NIST RT'05S evaluation

Pre-processing techniques and speaker diarization on multiple microphone meetings

CLIPS
Communication Langagière et Interaction Personne-Système
CNRS - INPG - UJF
BP 53 - 38041 Grenoble Cedex 9 - France

Laurent Besacier

lium

Sylvain Meignier

Laboratoire d'Informatique
Université d'Avignon

Dan Istrate
Corinne Fredouille
Jean-François Bonastre

Outline

- ✗ Context and objectives
- ✗ System overview
- ✗ Some results
- ✗ Conclusion and Perspective

Context & objectives (1)

✗ Collaboration between three French labs => LIA, CLIPS and LIUM (ELISA consortium):

- ✔ Speaker diarization in multiple microphone environment
- ✔ Participants at the French eval. campaign on BN (ESTER campaign – January 2005)
- ✔ New implementation for LIA and LIUM systems
- ✔ Systems tuned on BN data only
- ✔ All the systems ran at the LIA

✗ SAD system developed at the LIA

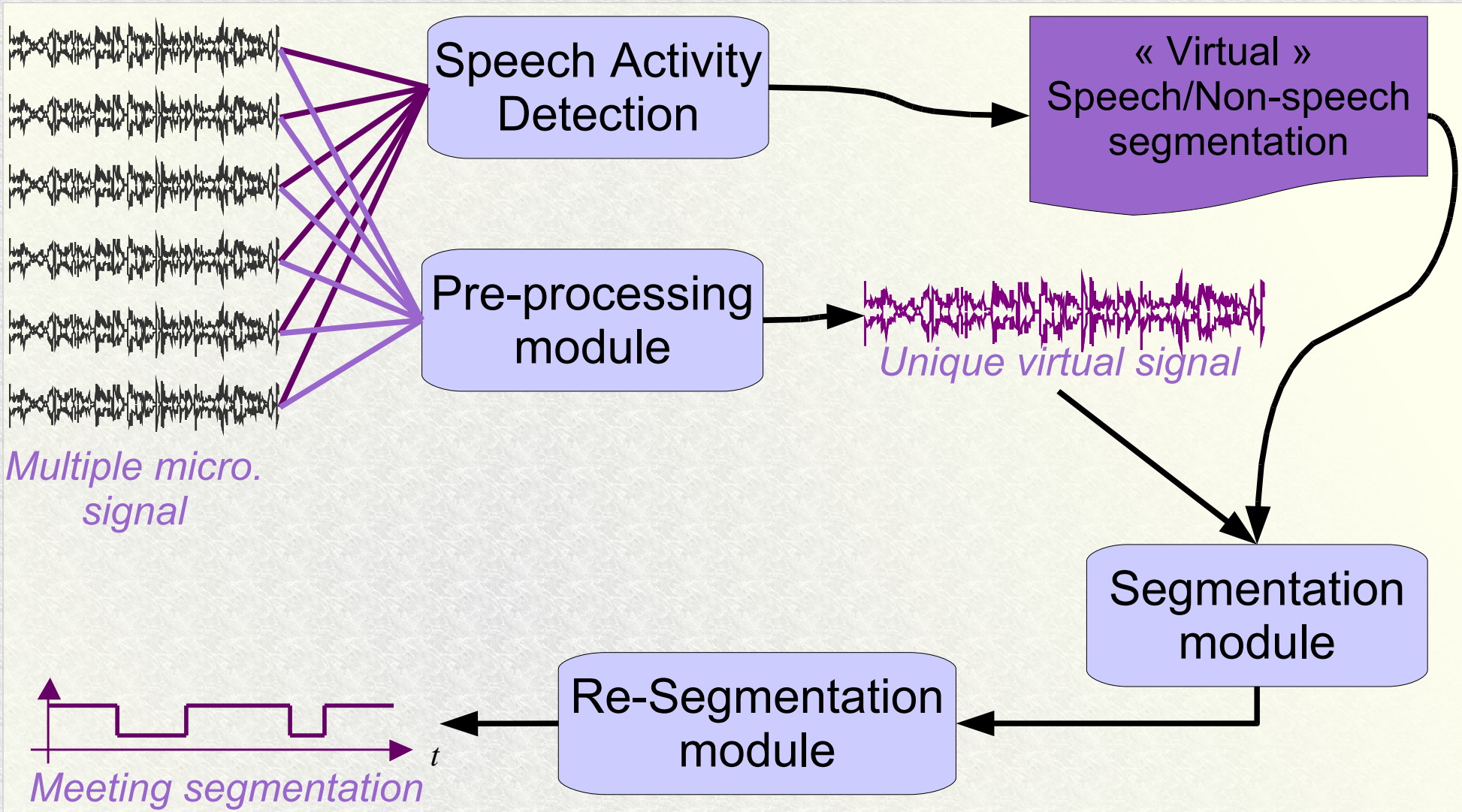
Context & objectives (2)

✗ Objectives for the meeting evaluation:

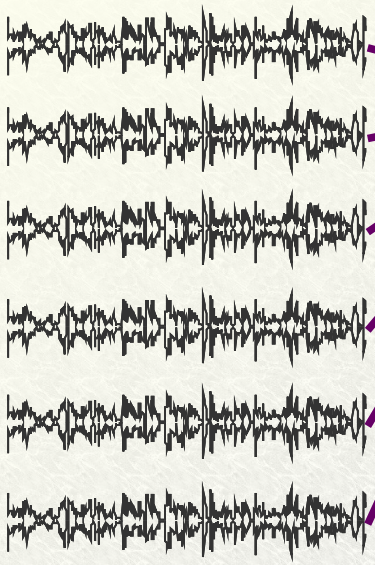
1. To investigate pre-processing techniques to handle multiple microphone signals in a transparent way for the speaker diarization systems
2. To test the robustness of BN speaker diarization systems

Answer the question: *Are pre-processing techniques sufficient to maintain the performance of BN speaker diarization systems ?*

System overview



Speech Activity Detection (1)



Multiple micro. signal

Speech Activity Detection

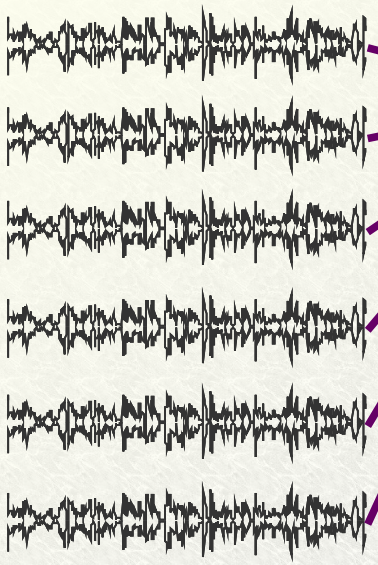
« Virtual » Speech/Non-speech segmentation

2 step SAD:

1. SAD applied on each individual channel
2. Merging based on commonly shared non-speech segments

=> « Virtual » speech/non-speech segmentation

Speech Activity Detection (2)



Multiple micro. signal

Speech Activity Detection

« Virtual » Speech/Non-speech segmentation

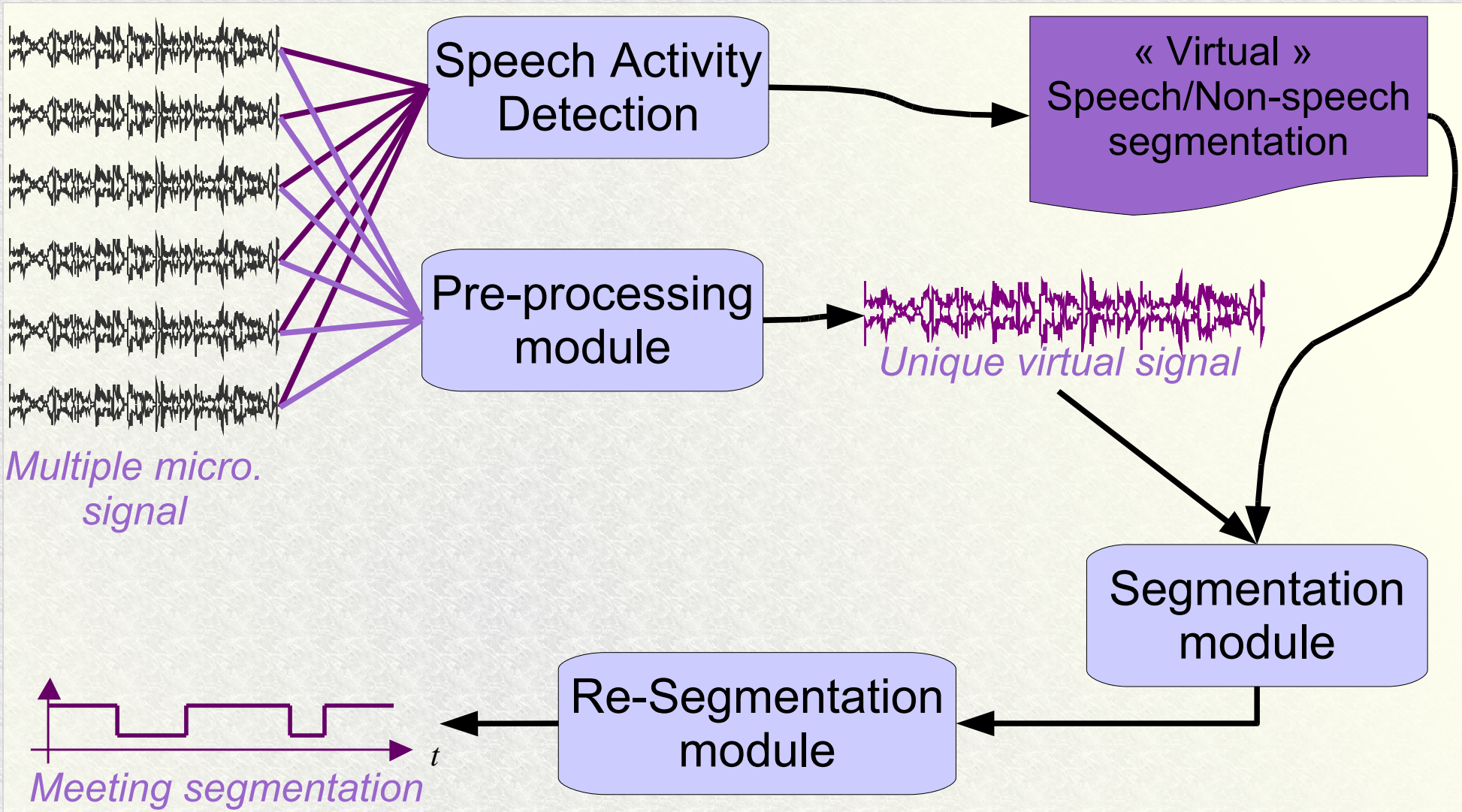
✗ Energy based SAD:

- ✔ ALIZE Open-Source platform
- ✔ freely distributed (LIA ASpkR package)

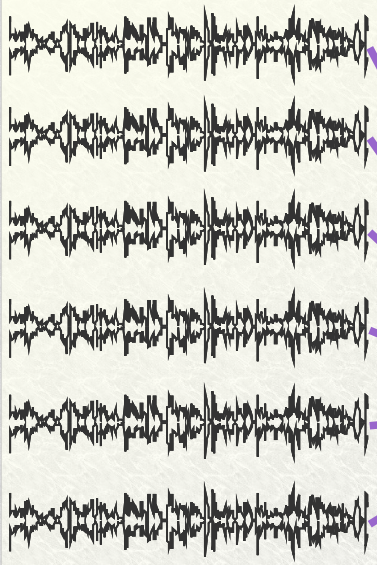
✗ Principle:

- ✔ 3 Gaussian comp. for energy modelling
- ✔ X% of the most energized frames labelled as speech + morphological rules to provide speech/non-speech segmentation

System overview



Pre-processing Module



Multiple micro. signal

Pre-processing module

Unique virtual signal

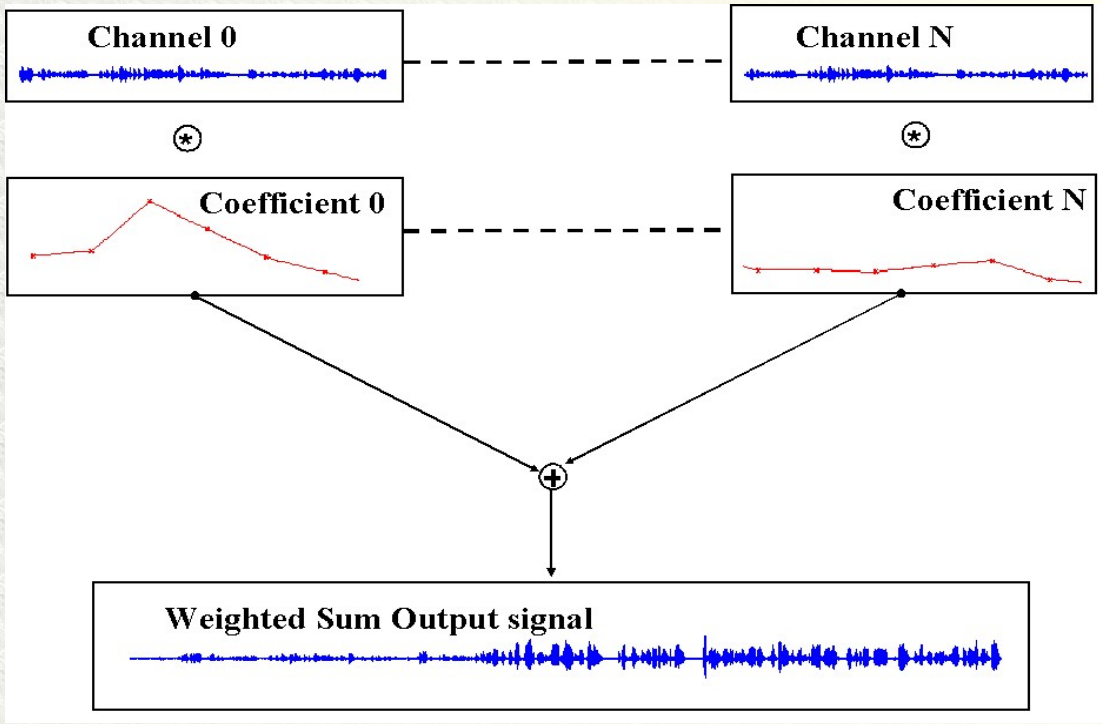
Goal: to provide one “virtual” signal from N “synchronized” signals

=> Weighted sum of N signals

Weighted Signal Sum (1)

- ✗ Weighted sum of N signals
  - ✔ Weight = Signal Quality Index => SNR
  - ✔ 2 strategies for estimating SNR

$$w_i = \frac{SNR_i}{\left(\sum_{j=1}^{n} SNR_j\right)}$$



Weighted Signal Sum (2)

- ✗ SAD based SNR estimate
  - ✔ SAD => Speech/Non-speech segmentation
  - ✔ Non-speech segments => global average of noise energy
  - ✔ 32ms rate on 64ms window (1024 samples)

$$SNR_{dB} = 10 \log_{10}\left(\left(\sum_{i=0}^{1023} s_i^2 - \mu_{E_{noise}}\right) / \mu_{E_{noise}}\right)$$

  - ✔ Drawbacks:
    - Dependent on SAD quality (misclassification error rate !)
    - Global noise energy estimate !

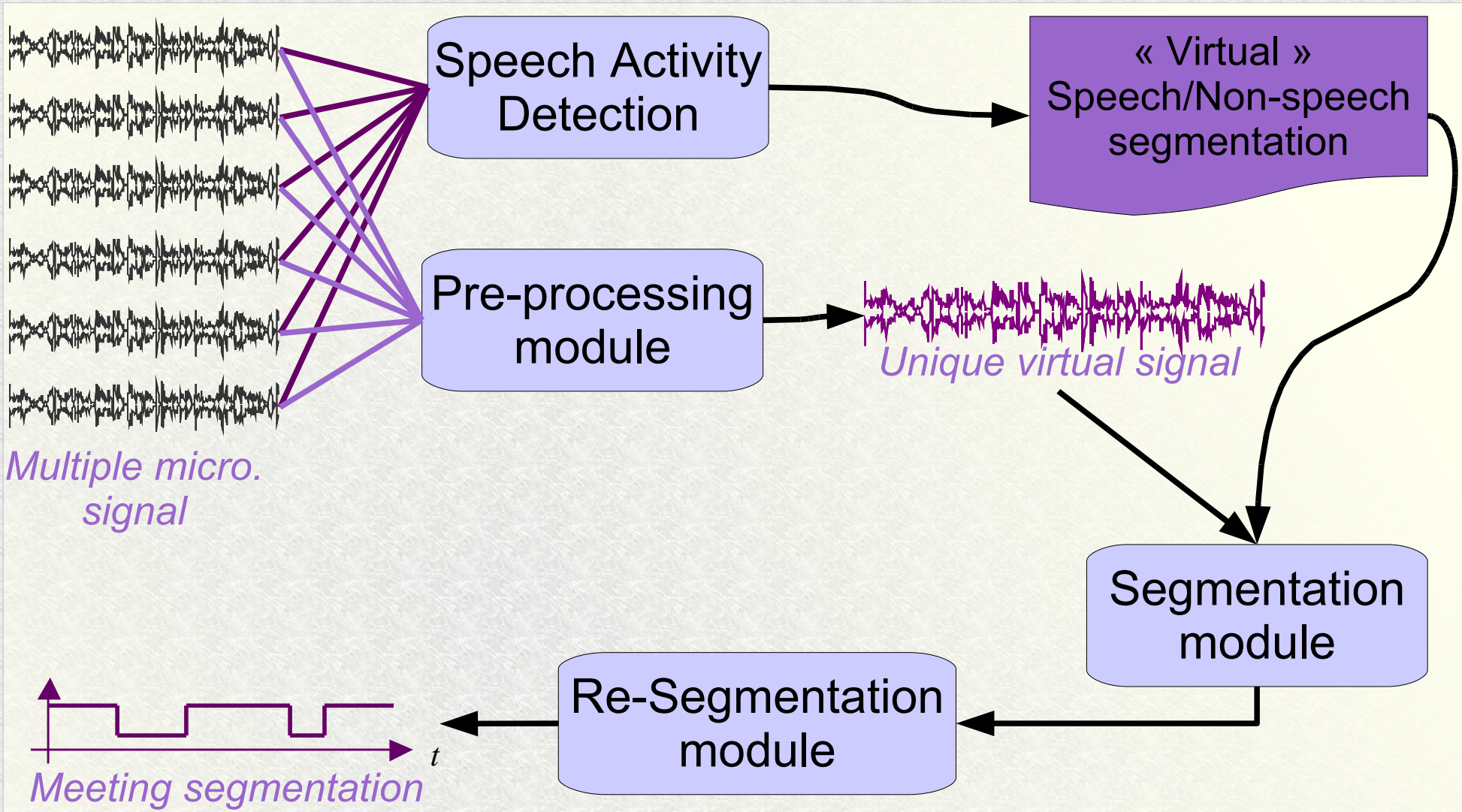
Weighted Signal Sum (3)

- ✗ Signal power spectrum based SNR estimate
  - ✔ Tracking signal power spectral minima [Cui & al, 03]
    - => background noise power estimate
    - => SNR estimate
  - ✔ Use of signal power spectrum histograms computed each 2s

$$SNR_{dB} = 10 \log_{10}\left(\sum_{i=0}^{M} \widetilde{S}_i \Big/ \sum_{i=0}^{M} \widetilde{N}_i\right)$$

Signal and Noise spectral amplitudes at frequency i

System overview



Speaker segmentation module

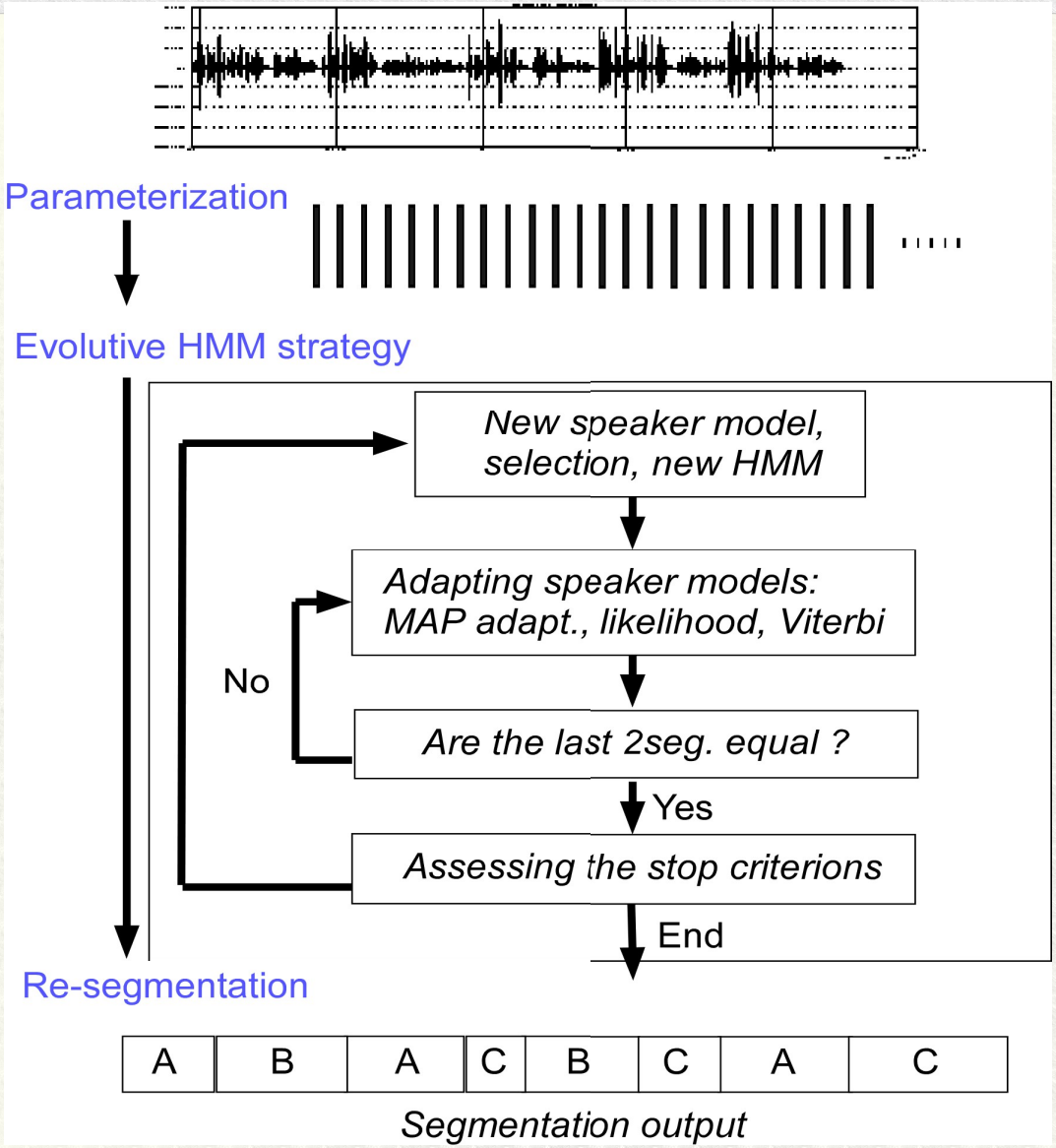
« Virtual »
Speech/Non-speech
segmentation

Unique virtual signal

1. LIA segmentation system
2. CLIPS segmentation system
3. LIUM segmentation system

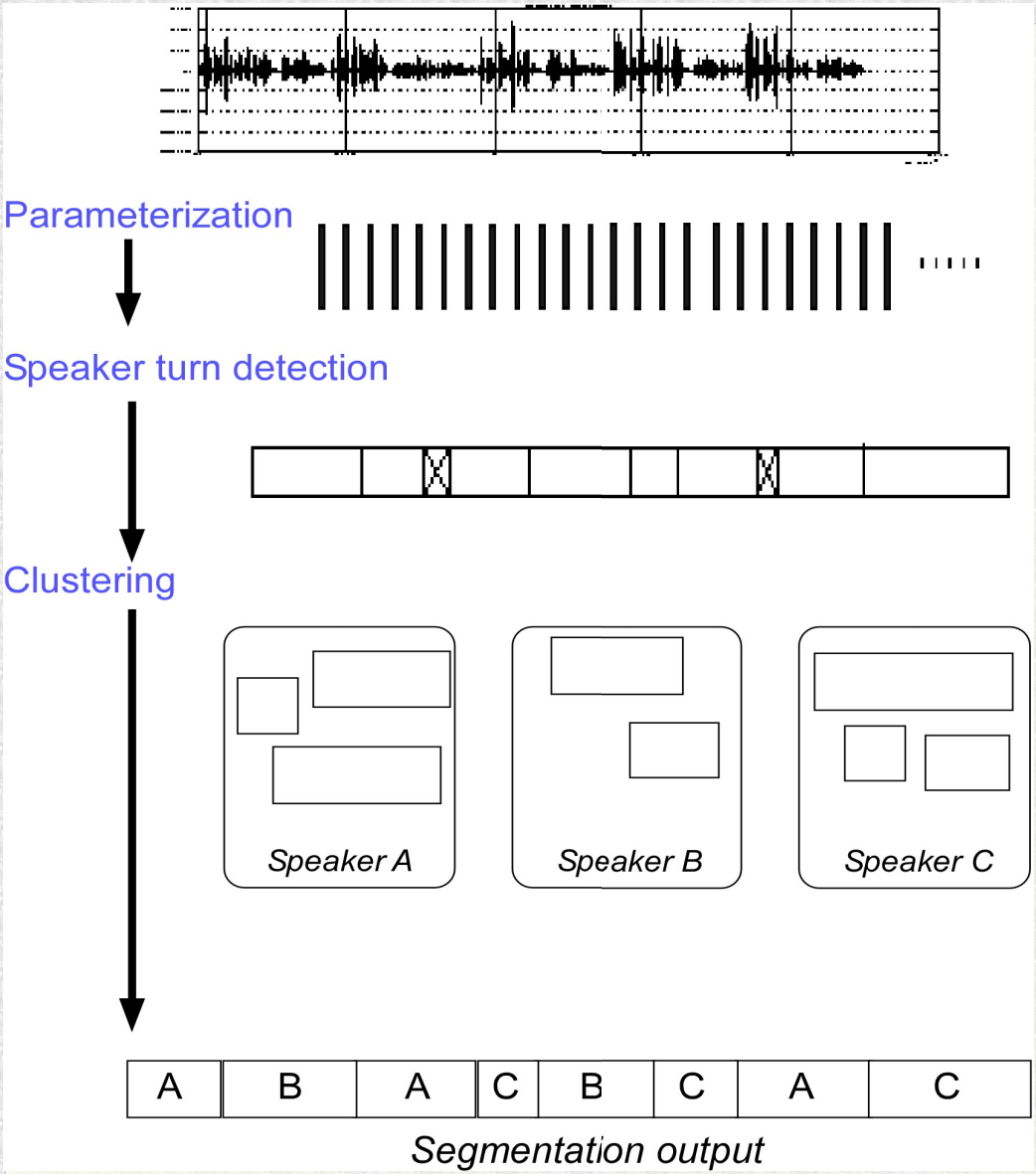
Segmentation
module

LIA Segmentation system



- ✗ E-HMM strategy
- ✗ New implementation on the ALIZE platform (End of 2004)
- ✗ Freely distributed on beta version
- ✗ Only tuned and tested for the French eval. on BN data (January 2005)

CLIPS & LIUM Segmentation syst.

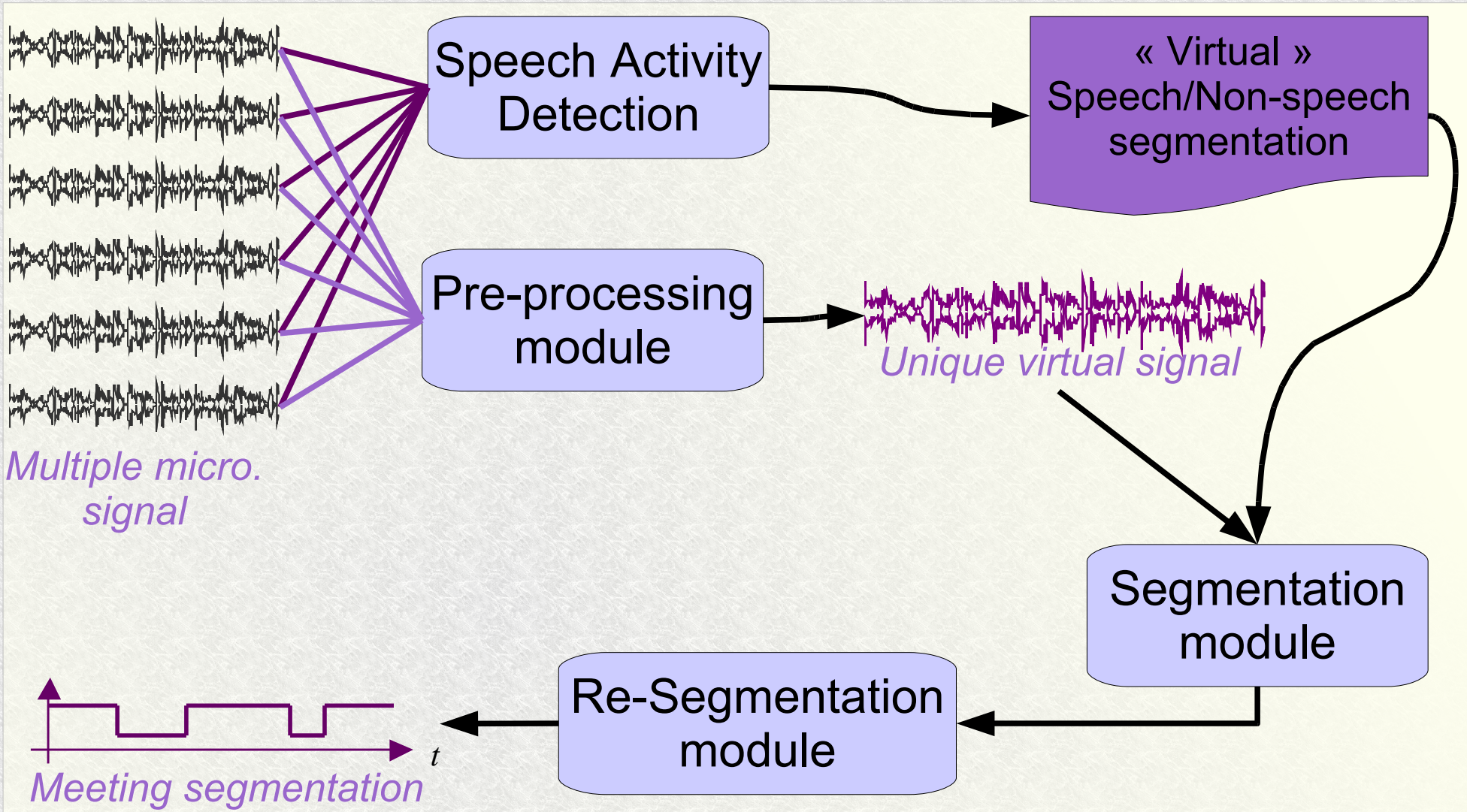


✗ Overview:

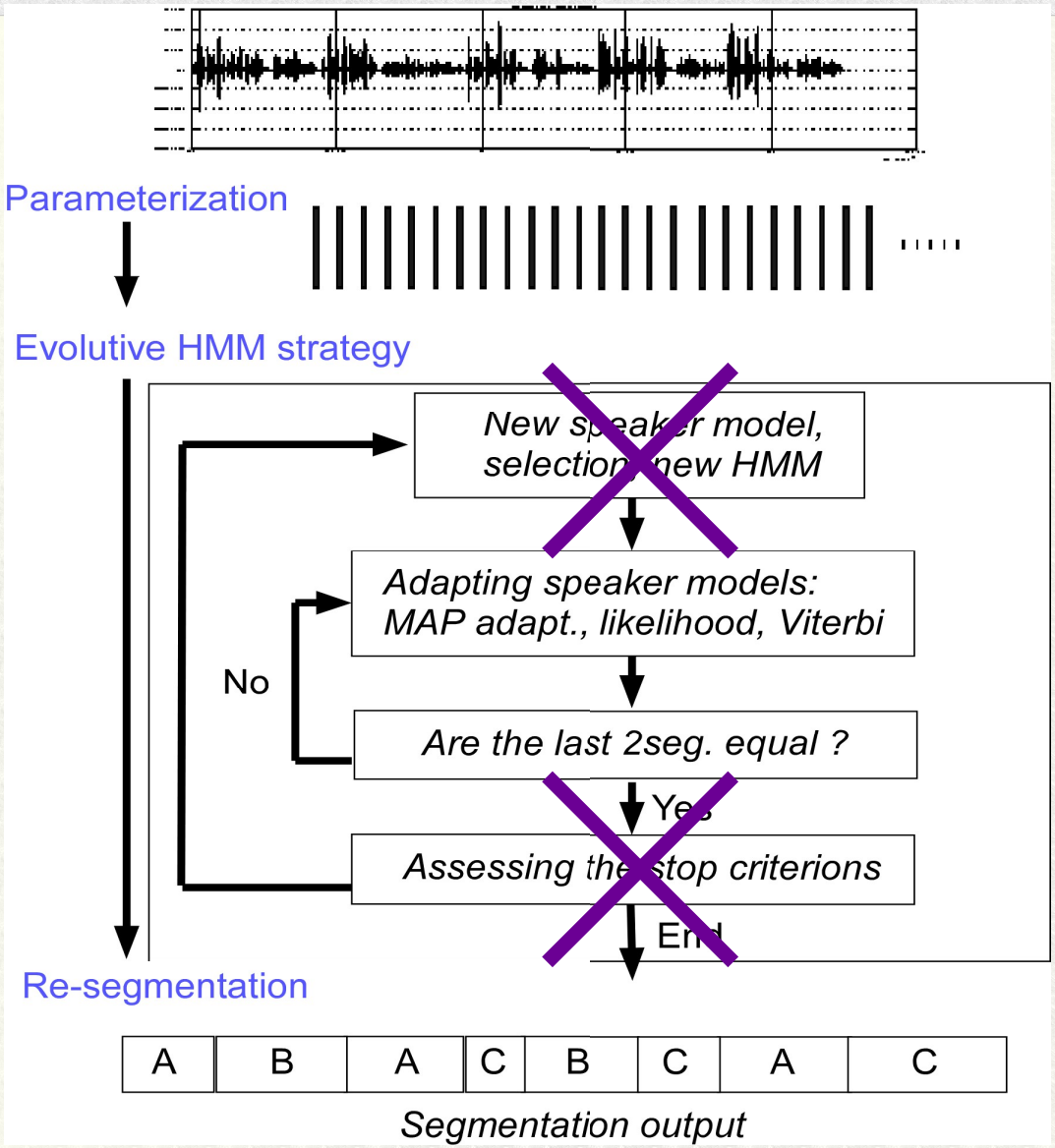
- ✔ Turn detection based on BIC
- ✔ Different strategies for speaker clustering

✗ Tuned and tested for the French eval. on BN data only

System overview



LIA Re-Segmentation system



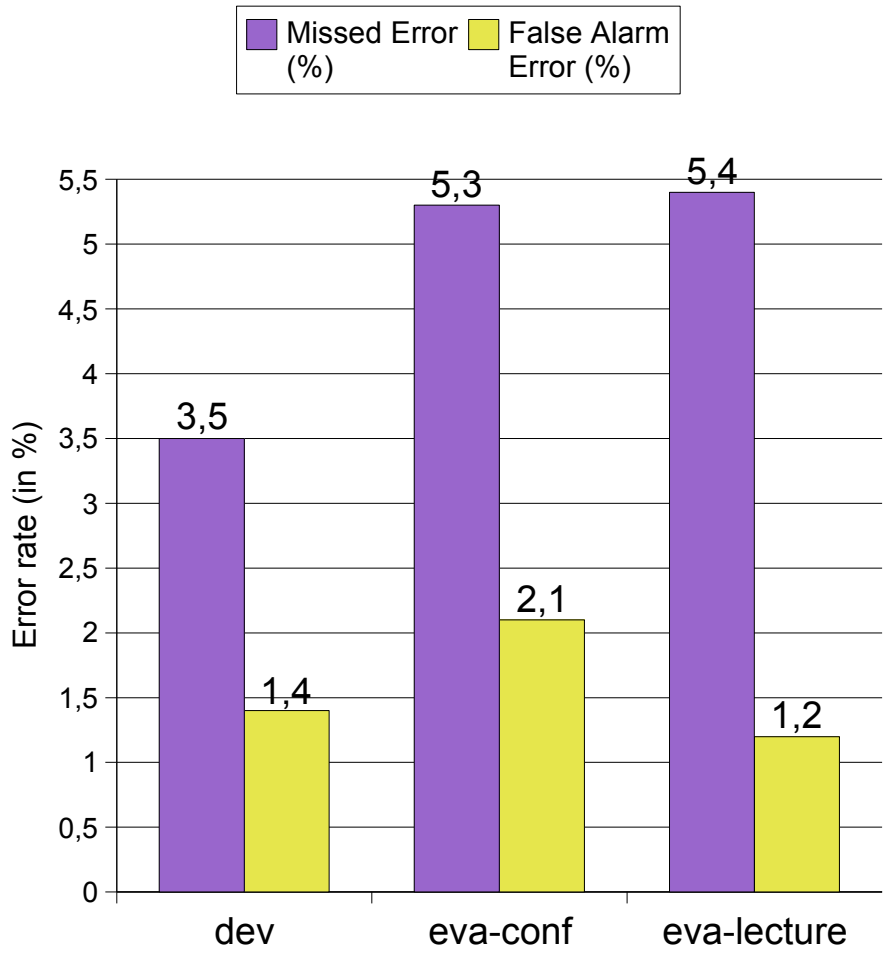
- ✗ To refine boundaries
- ✗ Applied on LIA, LIUM and CLIPS systems

- ✗ E-HMM strategy
  - ✔ Only the adaptation / decoding phase
  - ✔ Speaker deletion if irrelevant

Protocols

- ✗ Development corpus:
  - ✔ RT'04S development & evaluation
  - ✔ 12 meetings of 10 mn each
  - ✔ Short silence period, very low SNR (90% of file SNR < 3dB), strong overlap
- ✗ Evaluation corpus:
  - ✔ Conference data: short silence period, low SNR (13% of file SNR < 3dB), strong overlap
  - ✔ Lecture data: low SNR (20% of file SNR < 3dB), 1 speaker/meeting mostly

SAD Performance



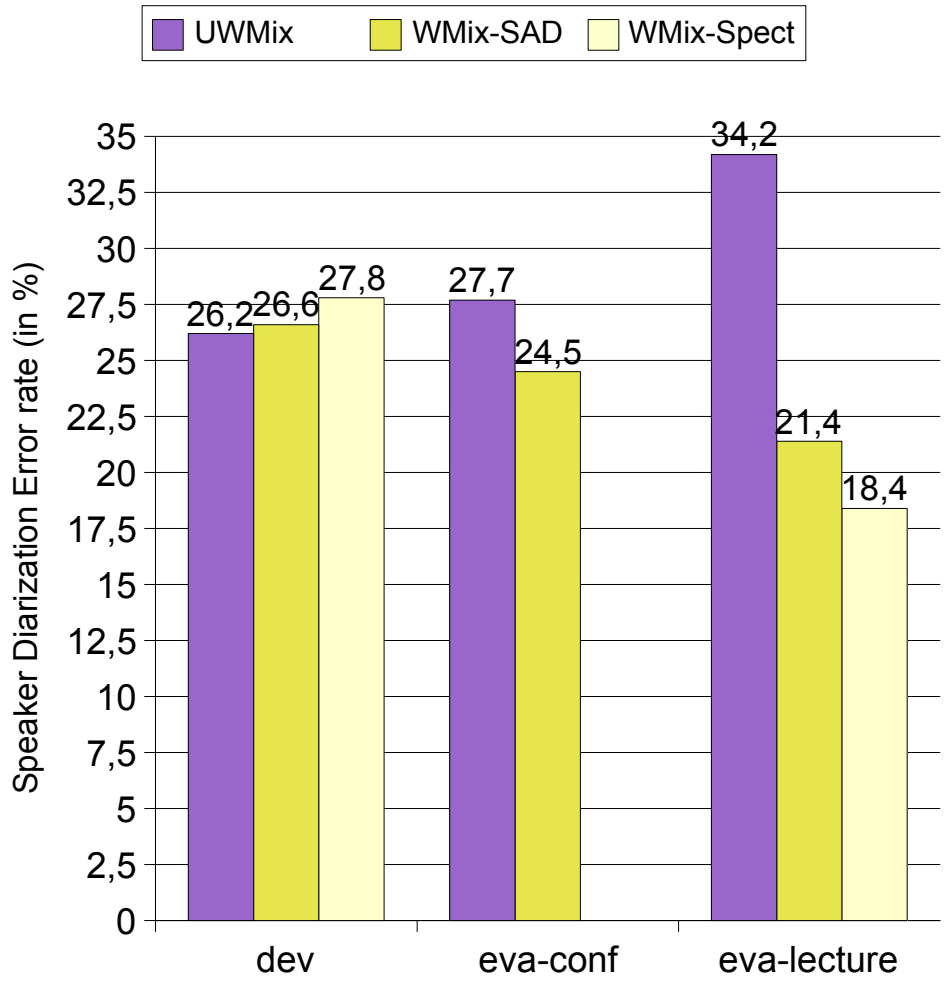
✗ Large Missed speech error rates

✗ Energy based-SAD applied on individual channels

=> difficulty to deal with background voices

=> difficulty to tune SAD parameters due to the variety of meeting sites

Pre-processing techniques (1)

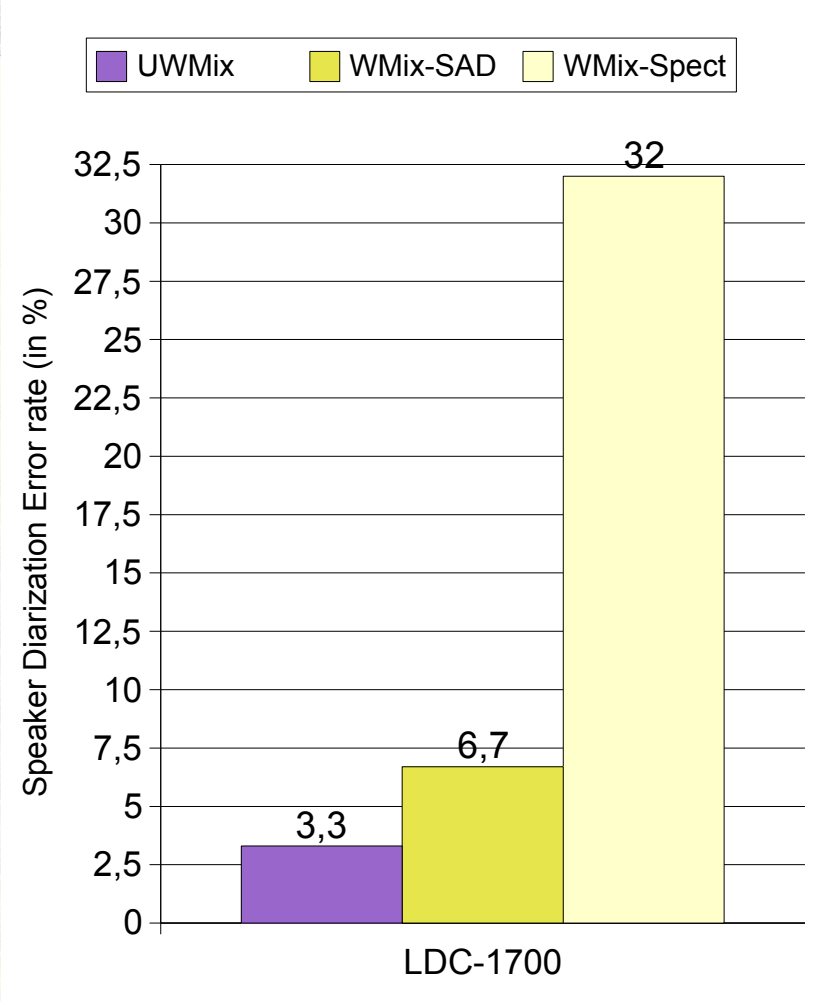


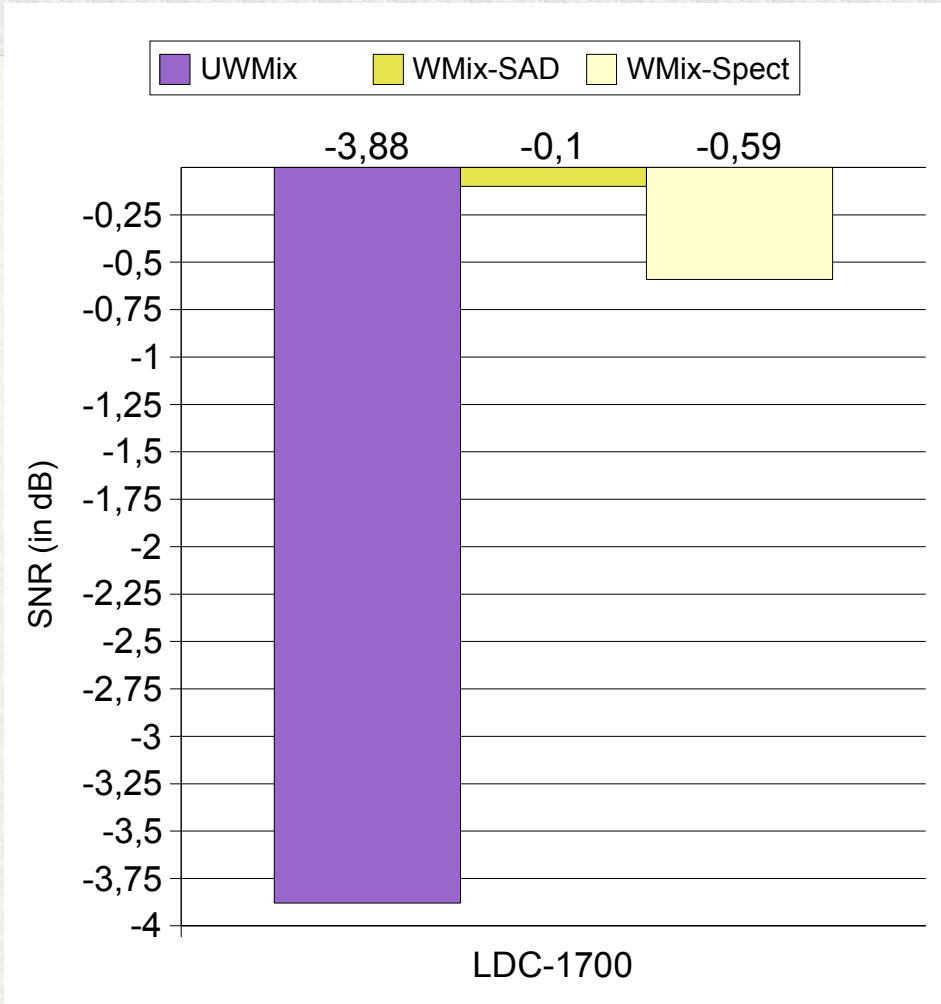
✗ LIA speaker diarization system perf. according to the pre-processing techniques

✗ Very contrastive behaviors:

- ✔ Probably due to the difficulty for estimating SNR (dev.)
- ✔ Interesting behavior of WMix-Spect technique on eva-lecture => retrieval of most significant channel !

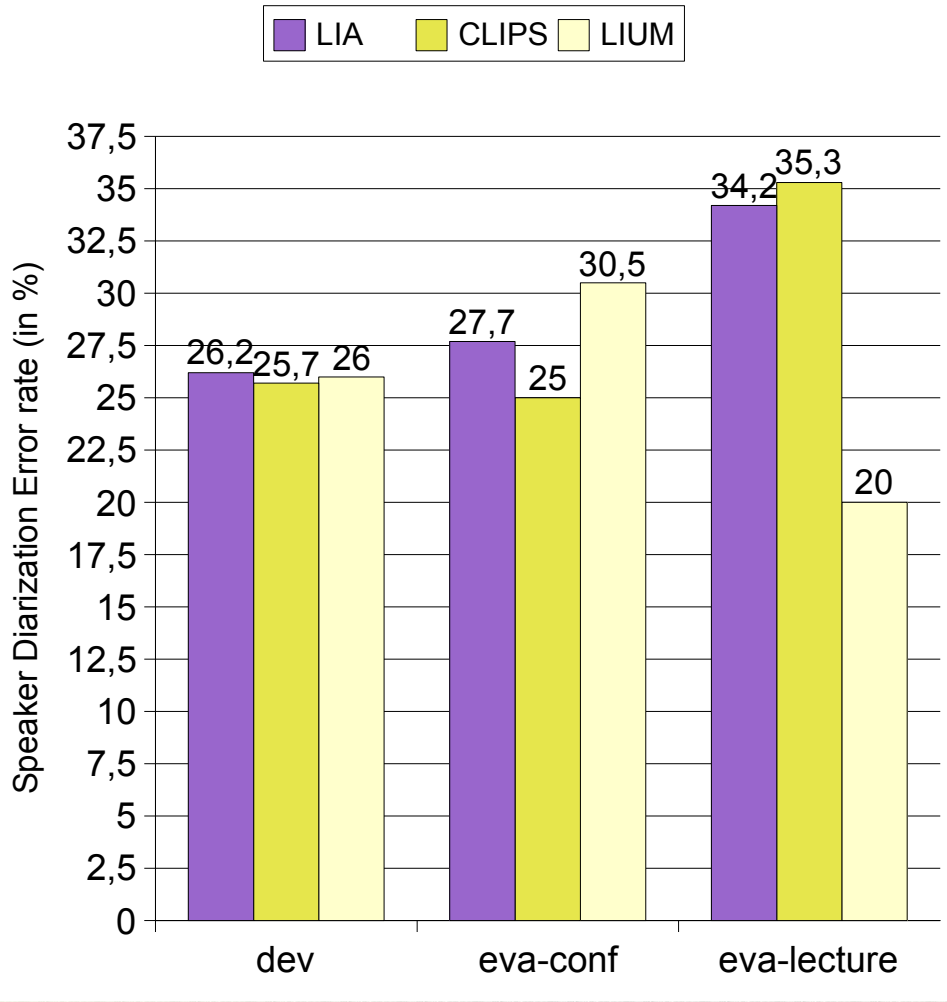
Pre-processing techniques (2)





SNR improvement does not mean speaker error rate improvement !!

Speaker diarization performance



✗ All based on the unweighted Mix

✗ Close for dev. corpus

✗ CLIPS/LIUM: similar strategies but very different behaviors

✗ Disappointing perf. compared with the French BN (ESTER): ~ 17% DER

Conclusion & perspective (1)

- ✗ Pre-processing technique:
  - ✔ From N microphone signals to 1 « virtual » one
  - ✔ Signal Mix based on SNR (signal quality index)
  - ✔ SNR estimate: SAD and signal power spectrum
  - ✔ Promising performance compared with unweighted Mix
- ✗ Robustness of BN speaker diarization
  - ✔ Unfortunately not, despite pre-processing techniques

Conclusion & perspective (2)

✗ Focusing on pre-processing => not sufficient

✗ Speaker diarization strategy has to be tuned on meeting data

✗ Further work:

- ✔ To better exploit the specificities of meeting environment (limited nb of speakers, overlaps, fast transitions, ...)
- ✔ To go on investigating pre-processing approach
- ✔ We promise to begin working earlier for the next evaluation to reach the ICSI performance !!

Contacts

Laboratoire d'Informatique
Université d'Avignon

dan.istrate@lia.univ-avignon.fr
corinne.fredouille@lia.univ-avignon.fr
jean-francois.bonastre@lia.univ-avignon.fr

CLIPS
Communication Langagière et Interaction Personne-Système
CNRS - INPG - UJF
BP 53 - 38041 Grenoble Cedex 9 - France

laurent.besacier@imag.fr

lium

sylvain.meignier@lium.univ-lemans.fr