Speaker Diarization

CS 655: Analyzing Sequences
CSLU, OHSU
11/10/2014

Intro.

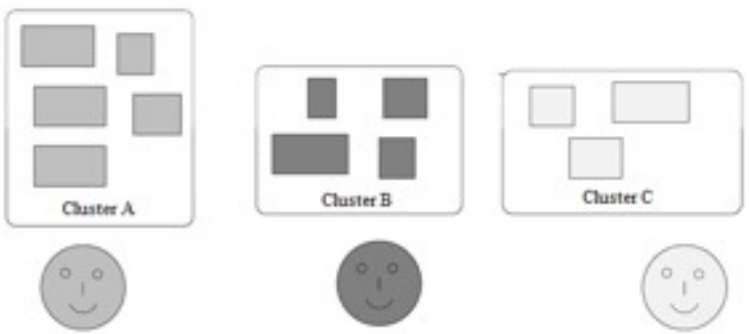
- What is Speaker Diarization?
  - Homogeneous segments
  - Speaker identity

- The output can be fed to other speech processing tasks.
-

A speaker diarization system

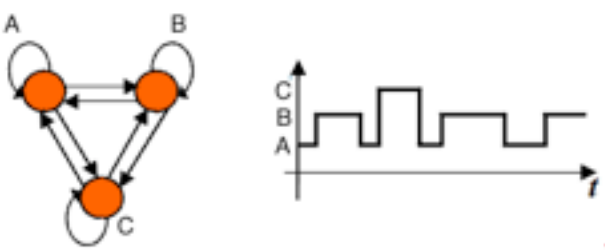
Speaker Segmentation

Speaker Clustering



HMM training

HMM decoding

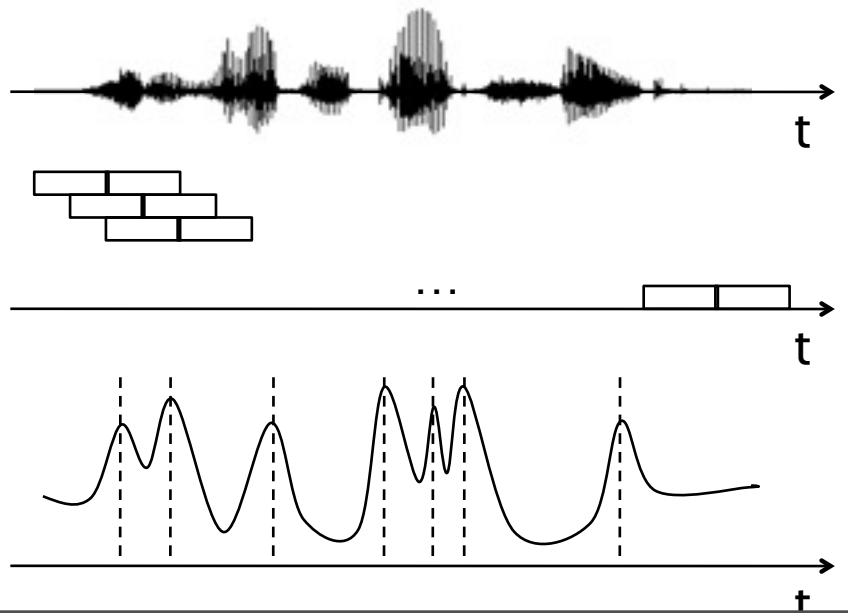


P. Delacourt, “SEGMENTATION ET INDEXATION PAR LOCUTEURS D’UN DOCUMENT AUDIO”, 1998.

Segmentation

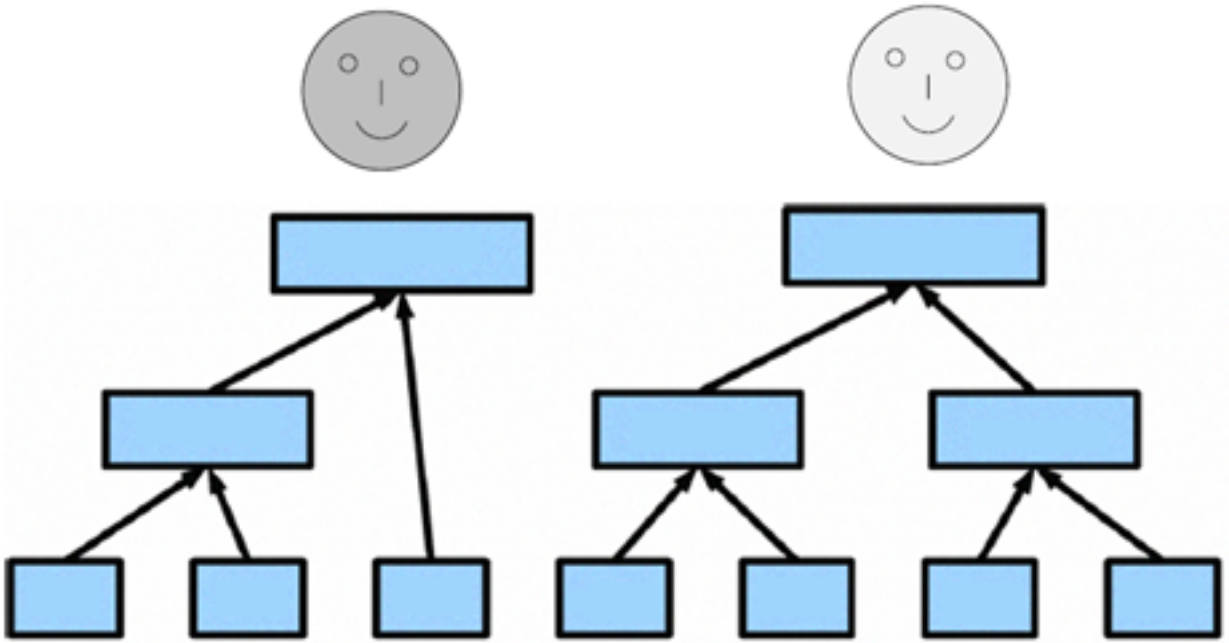
- Speaker Segmentation is the first essential part in speaker diarization systems.
- Speaker Segmentation: Tries to extract the longest possible homogenous segments in a conversation.

- Popular segmentation approach:
  I. Slide window over feature sequence,
  II. compute distance between two parts of window,
  III. Peaks in the curve are changing points



S.H. Mohammadi, Speaker diarization in adverse conditions, 2011.

Bottom-up clustering



S.H. Mohammadi, Speaker diarization in adverse conditions, 2011.

- Important question:
  - How do we measure the similarity (or dissimilarity) between speaker segments?
  - We need some distance measure for speaker segmentation and clustering
- Common distances used:
  - KL (Kullback-Leibler)
  - GLR (Generalized Likelihood Ratio)
  - BIC (Bayesian Information Criterion)

Kullback-Leibler divergence

- Given two probability distributions, X and Y, computes how far apart they are
-

$$KL(X;Y) = E_X(log\frac{P_X}{P_Y})$$

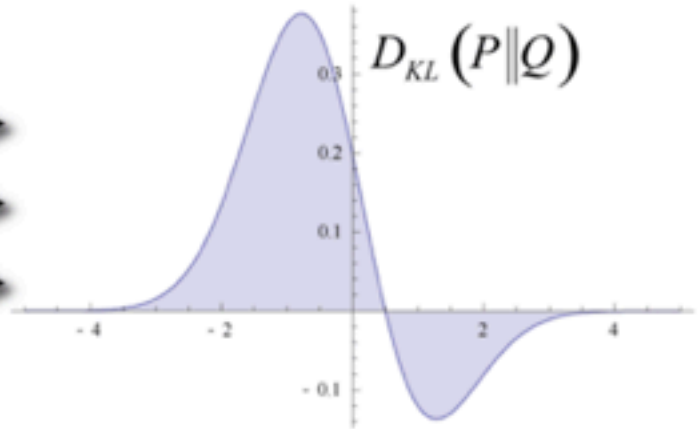
Single Gaussian Assumption

$$KL(X,Y) = \frac{1}{2}tr[(C_X - C_Y)(C_Y^{-1} - C_X^{-1})] + \frac{1}{2}tr[(C_Y^{-1} - C_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T]$$

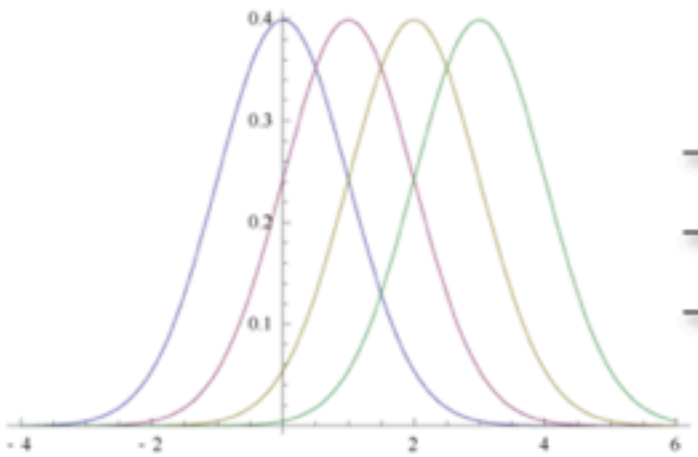
$$KL2(X;Y) = KL(X;Y) + KL(Y;X)$$

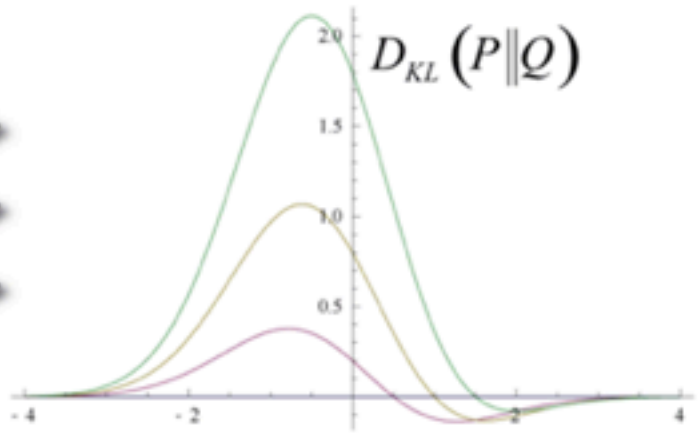
KL divergence

Original Gaussian PDF's



KL Area to be Integrated





http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

Generalized Likelihood Ratio

H0: two segments are from one source (speaker)

$$\mathcal{X} = \mathcal{X}_i \bigcup \mathcal{X}_j \sim M(\mu, \sigma)$$

H1: two segments are from two different sources (speakers)

$$\mathcal{X}_i \sim M_i(\mu_i, \sigma_i)$$

$$\mathcal{X}_j \sim M_j(\mu_j, \sigma_j)$$

$$GLR(i,j) = \frac{H_0}{H_1} = \frac{\mathcal{L}(\mathcal{X}, M(\mu,\sigma))}{\mathcal{L}(\mathcal{X}_i, M_i(\mu_i,\sigma_i))\mathcal{L}(\mathcal{X}_j, M_j(\mu_j,\sigma_j))}$$

$$D(i,j) = -log(GLR(i,j))$$

Single Gaussian Assumption

$$= \frac{N}{2} log|\Sigma_{\mathcal{X}}| - \frac{N_i}{2} log|\Sigma_{\mathcal{X}_i}| - \frac{N_j}{2} log|\Sigma_{\mathcal{X}_j}|$$

Multiple Gaussian (GMM) Assumption = More complex formula

http://www.xavieranguera.com/phdthesis/node12.html

Bayesian Information Criterion

BIC is Likelihood criterion penalized by model complexity
Its Goal: Model selection, or "which model better fits the data"?
for speech: Is modeling each segment using one model "M" is better or modeling them using two separate models "Mi and Mj"?

$$BIC(\mathcal{M}_i) = \log \mathcal{L}(\mathcal{X}_i, \mathcal{M}_i) - \lambda \frac{1}{2} \#(\mathcal{M}_i) log(N_i)$$

$$\Delta BIC(\mathcal{M}_i) = \log \mathcal{L}(\mathcal{X}, \mathcal{M}) - (\log \mathcal{L}(\mathcal{X}_i, \mathcal{M}_i) + \log \mathcal{L}(\mathcal{X}_j, \mathcal{M}_j)) - \lambda \Delta \#(i,j) log(N)$$

BIC

- If delta BIC is positive, that is considered a speaker changing point

Single Gaussian Assumption

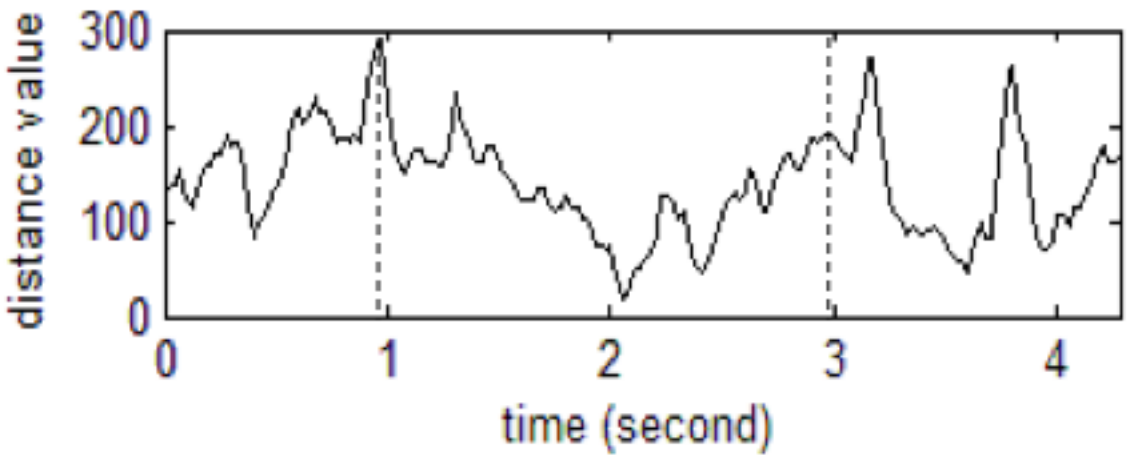
$$\Delta BIC = \frac{N_Z}{2}\log|\Sigma_Z| - \frac{N_X}{2}\log|\Sigma_X| - \frac{N_Y}{2}\log|\Sigma_Y| - \frac{\lambda}{2}\left(d + \frac{1}{2}d(d+1)\right)$$

GLR

Penalty Term

Sample Distance curve

- Sample distance curve for speaker segmentation
- Peaks are speaker changing point, the dashed lines are gold-standard



S.H. Mohammadi, et al, KNNDIST: A Nonparametric distance measure for speaker segmentation, 2012.