Global Speaker Clustering towards Optimal Stopping Criterion in 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]Telefonica Research, Barcelona, Spain
[3]University of Avignon, CERI/LIA, France

November 19, 2014

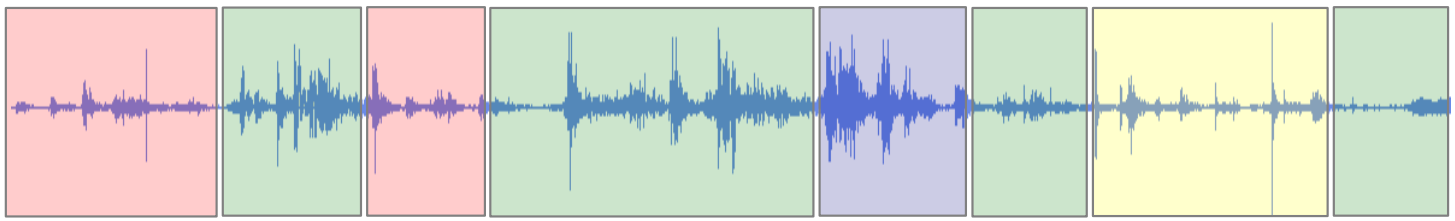
Centre d'Accessibilitat
i Intel·ligència Ambiental
de Catalunya

Universitat Autònoma
de Barcelona

Outline

Speaker diarization

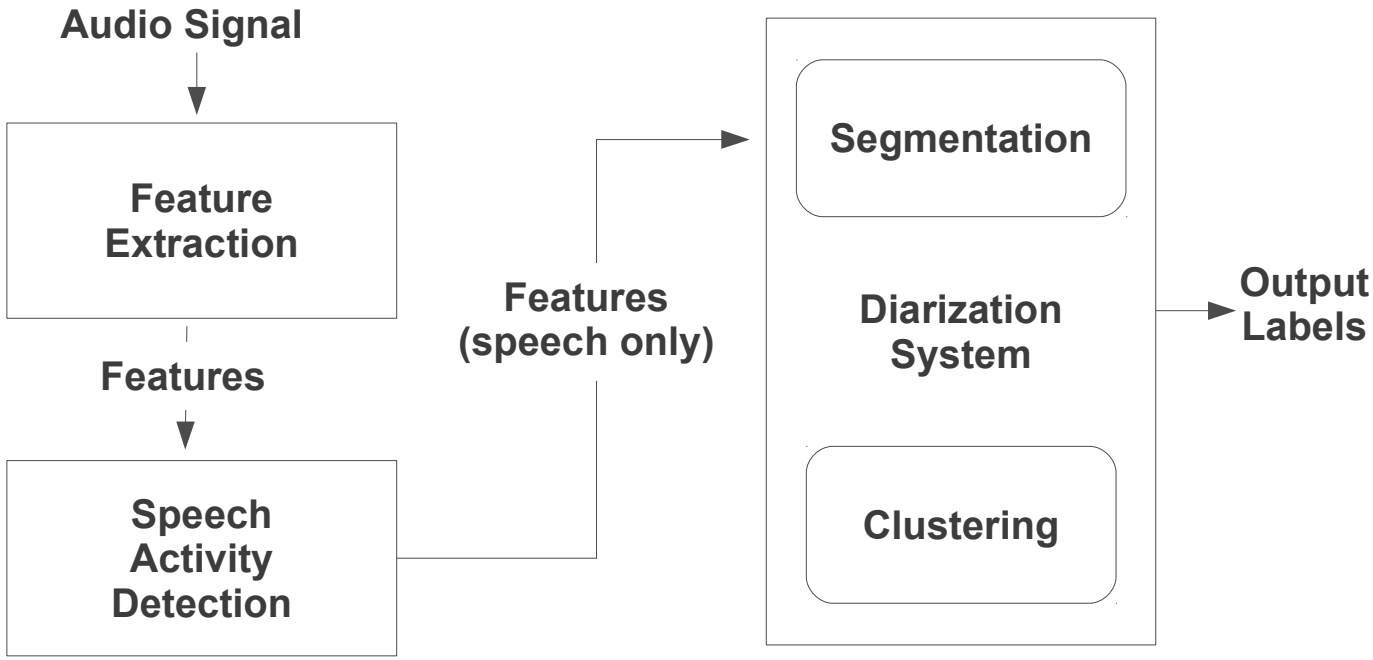
"Who spoke when?"



Segmenting an audio file into speaker turns

- Multi-speaker audio signal
- Unknown speakers
- Unknown number of speakers
- Unsupervised process

Speaker diarization process



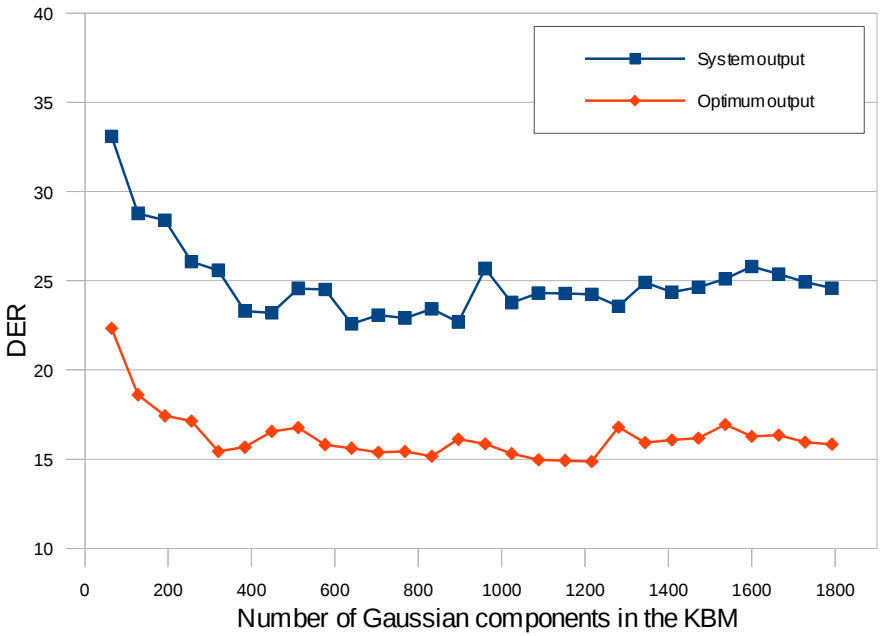
Binary Key Speaker Diarization

- 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 over 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.

Motivation



- Clustering selection does not return optimum number of clusters
- Explore alternative clustering selection algorithms
- **Global clustering** recently proposed vs AHC
- Implicitly finds optimum number of clusters

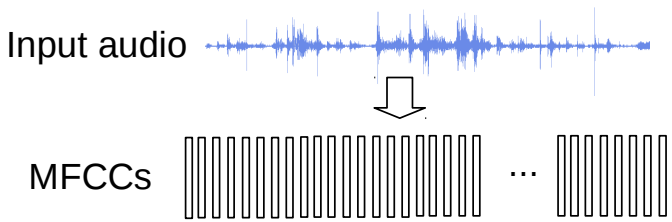
H. Delgado, C. Fredouille, J. Serrano. "Towards a Complete Binary Key System for the Speaker Diarization task," in Proc. Interspeech, 2014.

Outline

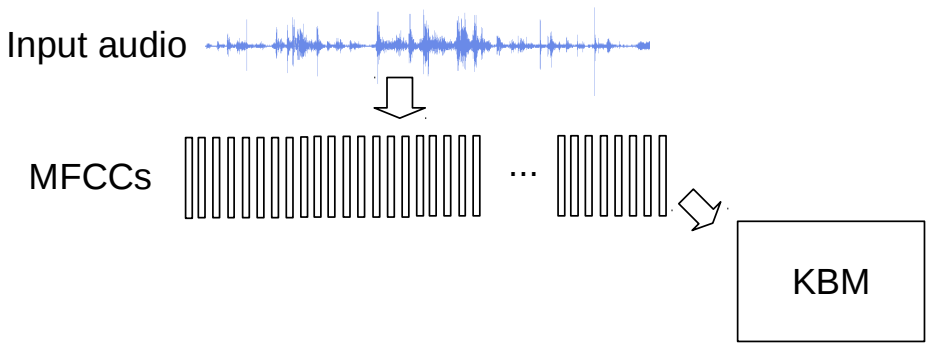
Binary Key Speaker Diarization System

Input audio 

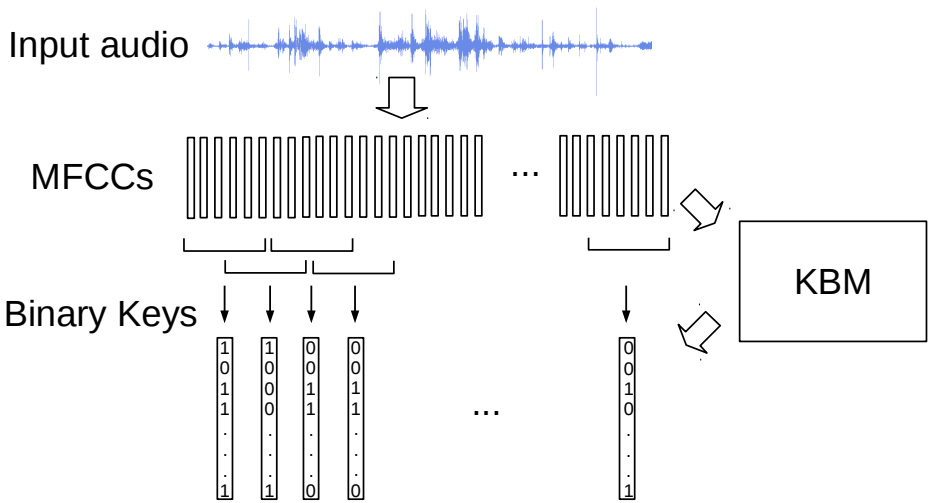
Binary Key Speaker Diarization System



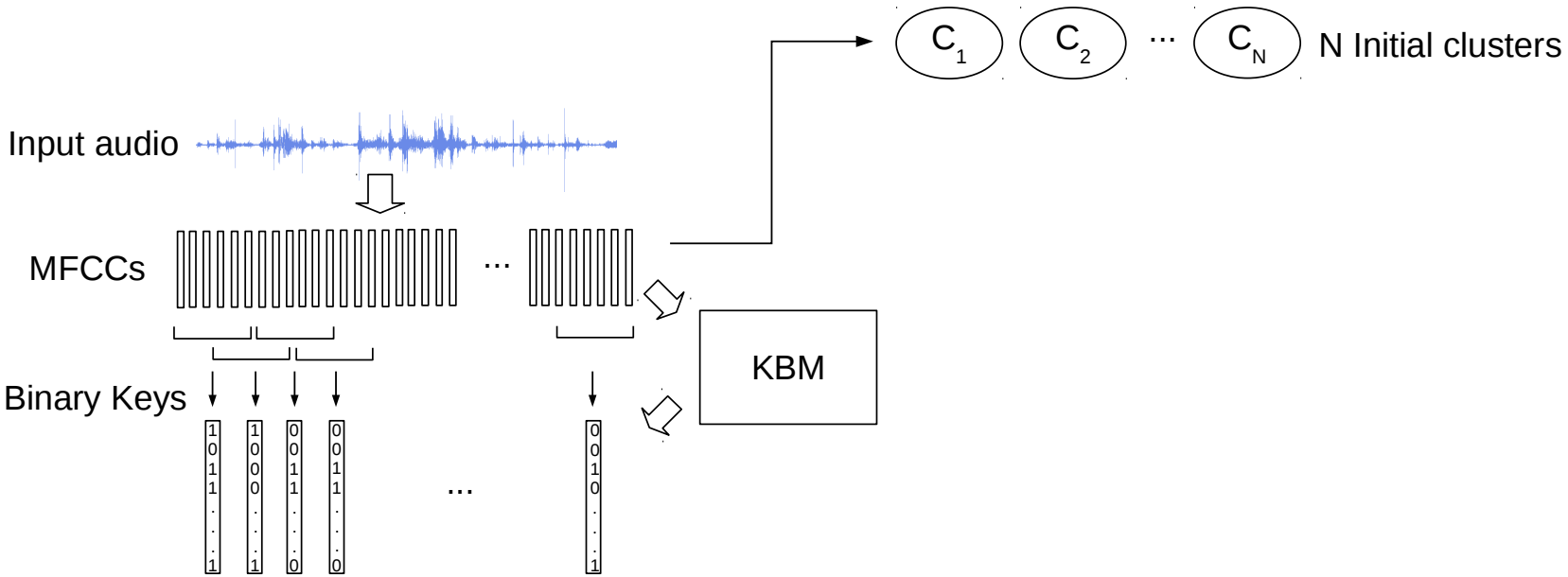
Binary Key Speaker Diarization System



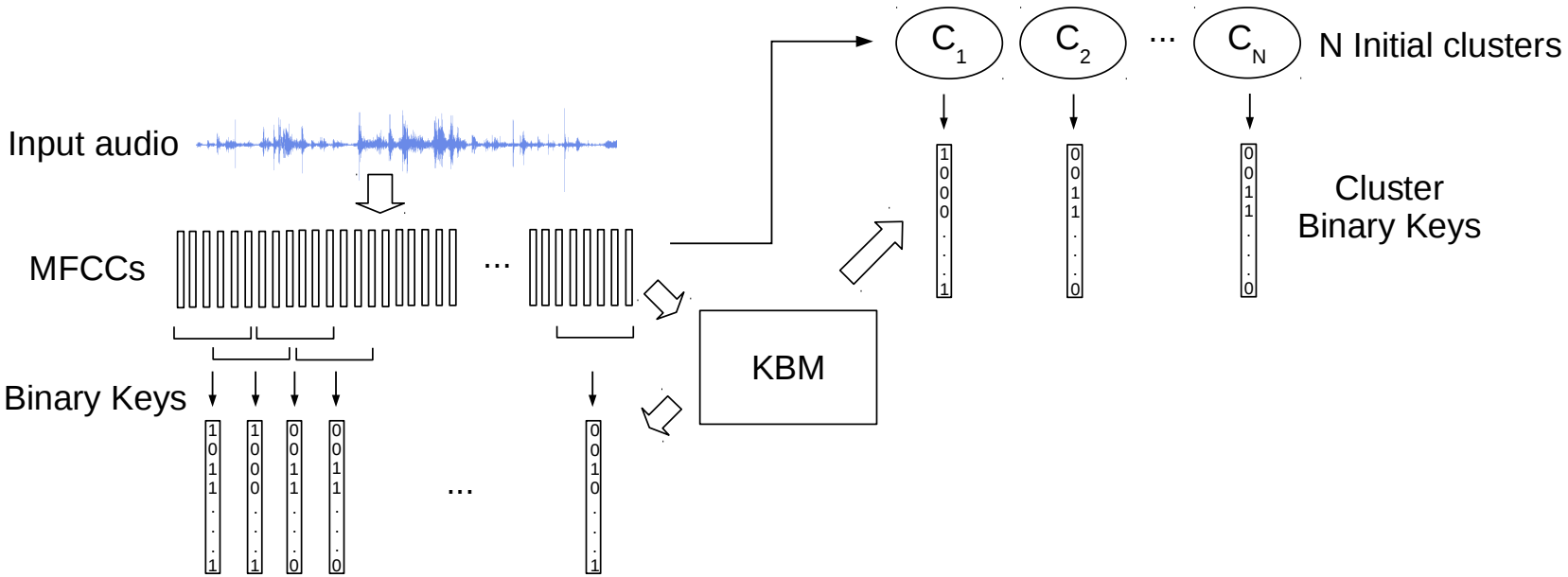
Binary Key Speaker Diarization System



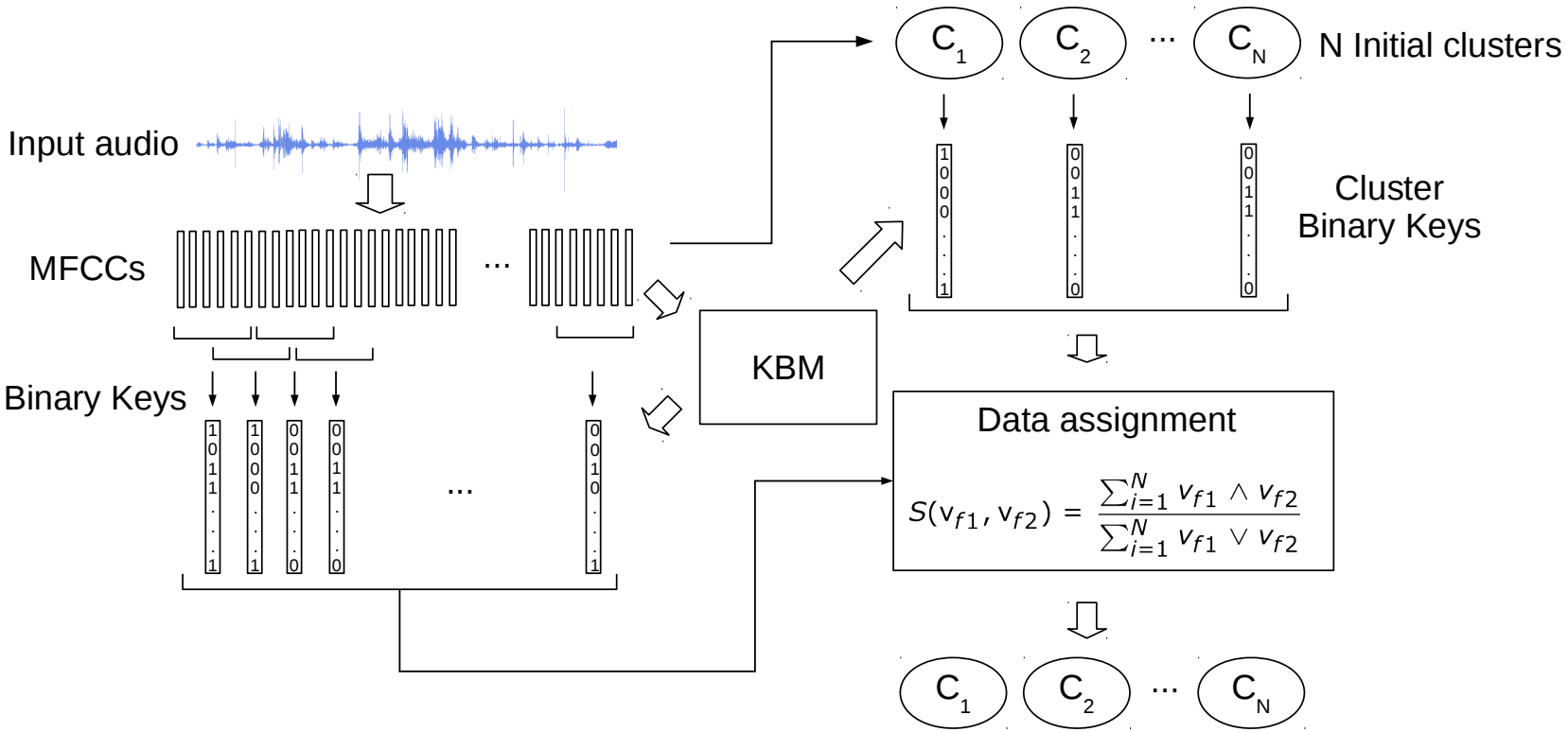
Binary Key Speaker Diarization System



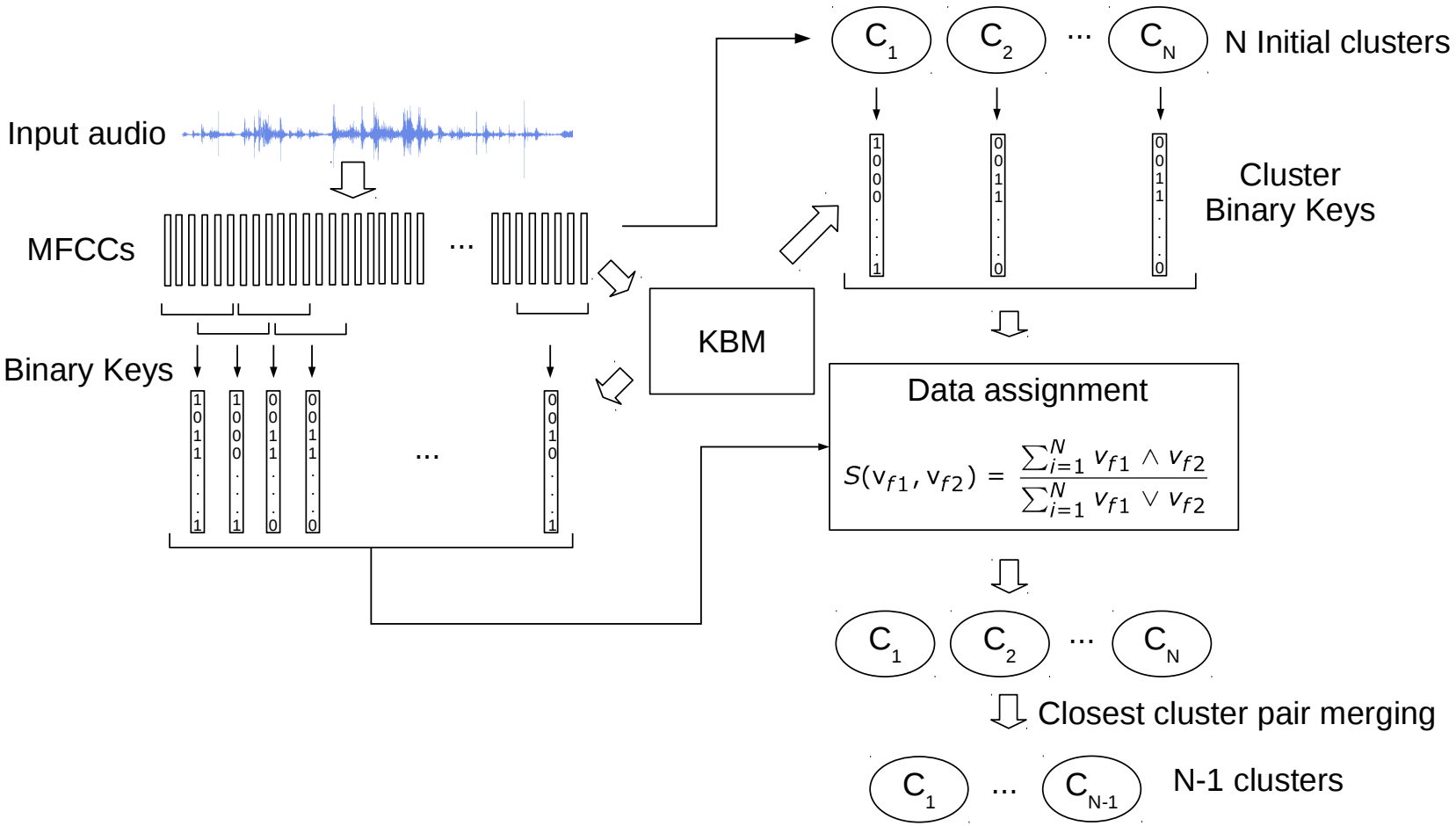
Binary Key Speaker Diarization System



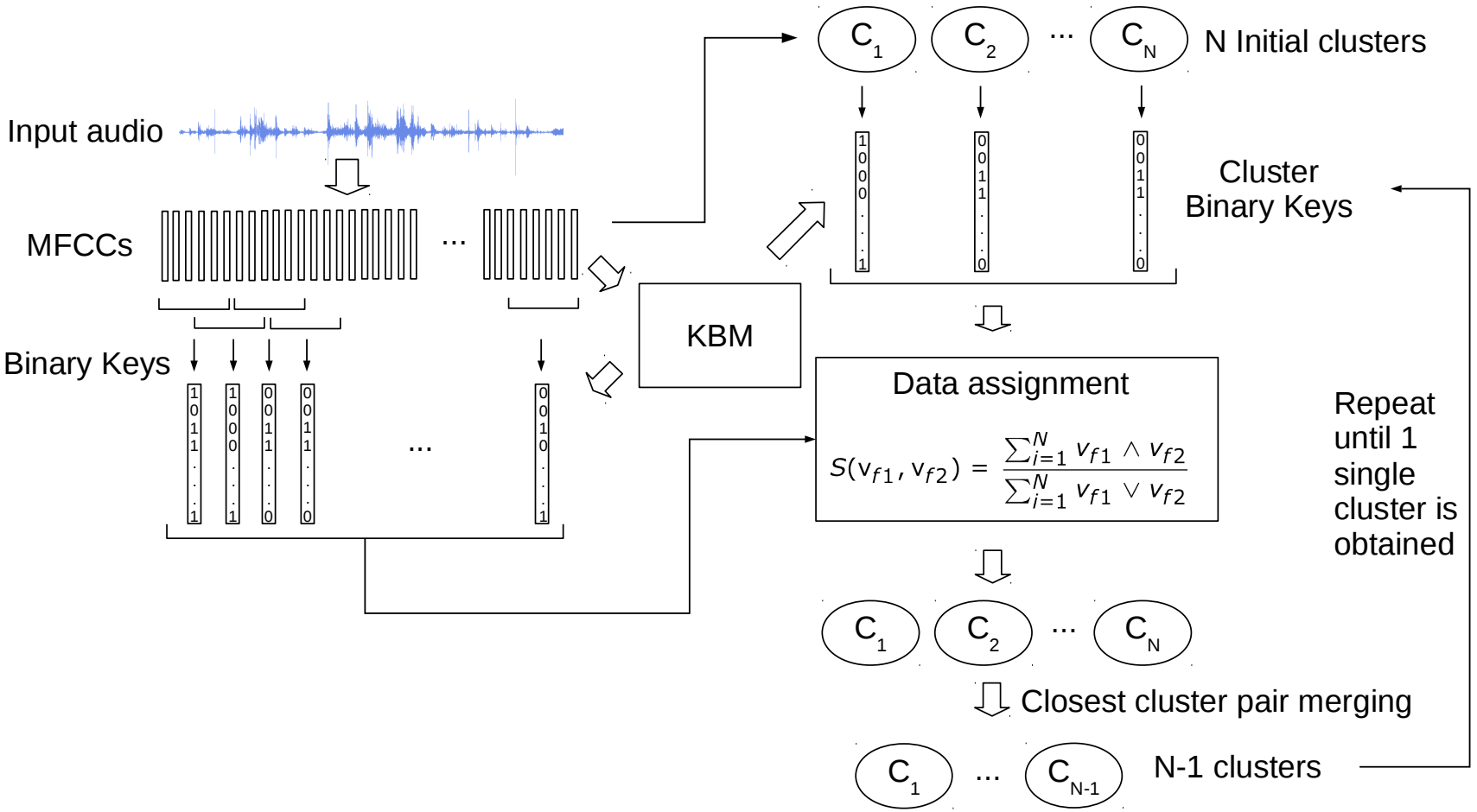
Binary Key Speaker Diarization System



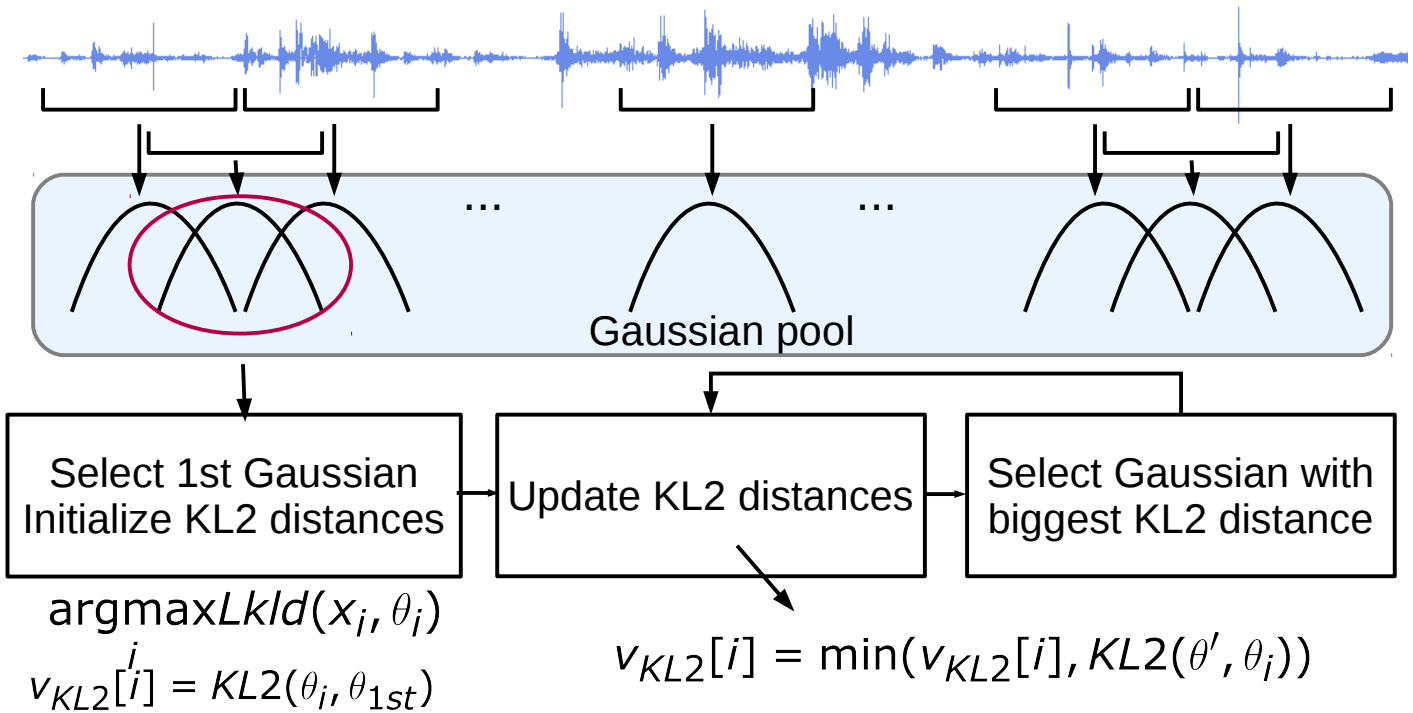
Binary Key Speaker Diarization System



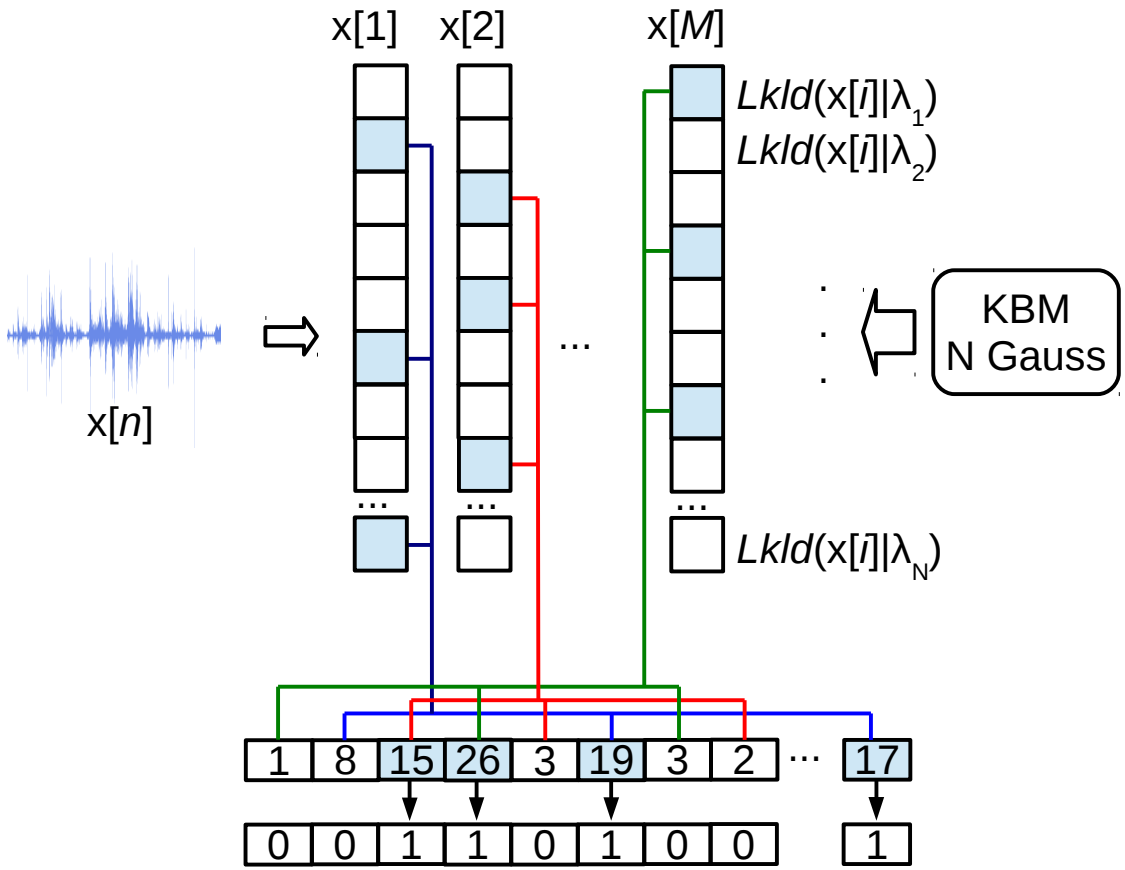
Binary Key Speaker Diarization System



KBM training



Binary key computation



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[4]
- 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

[4] Trung Hieu Nguyen, Eng Siong Chng, and Haizhou Li, "T-test distance and clustering criterion for speaker diarization," in Proc. Interspeech, 2008.

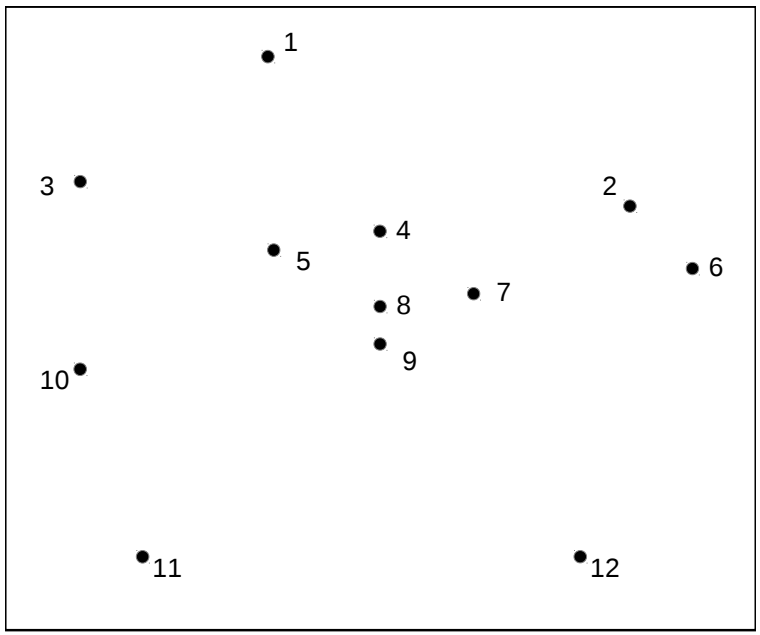
Outline

Global Speaker Clustering (I)

- Given an initial clustering with N clusters
  - Clusters should be highly pure
  - But each speaker may be distributed along several clusters
- Parametrize each cluster with an i-vector
- Find optimal clustering of the i-vectors
  - Minimize number of clusters
  - Minimize dispersion of i-vector within the clusters

G. Dupuy et al. "I-vectors and ILP lustering adapted to cross-show speaker diarization," in Proc. Interspeech, 2012.

Global Speaker Clustering (II)



N number of initial clusters

$x_{k,k}$ binary variable equal to 1 when i-vector k is a center

$x_{k,j}$ binary variable equal to 1 when i-vector j i assigned to center k

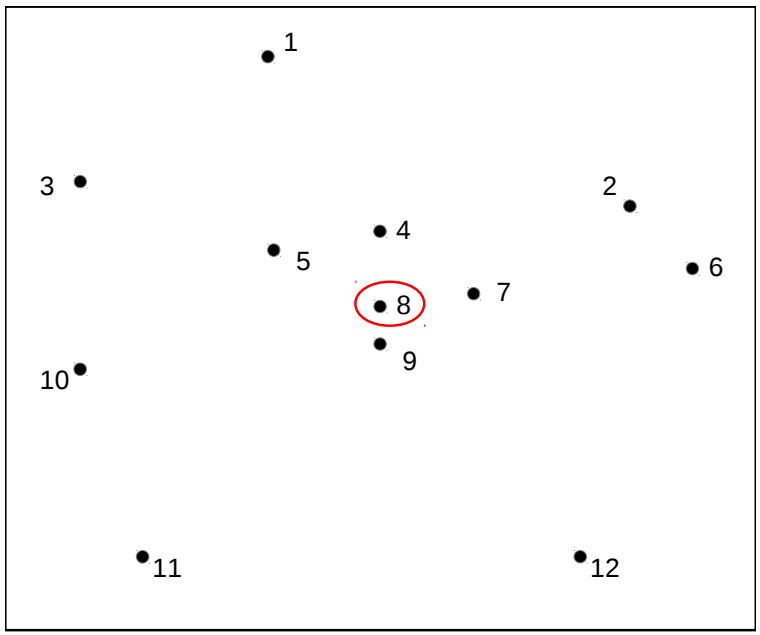
$d(k,j)$ distance between i-vector k and j

δ threshold

D normalization factor

G. Dupuy et al. "I-vectors and ILP clustering adapted to cross-show speaker diarization," in Proc. Interspeech, 2012.

Global Speaker Clustering (II)



N number of initial clusters

$x_{k,k}$ binary variable equal to 1 when i-vector k is a center

$x_{k,j}$ binary variable equal to 1 when i-vector j i assigned to center k

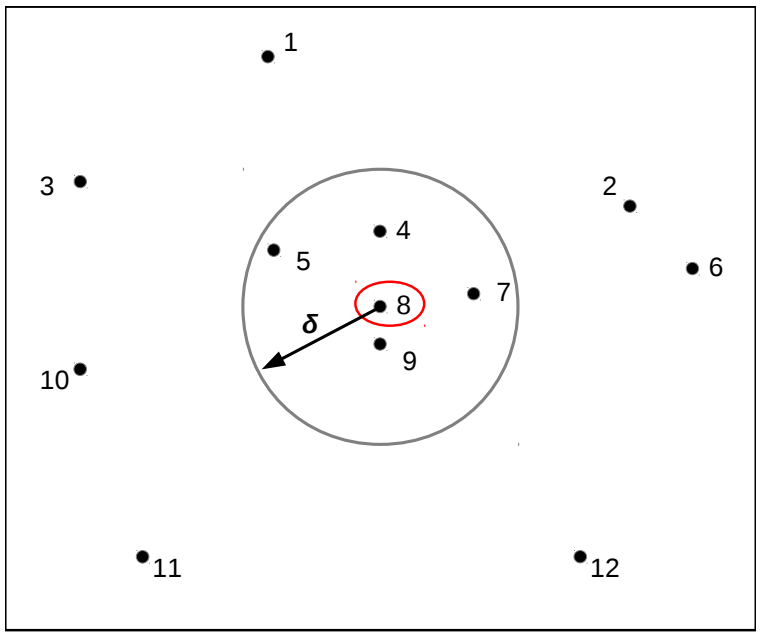
$d(k,j)$ distance between i-vector k and j

δ threshold

D normalization factor

G. Dupuy et al. "I-vectors and ILP clustering adapted to cross-show speaker diarization," in Proc. Interspeech, 2012.

Global Speaker Clustering (II)



N number of initial clusters

$x_{k,k}$ binary variable equal to 1 when i-vector k is a center

$x_{k,j}$ binary variable equal to 1 when i-vector j i assigned to center k

$d(k,j)$ distance between i-vector k and j

δ threshold

D normalization factor

G. Dupuy et al. "I-vectors and ILP clustering adapted to cross-show speaker diarization," in Proc. Interspeech, 2012.

Global Speaker Clustering (II)

Minimize

$$\sum_{k=1}^{N} x_{k,k} - \frac{1}{D} \sum_{k=1}^{N} \sum_{j=1}^{N} d(k,j) x_{k,j} \quad (1)$$

Subject to

$$x_{k,j} \in \{0,1\} \qquad \forall k, \forall j \quad (2)$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad \forall j \quad (3)$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad \forall k, \forall j \quad (4)$$

$$d(k,j) x_{k,j} \leq \delta \qquad \forall k, \forall j \quad (5)$$

- N number of initial clusters
- $x_{k,k}$ binary variable equal to 1 when i-vector k is a center
- $x_{k,j}$ binary variable equal to 1 when i-vector j i assigned to center k
- $d(k,j)$ distance between i-vector k and j
- δ threshold
- D normalization factor

G. Dupuy et al. "I-vectors and ILP clustering adapted to cross-show speaker diarization," in Proc. Interspeech, 2012.

Adaptation to Binary Key Speaker Diarization

Any other high-level features could be used to parametrize the clusters

- i-vector $\rightarrow$ **Binary Key**

Distance measure:

$$D(\mathrm{v}_{f1}, \mathrm{v}_{f2}) = 1 - S(\mathrm{v}_{f1}, \mathrm{v}_{f2}) = 1 - \frac{\sum_{i=1}^{N}(v_{f1}[i] \wedge v_{f2}[i])}{\sum_{i=1}^{N}(v_{f1}[i] \vee v_{f2}[i])}$$

Outline

Experiments

1. Initial clusters should be highly pure $\rightarrow$ **Purity analysis**
2. Take purer clusterings as input $\rightarrow$ **Perform the global clustering process**

Diarization system 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 of 896
- Binary key parameters
  - 5 top Gaussians at frame level
  - Top 20% Gaussians at segment/cluster level
- Clustering initialization with 25 and 50 initial clusters
- Data assignment using 1s segments (plus 1 second before and after)

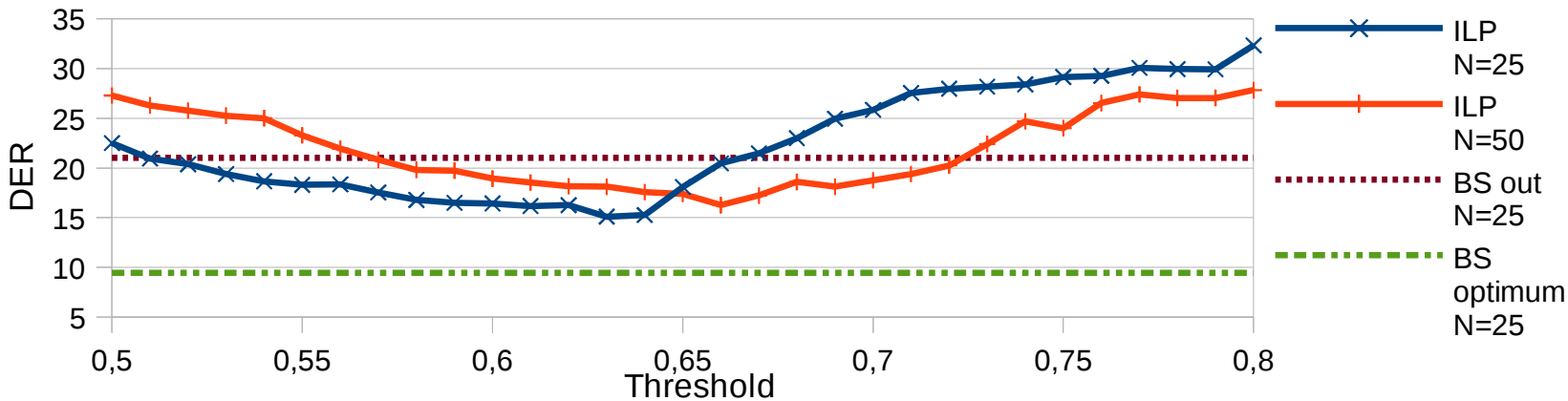
Cluster purity

Show ID	#spk	N_init = 25						N_init = 50					
		Highest purity			SysOut purity			Highest purity			SysOut purity		
		#C	Purity	DER	#C	Purity	DER	#C	Purity	DER	#C	Purity	DER
BFMTV_BFMStory_1	6	19	0.910	24.60	8	0.883	4.37	45	0.923	51.49	6	0.865	6.42
BFMTV_BFMStory_2	18	18	0.891	18.43	18	0.891	18.43	41	0.941	33.32	31	0.913	21.00
BFMTV_BFMStory_3	10	19	0.951	29.77	13	0.937	7.25	35	0.962	33.19	11	0.915	5.58
BFMTV_BFMStory_4	6	11	0.962	11.05	6	0.952	1.59	20	0.963	14.82	6	0.950	1.78
BFMTV_CultureEtVous_1	5	14	0.950	52.13	4	0.891	10.11	21	0.970	52.04	3	0.881	8.88
BFMTV_CultureEtVous_2	6	10	0.925	27.16	4	0.776	21.09	23	0.956	43.30	4	0.780	20.63
BFMTV_CultureEtVous_3	16	22	0.904	63.80	5	0.744	24.44	29	0.851	62.16	4	0.702	23.68
BFMTV_CultureEtVous_4	9	22	0.890	54.37	2	0.640	33.07	41	0.902	70.80	3	0.650	30.63
BFMTV_CultureEtVous_5	6	21	0.905	72.06	3	0.823	9.90	20	0.917	65.02	3	0.823	9.90
BFMTV_CultureEtVous_6	12	18	0.870	52.70	5	0.700	25.37	29	0.890	57.79	5	0.720	21.15
BFMTV_CultureEtVous_7	14	18	0.839	43.19	6	0.701	31.22	34	0.850	68.51	9	0.758	26.67
LCP_CaVousRegarde_1	7	23	0.925	70.01	4	0.823	15.46	48	0.957	82.51	6	0.883	12.11
LCP_CaVousRegarde_2	5	7	0.938	8.00	4	0.903	3.11	14	0.951	9.28	4	0.917	2.64
LCP_CaVousRegarde_3	5	21	0.950	40.21	6	0.836	19.28	37	0.950	55.75	15	0.886	21.94
LCP_EntreLesLignes_1	5	15	0.919	24.07	10	0.909	11.34	19	0.958	19.64	19	0.958	19.64
LCP_EntreLesLignes_2	5	27	0.932	28.44	6	0.891	4.92	26	0.949	24.37	6	0.891	4.44
LCP_EntreLesLignes_3	5	15	0.945	29.45	3	0.823	14.74	26	0.935	29.45	3	0.823	14.74
LCP_LCPInfo13h30_1	16	21	0.890	23.69	14	0.882	10.04	39	0.938	26.83	20	0.902	13.19
LCP_LCPInfo13h30_2	12	18	0.951	16.77	11	0.890	10.87	41	0.953	34.64	23	0.918	17.40
LCP_LCPInfo13h30_3	10	13	0.905	24.55	7	0.871	12.28	27	0.921	24.67	11	0.841	17.10
LCP_PileEtFace_1	3	14	0.921	25.17	3	0.821	8.24	16	0.955	19.52	6	0.921	10.91
LCP_PileEtFace_2	3	15	0.954	29.25	3	0.864	8.11	31	0.960	72.58	2	0.853	8.28
LCP_PileEtFace_3	3	9	0.910	11.18	3	0.797	6.89	24	0.932	37.36	3	0.808	6.89
LCP_PileEtFace_4	3	7	1.000	6.81	6	0.988	3.95	17	1.000	21.56	8	0.988	7.59
LCP_PileEtFace_5	3	18	0.936	53.94	3	0.912	4.20	4	0.936	3.67	4	0.936	3.67
LCP_TopQuestions_1	8	22	0.987	35.89	8	0.976	1.36	12	0.989	7.42	9	0.981	2.11
LCP_TopQuestions_2	5	15	0.985	23.41	3	0.914	8.13	20	0.985	29.11	6	0.959	4.09
LCP_TopQuestions_3	6	11	0.973	21.53	5	0.948	5.17	14	0.973	13.34	5	0.948	5.17
Overall	-	-	0.929	-	-	0.857	9.47	-	0.942	-	-	0.870	10.60

Cluster purity

Show ID	#spk	N_init = 25						N_init = 50					
		Highest purity			SysOut purity			Highest purity			SysOut purity		
		#C	Purity	DER	#C	Purity	DER	#C	Purity	DER	#C	Purity	DER
BFMTV_BFMStory_1	6	19	0.910	24.60	8	0.883	4.37	45	0.923	51.49	6	0.865	6.42
BFMTV_BFMStory_2	18	18	0.891	18.43	18	0.891	18.43	41	0.941	33.32	31	0.913	21.00
BFMTV_BFMStory_3	10	19	0.951	29.77	13	0.937	7.25	35	0.962	33.19	11	0.915	5.58
BFMTV_BFMStory_4	6	11	0.962	11.05	6	0.952	1.59	20	0.963	14.82	6	0.950	1.78
BFMTV_CultureEtVous_1	5	14	0.950	52.13	4	0.891	10.11	21	0.970	52.04	3	0.881	8.88
BFMTV_CultureEtVous_2	6	10	0.925	27.16	4	0.776	21.09	23	0.956	43.30	4	0.780	20.63
BFMTV_CultureEtVous_3	16	22	0.904	63.80	5	0.744	24.44	29	0.851	62.16	4	0.702	23.68
BFMTV_CultureEtVous_4	9	22	0.890	54.37	2	0.640	33.07	41	0.902	70.80	3	0.650	30.63
BFMTV_CultureEtVous_5	6	21	0.905	72.06	3	0.823	9.90	20	0.917	65.02	3	0.823	9.90
BFMTV_CultureEtVous_6	12	18	0.870	52.70	5	0.700	25.37	29	0.890	57.79	5	0.720	21.15
BFMTV_CultureEtVous_7	14	18	0.839	43.19	6	0.701	31.22	34	0.850	68.51	9	0.758	26.67
LCP_CaVousRegarde_1	7	23	0.925	70.01	4	0.823	15.46	48	0.957	82.51	6	0.883	12.11
LCP_CaVousRegarde_2	5	7	0.938	8.00	4	0.903	3.11	14	0.951	9.28	4	0.917	2.64
LCP_CaVousRegarde_3	5	21	0.950	40.21	6	0.836	19.28	37	0.950	55.75	15	0.886	21.94
LCP_EntreLesLignes_1	5	15	0.919	24.07	10	0.909	11.34	19	0.958	19.64	19	0.958	19.64
LCP_EntreLesLignes_2	5	27	0.932	28.44	6	0.891	4.92	26	0.949	24.37	6	0.891	4.44
LCP_EntreLesLignes_3	5	15	0.945	29.45	3	0.823	14.74	26	0.935	29.45	3	0.823	14.74
LCP_LCPInfo13h30_1	16	21	0.890	23.69	14	0.882	10.04	39	0.938	26.83	20	0.902	13.19
LCP_LCPInfo13h30_2	12	18	0.951	16.77	11	0.890	10.87	41	0.953	34.64	23	0.918	17.40
LCP_LCPInfo13h30_3	10	13	0.905	24.55	7	0.871	12.28	27	0.921	24.67	11	0.841	17.10
LCP_PileEtFace_1	3	14	0.921	25.17	3	0.821	8.24	16	0.955	19.52	6	0.921	10.91
LCP_PileEtFace_2	3	15	0.954	29.25	3	0.864	8.11	31	0.960	72.58	2	0.853	8.28
LCP_PileEtFace_3	3	9	0.910	11.18	3	0.797	6.89	24	0.932	37.36	3	0.808	6.89
LCP_PileEtFace_4	3	7	1.000	6.81	6	0.988	3.95	17	1.000	21.56	8	0.988	7.59
LCP_PileEtFace_5	3	18	0.936	53.94	3	0.912	4.20	4	0.936	3.67	4	0.936	3.67
LCP_TopQuestions_1	8	22	0.987	35.89	8	0.976	1.36	12	0.989	7.42	9	0.981	2.11
LCP_TopQuestions_2	5	15	0.985	23.41	3	0.914	8.13	20	0.985	29.11	6	0.959	4.09
LCP_TopQuestions_3	6	11	0.973	21.53	5	0.948	5.17	14	0.973	13.34	5	0.948	5.17
Overall	-	-	0.929	-	-	0.857	9.47	-	0.942	-	-	0.870	10.60

Global Clustering Results (1)



Global Clustering Results (2)

Show ID	#spk	N_init = 25			N_init = 50		
		θ_{opt}	#C	DER	θ_{opt}	#C	DER
BFMTV_BFMStory_1	6	0.63	5	4.52	0.78	6	7.17
BFMTV_BFMStory_2	18	0.50	18	18.17	0.61	24	21.44
BFMTV_BFMStory_3	10	0.57	11	6.42	0.63	11	4.51
BFMTV_BFMStory_4	6	0.62	7	1.78	0.65	7	1.69
BFMTV_CultureEtVous_1	5	0.77	2	18.57	0.82	3	20.35
BFMTV_CultureEtVous_2	6	0.81	3	13.79	0.73	5	17.42
BFMTV_CultureEtVous_3	16	0.77	4	30.31	0.77	6	55.28
BFMTV_CultureEtVous_4	9	0.83	4	39.64	0.89	14	30.56
BFMTV_CultureEtVous_5	6	0.77	4	38.63	0.75	4	33.24
BFMTV_CultureEtVous_6	12	0.76	4	25.74	0.85	4	22.68
BFMTV_CultureEtVous_7	14	0.77	4	29.73	0.82	4	26.48
LCP_CaVousRegarde_1	7	0.69	4	9.61	0.79	5	12.05
LCP_CaVousRegarde_2	5	0.63	4	3.11	0.73	4	2.60
LCP_CaVousRegarde_3	5	0.69	9	19.58	0.74	7	19.31
LCP_EntreLesLignes_1	5	0.63	11	10.08	0.50	19	19.64
LCP_EntreLesLignes_2	5	0.68	5	3.41	0.75	4	9.27
LCP_EntreLesLignes_3	5	0.77	3	15.39	0.76	3	15.00
LCP_LCPInfo13h30_1	16	0.52	15	10.07	0.65	13	10.42
LCP_LCPInfo13h30_2	12	0.53	11	15.26	0.56	22	11.40
LCP_LCPInfo13h30_3	10	0.57	6	13.76	0.57	17	13.59
LCP_PileEtFace_1	3	0.77	2	12.28	0.76	4	10.32
LCP_PileEtFace_2	3	0.74	3	7.15	0.77	2	8.28
LCP_PileEtFace_3	3	0.70	5	8.22	0.73	4	8.82
LCP_PileEtFace_4	3	0.63	6	2.94	0.77	3	1.35
LCP_PileEtFace_5	3	0.74	2	4.98	0.50	4	3.67
LCP_TopQuestions_1	8	0.63	8	1.60	0.63	9	1.01
LCP_TopQuestions_2	5	0.72	4	2.89	0.77	4	2.36
LCP_TopQuestions_3	6	0.66	5	2.79	0.64	7	4.30
Overall	-	-	-	9.57	-	-	10.20

Diarization results: Summary

Baseline system output (faulty stopping criterion)	ILP $N_{=}25$, $\theta = 0.63$
21.0%	**15.1%**

Baseline optimum output	ILP optimum $N_{=}25$
9.47%	9.57%

Outline

Conclusions and future work

- Global clustering provides an alternative for cluster selection in Binary Key speaker diarization
- The approach outperforms the original stopping criterion but...
- It is still far from being optimal (audio-dependent parameters)

Future work

- Get a method to automatically select the threshold per audio file
- Optimum ILP does not outperform optimum AHC
  - Try something else...
  - Session variability compensation?

Thank you!

hector.delgado@uab.cat