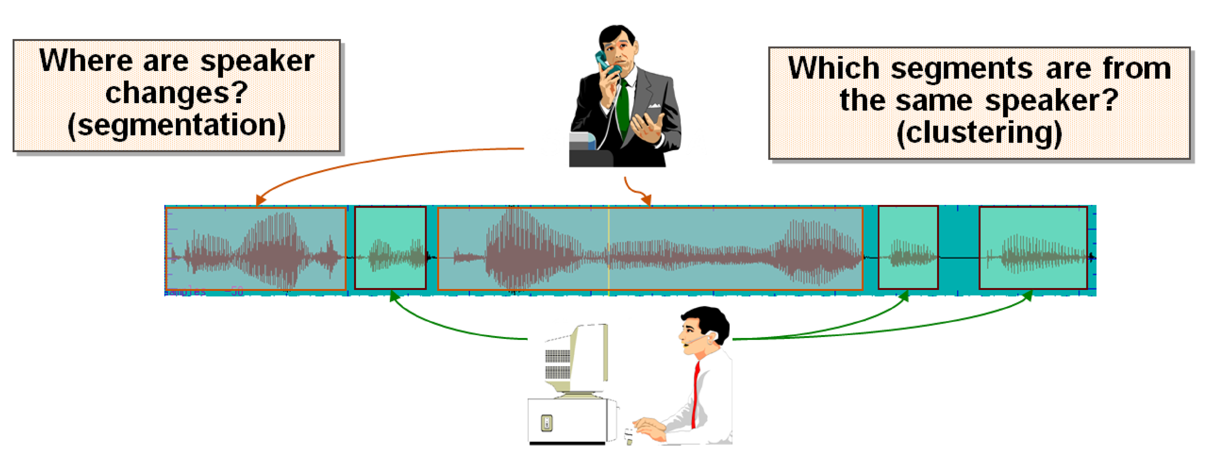
Unsupervised Methods for Speaker Diarization

Stephen Shum

MIT Masters of Science Thesis, Advisors: Jim Glass & Najim Dehak

Introduction

- Audio Diarization
  - ✓ The task of marking and categorizing the different audio sources within an unmarked audio sequence [1].
  - ✓ Originally built for desktop and tablet interfaces.
- Speaker Diarization
  - ✓ "Who is speaking when?"
  - ✓ Segmentation + Clustering
- 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

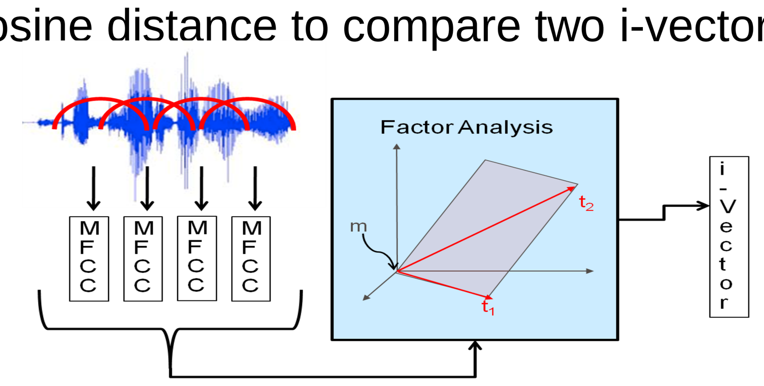


Speaker Representation

- From GMMs to Factor Analysis
  - ✓ Associated with each speaker is a distribution of acoustic features (AF) that can be modeled by a Gaussian Mixture Model (GMM).
  - ✓ A speaker **supervector** is created by concatenating all mixture mean components in a GMM.
    - ➢ 20 dim (AF) x 1024 mix (GMM) ≈ 20,000 dim
  - ✓ Assume all pertinent speaker variabilities lie in some low-dimensional subspace T of the supervector space
    - ➢ Rank(T) set between 100 and 600
- Total Variability Subspace [2]

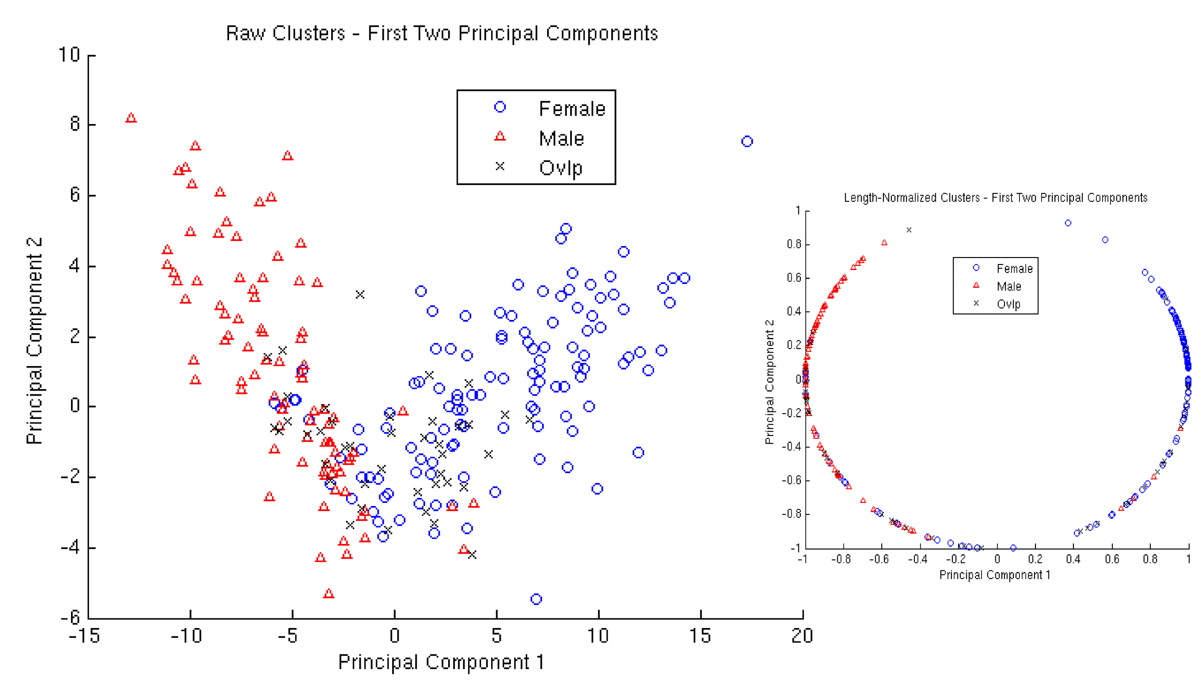
$$M = m + Tw$$

- ✓ w is vector of total factors (Identity Vector, i-vector)
- ✓ Use cosine distance to compare two i-vectors



Intra-Conversation Variability

- Use PCA to find prominent directions of intra-conversation variability

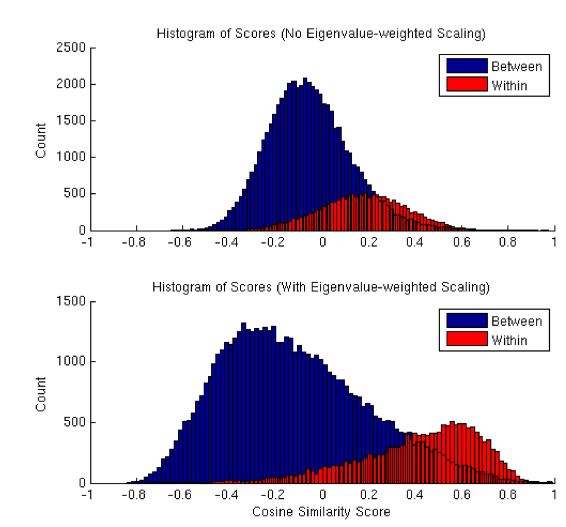


- Further emphasize principal directions
  - ✓ i.e. the most principal components have largest eigenvalues

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

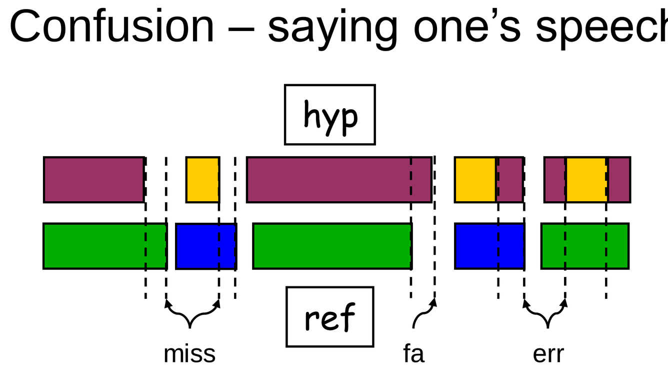
w_i' : PCA-projected i-vector

Λ : Diagonal matrix of eigenvalues

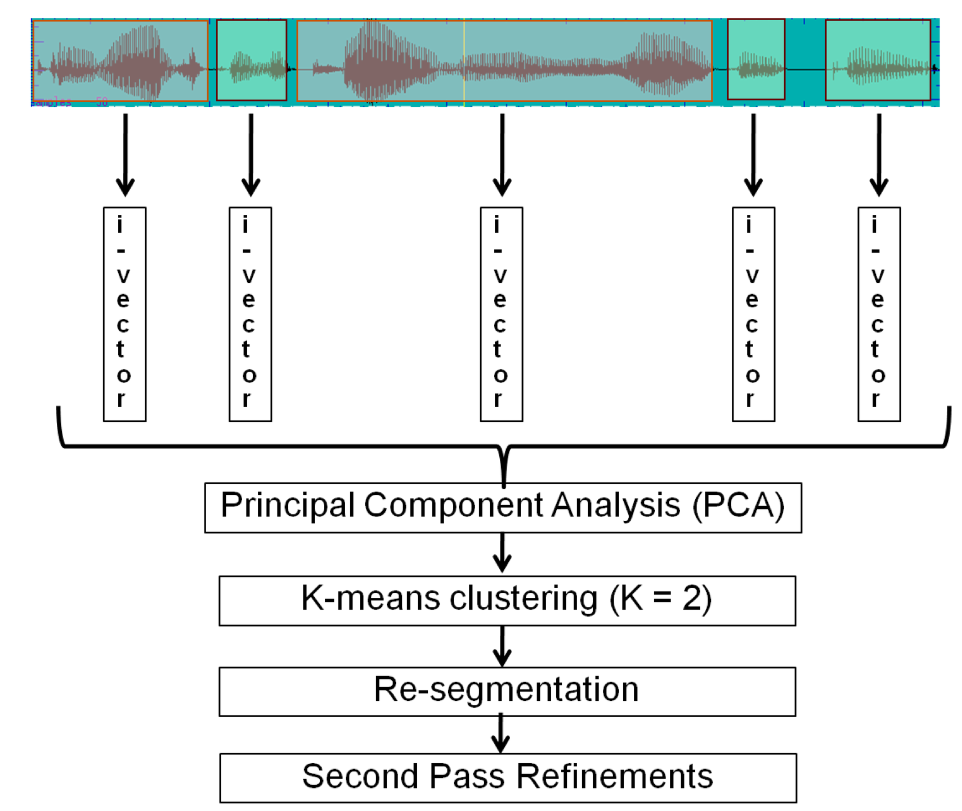


Experimental Framework

- Summed-channel Telephone Speech
  - ✓ 2008 NIST Speaker Recognition Evaluation Test Data
  - ✓ 2215 two-speaker conversations (~5min each)
  - ✓ Obtain a reference diarization by applying Automatic Speech Recognition (ASR) or Voice Activity Detection (VAD) on each channel separately.
  - ✓ Scoring ignores overlapped speech.
- Diarization Error Rate (DER)
  - ✓ Miss – speaker in reference but not in hypothesis
  - ✓ False Alarm – speaker in hyp but not in ref
  - ✓ Speaker Confusion – saying one's speech is another's



System Summary



- First Pass K-means (K=2) Clustering
- Viterbi Re-segmentation
  - ✓ Apply the Viterbi algorithm at the acoustic feature level to re-formulate segment boundaries and re-assign frames to each cluster (Speaker A, Speaker B, Non-speech N).
- Second Pass Refinements
  - ✓ Extract a single i-vector for each respective speaker based on the re-segmentation assignments.
  - ✓ Re-assign each segment i-vector to the speaker whose i-vector is closer in cosine distance
  - ✓ Essentially another pass of K-means, where the "means" are computed via i-vector estimation.

Experiment Results

- Using our own segmentation

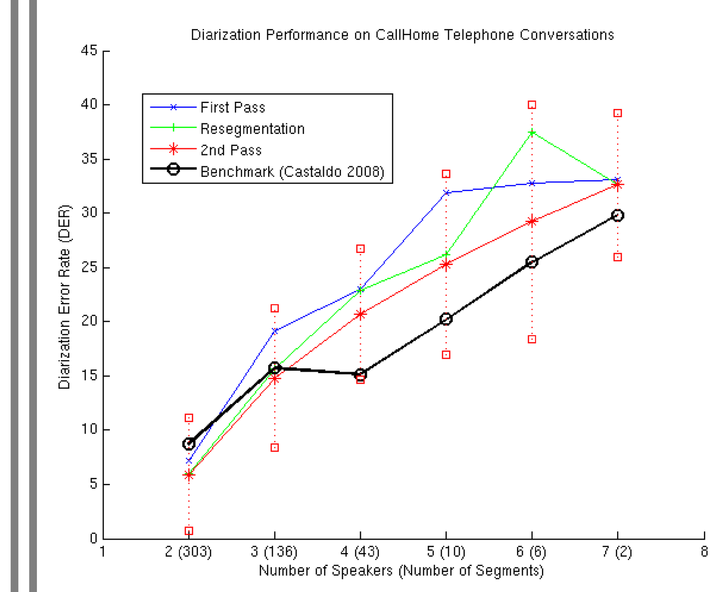
	Error Breakdown				
	Miss	False Alarm	Confusion	**DER (%)**	**σ (%)**
First Pass	7.7	2.0	2.8	**12.5**	**8.2**
Re-segmentation	0.3	2.3	2.6	**5.2**	**8.2**
Second Pass	0.3	2.3	1.1	**3.7**	**6.4**

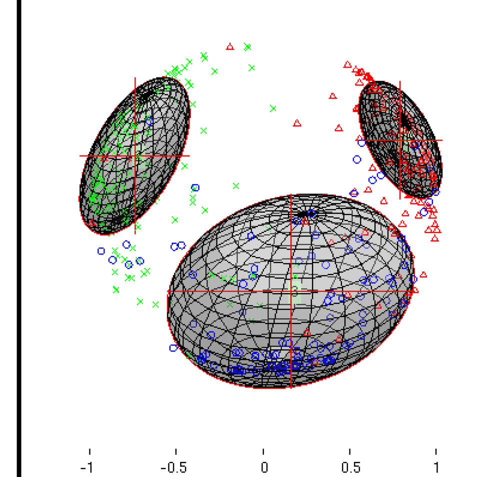
- Using reference segmentation
  - ✓ Removes all errors attributed to Miss & False Alarm
  - ✓ Can then focus solely on Speaker Confusion error
  - ✓ Allows for direct comparison with other systems [3].

	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

Towards K-speaker Diarization

- Speech with more than two speakers
  - ✓ How well does our system generalize when given the number of speakers K?
  - ✓ Evaluate on 500 CallHome telephone conversations
    - ➢ Each call contains 2-7 speakers, length of 1-5min





- Estimating the Number of Speakers (K)
  - ✓ Exploring Variational Bayesian GMMs
  - ✓ But because data lies on the unit hypersphere, ought to consider a more suitable distribution.
    - ➢ E.g. Mixtures of von Mises-Fisher Distributions [4]

Future Work

- Processing of overlapped speech
  - ✓ Segments containing overlapped speech potentially corrupt our PCA and subsequent speaker modeling.
- Clustering on sparse data
  - ✓ Some speakers speak relatively little in a conversation
  - ✓ Want to be able to handle these sorts of imbalance.
- Temporal modeling of conversation

References

[1] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarisation Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, September 2006.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2010.

[3] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of Telephone Conversations Using Factor Analysis," *IEEE Journal of Selected Topics in Signal Processing*, December 2010.

[4] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions," *Journal of Machine Learning Research*, September 2005.