The LIA RT'07 speaker diarization system

Corinne Fredouille and Nick Evans
LIA - Computer Science Lab. of Avignon (France)
University of Avignon
(corinne.fredouille,nicholas.evans)@univ-avignon.fr

Context

Tasks	SDM	MDM	ADM
Conference	X	X	
Lecture	X	X	X
Coffee Break	X	X	X

- Speaker diarization task only
- System optimised for the conference subdomain
- Unchanged for lecture meetings and coffee breaks

Outline

- Baseline SAD and speaker diarization systems
- Post-evaluation experiments
- Between-channel delay features
- Conclusion and future work

Baseline SAD and Speaker diarization systems

Multiple distant microphones

- Still a simple sum of the multiple signals to get a unique signal to segment
- Attempts to use between-channel delay features, but ...

Speech Activity Detection

- Two-state HMM characterising speech and non-speech information :
  - 12MFCC+energy+Δ+ΔΔ, no normalization
  - 32 Gaussian components per state, trained on 2004 NIST/RT and ISL data
  - Transition probabilities equally balanced
- Iterative process => decoding and model adaptation until stability
- Minimum duration rules for non-speech segments only:
  - Primary systems => 0.3 seconds
  - Constrastive systems => 0.6 seconds

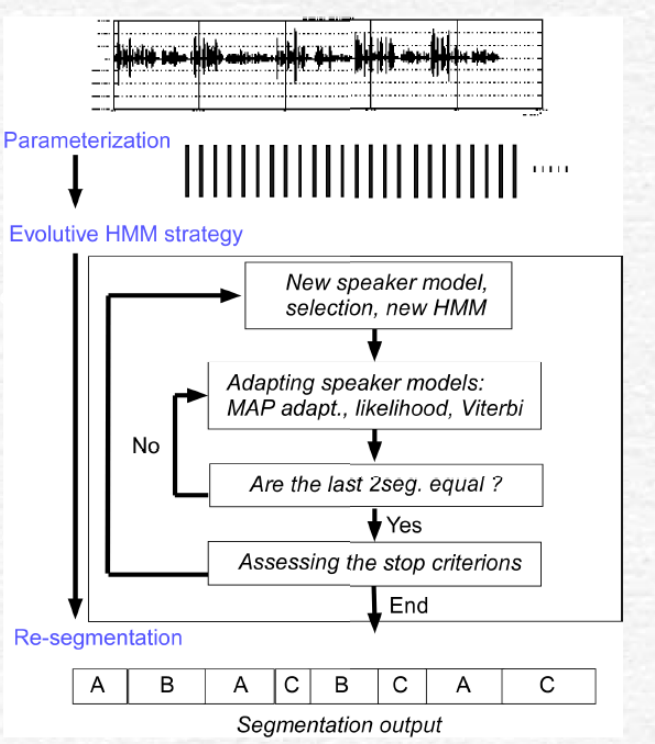
Baseline E-HMM system 1/3

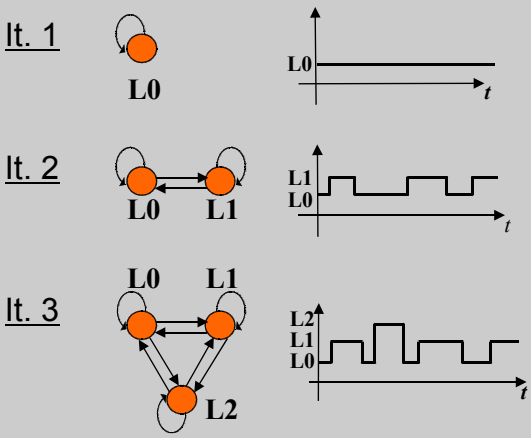
- LIA core system : **still** based on an E-HMM = integrated approach (1 step) :
  - a HMM representing the discussion between speakers
    - State = speakers
    - Transition = turn changes in discussion
- Step 1: Segmentation phase:
  - Iterative process permitting to build the E-HMM
- Step 2: Resegmentation phase:
  - Iterative process permitting to refine the segmentation output by deleting irrelevant speakers
- Step 3: Normalisation and resegmentation phase:
  - 16 LFCC+log Energy+Δ associated with a segmental 0-mean and 1-variance normalisation followed by a second resegmentation phase

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline
Post-eval ex
Delay feat
Conclusion

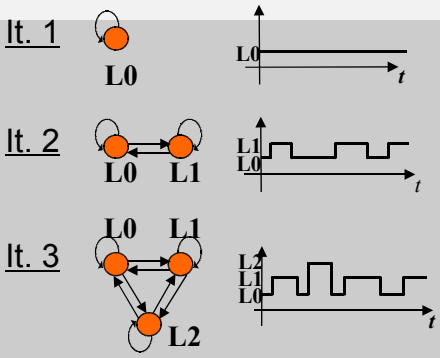
Baseline E-HMM system 2/3





- Add a new speaker (state) to the E-HMM at each iteration according to a selection technique
- GMM model adaptation / Viterbi decoding => evolutive segmentation

Baseline E-HMM system 3/3



- Parameterisation:
  - 20 LFCC + log. energy
  - No parameter normalization
- Adaptation model:
  - 128 Gaussian components for GMM speaker model
  - GMM Model adaptation from a generic model (world model)
  - MAP adaptation scheme
- Viterbi decoding
  - 30 frame minimum duration constraint decoding
- Selection technique for speaker addition
  - Return to Likelihood maximum criterion

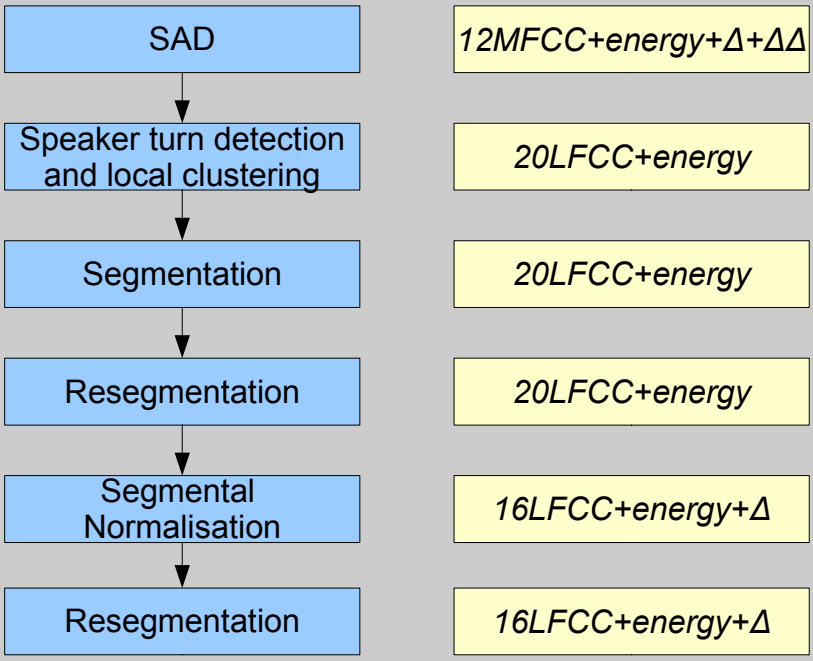
Complete speaker diarization 1/3

- Addition of a preliminary phase implementing a speaker turn detection and a local speaker clustering
- Objective: to provide an approximate pre-segmentation to initialise and speed-up the core segmentation phase
- Consequently, compared with previous LIA systems, here:
  - the segmentation phase is constrained to the boundaries present in the pre-segmentation output
  - the selection method handles the segments available in the pre-segmentation output (3s min.) => variable length !
  - the resegmentation remains unchanged => boundaries and speaker labels entirely re-examined

Complete speaker diarization 2/3

- Speaker turn detection:
  - Classical GLR criterion applied on 2 consecutive 0.5s long windows (0.05s step)
  - Speaker changes = relevant maximum peaks on the GLR curve
  - Single diagonal matrix Gaussian components
- Local clustering:
  - Aggregation of consecutive segments, still based on the GLR criterion associated with a decision threshold
  - Single diagonal matrix Gaussian components

Complete speaker diarization 3/3



System evaluation

- SAD and speaker diarization systems developed on the RT'05 and RT'06 datasets (used separately)
- Speaker diarization system optimised:
  - on the conference subdomain
  - on the forced alignment references provided by ICSI mainly
  - Without taking overlapping segments into account

Development results

Show	Missed	FAlarm	Speaker	Overall
RT'05				
AMI_20041210	1.0	0.9	1.3	3.2
AMI_20050204	3.4	0.9	33.3	37.7
CMU_20050228	11.1	0.9	5.7	17.7
CMU_20050301	3.3	1.8	13.0	18.1
ICSI_20010531	6.3	3.0	13.0	22.4
ICSI_20011113	8.0	2.5	29.1	39.6
NIST_20050412	6.8	3.8	1.9	12.4
NIST_20050427	2.9	6.1	6.9	15.9
VT_20050304	0.7	1.1	8.9	10.7
VT_20050318	3.2	2.2	25.8	31.2
RT'05_average	4.6	2.3	13.3	20.2

Show	Missed	FAlarm	Speaker	Overall
RT'06				
CMU_20050912	11.1	6.4	10.0	27.5
CMU_20050914	9.8	3.0	4.3	17.1
EDI_20050216	5.0	1.5	21.6	28.1
EDI_20050218	4.4	2.5	10.7	17.6
NIST_20051024	6.6	1.7	8.7	17.0
NIST_20051102	5.1	3.5	21.3	29.9
VT_20050623	4.6	7.4	3.5	15.5
VT_20051027	3.2	2.9	11.0	17.13
RT'06_average	6.4	3.6	11.6	21.5

- Stable performance for SAD
- Much greater variation for speaker error rates
- Average overall DER more stable

Baseline
Post-eval exp.
Delay features
Conclusion

Evaluation results 1/2

Show	Missed	FAlarm	Speaker	Overall
RT'07				
CMU_20061115-1030	7.4	4.6	9.7	21.8
CMU_20061115-1530	3.3	5.1	14.5	23.0
EDI_20061113-1500	8.9	0.8	22.8	32.5
EDI_20061114-1500	3.2	1.8	23.3	28.4
NIST_20051104-1515	3.8	0.9	7.6	12.2
NIST_20060216-1347	2.5	1.4	20.9	24.8
VT_20050408-1500	1.5	0.6	36.9	39.0
VT_20050425-1000	5.5	0.7	3.7	9.9
RT'07_average	4.5	2.0	17.7	24.2

- Stable performance for SAD compared with dev.
- Sill greater variation for speaker error rates depending on the meeting
- Overall speaker error rate greater than dev. results

Evaluation results 2/2

Subdomain	Mic. Cond.	Missed	FAlarm	Speaker	Overall
Conference meeting	MDM	4.5	2	17.7	24.2
	SDM	4.7	2.1	17.7	24.5
Lecture meeting	ADM	4.1	7.2	19.3	30.5
	MDM	3.4	6.9	20.9	31.2
	SDM	3.6	6.5	19.4	29.5
Coffee break	ADM	3.5	3.6	19.2	26.4
	MDM	3	5	17.5	25.5
	SDM	3.3	4.6	18.4	26.3

- Small difference in performance between the different microphone conditions
- Speaker diarization system not effective in utilising additional information providing by the multiple channels

Post-evaluation experiments

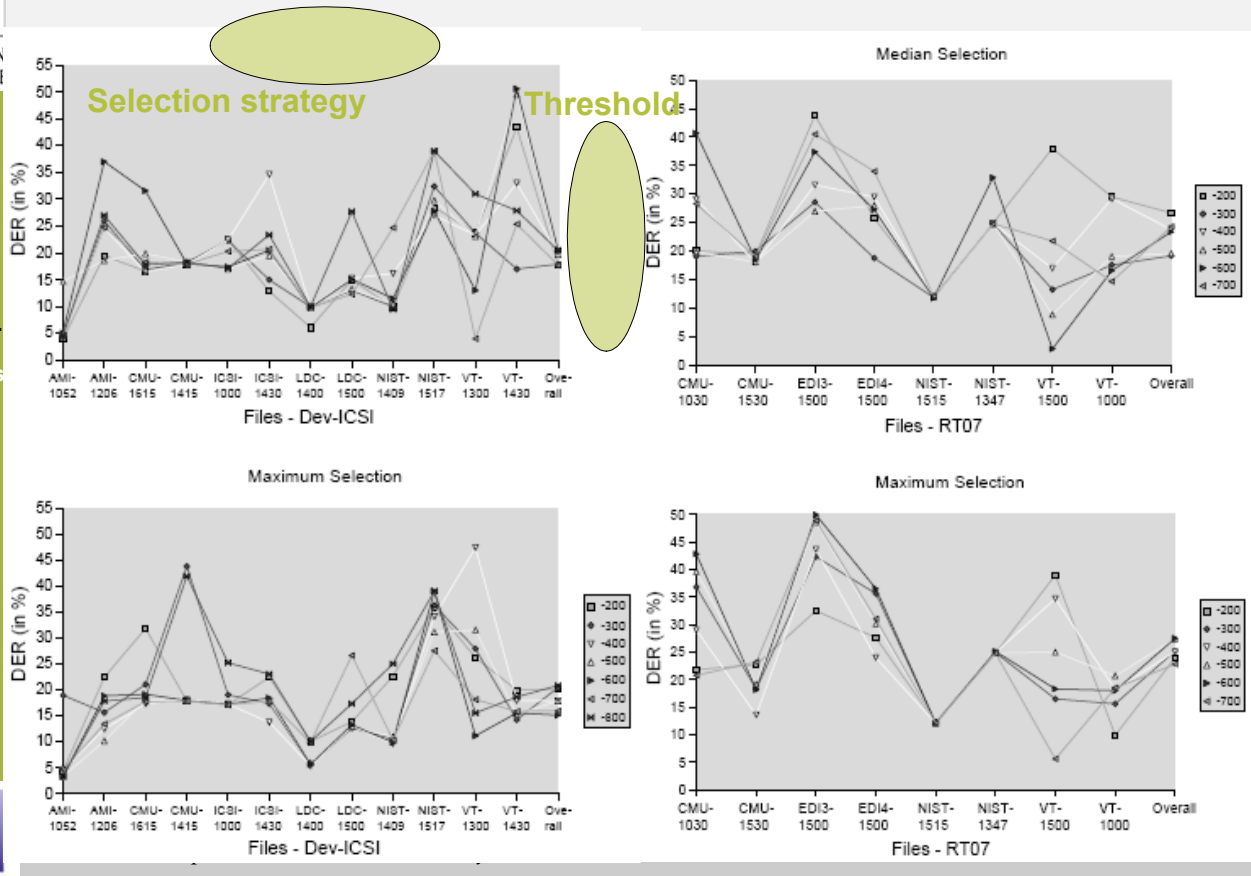
Result analysis 1/2

- Changes between 2006 and 2007 => mainly the addition of the pre-segmentation phase: speaker turn detection + clustering
- Clustering may influence largely the segmentation (and resegmentation) steps:
  - Boundaries issued from clustering kept for the segmentation phase
  - Segments (min. 3s long) issued from clustering directly handled by the selection method used to add speakers

Result analysis 2/2

- Post-evaluation experiments:
  - Clustering threshold variation
  - Selection strategy:
    - Maximum likelihood criterion
    - « Median selection » => selection based on the segment for which the likelihood is close to the likelihood mean computed along all the segments available (Min. 3s long)
  - Experimental datasets:
    - RT'07: conference subdomain (MDM)
    - Dev-ICSI: Development dataset of ICSI (ICSLP'2006) => comparable results and « more stable » dataset !

Result comparison



Baseline
Post-eval exp.
Delay features
Conclusion

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline

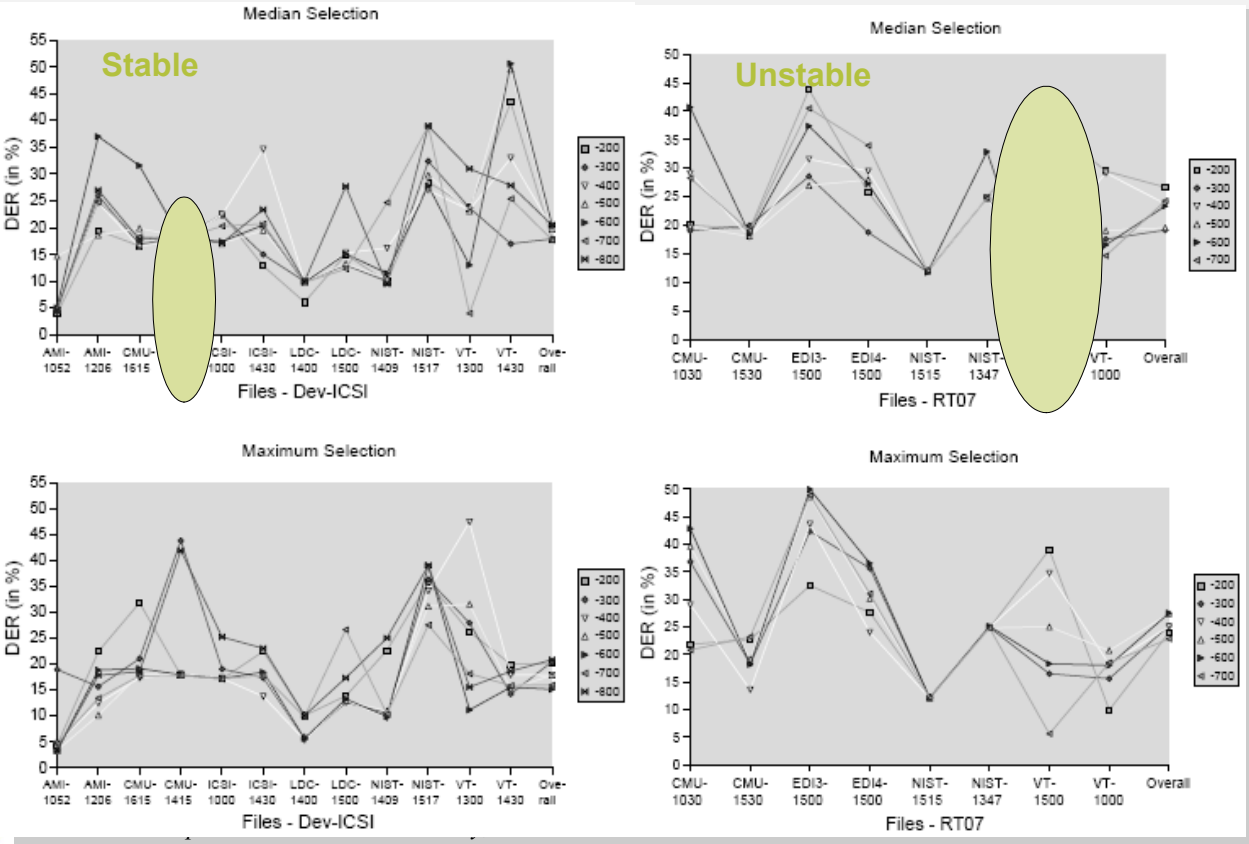
Post-eval exp.

Delay features

Conclusion

Laboratoire d'Informatique Université d'Avignon

Result comparison: sensitivity to threshold variation



UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline

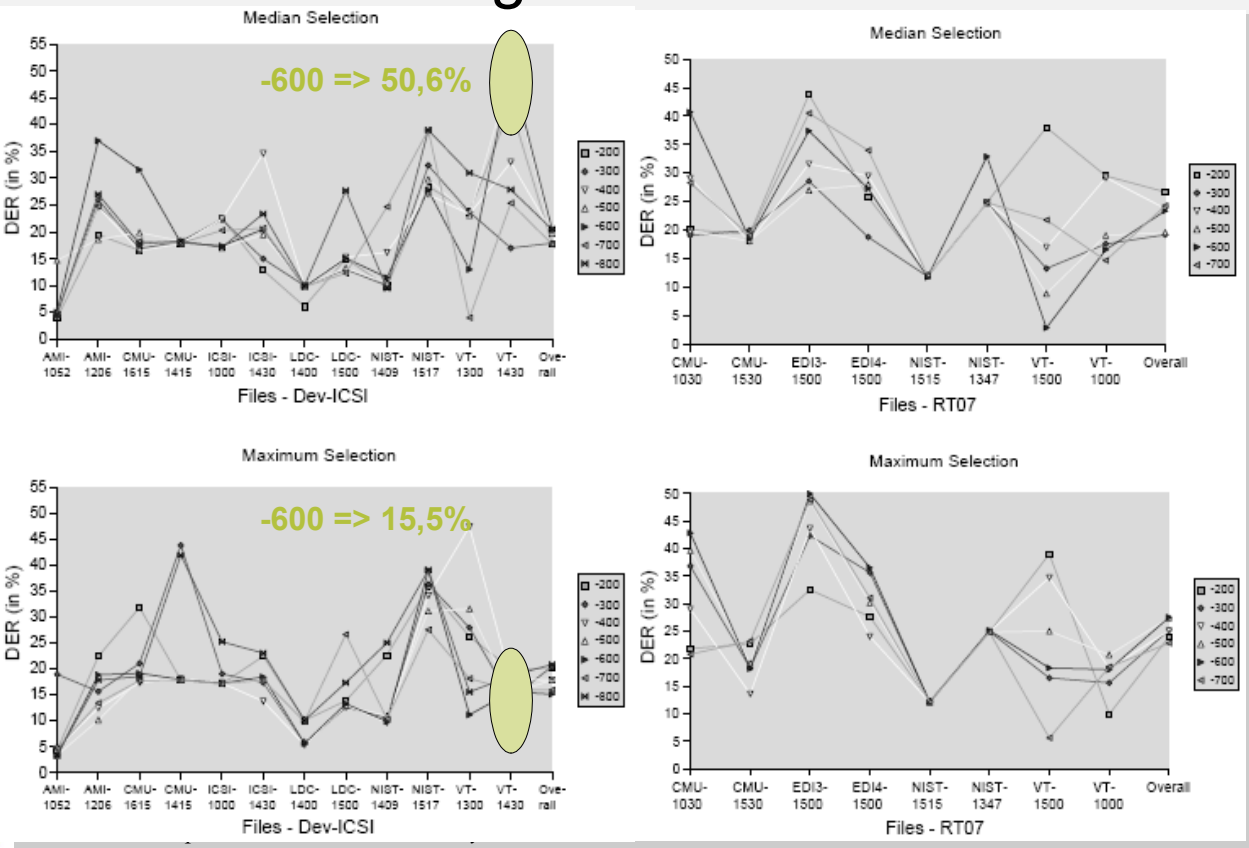
Post-eval exp.

Delay features

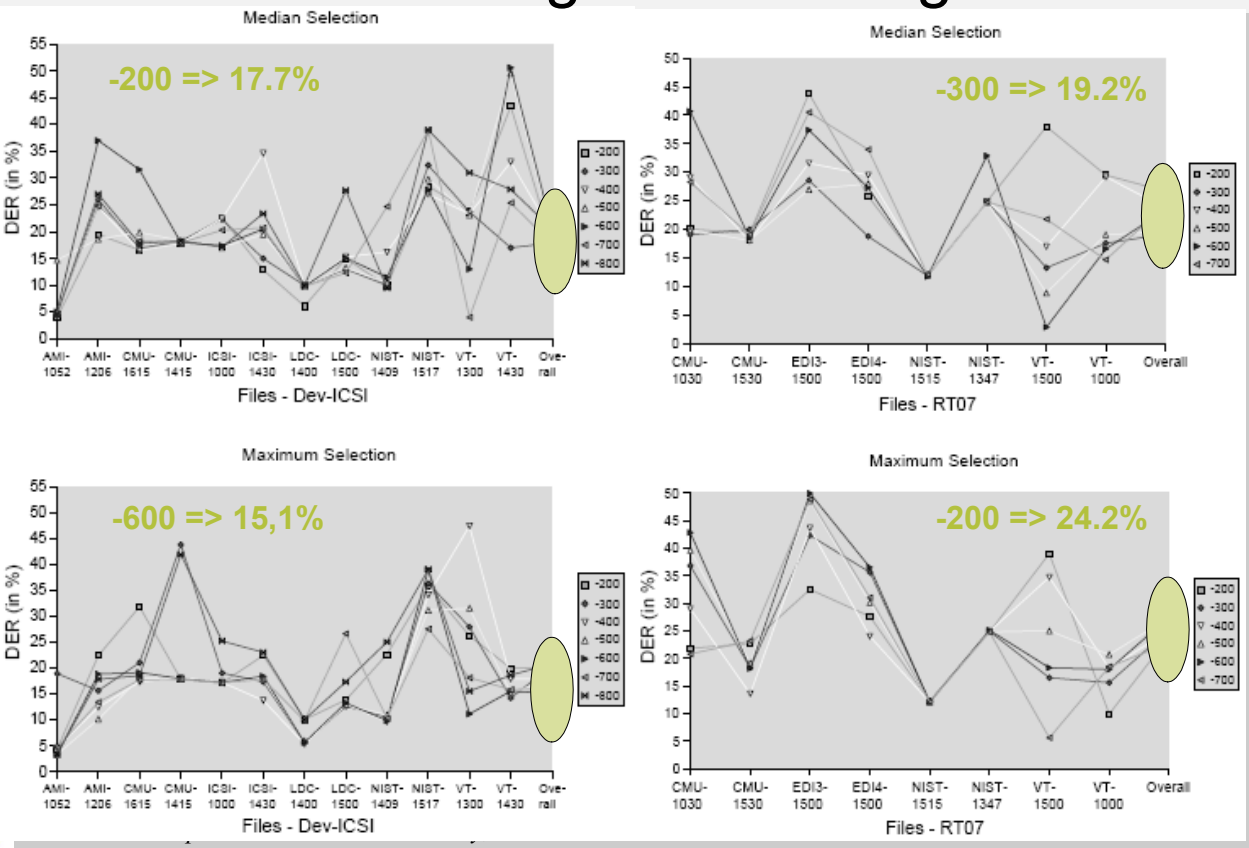
Conclusion

Laboratoire d'Informatique Université d'Avignon

Result comparison: Behaviour of selection strategies



Result comparison: Optimal threshold variable according to the configuration



UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline
Post-eval exp.
Delay features
Conclusion

Summary

- Objective of the pre-segmentation step => to speed-up the segmentation phase
  - Previous system=3*RT, current system=0,25*RT
  - Goal reached !
- Experiments performed on development datasets (RT'05 and RT'06) => no loss of performance
- On RT'07 and Dev-ICSI, overall speaker diarization system unstable according to the clustering threshold, datasets and selection strategies !

Between-channel delay features

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline
Post-eval exp.
Delay features
Conclusion

Experiments with delay features

Show	SDM ref fake	Auto ref fake	D real	AD real
RT'06				
CMU_20050912	7.8	7.8	55.6	33.8
CMU_20050914	3.3	3.3	28.0	22.1
EDI_20050216	25.0	25.0	48.1	26.0
EDI_20050218	43.7	43.7	50.4	17.6
NIST_20051024	10.8	2.2	24.4	46.3
NIST_20051102	2.3	4.8	42.7	46.6
VT_20050623	6.5	13.7	43.5	15.3
VT_20051027	22.7	22.7	33.7	29.6
Average	15.3	15.2	40.8	30.5

Conclusion

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Baseline
Post-eval exp.
Delay features
Conclusion

Conclusion

- SDM => performance not far from the best system !
- MDM => huge performance gap with the best system !

- LIA speaker diarization system:
  - Not effective in utilising multiple channel information
  - Not designed for utilising between-channel delay features

- Future work ???!!!! => to deal with these issues !