EURECOM
Sophia Antipolis

UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

The LIA-EURECOM RT'09 Speaker Diarization System

LIA – Laboratoire Informatique d'Avignon (France)
EURECOM - Sophia Antipolis (France)

Corinne Fredouille,
Simon Bozonnet and Nick Evans
(corinne.fredouille)@univ-avignon.fr
(simon.bozonnet,nick.evans)@eurecom.fr

Collaboration work

- **Joint work with:**
  - University of Avignon, ***Laboratoire Informatique d'Avignon*** (**L.I.A.**)
  - ***EURECOM*** – Sophia Antipolis

→ A **common submission** for RT'09 evaluation

Outline

- **Evaluation context**
- **System description**
  - Speech activity detection system
  - Speaker diarization system
- **Development Results**
- **Validation Results**
- **MDM Results**
- **SDM Results (unofficial)**
- **Post-eval Analysis**
- **Top-down system, comparison with Bottom-up**
- **Conclusion & Future work**

Evaluation Context

- **tasks:**

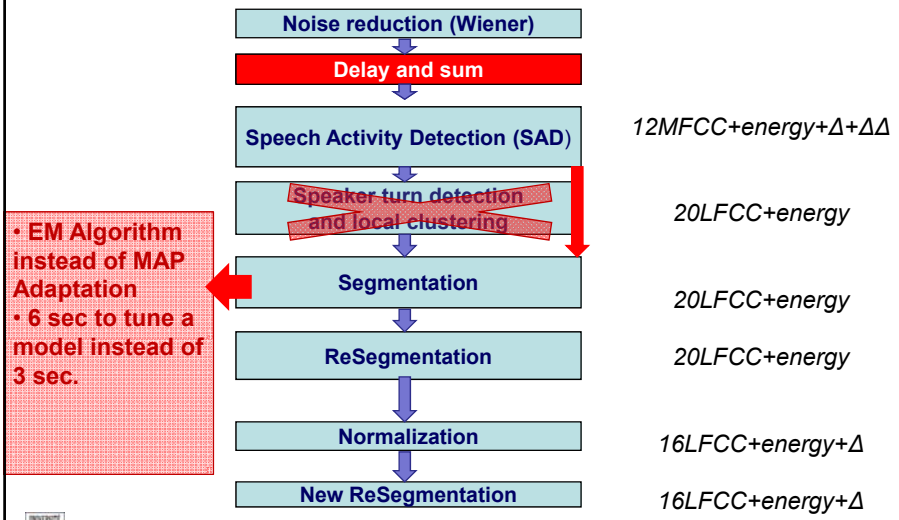
Task	ADM	MDM	MM3A	SDM	AV
Done:		X		X	

- **No delay features used except for beamforming**

[N. Evans, C. Fredouille, JF. Bonastre, *Speaker diarization using unsupervized discriminant analysis of inter-channel delay features ICASSP 2009 - April 19-24, 2009*]

Speaker diarization systems

- **Compared to RT07's system:**



Pre-treatment

- **Wiener filter noise reduction on all the files**
  - Qualcom-ICSI-OGI frontend

[Adami, A., Burget, L. Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P. Kajarekar, S., Morgan, N. Sivadas, S., Qualcomm-ICSI-OGI features for ASR. In: Proc. ICSLP. (2002) 21-24]

- **Single virtual channel for each show: Delay and Sum**
  - BeamformIt toolkit, 500ms analysis window, 250ms frame rate.

[X. Anguera, C. Wooters, J. Hernando, *Acoustic Beamforming for Speaker Diarization of Meetings, IEEE September 2007]*

[X. Anguera, BeamformIt , the fast and robust acoustic beamformer, http://www.xavieranguera.com/beamformit]

Speech activity detection systems (SAD)

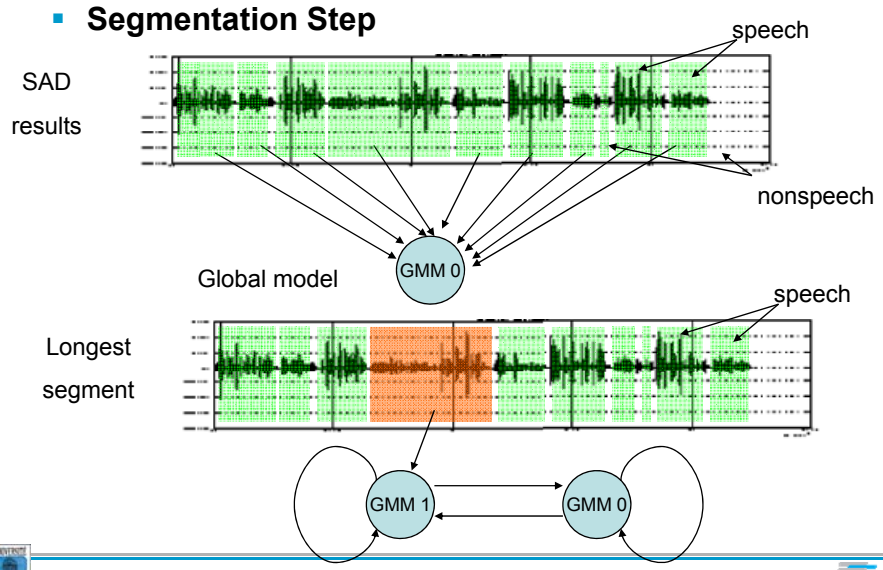
- **Same as for RT'07:**
  - Two-state HMM characterising speech and nonspeech
    - 12MFCC+10energy+9Δ+8ΔΔ, no normalization (39 coefficients)
    - 32 Gaussian components per state, trained on separate meeting data
  - Transition probabilities equally balanced
  - Iterative process => decoding and model adaptation
    - ☞until stability
  - Minimum duration rules for non-speech segments:
    - ☞Length > = 0.3 seconds

Speaker diarization systems

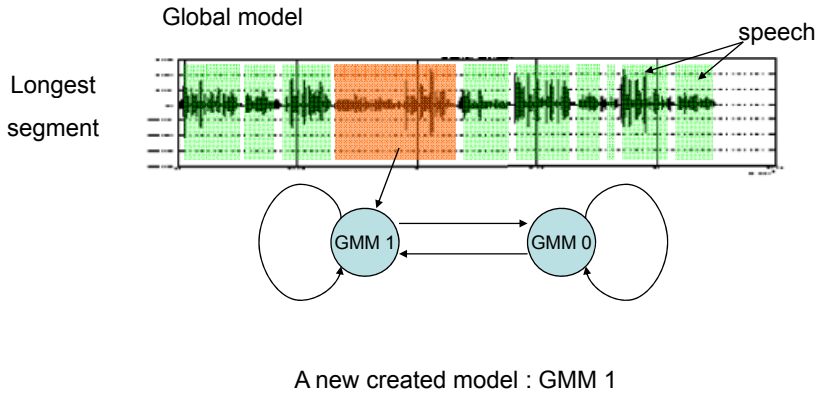
- **LIA-EURECOM's system:**
  - based on an E-HMM = top-down approach:
    - ☞An HMM = discussion between speakers
      - State = speaker
      - Transition = turn changes in discussion
  - Step 1: **Segmentation step**:
    - Iterative process to build the E-HMM
  - Step 2: **Resegmentation step**:
    - Iterative process to refine the segmentation output by deleting irrelevant speakers
  - Step 3: **Normalization and resegmentation phase**:
    - normalisation followed by a second resegmentation phase

Speaker diarization systems

- **Segmentation Step**

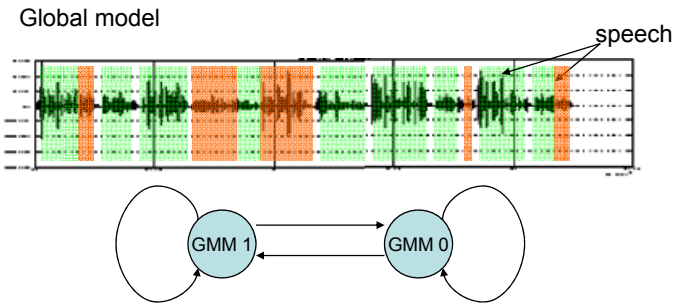


Speaker diarization systems



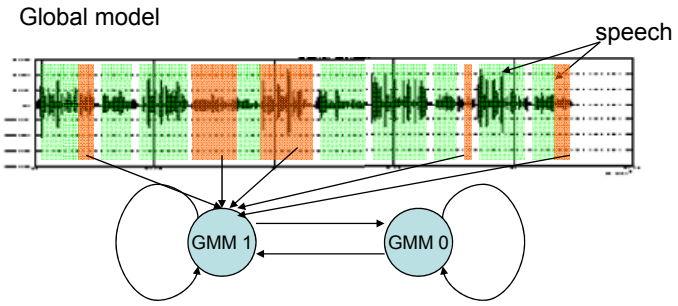
A new created model : GMM 1

Speaker diarization systems



Viterbi Decoding

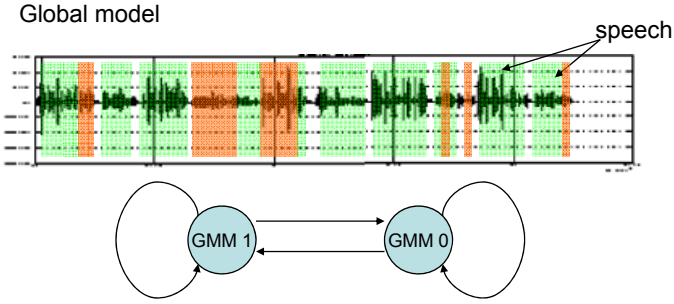
Speaker diarization systems



Update models with EM Algorithm

Speaker diarization systems

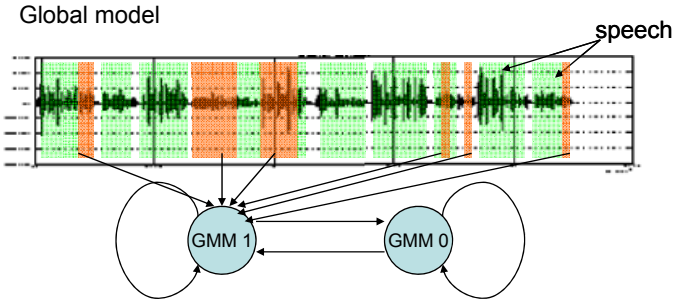
Global model



2nd Viterbi Decoding

Speaker diarization systems

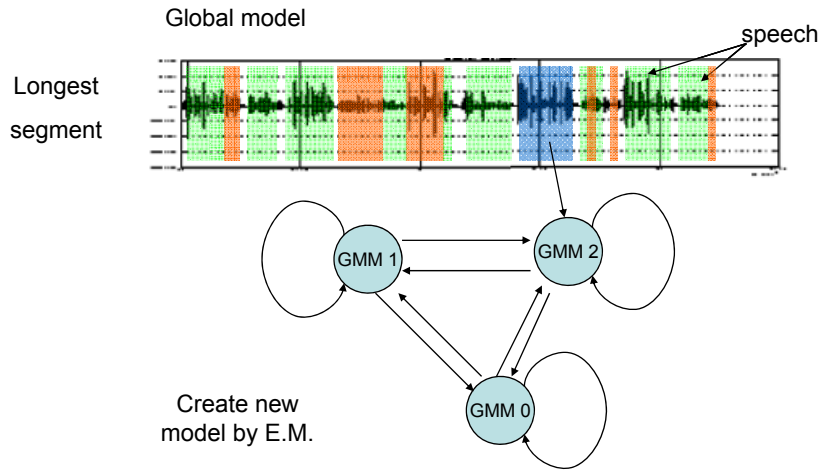
Global model



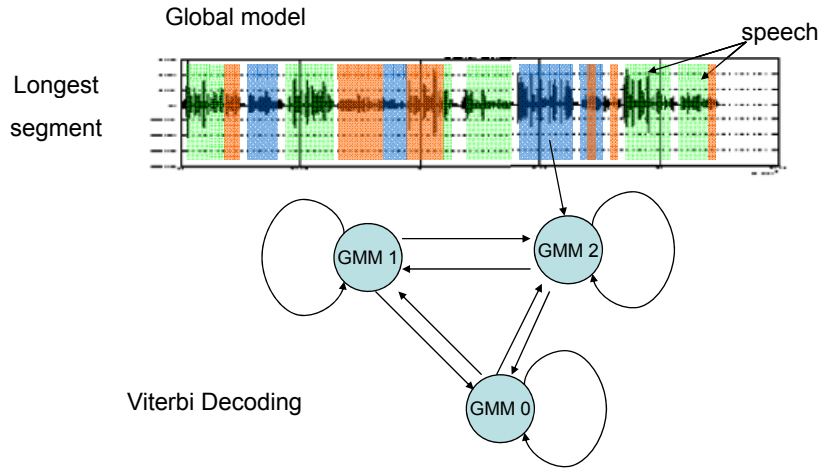
Update models with EM Algorithm

… and so on… …until there are no more changes

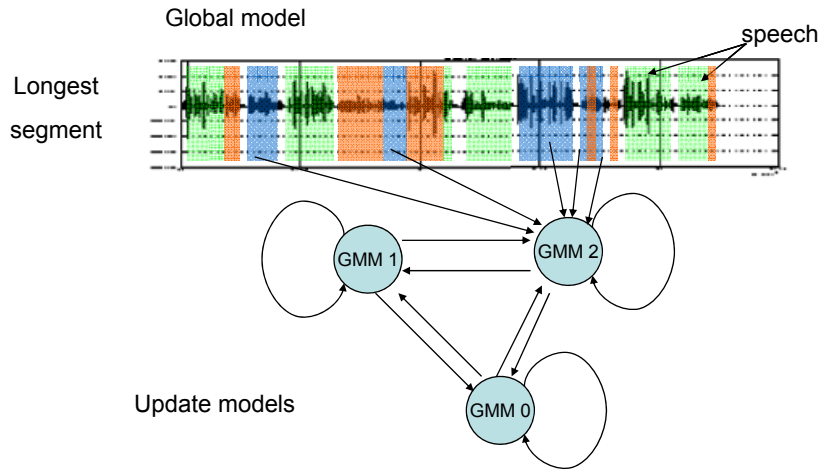
Speaker diarization systems



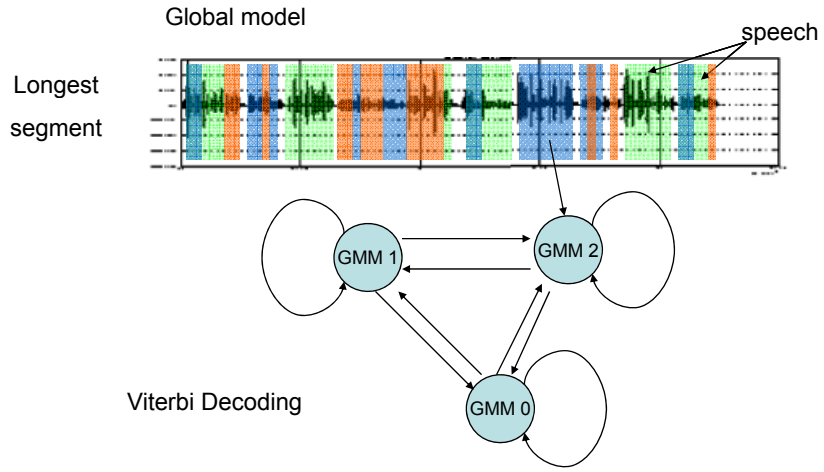
Speaker diarization systems



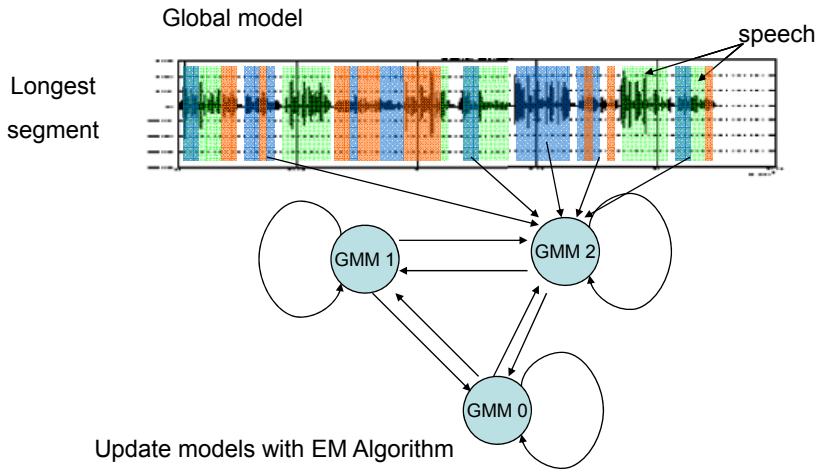
Speaker diarization systems



Speaker diarization systems



Speaker diarization systems

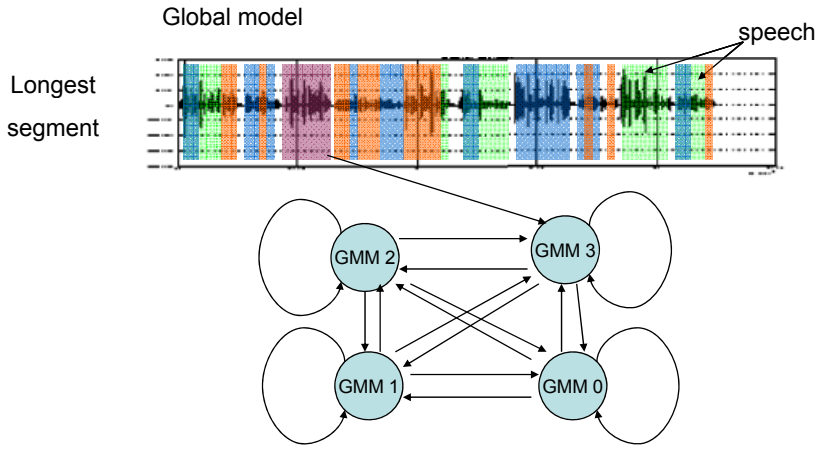


Update models with EM Algorithm

… and so on… …while there are still changes

Speaker diarization systems

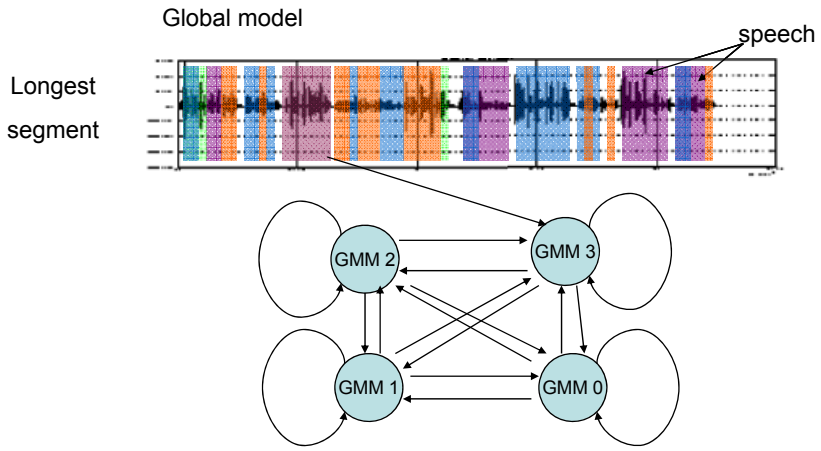
- **Add a new model**



Create a new model

Speaker diarization systems

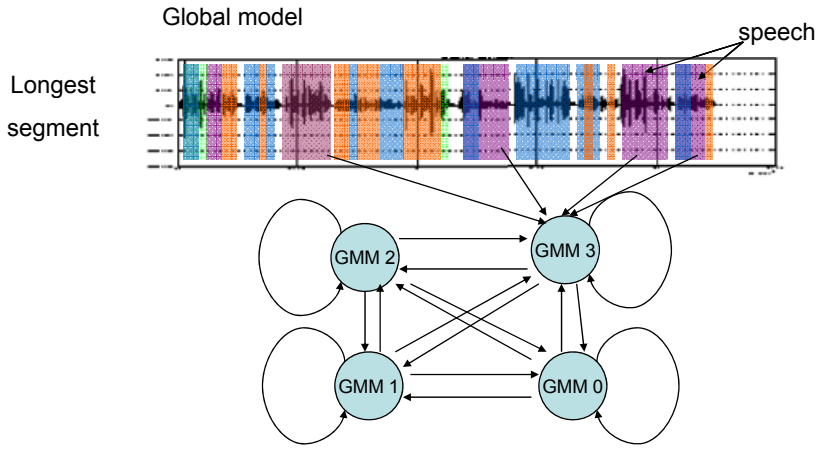
- **Add a new model**



Viterbi Decoding…

Speaker diarization systems

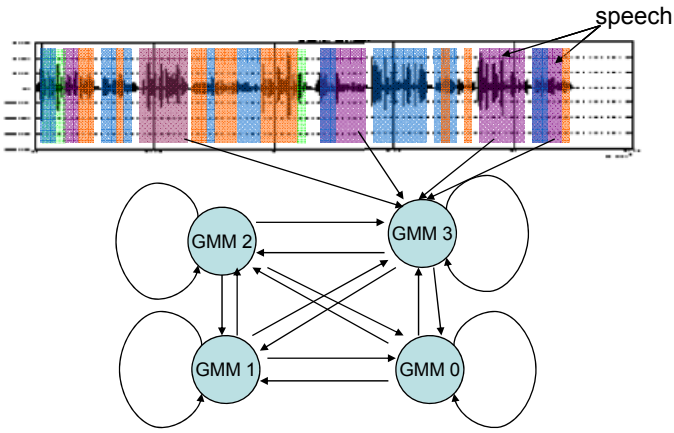
- **Add a new model**



Update… Decode…

Speaker diarization systems

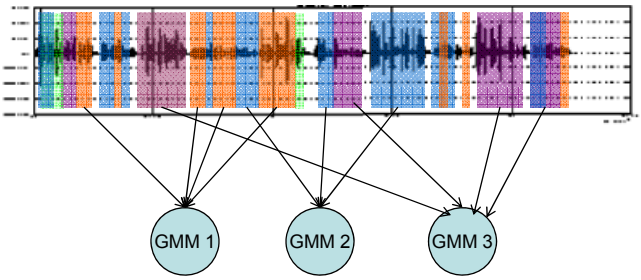
- **Stop criterion**



No more sufficiently long segments belonging to GMM 0

Speaker diarization systems

- **ReSegmentation Step**



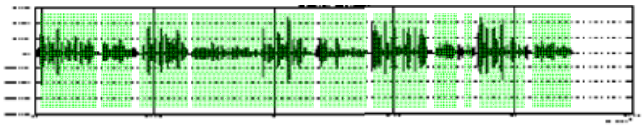
Creation of models by MAP adaptation coupled with a generic speech model

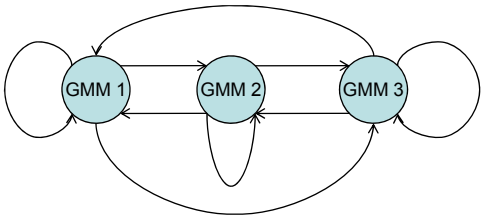
No more global model GMM0

Speaker diarization systems

- **ReSegmentation Step**

SAD results

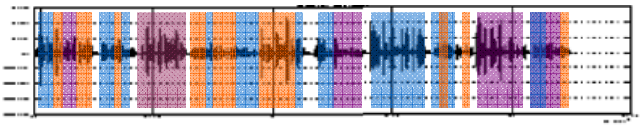


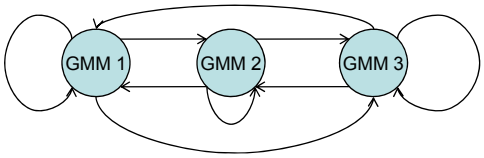


Viterbi Decoding

Speaker diarization systems

- **ReSegmentation Step**

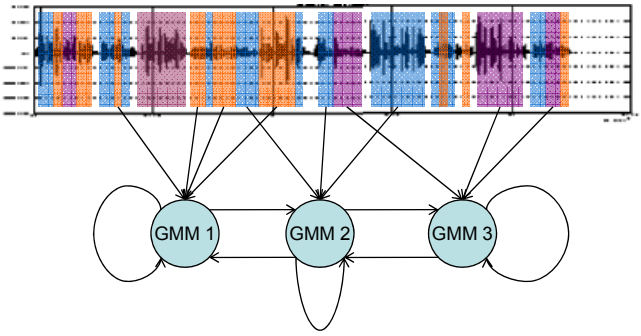




Viterbi Decoding

Speaker diarization systems

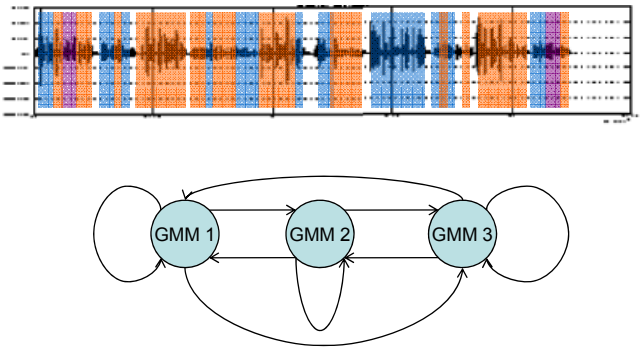
- **ReSegmentation Step**



Update models by MAP adaptation

Speaker diarization systems

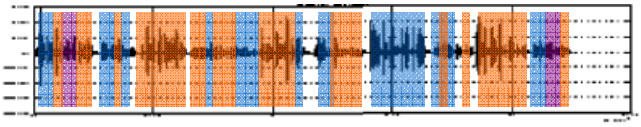
- **ReSegmentation Step**

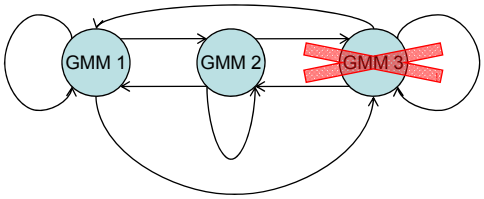


Viterbi decoding

Speaker diarization systems

- **ReSegmentation Step**

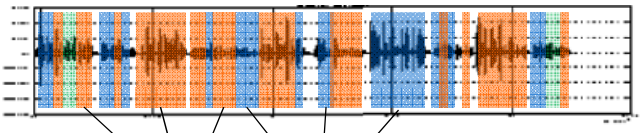


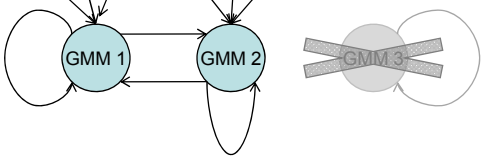


Model GMM 3: less than 6 sec. ⇨ removed

Speaker diarization systems

- **ReSegmentation Step**

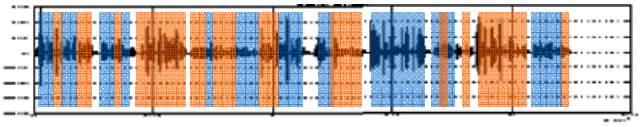


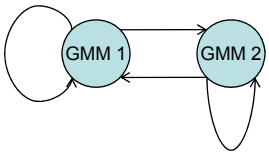


Remove GMM 3 and reassign segments

Speaker diarization systems

- **ReSegmentation Step**

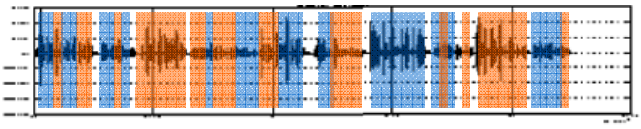


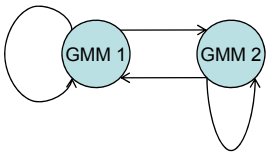


Viterbi decoding

Speaker diarization systems

- **ReSegmentation Step**



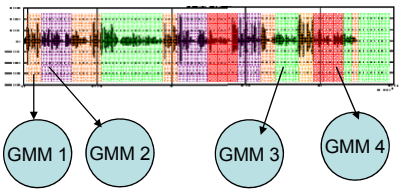


Update models with MAP adaptation

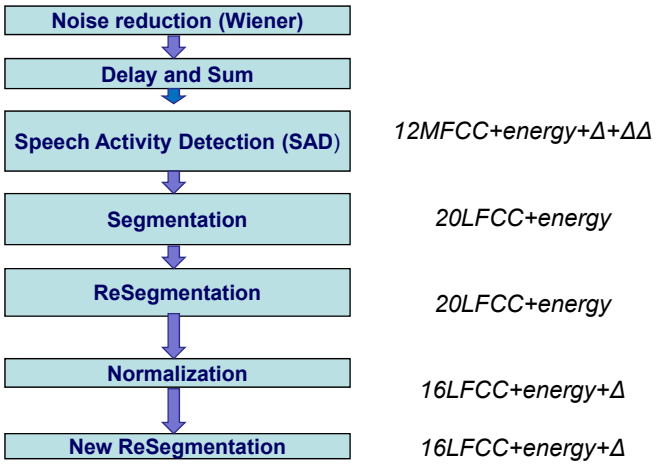
… and so on… …while there are still changes

Speaker diarization systems

- **Normalization and resegmentation phase**
  - The feature vectors are normalized, segment-by-segment
    - fit a zero-mean and unity-variance distribution.
  - New Resegmentation with:
    - 16LFCCs, energy, and their first and second derivatives



Algorithm Overview (Reminder)



Development results

In % of scored speaker time	MS with ovlp	FA with ovlp	SpkERR with ovlp	DER no ovlp	DER with ovlp
AMI_20041210-1052	0.3	0.3	0.4	0.8	1
AMI_20050204-1206	1.3	2.1	6.4	8.9	9.8
CMU_20050228-1615	4	0.8	5.9	7.9	10.7
CMU_20050301-1415	1.4	1.7	6	7.8	9.1
CMU_20050912-0900	9.5	8.8	8.5	18.4	26.8
CMU_20050914-0900	7	6.3	6.3	14.2	19.6
EDI_20050216-1051	3.5	2.9	22.8	27.4	29.2
EDI_20050218-0900	3.1	3.1	6.4	10.2	12.6
ICSI_20000807-1000	4.3	0.4	21.5	23.6	26.2
ICSI_20010208-1430	2.8	1.1	29.8	32.6	33.7
ICSI_20010531-1030	1.8	2.6	11.1	15	15.5
ICSI_20011113-1100	4.3	3.3	12.9	15.9	20.5
LDC_20011116-1400	2	3.2	1.4	5.4	6.6
LDC_20011116-1500	6	1.3	9.3	11.3	16.6
NIST_20030623-1409	2.2	0.3	3.9	5.5	6.4
NIST_20030925-1517	9.4	3.6	18.3	27.6	31.3
NIST_20050427-0939	2	2.3	1.3	3.8	5.6
NIST_20051024-0930	3.9	1.1	6.5	8.2	11.5
NIST_20051102-1323	3.7	2.8	3.3	7	9.8
VT_20050304-1300	0.3	0.8	4.7	5.7	5.8
VT_20050318-1430	2	1.8	38.2	41.9	42
VT_20050623-1400	4.9	4.7	25.1	32.4	34.7
VT_20051027-1400	3.1	2.4	10.5	14.4	16
TOTAL (5 hours 40 min 23 files)	3.9 (0.7)	2.8 (3)	11.2	14.86	17.68

Files issued from RT'04 RT'05, RT'06

EURECOM

Evaluation results

- **Results on RT'07 (with/without overlap)**

System	Mic. Cond.	MissSp	FA	SpkErr	DER
Old	MDM	4.5/0.8	2.0/2.2	17.7/18.6	24.2/21.5
	SDM	4.7/1.1	2.1/2.3	17.7/17.9	24.5/21.3
New	MDM	4.1/0.4	1.8/1.9	11.8/11.9	17.7/14.3
	SDM	4.7/1.1	2.1/2.3	11.4/11.6	18.3/15.0

- **relative improvements in DER of**
  - 27% for MDM
  - 25% for SDM

EURECOM

Official Eval09 Results

- **MDM**

Show	SAD		Speaker Err. (no ovl)	DER (no overlap.)	DER (with overlap)
	Missed Sp.	False Alarm			
EDI_20071128-1000	1.0	1.6	0.8	3.41	6.28
EDI_20071128-1500	1.0	4.5	27.2	37.13	39.97
IDI_20090128-1600	0.5	0.4	13.0	14.16	17.58
IDI_20090129-1000	1.3	4.4	7.8	10.85	15.13
NIST_20080201-1405	1.5	1.0	34.8	43.98	51.99
NIST_20080227-1501	0.3	0.4	18.5	19.80	27.43
NIST_20080307-0955	0.2	1.6	17.9	17.90	22.84
TOTAL	**2.79 %**		**15.5 %**	**18.47 %**	**23.51 %**

SDM set

- **SDM**

Show	Missed Sp. (with ovlp)	False Alarm (with ovlp)	Speaker Err. (with ovlp)	DER (no overlap.)	DER (with overlap)
EDI_20071128-1000 ci01_NONE	8.4	0.3	2.2	8.11	10.85
EDI_20071128-1500 ci01_NONE	10.1	1.4	31.2	40.30	42.72
IDI_20090128-1600 ci01_NONE	4.1	0.5	11.1	12.24	15.65
IDI_20090129-1000 ci01_NONE	4.7	2.5	10.2	13.92	17.35
NIST_20080201-1405 d03_NONE	15.8	1.5	44.9	59.21	62.22
NIST_20080227-1501 d03_NONE	8.5	1.1	11.1	12.62	20.68
NIST_20080307-0955 d03_NONE	3.5	1.8	26.6	28.92	31.98
TOTAL	**7.2 %**	**1.2 %**	**17.6%**	**21.46 %**	**26.02 %**

Comparison old/new system

- **Results on RT'09 dataset**

System	Mic. Cond.	MissSp	FA	SpkErr	DER
Old	MDM	6.6/1.1	1.6/1.8	28.1/29.0	36.3/32.0
	SDM	7.2/1.8	1.2/1.4	21.0/21.5	29.5/24.7
New	MDM	6.1/0.6	1.9/2.1	15.5/15.8	23.5/18.5
	SDM	7.2/1.8	1.2/1.4	17.6/18.3	26.0/21.5

- **relative improvements in DER of**
  - 35% for MDM
  - 12% for SDM

Post-eval Analysis MDM

- **Missed Speaker/Extra Speaker**

file	Real number of speakers (+time)	detected speakers MDM		DER (no ovlp)
EDI_20071128-1000 ci01_NONE	4 (186 to 550 sec)	4	+0	3.41
EDI_20071128-1500 ci01_NONE	4 (103 to 587 sec)	6	+2	37.13
IDI_20090128-1600 ci01_NONE	4 (201 to 820 sec)	6	+2	14.16
IDI_20090129-1000 ci01_NONE	4 (187 to 766 sec)	5	+1	10.85
NIST_20080201-1405 d03_NONE	5 (210 to 456 sec)	5	+0	43.98
NIST_20080227-1501 d03_NONE	6 (125 to 317 sec)	8	+3	19.8
NIST_20080307-0955 d03_NONE	11 (5 to 240 sec)	9	-2	17.9

Post-eval Analysis SDM

- **Missed Speaker/Extra Speaker**

file	Real number of speakers (+time)	detected speakers SDM		DER (no overlap.)
EDI_20071128-1000 ci01_NONE	4 (186 to 550 sec)	5	+1	8.11
EDI_20071128-1500 ci01_NONE	4 (103 to 587 sec)	5	+1	40.30
IDI_20090128-1600 ci01_NONE	4 (201 to 820 sec)	6	+1	12.24
IDI_20090129-1000 ci01_NONE	4 (187 to 766 sec)	5	+1	13.92
NIST_20080201-1405 d03_NONE	5 (210 to 456 sec)	6	+1	59.21
NIST_20080227-1501 d03_NONE	6 (125 to 317 sec)	9	+3	12.62
NIST_20080307-0955 d03_NONE	11 (5 to 240 sec)	6	-5	28.92

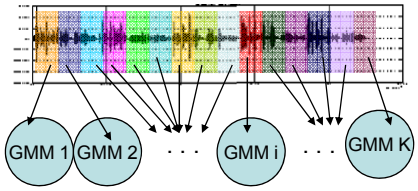
20/05/2009 - 41

EURECOM

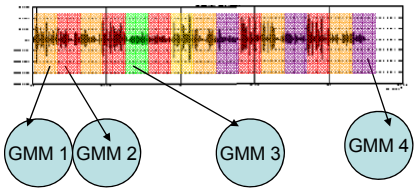
Some thoughts

- **Top Down and Bottom Up systems**

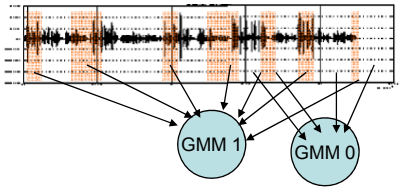
Bottom Up



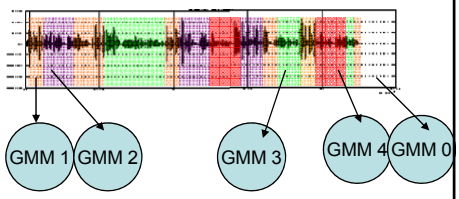
All models added together
Merge models



Top Down



Add models ITERATIVELY



20/05/2009 - 42

EURECOM

Some thoughts

- **Disavantages:**

Bottom Up System

- fix the maximum number of speakers **at the beginning**
- difficult to improve the initialization
- impurity shared among all models

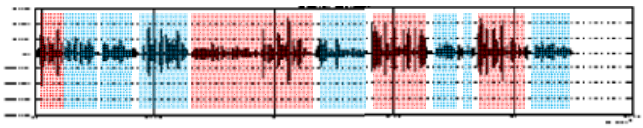
Top Down System

- purity of the first model affects the quality of other, sequentially added models
- possibility that subsequent model initializations could benefit from the removal of previous speakers' data.

Some thoughts

- **1st experiment:**

Original
Labels
(2 speakers)

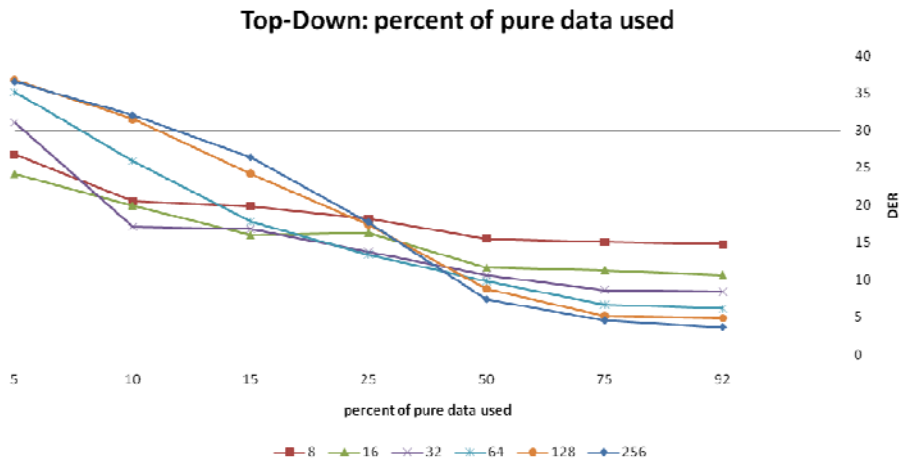


Tune each model with X % of the total speaker time

GMM 1 GMM 2

Example: (X =10 %)
Total speaker 1 time = 100 sec.
10 sec from speaker 1
To tune GMM 1

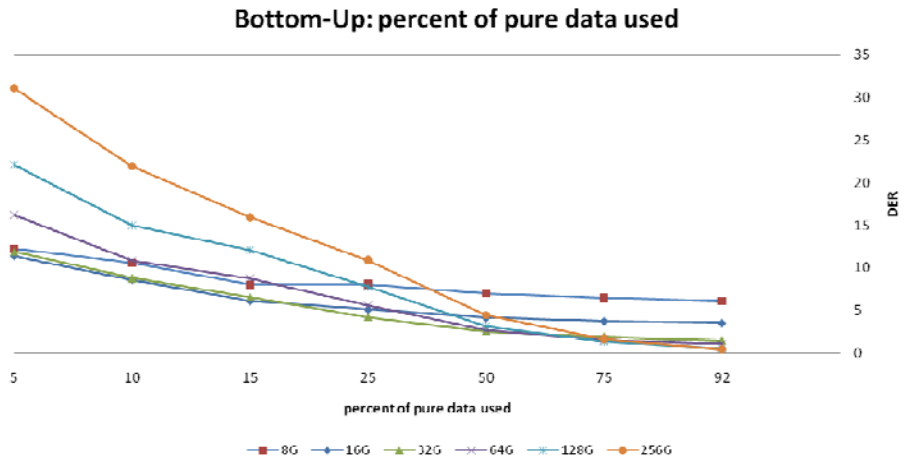
Theoretical Performance : Pure segments for initialization



20/05/2009 - 45

EURECOM

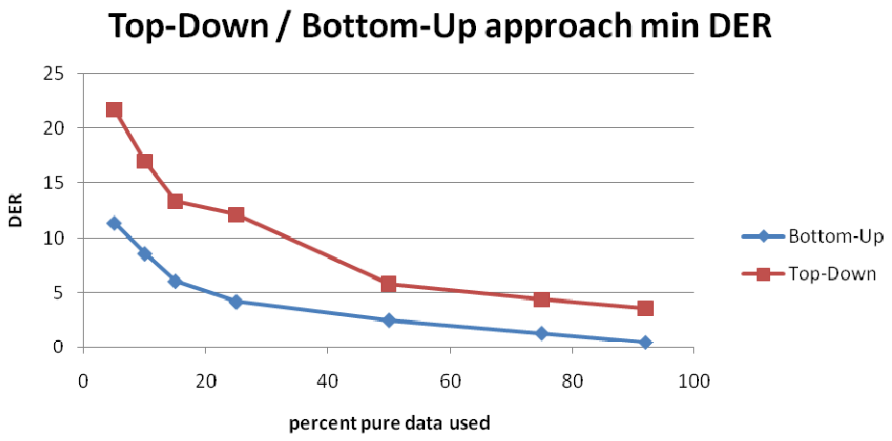
Theoretical Performance : Pure segments for initialization



20/05/2009 - 46

EURECOM

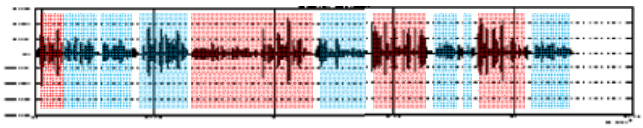
Theoretical Performance : Pure segments for initialization



Some thoughts

- **2nd experiment:**

Original Labels (2 speakers)

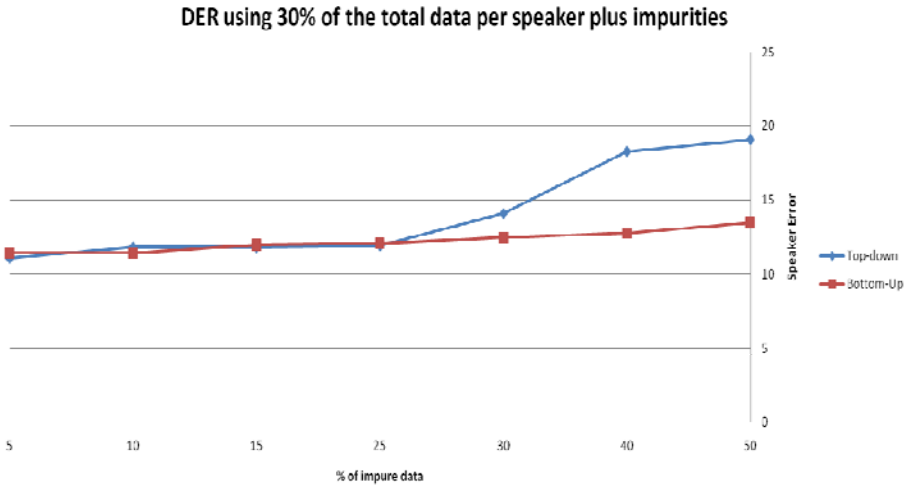


Tune each model with 30 % of the total speaker time + Y % impurities

GMM 1 GMM 2

Example: (10 % impurities)
Total speaker 1 time = 100 sec.
30 sec from speaker 1
0.1*30=3 sec from speaker 2
To tune GMM 1

Robustness to impurities



Top Down and Bottom Up systems Conclusion

- **Hypothesis:**
  - Pure initial data to tune the model
  - K number of speakers: known
- **Results:**
  - best theorical performance with **Bottom Up system** if we assume we know **the real number of speakers and when each speaks**
  - Up to 25 % of impure data**: same performance for both systems,**
  - beyond 25% : **Bottom Up is better.**
- But: we work with the references, the real number of speaker
  - theoretical experiments: the top-down approach may be « more realistic » : we learn something at each iteration
- SDM results show: top down leads to state-of-the-art results

Conclusion & future work

- **Absolute improvements:**
  - 35% relative DER for MDM
  - 13% relative DER for SDM
    - Top-down approach warrants continued attention
    - Expect better performance for MDM with incorporated delay features

- **Future Work:**
  - System quite unstable to different meetings and sensitive to speaker model initialization

→ Improved initialization strategies through top-down approach

EURECOM Sophia Antipolis

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

The LIA-EURECOM RT'09 Speaker Diarization System

LIA – Laboratoire Informatique d'Avignon (France)
EURECOM - Sophia Antipolis (France)

Corinne Fredouille,
Simon Bozonnet and Nick Evans
(corinne.fredouille)@univ-avignon.fr
(simon.bozonnet,nick.evans)@eurecom.fr