Unsupervised Speaker Diarization Using Riemannian Manifold Clustering

Che-Wei Huang, Bo Xiao, Panayiotis Georgiou, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL), Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089
{cheweihu,boxiao}@usc.edu, {georgiou, shri}@sipi.usc.edu

Motivations & Objectives

- To perfectly cluster short segments
  - Likelihood-based clustering not good on short segments
- Geometric point of view
  - Non-convexity of speaker clusters
  - Faithful geodesic metric
- Speaker clustering by Riemannian manifold clustering
- Suppress sparsity issue in local samples
- Stabilize performance over parameter tuning
- **Hypothesis**: Speech segments from different speakers form distinct manifolds

Riemannian Locally Linear Embedding

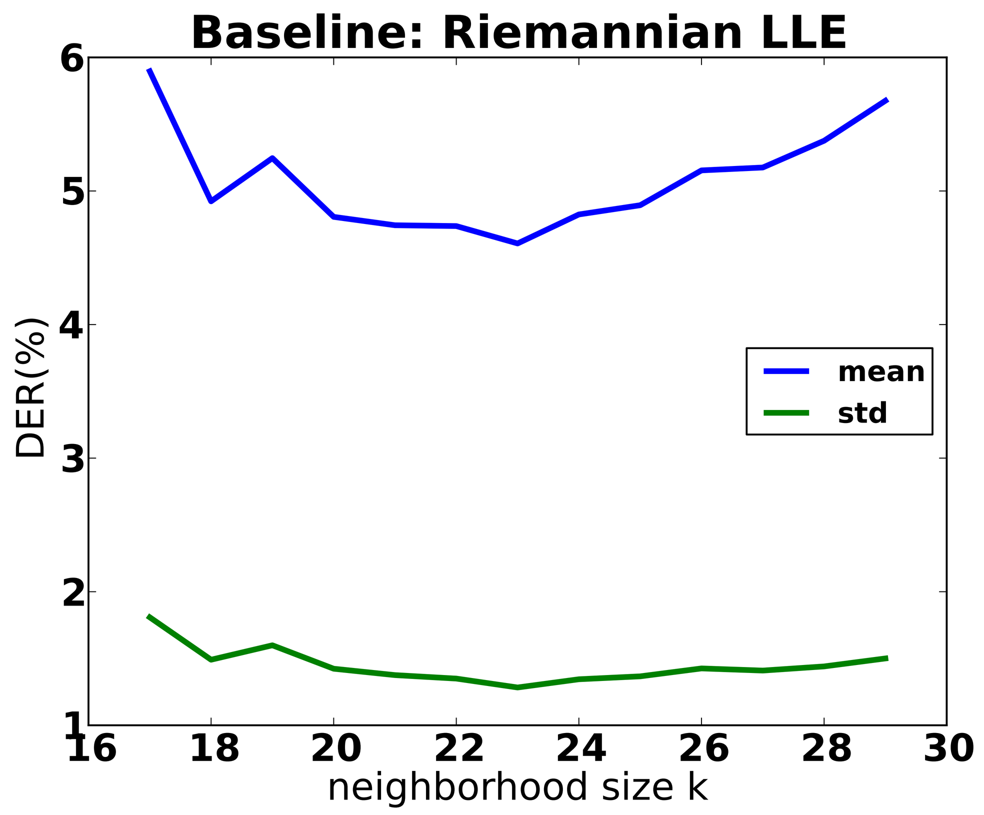
- Generalization of spectral clustering
- Built-in geodesic metric
  - Data samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$
  - Known number of clusters m
  - Riemannian geodesic metric $\|\cdot\|_{\mathbf{x}_i}$
  - $N(i)$ index set of $\mathbf{k}$-NN of $\mathbf{x}_i$
  - $\mathbf{c}_i(\mathbf{w}_i) = \|\sum_{j=1}^{n} w_{ij}\mathbf{x}_j - \mathbf{x}_i\|^2_{\mathbf{x}_i}$
  - $\text{argmin}_{\mathbf{w}_i}\, \mathbf{c}_i(\mathbf{w}_i)$ s.t. $\sum_{j=1}^{n} w_{ij} = 1$ and $w_{ij} = 0$ if $j \notin \text{N}(i)$
  - Similarity matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_n]^T$
  - Graph Laplacian $\mathcal{L} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$
  - 2nd to $(m+1)$th smallest eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_m$ of $\mathcal{L}$
  - Embedded coord.s $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_m]^T$
  - Kmeans clustering for rows of $\mathbf{X}$ with m centroids
  - Label of ith row = label of $\mathbf{x}_i$
- Issues:
  - Unknown choice of $\mathbf{k} \sim f(\mathit{intri.dim})$
  - GMM vs Single multivariate Gaussian

Dataset & Experiment

- Microphone interview from NIST 2010 SRE
- 2477 5-min sections $\sim$ 206 hours
- Oracle segmentation
- Known # of speakers
- 20 MFCC w/ frame size 40ms & frame step 20ms, w/o normalization
- Overlapped speeches clustered, not evaluated
- Segments as single multivariate Gaussians

Baseline & Proposal

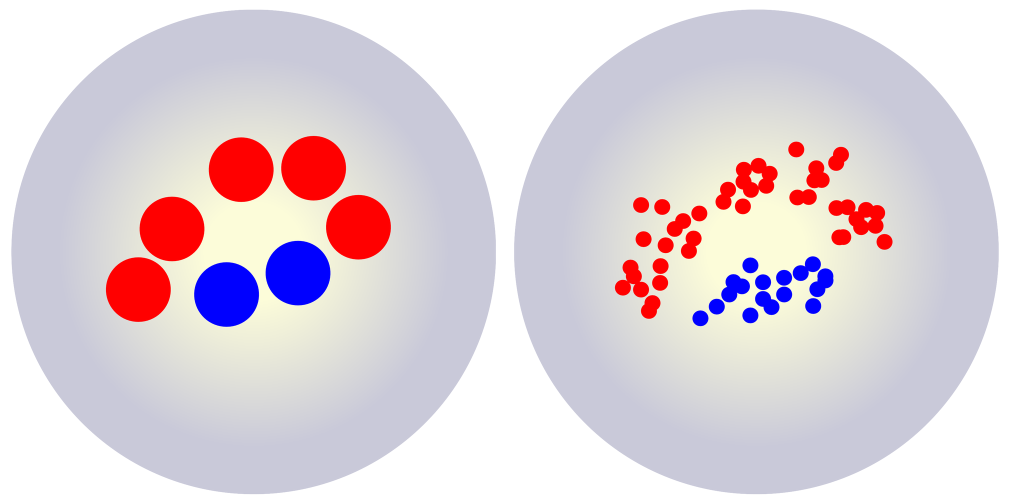
- Riemannian LLE as the baseline



- Best at $k = 23$ w/ DER$= 4.607\%$
- Difficulties:
  - High intrinsic dimensionality $\rightarrow$ high $\mathbf{k}$
  - Sparse local samples
  - Need GMM for long segments
  - Narrow range of optimal $\mathbf{k}$
- **Proposal:**
  Impose length constraint on segments
- Advantages:
  Higher local density $\rightarrow$ safer for high $\mathbf{k}$
- Disadvantage:
  Higher computational loading

Baseline & Proposal

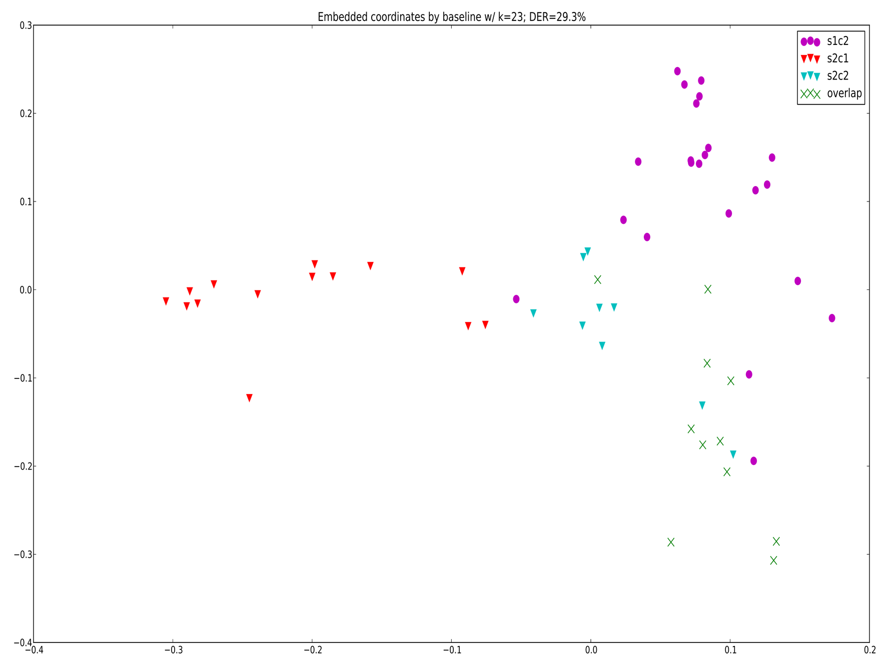
- Manifold of Gaussian pdfs = sphere in Hilbert space
- **Schematic diagrams before and after length constraint**

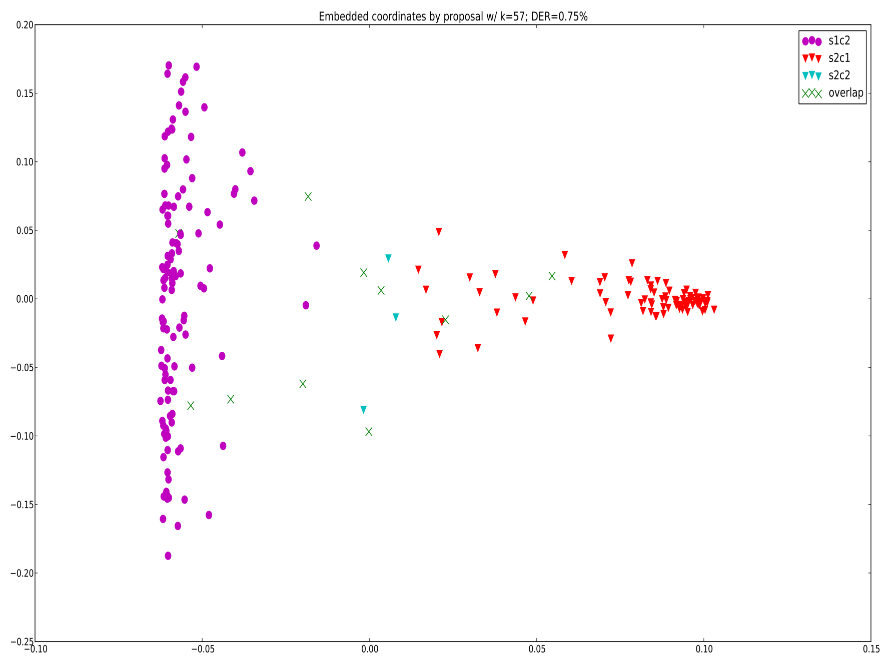


- Data's **self-expressiveness**

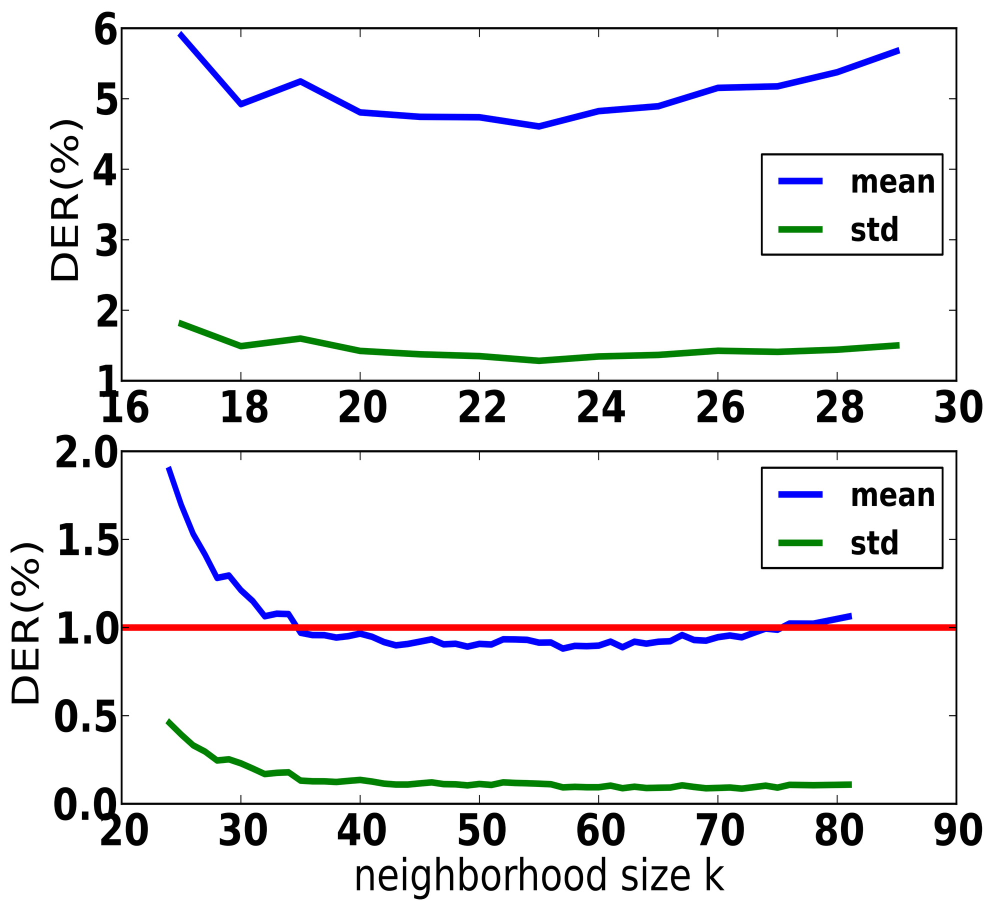
Comparisons

- Embedded coordinates of samples in the same section
- Detectable clusters w/ dense samples





Comparisons



- $1s$ length constraint
- Best at $k = 57$ w/ DER$= 0.88\%$
- Stable under 1% for wide range of k

Conclusions

- Effective Riemannian manifold modeling
- Performance less sensitive to the parameter
- Potentially higher computational complexity

Future work

- Performance with imperfect VAD
- Performance with imperfect segmentation
- Number of clusters estimation
- Optimal length constraint
- Automate choice of $\mathbf{k}$
- Deal with overlapped speeches
- Originally mono channel data

Acknowledgement

This research is supported by NSF, NIH and DARPA.