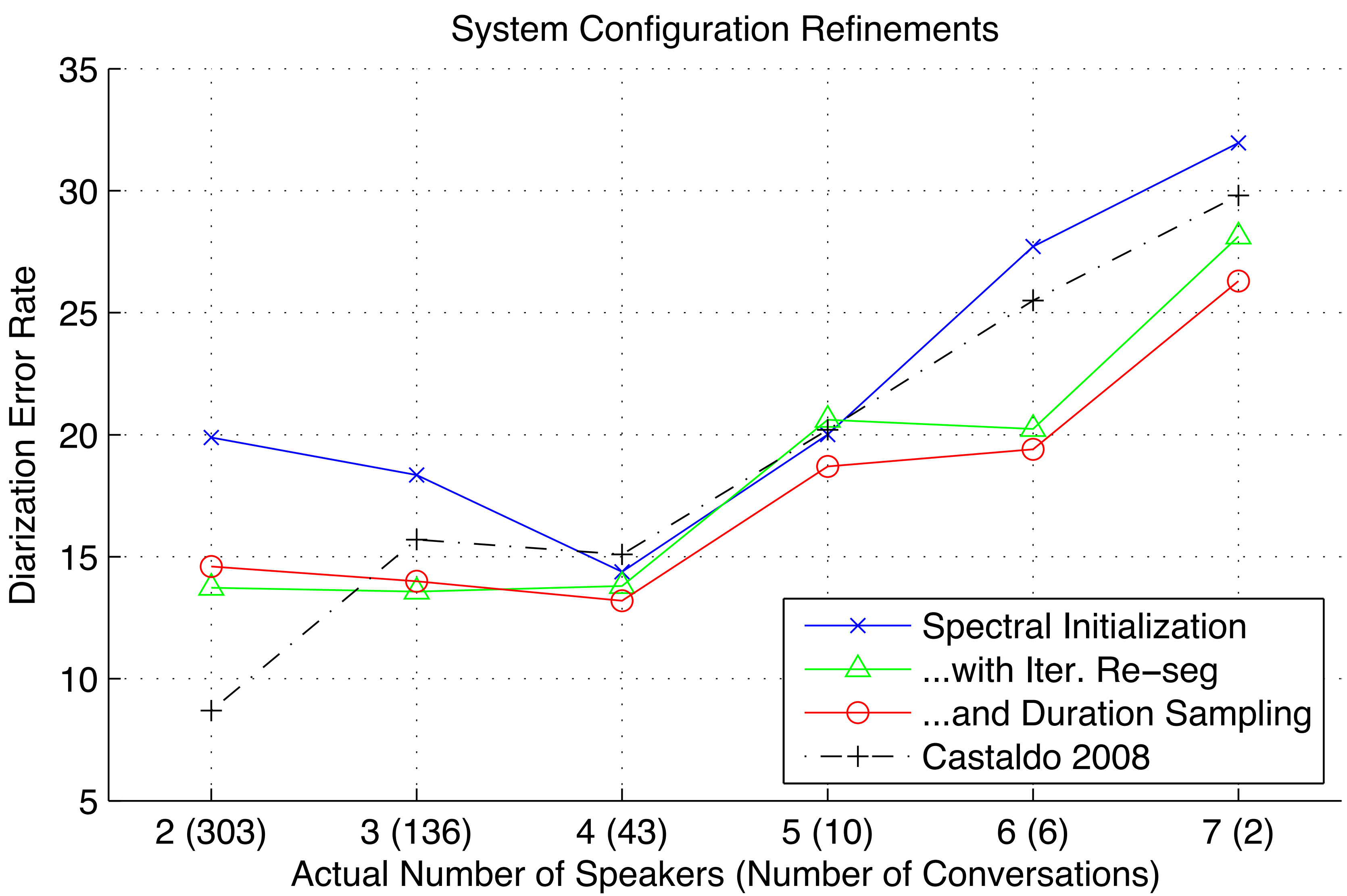
A Simplified Cartoon



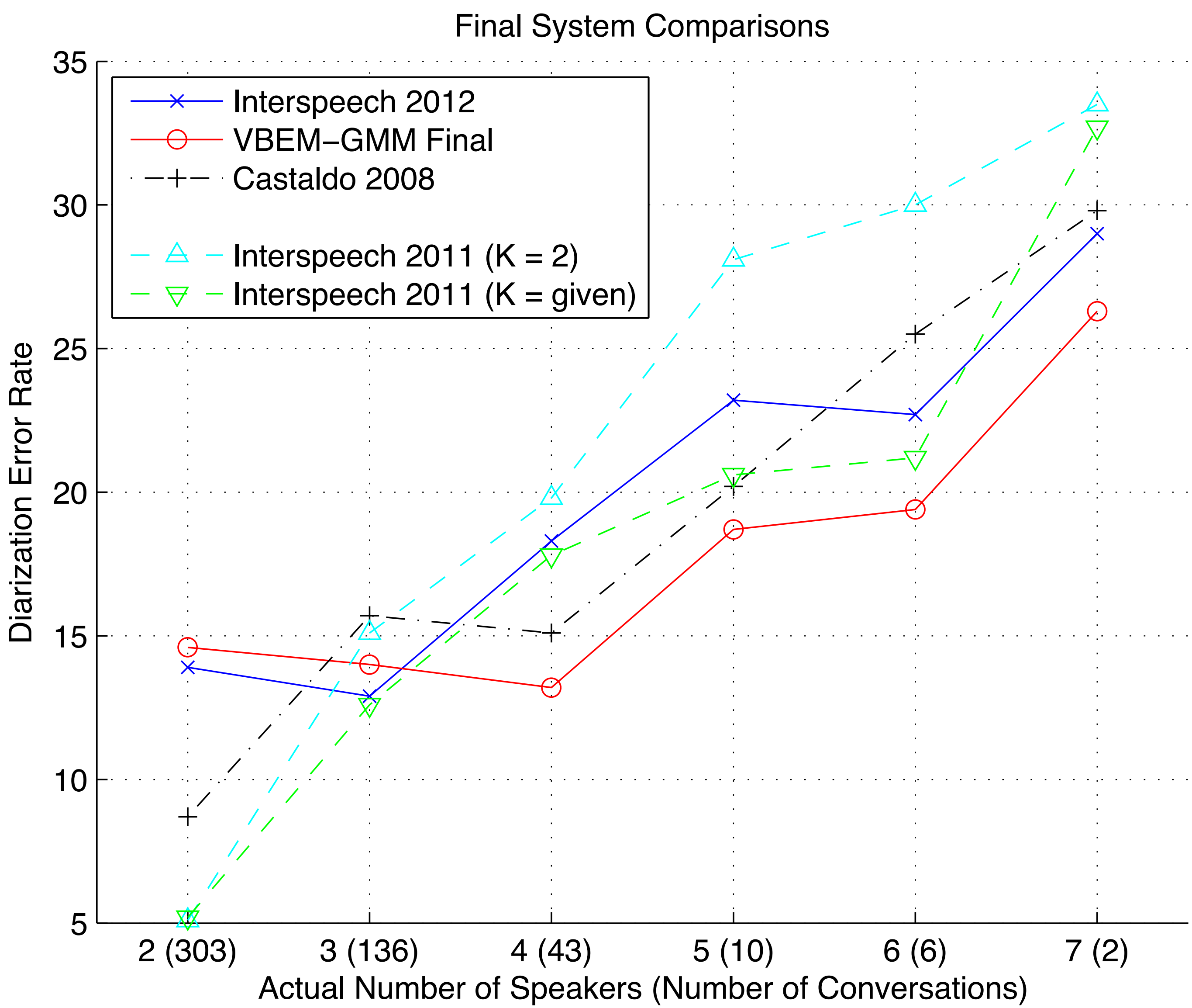
Final System Diagram



Proposed System Refinements



Final System Comparisons



Reconciling Our 2-Speaker Results

- **Interspeech 2011 vs. Kenny 2010 vs. Castaldo 2008**
  - State-of-the-art results on diarization on two-speaker telephone calls (number of speakers given)
- **Interspeech 2012**
  - On the CallHome corpus, when it is known that the conversation contains only two participants
    - **DER = 5.2% vs. 8.7% (Castaldo 2008)**

DER Observations

- **Over-detecting the number of speakers**
  - In the conversations where we correctly detect two speakers (136/303),
    - **DER = 6.5% vs. 8.7% (Castaldo 2008)**
  - But DER is unforgiving towards overestimation

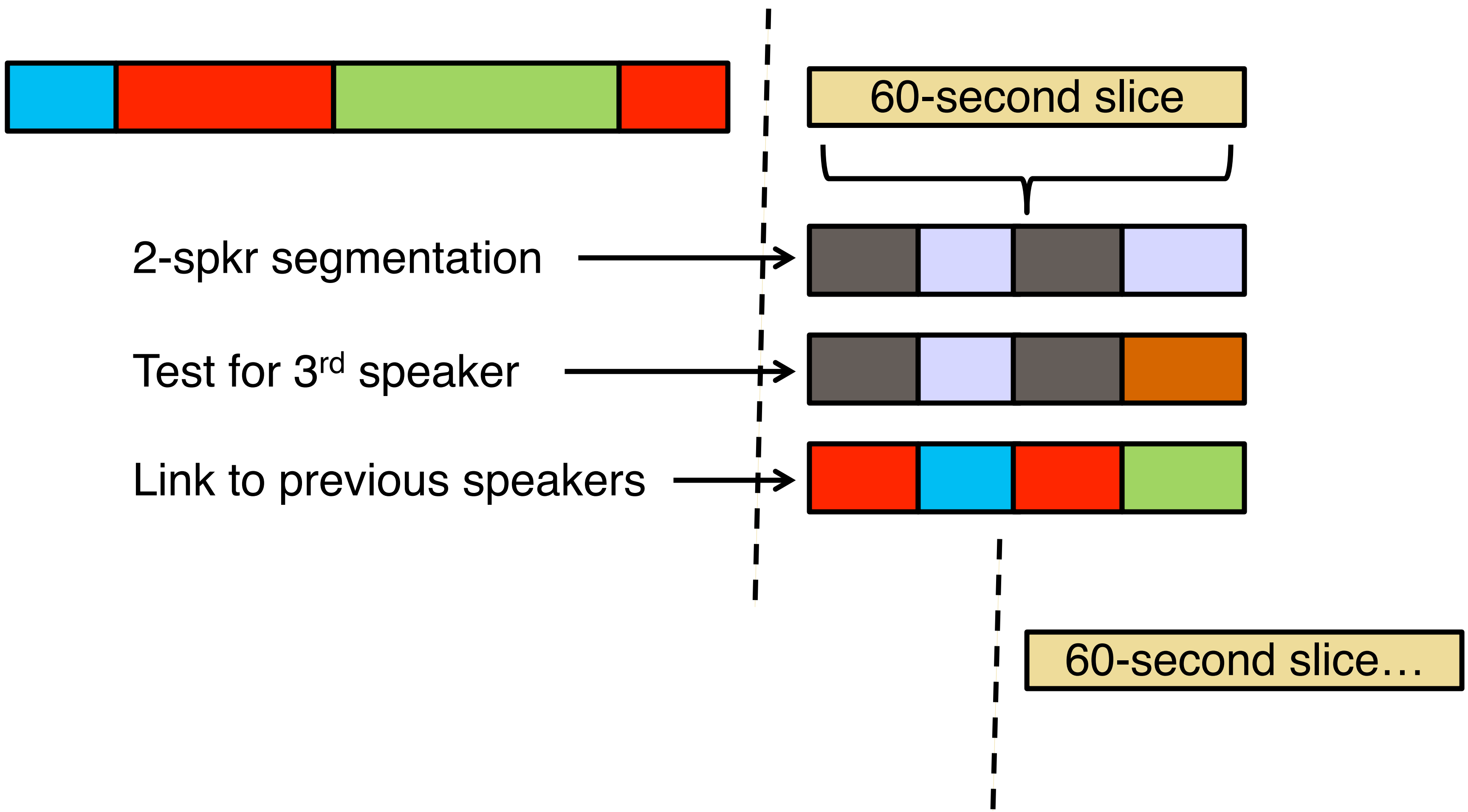
- **Conversely, underestimation**

Roadmap

- **Introduction**
  - Summary of Contributions
- **Background**
  - Diarization System Overview
  - Speaker Modeling with Factor Analysis
- **Our Incremental Approach**
  - K-means and Spectral Clustering (Interspeech 2011, 2012)
  - Towards Probabilistic Clustering Methods
  - Iterative System Optimization (Re-segmentation/Clustering)
  - Duration-Proportional Sampling
- **Analysis and Discussion**
  - Benchmark Comparison (Castaldo 2008)
- **Conclusion**

Explaining (Castaldo 2008)

- Causal system with fixed output delay
- Stream of factor analysis-based features (every 10ms)



Summary of Differences

- **Castaldo 2008**
  - Exploits structure of telephone conversations
    - **Assumes no more than 3 speakers exist in any 60-second slice**
  - Explicit use of speaker recognition system
    - **Links speakers from current slice to previous slices**

- **Our “bag of i-vectors”**
  - More general approach to clustering
    - **Can handle any number of speakers, regardless of temporal conversation dynamics**
    - **Prone to missing speakers that seldom participate**
    - **Prone to separate speakers that participate often**

Future Work

- **Dimensionality Reduction**
  - So far, only using first 3 principal components
  - t-SNE (Stochastic Neighbor Embedding)
    * **van der Maaten 2008**

- **Within-utterance Factor Analysis**
  - Is there some way to directly exploit variabilities within the acoustic features of a particular conversation?

- **Temporal Modeling and Bayesian Nonparametric Inference**
  - Hierarchical Dirichlet Process – Hidden Markov Model (HDP-HMM)
    * **Fox 2008, Johnson 2010**

Summary

- **Extended previous work in applying factor analysis-based speaker modeling to speaker diarization**
  - Castaldo 2008, Kenny 2010, Interspeech 2011/2012
- **Integrated variational inference into speaker clustering**
  - Valente 2005, Kenny 2010, SM Thesis 2011
- **Validated an iterative optimization procedure to refine clustering and segmentation hypotheses**
  - Interspeech 2012
- **Proposed a duration-proportional sampling scheme to combat issues of i-vector underrepresentation**
  - SM Thesis 2011

Thanks!

CSAIL

- **Questions?**
  - sshum @ csail.mit.edu