KNNDIST: A Non-Parametric Distance Measure for Speaker Segmentation

Seyed Hamidreza Mohammadi[1], *Hossein Sameti*[2], *Mahsa Sadat Elyasi Langarani*[2], *Amirhossein Tavanaei*[2]

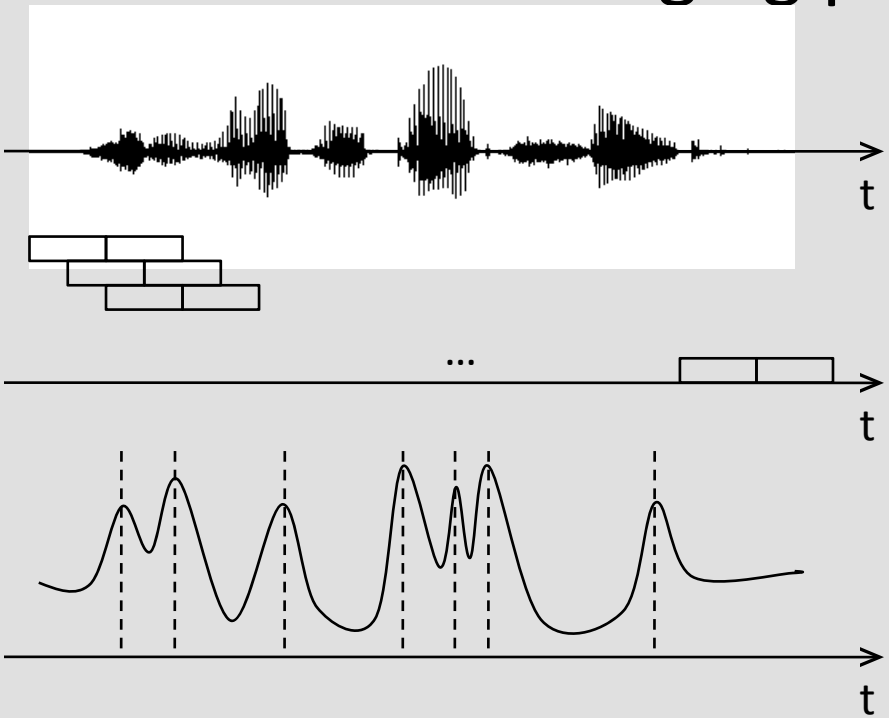
[1] Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR

[2] Speech Processing Laboratory, Sharif University of Technology, Tehran, Iran

Introduction

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

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



- In the preprocessing phase, **silence** is usually removed completely.
- In this study, it is proposed that **non-vowel** should also be removed.
- The motivation was to increase the accuracy of parameter estimation in distance computation.

- Common distances used:
  - GLR (Generalized Likelihood Ratio)
  - KL (Kullback-Leibler)
  - BIC (Bayesian Information Criterion)
- Advantage:
  - Parametric (Gaussian assumption) so easy to compute
- Disdvantages:
  - Parametric so not so accurate parameter estimation when data window is small

KNNDIST

- In parametric distances, the distance is computed between probabilistic distributions.
- To overcome parametric distances, a non-parametric method is used.
- The idea is taken from k-nearest-neighbor classification.
- The distance is computed from the mean distances of the k-nearest-neighbor.

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

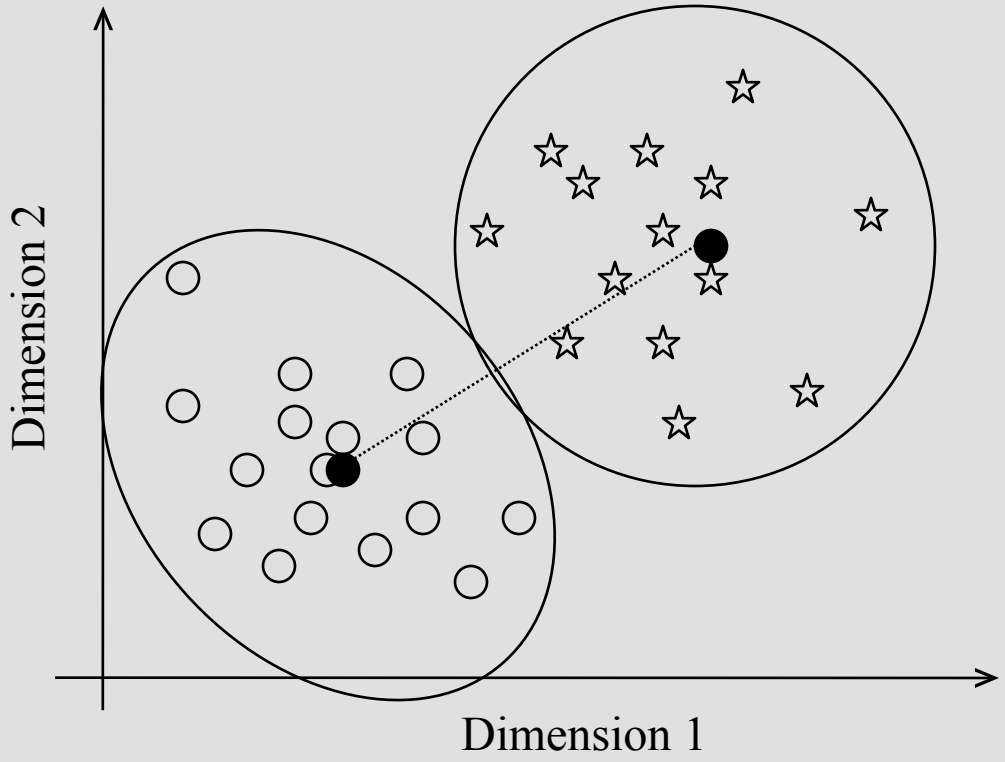


Figure - Parametric distances

$$KNNDIST(X,Y,k) =$$
$$\sum \min_k \left(euclid(x_m, y_n) \right), m \in M, n \in N$$

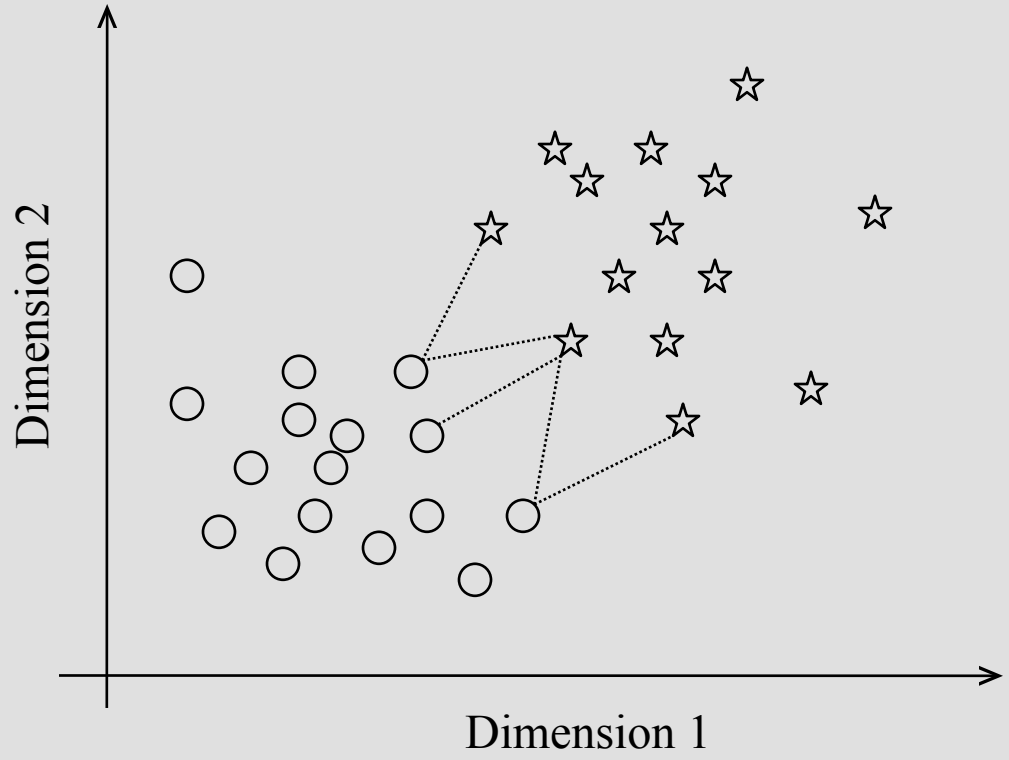


Figure - KNNDIST

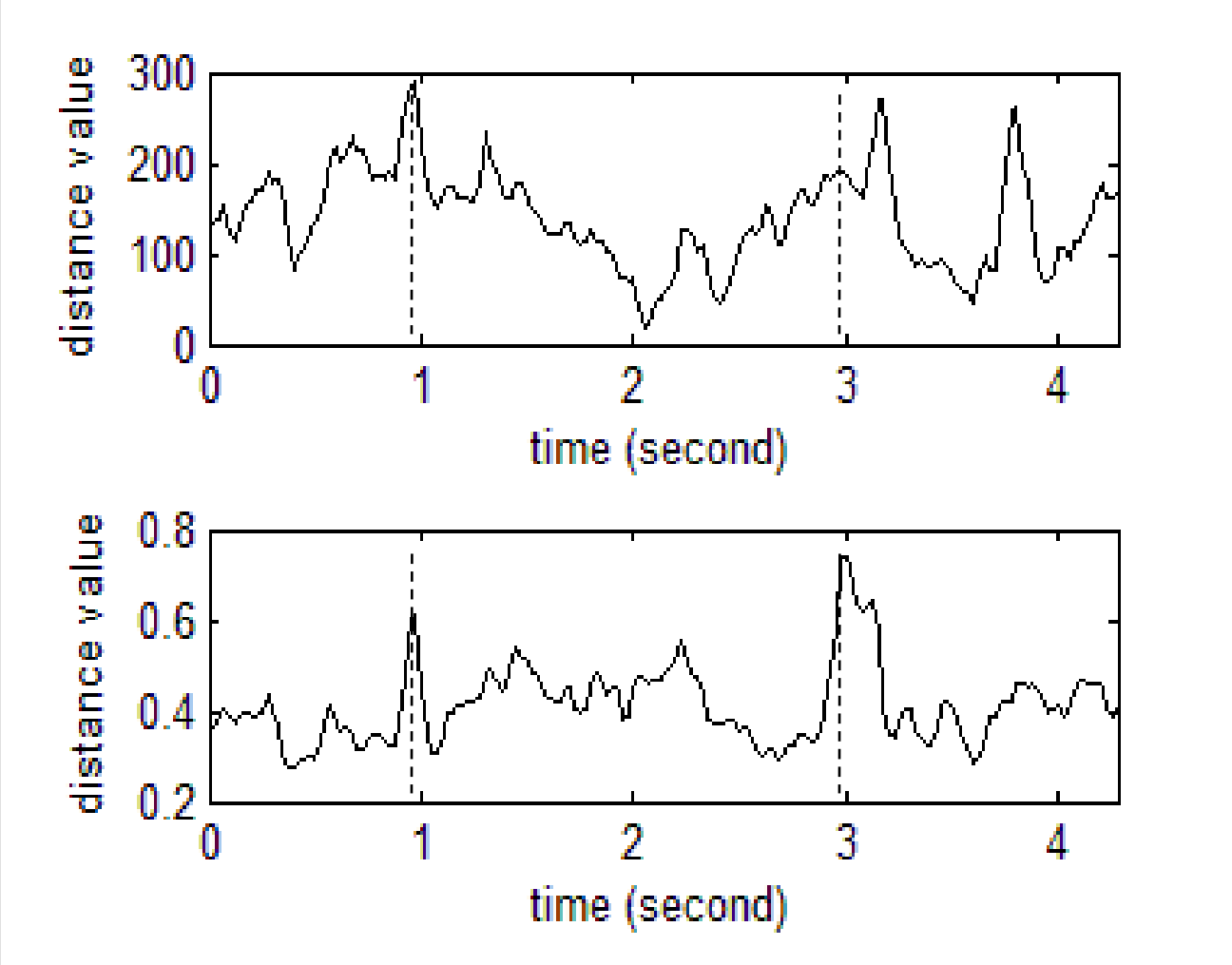


Figure – computed curves GLR (top) KNNDIST (bottom)

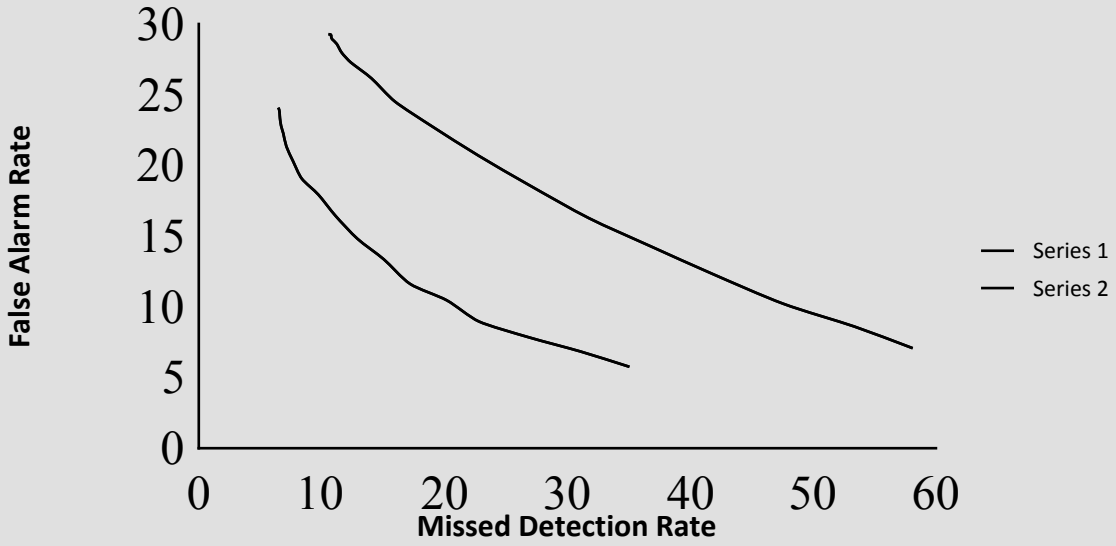


Figure – DET curve for BIC vs. KNNDIST

Table – Results (False Alarm Rate and Missed Detection Rate)

Window Length (seconds)	**Database**	**Results**			
		BIC		**KNNDIST**	
		MDR	*FAR*	*MDR*	*FAR*
0.75	TIMIT	23.35%	37.67%	8.15%	37.88%
1	TIMIT	14.78%	33.05%	5.80%	34.27%
1.5	TIMIT	10.73%	29.25%	6.53%	25.29%
2	TIMIT	8.15%	20.80%	7.24%	17.21%
3	TIMIT	13.01%	14.42%	14.14%	17.56%
2	AMI IS1008a	26.13%	48.84%	26.19%	37.33%
2	AMI ES2008b	29.72%	53.98%	28.83%	39.69%

Conclusion

- GOOD for applications with small window length (like speaker segmentation).
- NOT GOOD for applications that use bigger window length (like speaker clustering) because of higher computational cost and lower accuracy.
- TODO:
  - determine best “k” automatically
  - Use another distance instead of Euclidian to find *k* nearest neighbors