Speaker diarization for meeting recordings

Corinne Fredouille
LIA - Computer Science lab. of Avignon (France)
University of Avignon
(corinne.fredouille@univ-avignon.fr)

Context

- Who spoke when ?
- No information on speaker identities nor on the number of speakers
- LIA lab. involved in this task since 2000 (Sylvain Meignier thesis):
  - Telephone conversations
  - Broadcast News shows
  - Meetings

Context

- Eval. 2006 => **Speaker diarization** and **Speech Activity Detection** tasks on **multiple distant microphones** (MDM) only !
- SAD task because it is required for speaker diarization system
- System developments mainly done on conference data, on non-overlap areas
- Just few run on lecture data
- A first proposal to handle overlap areas

Outline

- Baseline SAD and speaker diarization systems
- Technical progress from 2005
- Attempt to deal with overlap areas
- Conclusion

Baseline Speaker Diarization and SAD systems

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Multiple distant microphones

- Still a simple sum of the multiple signals to get a unique signal to segment
- Use of multi-channel information only in the system devoted to overlap areas

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Speech Activity Detection

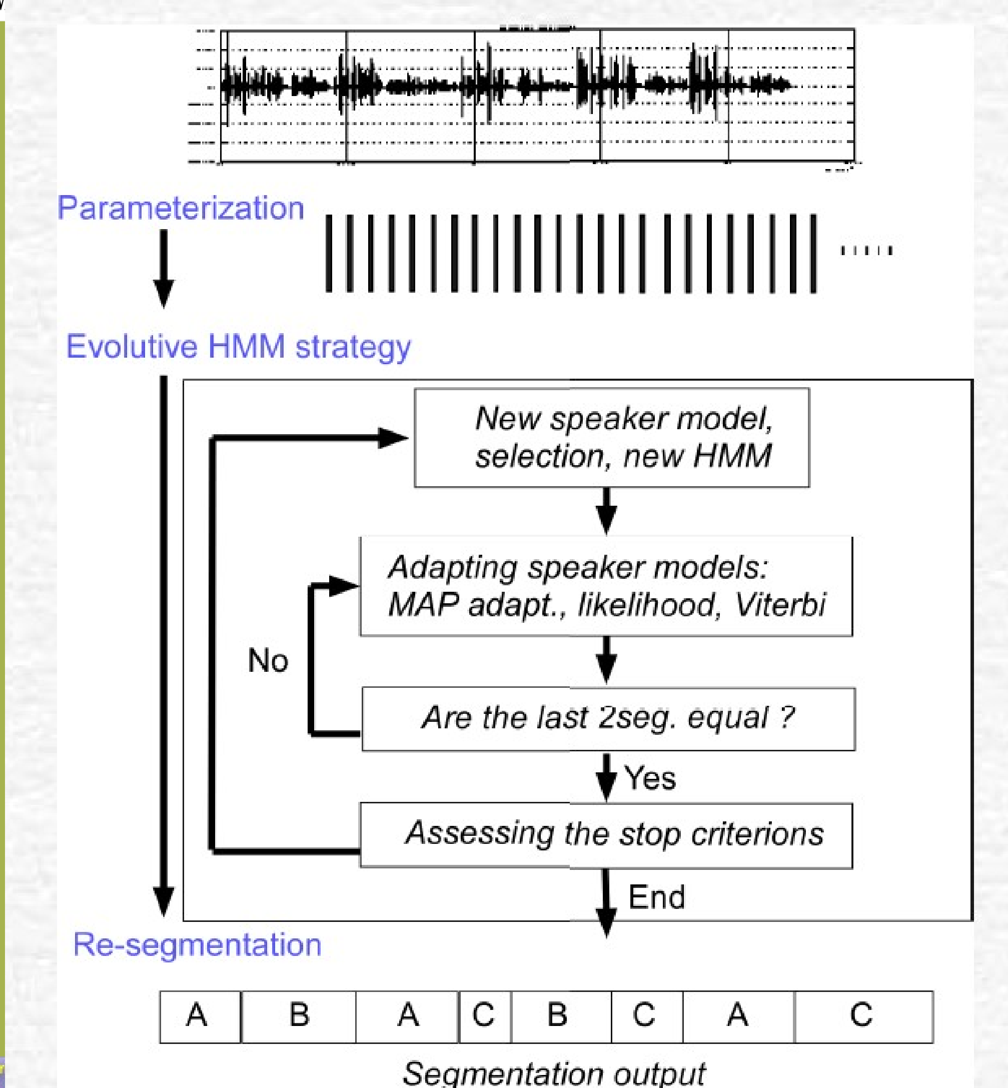
- Simpler technique than 2005
  - In 2005 => Energy based detection applied on each individual microphone signal + merging algorithm
- In 2006, two-state HMM representing speech/non speech information :
  - 13MFCC+Δ+ΔΔ, no normalization
  - 64 Gaussian components per state, trained on 2004 NIST/RT and ISL data
  - Transition probability equally balanced
  - Viterbi decoding (5 frame duration constraint)
  - Rules on min. segment length for both speech and non speech
- Tuned on 2005 conference eval. data only

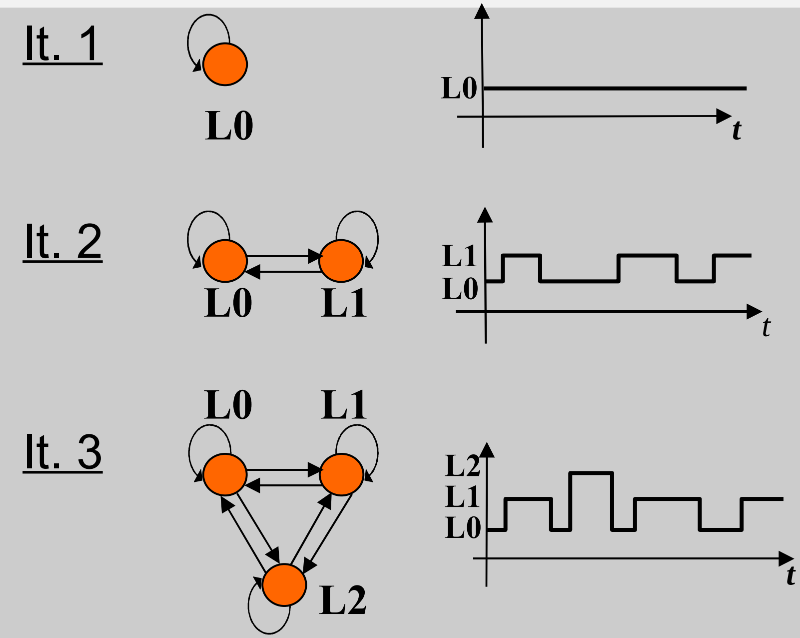
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

LIA Speaker diarization system

- Classically, 2 steps:
  - Speaker turn detection
  - Speaker clustering
- LIA system : E-HMM = integrated approach (1 step) based on:
  - a HMM representing the discussion between speakers
    - State = speakers
    - Transition = turn changes in discussion
  - Iterative process permitting to build the HMM

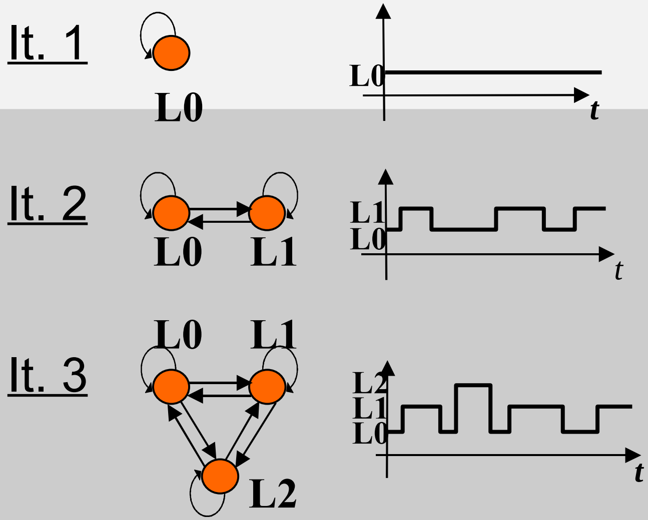
Baseline E-HMM system





- 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



- Parameterization:
  - 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 *(see just later)*

Baseline E-HMM system (cont'd)

- 2005 E-HMM based system:
  - Still unstable: strongly dependent on the quality of data due to adaptation scheme
  - Tends to an under-segmentation:
    - Detects the largest speakers, but misses the smallest ones (largest speakers may include other smaller speakers)
  - Reasons ?
    - Adaptation technique ?
    - Selection technique ?
    - Not enough control ?

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Technical progress from 2005

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Selection technique

- Used to add new speaker in the E-HMM
- Bad selection leads to bad speaker
- Initially,
  - involving L0 speaker (multi-speaker) only
  - based on the maximization of the likelihood ratio over all the 6s long segments, issued from L0

$$Max_{(all\,S_x)}[LogL(S_i/M_{L_0}) - LogL(S_i/M_{World})]$$

<u>Idea : to use speakers present in the E-HMM (other than L0) in the selection scheme</u>

Laboratoire d'Informatique Université d'Avignon

Selection technique (cont'd)

- Involving **ALL** the existing speakers
- Selection of segments close to L0 and far from Lx => « Discriminant » selection

$$Max_{(all\, S_x)}[LogL(S_i/M_{L_0}) - Mean_{(all\, L_x)}(LogL(S_i/M_{L_j}))]$$

- Also, selection of best frames in 6s long segments

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Purification of segments

- Viterbi decoding/model adaptation iterations => impure segments in terms of speaker homogeneity
- <u>Idea: purify segments before adding a new speaker</u>

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Purification of segments (cont'd)

- Purification based on modified BIC criterion inspired from ICSI segment purification technique
- Applied before adding a new speaker
- For each speaker Lx (not L0):
  - Find the best segment (LLR maximization): S_{bestx}
  - Compare this segment with all other ones according to the BIC criterion
  - If BIC value between segment S_{bestx} and $S_{xi} > 0$, keep S_{xi} in Lx
  - Else move S_{xi} in L0

Laboratoire d'Informatique Université d'Avignon

Purification of segments (cont'd)

- BIC criterion:
  - 5 Gaussian components for GMM representing separate sources (segments)
  - 10 Gaussian components for GMM representing both sources together

  => no complexity model penality required !

  - Purification scheme applied on 2s minimum duration segments only

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Normalization of parameters

- Basically, parameterization is :
  - 20 LFCC + log Energy
  - No derived coefficient (**Δ nor ΔΔ)**
  - No normalization of coefficients since channel information may be useful for segmentation process

- <u>Idea: Normalize the coefficients using the segmentation issued from the E-HMM, but after the re-segmentation phase</u>

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Normalization of parameters (cont'd)

- For each segment issued from the segmentation output:
  - Compute the mean and variance of coefficients associated with
  - Normalize these coefficients (0-mean, 1-variance)
- Apply once again the re-segmentation phase
- Also, 16 LFCC+log Energy+**Δ** => 34 coeff.

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Protocols

- 2006 Spring NIST/RT evaluation campaign
  - Meetings data collected at numerous sites equipped with different kinds of audio devices: head micro., **table micro.**, micro. arrays...
  - Three main tasks: Speech-To-Text, **Speaker Diarization**, **Speech Activity Detection**
  - Two sub-evals:
    - Conference room: 9 meetings of about 18mn each collected at 6 different sites
    - Lecture/seminar room: 38 seminars of 5mn each collected at 5 different sites

Laboratoire d'Informatique Université d'Avignon

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

Protocols (cont'd)

- Our development set :
  - Issued from the 2005 conference room sub-eval
  - 10 meetings of 12mn each, collected at 5 different sites (AMI, CMU, ICSI, NIST, VT)
  - Focused task : distant table microphones without taking the overlap areas into account (except for the dedicated system obviously)

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

The Speaker diarization system

- Speech activity detection based on a simple two-state HMM (64 Gaussian, 13MFCC+Δ+ΔΔ) trained on speech and non-speech signals
- Simple sum of signals issued from the different table microphones => **only one signal to segment**
- Baseline E-HMM with the different improvements (used separately or not)

Laboratoire d'Informatique Université d'Avignon

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

SAD results

2 state HMM	Mis. Speech Err.	FA Speech Err.	Overall Err.
Conf. Dev.	2.0	2.8	4.8
Conf. Eval.	0,5	4,2	4,7
Lecture Eval.	0	13,0	13,0

- LIA system tuned on Conference data (dev. set)
- More non-speech portions for the eval. than for the development set especially for CHIL data (more than 50% of non speech for one lecture file)
- Disturbing for the lecture eval since the speaker diarization scoring is strongly dependent on the SAD performance

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Dev. Set – Conference room
(no overlap areas)

Approach	Mis. Spk Err.	FA Spk Err.	Spk Err.	Spk Diariz. Err.
2005 System	*4,0*	*3,0*	*20,6*	*27,6*
Baseline (bug fixed)	2,0	2,8	17,7	22,5
Bas.+Normalization	2,0	2,8	13,3	18,1
Bas.+Selection	2,0	2,8	14,3	19,1
Bas.+Selection+Norm.	2,0	2,8	11,3	16,1
Bas.+BIC purif.	2,0	2,8	16	20,8
Bas.+BIC purif.+Norm.	2,0	2,8	13,1	17,9
Bas.+Sel.+BIC purif.	2,0	2,8	19,8	24,6
Bas.+Sel.+BIC purif.+Norm.	2,0	2,8	15,9	20,7

- Selection + Normalization => Best improvement
- Purification based system (BIC) outperforms the Baseline, but not the Baseline+Selection
- Unfortunately, no improvement with combination

Laboratoire d'Informatique Université d'Avignon

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

Eval. Set – Conference room (no overlap areas)

Approach	Mis. Spk Err.	FA Spk Err.	Spk Err.	Spk Diariz. Err.
2005 System	X	X	X	X
Baseline	0,6	6,9	31,9	39,4
Bas.+Normalization	0,6	6,9	24,5	32,0
Bas.+Selection	0,6	6,9	24,1	31,6
Bas.+Selection+Norm.	0,6	6,9	19,0	26,5
Bas.+BIC purif.	0,6	6,9	29,7	37,2
Bas.+BIC purif.+Norm.	0,6	6,9	20,2	27,7
Bas.+Sel.+BIC purif.	0,6	6,9	25,5	33,0
Bas.+Sel.+BIC purif.+Norm.	0,6	6,9	19,7	27,2

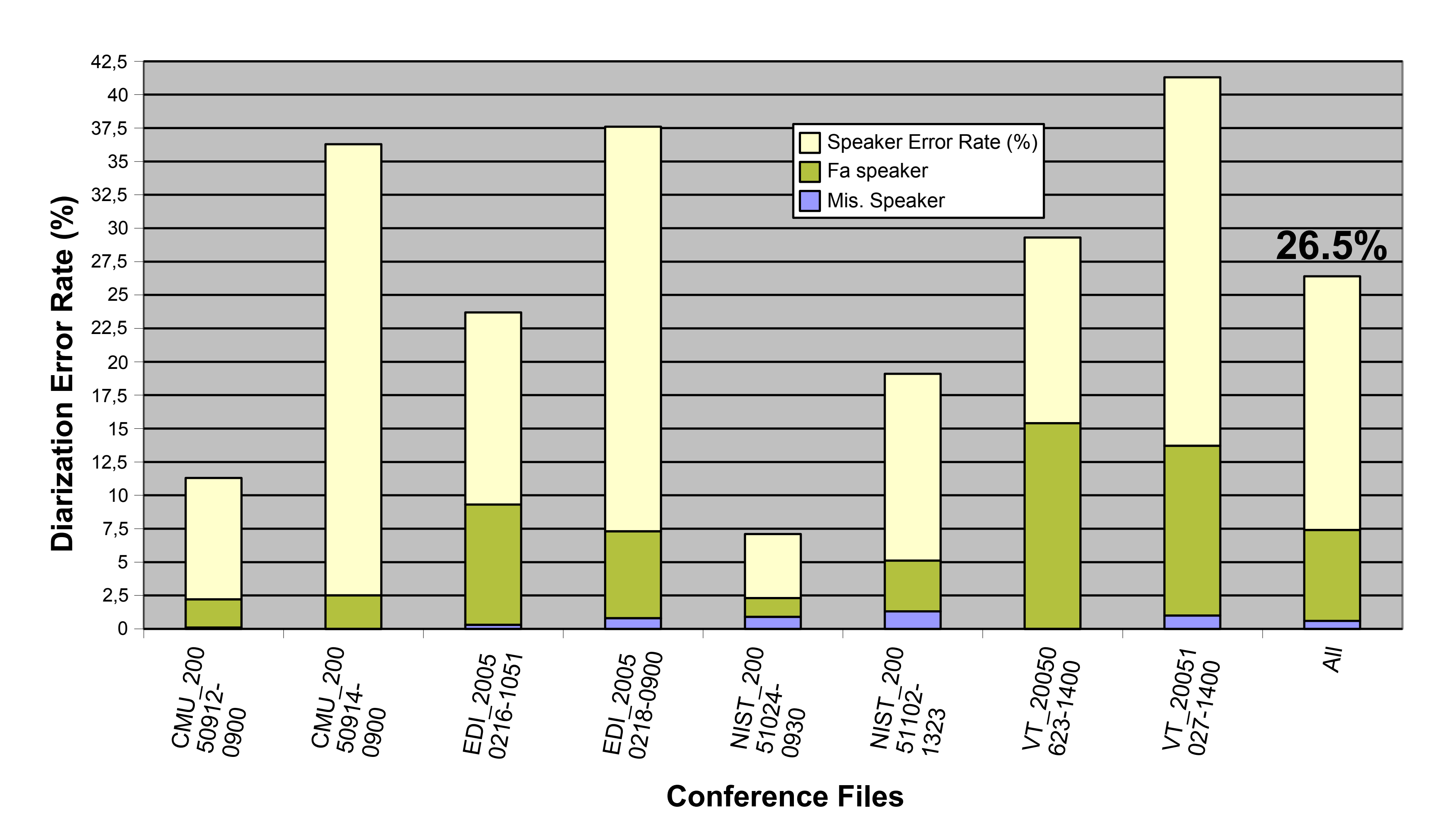
- Same remarks as with the dev set
- But a strong decrease of performance
- More difficult data ? Overfitting ? => difficult to answer yet !

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Eval. Set – Conference room (no overlap areas) - Bas.+Sel.+Norm

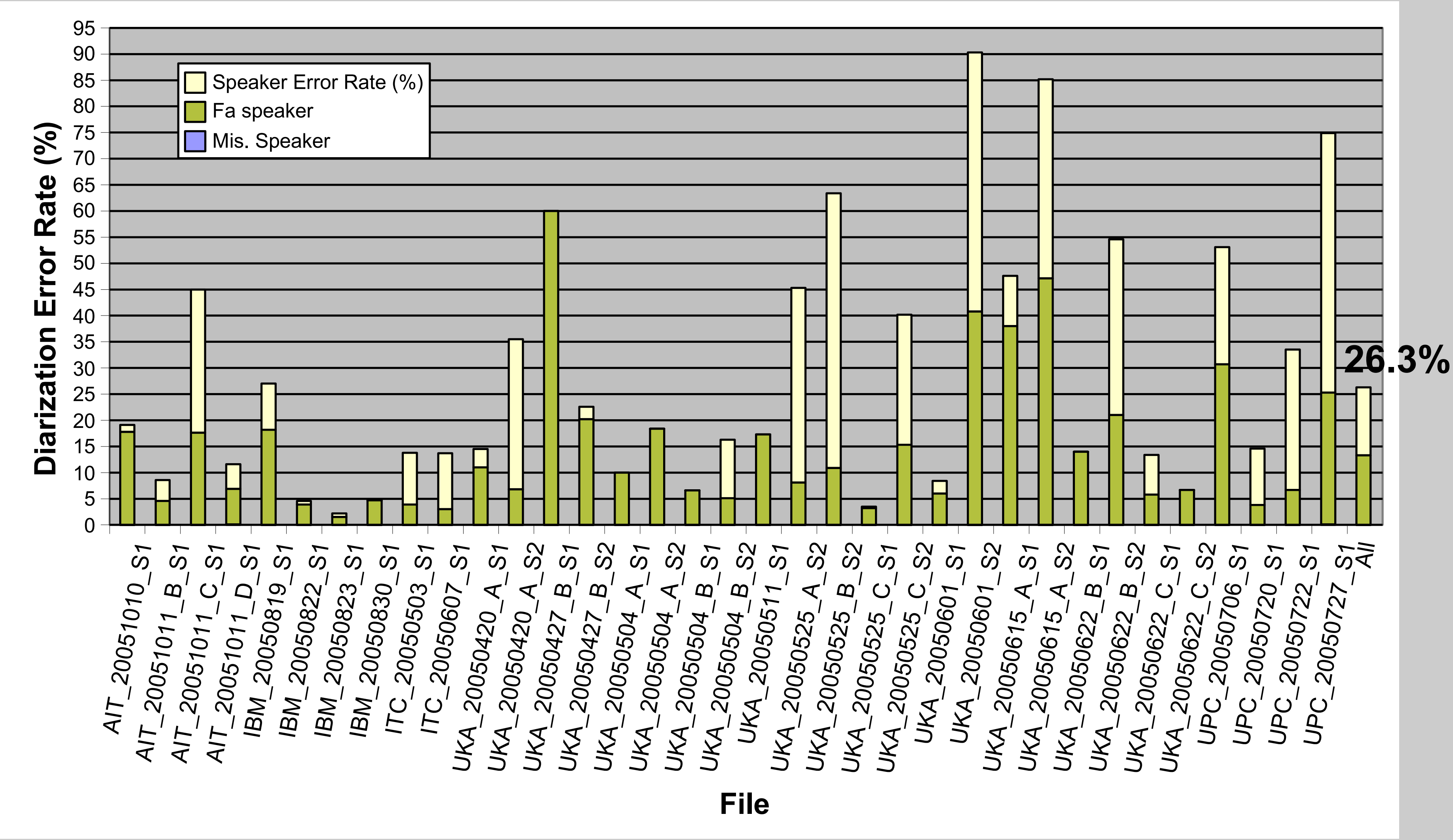
- Large difference even on a same site !!!



UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Eval. Set – Lecture room
(no overlap areas) - Bas.+Sel.+Norm



Laboratoire d'Informatique Université d'Avignon

Attempt to deal with overlap areas

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Context

- Primary condition: to deal with overlap areas
- Very challenging task: overlap areas look like new different speakers for the automatic systems (the LIA system does !) !

- <u>Idea : to use the output segmentation yielded on the unique signal to look for overlap areas over the multiple distant microphone signals</u>

Algorithm

- Assumption: all the speakers are already present in the E-HMM. Processing individual channels may help to distinguish people speaking together but near to different microphones (not always applicable)
- For each meeting:
  - Apply the speaker diarization system on the summed channel signal (Segmentation+ReSeg.)
  - Apply a resegmentation step on each individual channel signal, initiated from the unique signal segmentation
  - Merge the different segmentations by discarding redundant segments

Overlap area devoted system

Approach	Mis. Spk Err.	FA Spk Err.	Spk Err.	Spk Diariz. Err.
Bas.+Selection+Norm.	19,9	4,4	14,5	38,8
Bas.+Selection+Norm.+Overlap	17,6	8,9	14,5	41,0

- Decrease of mis. speaker error hidden by a strong increase of false alarm speaker error:
  - Due to speech/non speech detection issue
  - Non speech zones (misclassified by SAD system) are unfortunately not attributed to the same speaker depending on the individual signal processed

Conclusion

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

Conclusion

- RT'06: disappointing evaluation regarding the progress observed on the dev. Set
  - *even if we have been first on the lecture eval. on both SAD and speaker diarization tasks for three weeks thanks to corrupted references !*
- Speaker diarization improvement proposal (selection technique, purification, normalization) are rather promising especially when the combination will succeed

Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Conclusion (cont'd)

- Still a lot of work !
  - To make the LIA system more robust (adaptation techniques !!)
  - To work on more robust SAD techniques
  - To improve the LIA approach to deal with overlap zones (currently, worse than the standard system !):
    - By improving SAD
    - By taking a better benefit of multi-channels (Still !)
    - By incorporating external information (source localization ?)

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

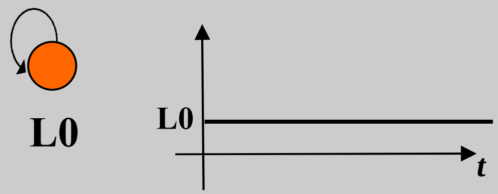
Thank you very much

Any questions ?

E-HMM steps

Stage 1: adding speaker L0

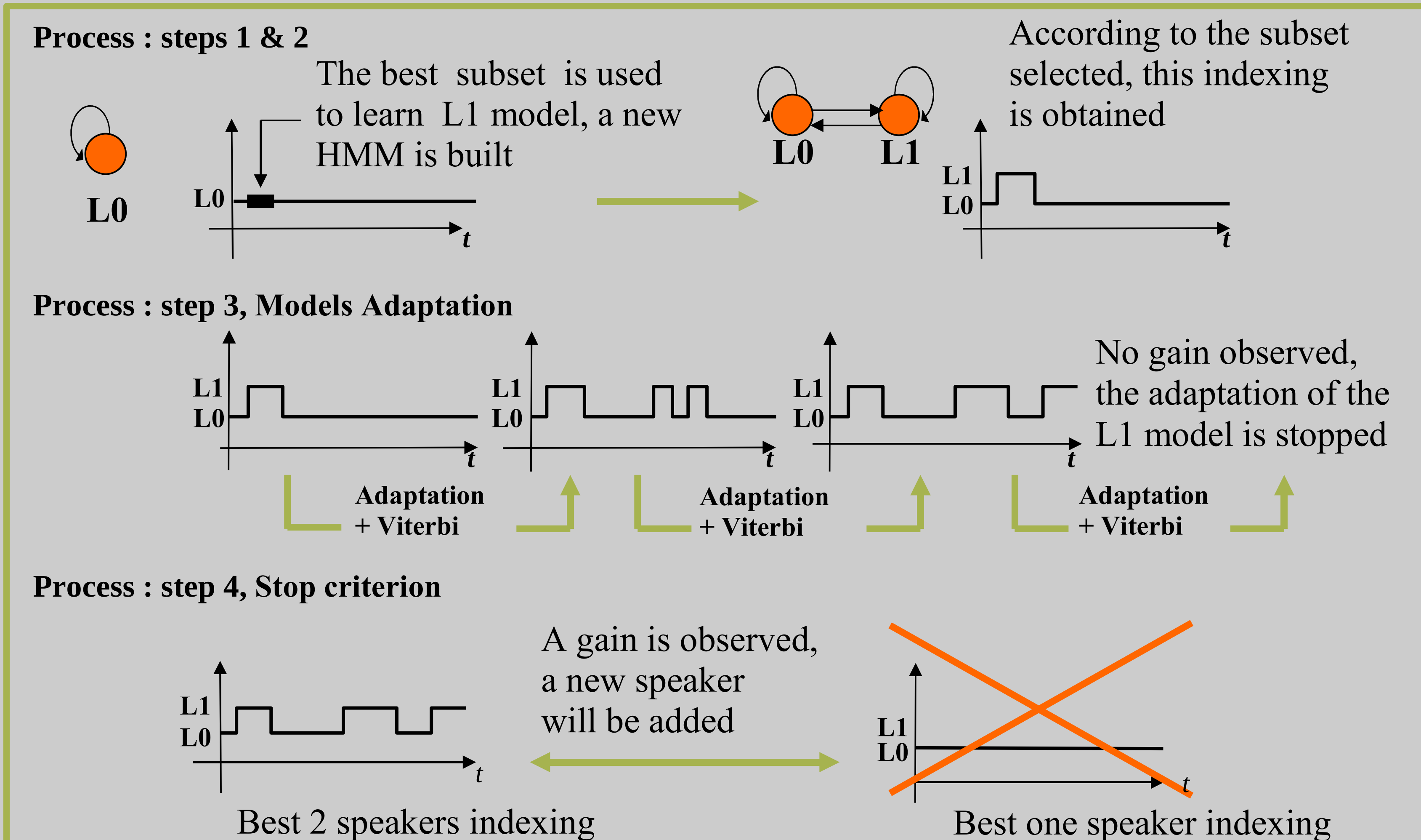
Process initialization



UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

E-HMM steps (cont'd)

Stage 2: adding speaker L1

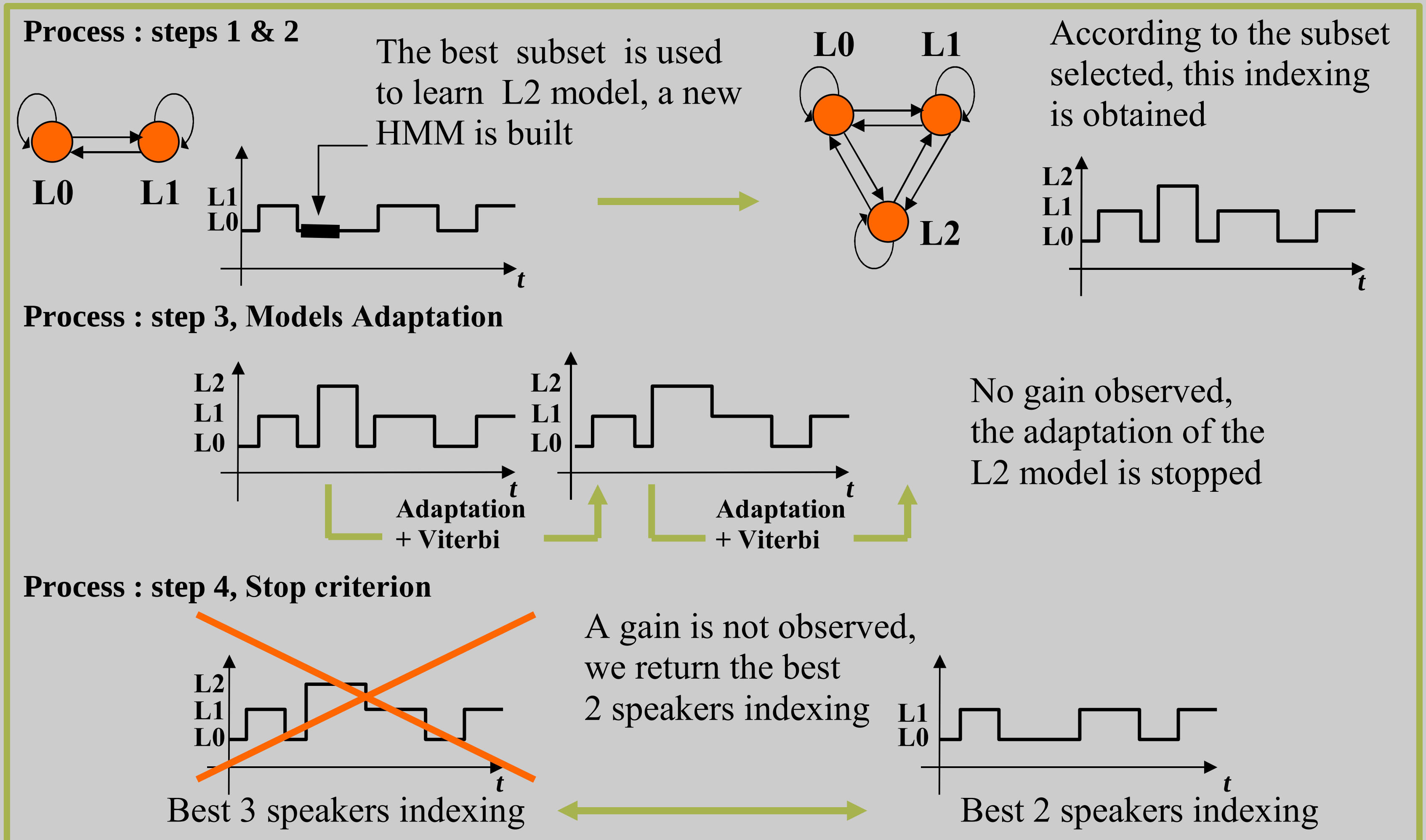


Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

E-HMM steps (cont'd)

Stage 3: adding speaker L2



Laboratoire d'Informatique Université d'Avignon

Results on eval. Lecture set (no overlap areas) (cont'd)