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

Dataset & Experiment
- Microphone interview from NIST 2010 SRE
- 2477 5-min sections ~ 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
  - Generalization of spectral clustering
  - Built-in geodesic metric
    - Data samples $x_1, \ldots, x_n$
    - Known number of clusters $m$
    - Riemannian geodesic metric $\| \cdot \|_{x_i}$
    - $N(i)$ index set of $k$-NN of $x_i$
    - $c_i(w_i) = \| \sum_{j=1}^{w_i} w_{ij} x_j - x_i \|_{x_i}^2$
    - $\arg\min_{w_i} c_i(w_i)$ subject to $\sum_{j=1}^{w_i} w_{ij} = 1$ and $w_{ij} = 0$ if $j \not\in N(i)$
    - Similarity matrix $W = [w_1, \ldots, w_n]^T$
    - Graph Laplacian $L = (I - W)/(I - W)$
    - 2nd to $(m + 1)$th smallest eigenvectors $x_1, \ldots, x_m$ of $L$
    - Embedded coord.s $X = [x_1, \ldots, x_m]^T$
    - Kmeans clustering for rows of $X$ with $m$ centroids
    - Label of $i$th row = label of $x_i$
  - Issues:
    - Unknown choice of $k \sim f(\text{intri.dim})$
    - GMM vs Single multivariate Gaussian

Baseline & Proposal
- Manifold of Gaussian pdfs = sphere in Hilbert space
- Schematic diagrams before and after length constraint
  - Data’s self-expressiveness

Baseline: Riemannian LLE
- Best at $k = 23$ w/ DER = 4.607%
- Difficulties:
  - High intrinsic dimensionality $\rightarrow$ high $k$
  - Sparse local samples
  - Need GMM for long segments
  - Narrow range of optimal $k$

Proposal:
- Impose length constraint on segments
- Advantages:
  - Higher local density $\rightarrow$ safer for high $k$
- Disadvantage:
  - Potentially higher computational complexity

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 $k$
- Deal with overlapped speeches
- Originally mono channel data

Acknowledgement
This research is supported by NSF, NIH and DARPA.