Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization

Héctor Delgado[1], Xavier Anguera[2],
Corinne Fredouille[3], Javier Serrano[1]

[1]CAIAC, Universitat Autònoma de Barcelona, Barcelona, Spain
[2]Sinkronigo S.L., Barcelona, Spain
[3]CERI/LIA, University of Avignon, France

September 10, 2015

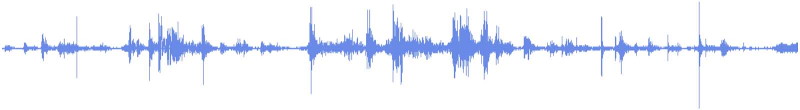
INTERSPEECH 2015
September 6 - 10
Dresden, Germany

Outline

Outline

Speaker diarization

Who spoke when?

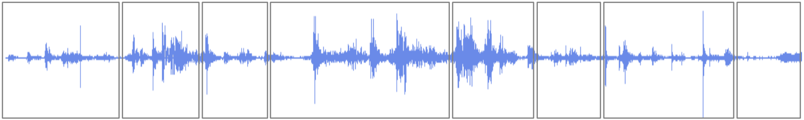


No prior information about

- Number of speakers
- Speaker identities

Speaker diarization

Who spoke when?

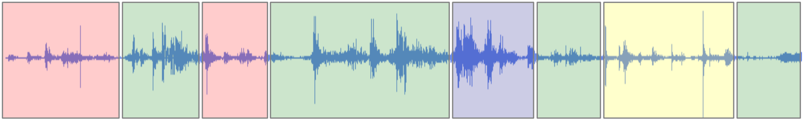


No prior information about

- Number of speakers
- Speaker identities

Speaker diarization

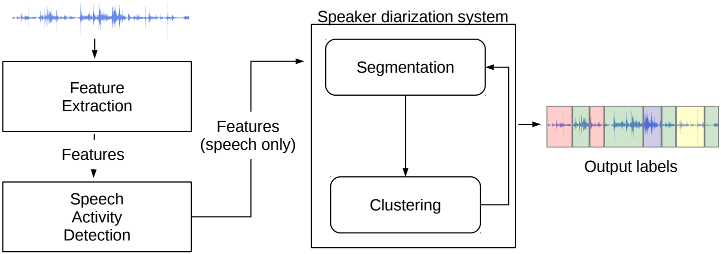
Who spoke when?



No prior information about

- Number of speakers
- Speaker identities

Generic speaker diarization scheme



- Iterative scheme
- Intensive computation: Speaker models re-training or adaptation, Viterbi re-assignment...
- Long processing times

Binary Key speaker diarization evolution

- A fast diarization system based on binary keys speaker modeling was presented[1]
- Based on a speaker modeling based on binary keys[2]
- Promising results and important speed gain on the NIST meeting audio databases
  - Average DER: 25.06 %
  - Speed $\simeq$ 0.1 xRT

[1] Anguera, X.; Bonastre, J.-F. "Fast speaker diarization based on binary keys," in Proc. Acoustics, Speech and Signal Processing (ICASSP), 2011

[2] Anguera, X.; Bonastre, J.-F. "A novel speaker binary key derived from anchor models," in Proc. Interspeech, 2010.

Binary Key speaker diarization evolution (2)

- Applied to broadcast TV data (REPERE dataset) with similar results ($\simeq$ 23% DER)[3]
- Further improved by using Cumulative Vectors and Cosine Distance: ($\simeq$ 19% DER)[4]

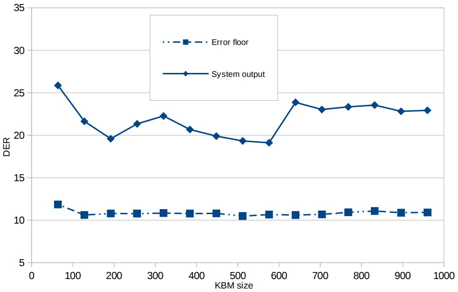
[3] Delgado, H.; Fredouille, C., Serrano, J. "Towards a complete binary key system for the speaker diarization task," in Proc. Interspeech, 2014.

[4] Delgado, H.; Anguera, X., Fredouille, C., Serrano, J. "Improved binary key speaker diarization system," in Proc. EUSIPCO, 2015.

Motivation

Main goal: **Fast and accurate speaker diarization system which does not require external training data**

But...



- Clustering selection still does not return near-optimum solution
  - A suitable final clustering selection criterion is required
- Can execution time be further reduced? (< 0.1xRT)

Delgado, H., Fredouille, C., Serrano, J. "Towards a Complete Binary Key System for the Speaker Diarization task," in Proc. Interspeech, 2014.

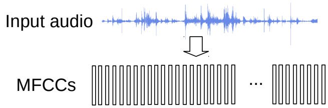
Outline

1. Introduction
2. Binary Key Speaker Diarization
3. Speeding-up binary key speaker diarization
4. Proposed final clustering selection criterion for binary key speaker diarization
5. Experiments and results
6. Conclusions

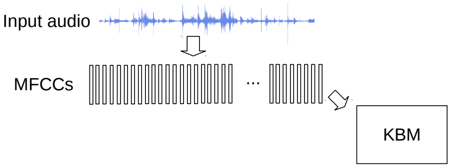
Binary Key Speaker Diarization System

Input audio 

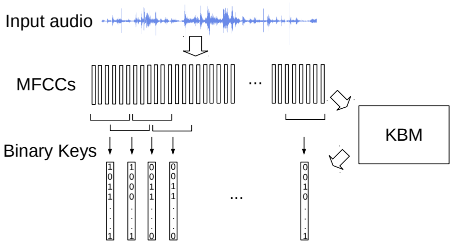
Binary Key Speaker Diarization System



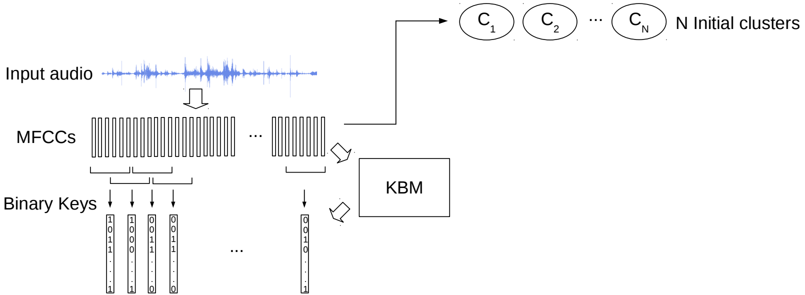
Binary Key Speaker Diarization System



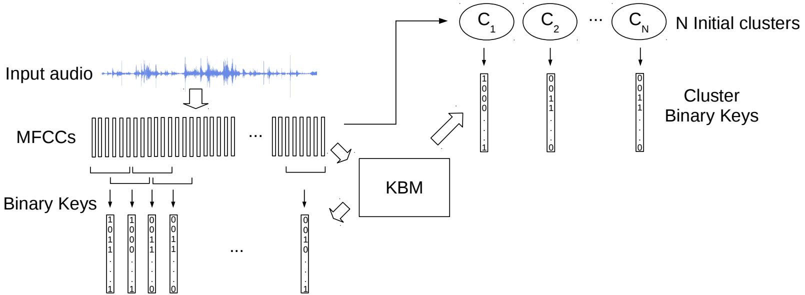
Binary Key Speaker Diarization System



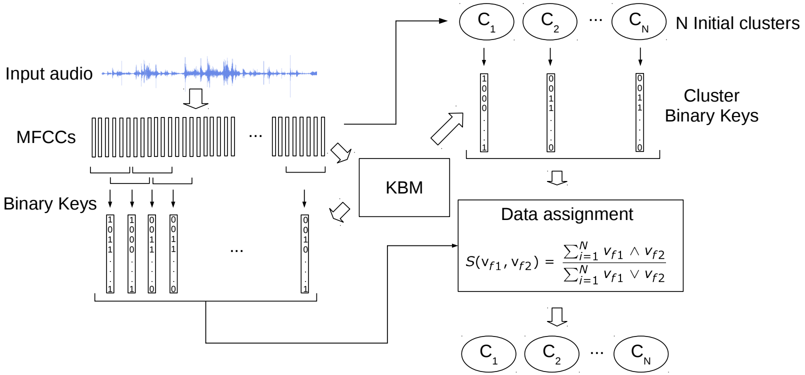
Binary Key Speaker Diarization System



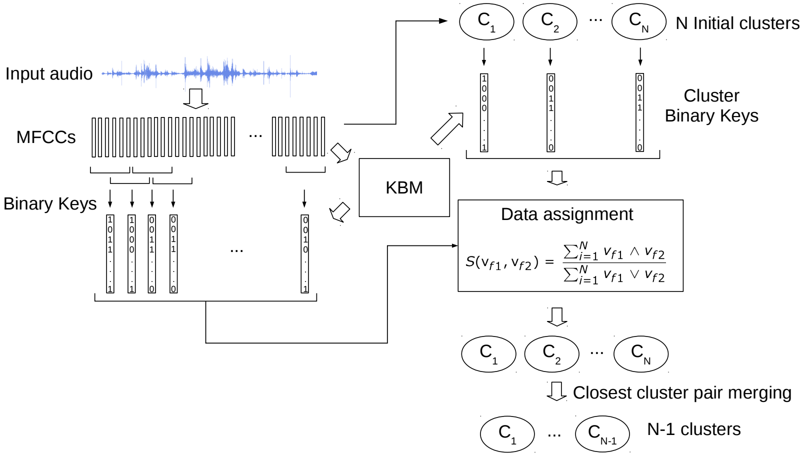
Binary Key Speaker Diarization System



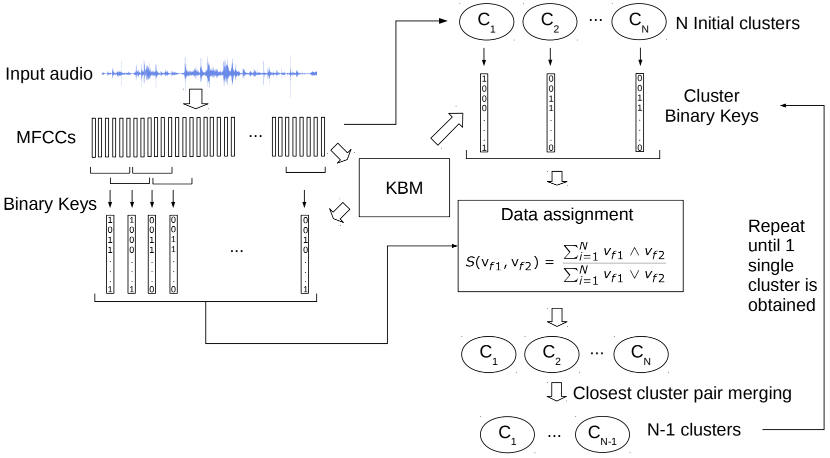
Binary Key Speaker Diarization System



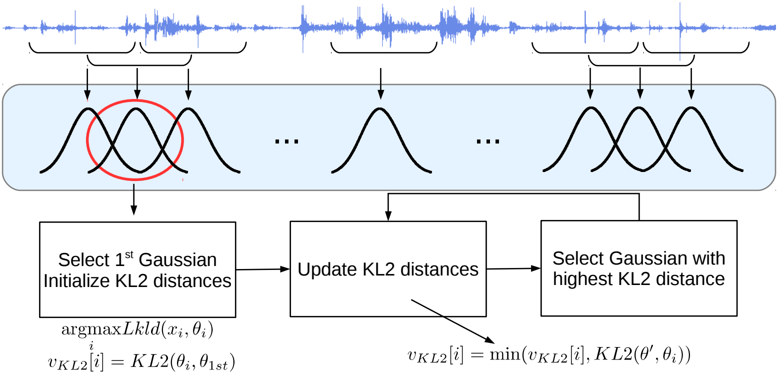
Binary Key Speaker Diarization System



Binary Key Speaker Diarization System

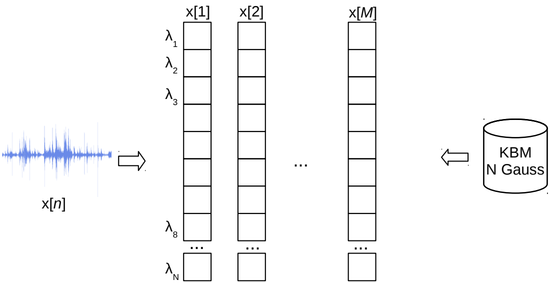


KBM training

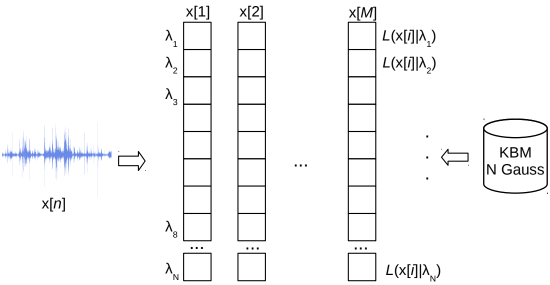


Binary key computation

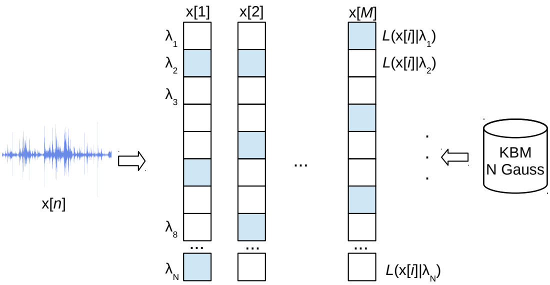
Binary key computation



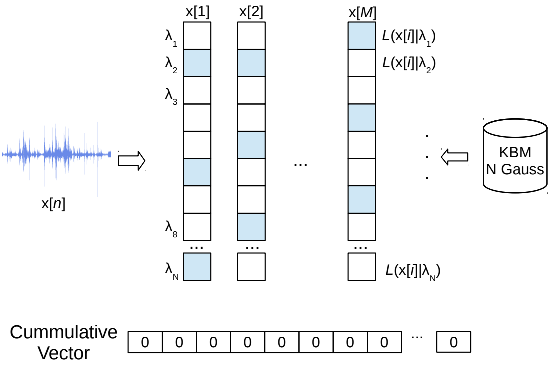
Binary key computation



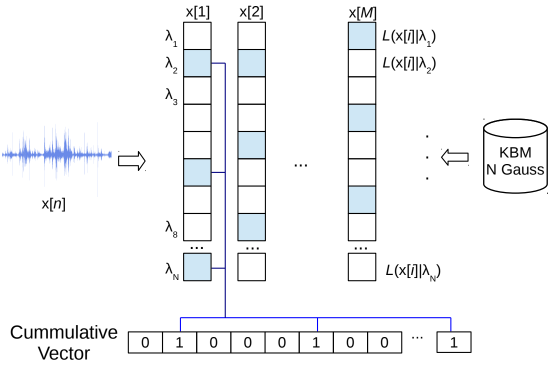
Binary key computation



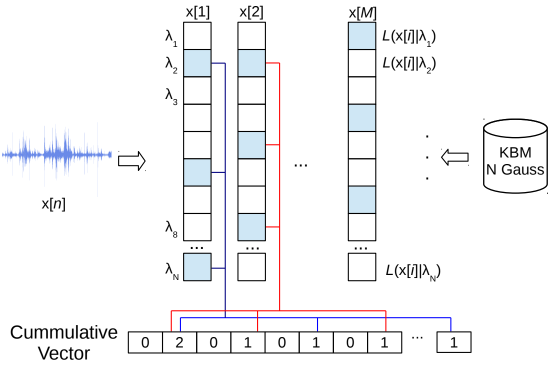
Binary key computation



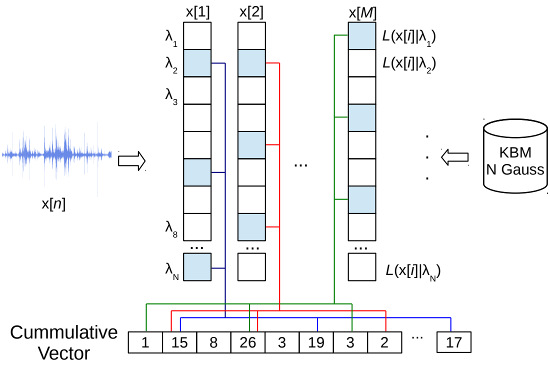
Binary key computation



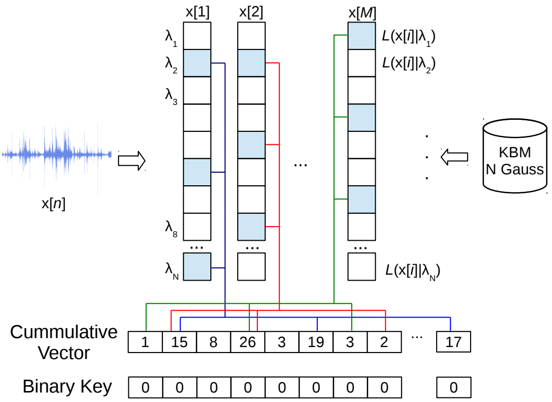
Binary key computation



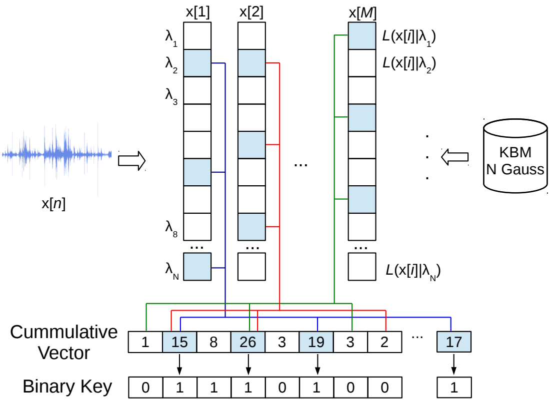
Binary key computation



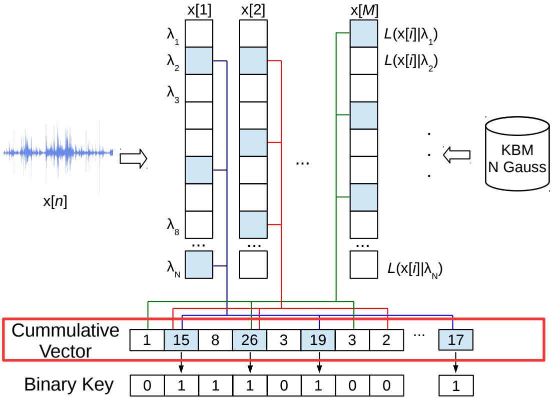
Binary key computation



Binary key computation



Binary key computation



For speaker verification: Hernandez-Sierra, G., Calvo, J.R., Bonastre, J.F, "Session compensation using binary speech representation for speaker recognition," Pattern Recognition Letters, 2014

For speaker diarization: Delgado, H., Anguera, X., Fredouille, C., Serrano, J. "Improved binary key speaker diarization system," in Proc. EUSIPCO, 2015.

Final clustering selection

$$T_s = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Select the optimum clustering by using the T-test metric[8]
- m_1, σ_1, n_1, m_2, σ_2 and n_2 are the mean, standard deviation and size of intra-cluster and inter-cluster distance distributions, respectively

Not very accurate. Main system's drawback!

[8] Nguyen, T. H., Chng, E. S., Li, H., "T-test distance and clustering criterion for speaker diarization," in Proc. Interspeech, 2008.

Outline

Speeding up KBM training: KL2 distance

Current Gaussian selection process based on the **Symmetrized Kullback-Leibler** (KL2) distance

$$D_{KL2} = D_{KL}(P||Q) + D_{KL}(Q||P)$$

KL divergence for normal multivariate distributions:

$$KL(P||Q) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^t\Sigma_P^{-1}(\mu_Q - \mu_P) - k - \ln\left(\frac{\det\Sigma_P}{\det\Sigma_Q}\right)\right)$$

Involves matrix operations like traces, inversions and determinants

Speeding up KBM training: Cosine distance

- How to lighten Gaussian comparison?
- Can we do without the Gaussian covariance matrices and use the means only?

Cosine distance:

$$D_{cos}(a, b) = 1 - S_{cos}(a, b)$$

where

$$S_{cos}(a, b) = \frac{a \cdot b}{||a||\,||b||}$$

- Does not take into account the vector's magnitude, but the direction defined
- Measures the cosine of the angle between the two vectors being compared

Outline

Final clustering selection

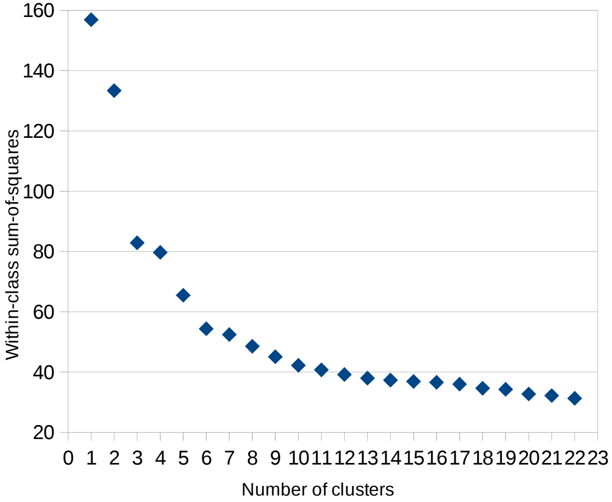
- Given a clustering solution C_k of k clusters $c_1, c_2, \ldots, c_k$, each cluster containing CVs representing speech segments
- **Within-Cluster Sum of Squares** (WCSS):

$$W(C_k) = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2$$

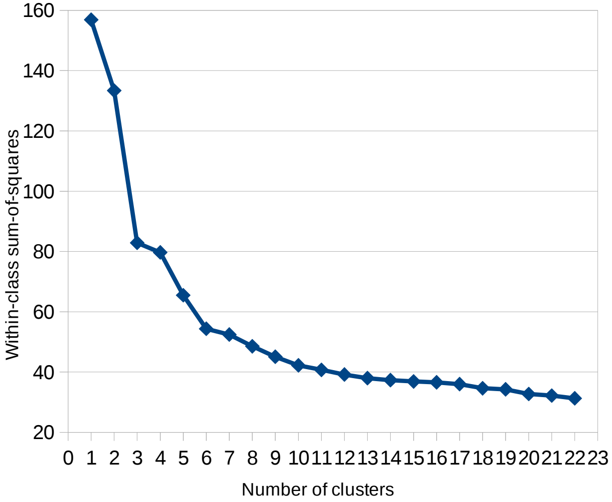
where μ_i is the mean of the points of cluster c_i (centroid)

- Set of clustering solutions $C = (C_1, \ldots, C_{N_{init}})$ with a decreasing number of clusters (from a single cluster to N_{init} clusters)
- Calculate WCSS for all C_i in C using the cosine distance

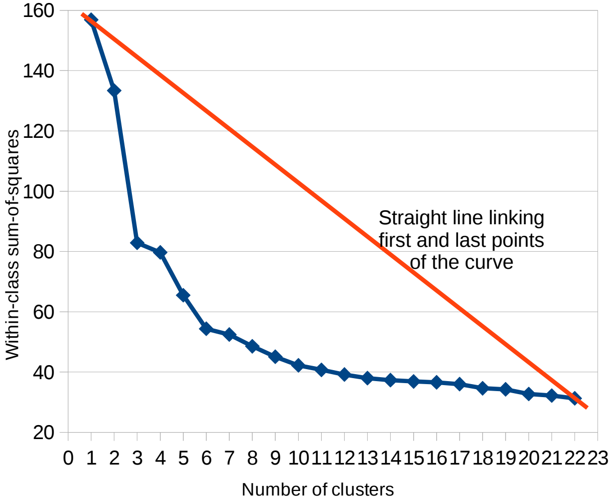
Final clustering selection: Elbow criterion



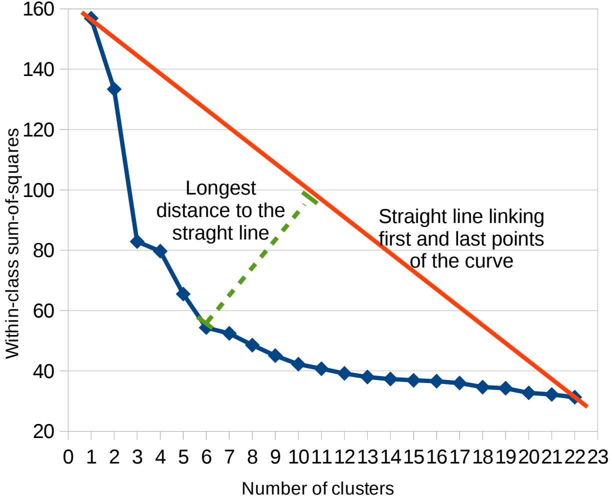
Final clustering selection: Elbow criterion



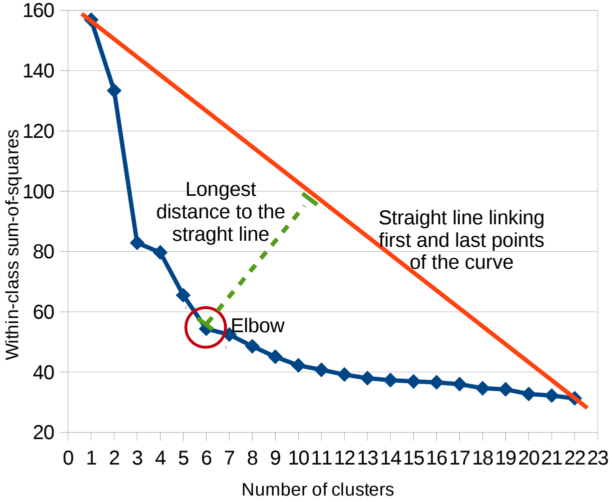
Final clustering selection: Elbow criterion



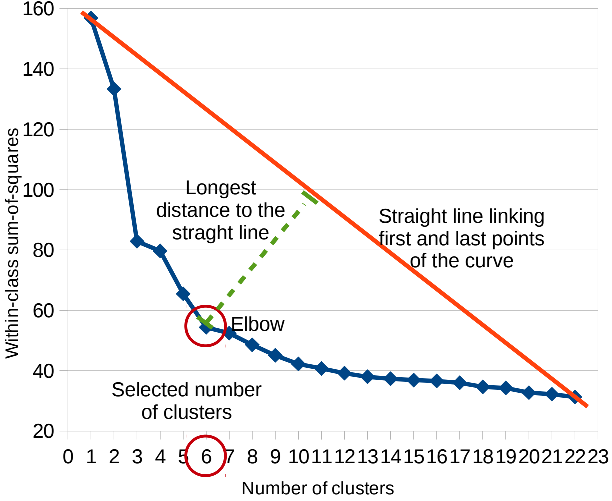
Final clustering selection: Elbow criterion



Final clustering selection: Elbow criterion



Final clustering selection: Elbow criterion



Outline

Evaluation setup

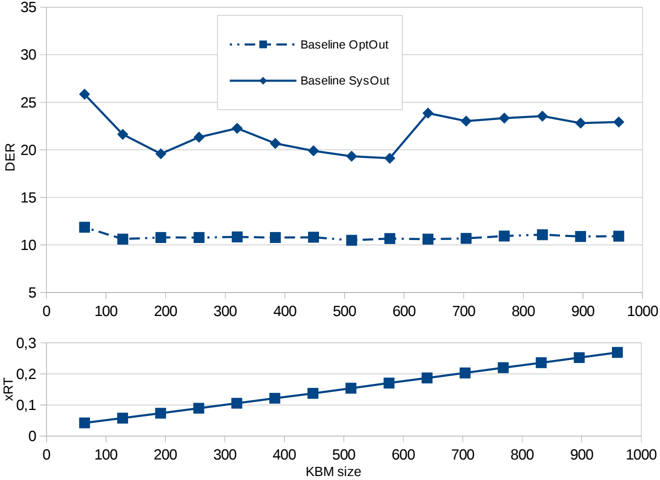
Database: REPERE Phase 1 test set (broadcast TV shows)

- Standard 19-order MFCCs
- Ground-truth SAD labels
- KBM settings
  - 2s window
  - Shift adjusted to get a pool of around 2000 Gaussians
  - KBM size: free parameter
- Cumulative Vectors: 5 top Gaussians at frame level
- Clustering initialization: 25 uniform-initialized clusters
- Data mapping: 1s segments (adding 1s after and before = 3s)

Evaluation metrics:

- DER with 0.25s forgiveness collar, overlapping speech is accounted
- Real-time factor (xRT): $xRT = t_{system}/dur_{speech}$ (excluding feature extraction)

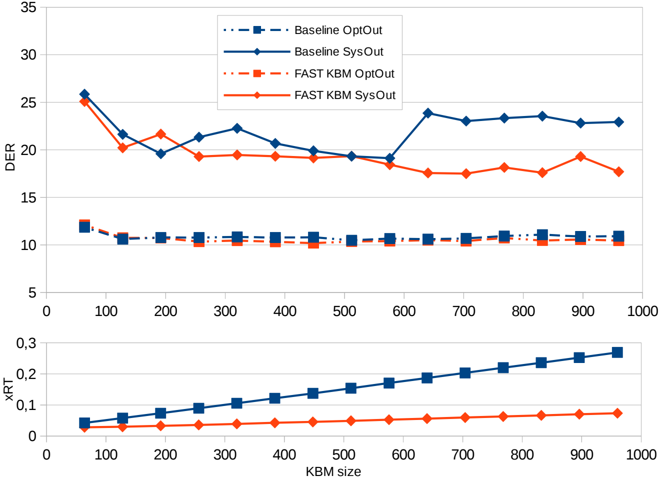
New KBM training results



SysOut: System output

OptOut: Result of the best clustering selected manually

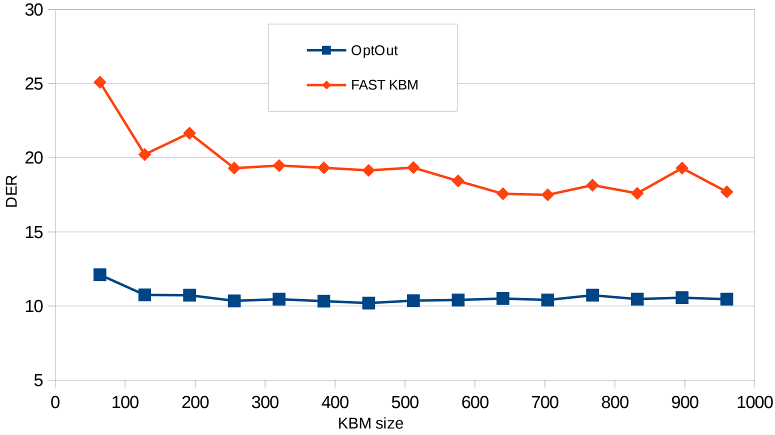
New KBM training results



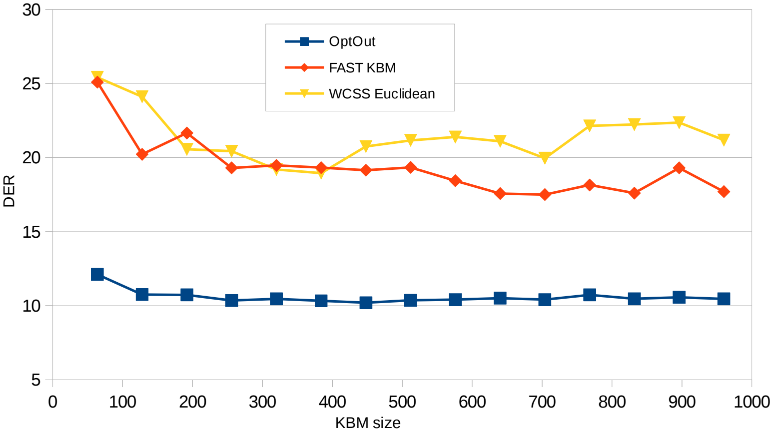
SysOut: System output

OptOut: Result of the best clustering selected manually

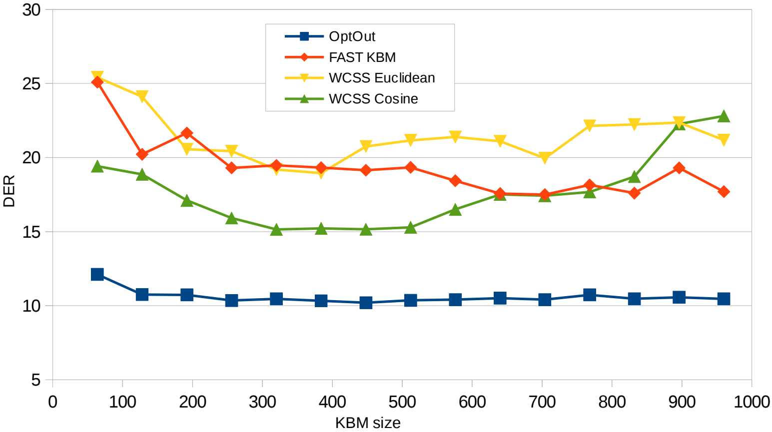
Final clustering selection results



Final clustering selection results



Final clustering selection results



Results summary

	KBM size	DER (%)	Abs. red.	Rel. red.	xRT
Baseline system	576	19.12	-	-	0.17
+ Fast KBM	704	**17.5**	**1.62**	**8.47**	**0.06**
+ New criterion	320	**15.15**	**3.97**	**20.76**	**0.037**

Outline

Conclusions and future work

- New KBM training based on cosine distance significantly speeds up the process
- New clustering selection criterion based on WCSS improves DER of output solution
- Obtained performance very near to state-of-the-art in this database, while being very fast

Future work

- The error floor has not been reached yet
- Further refine final clustering selection
- More speed ups? xRT < 0.01? (100 times faster-than-real-time)

Thank you!

hecdelflo@gmail.com

Download the speaker diarization system Matlab code from:

http://hectordelgado.me/software