IRENE...

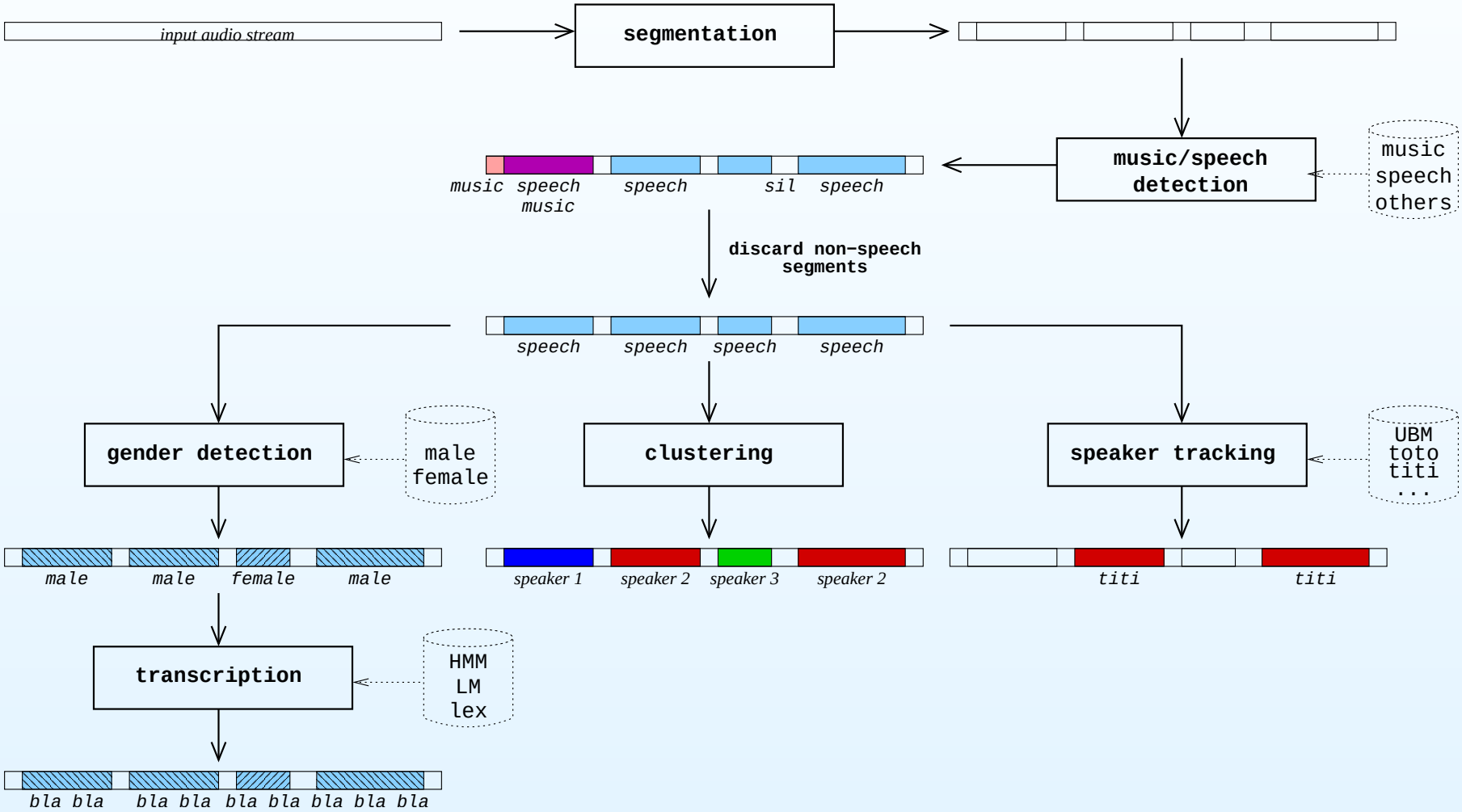
Mathieu Ben
Guillaume Gravier

Michaël Betser
François Yvon

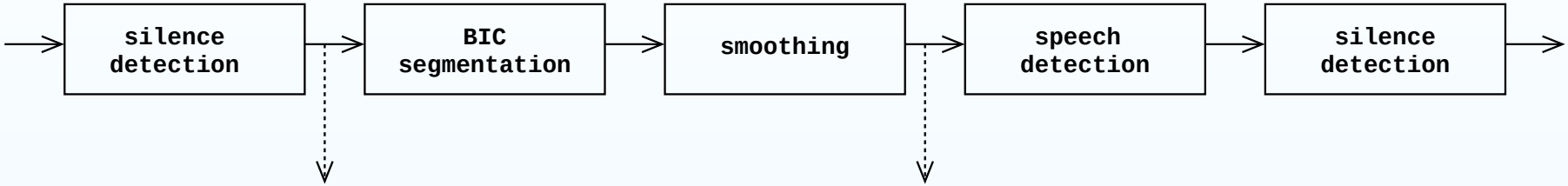
IRISA

ENST

Architeture du système



Segmentation



1. détection de silence (> 0.3 s)
2. détection de silence (> 0.7 s) + BIC
3. détection plage de parole + détection de silence (> 0.3 s)

source	SAD	BIC	BIC+SAD	REF
France-Inter	43.3	44.6	44.7	40.4
RFI	56.3	55.7	–	51.8
moyenne	49.1	49.6	–	45.5

Détection parole/musique

Détermination des classes présentes avec un critère MAP

$$\begin{aligned} \hat{x} &= \arg\max_x P(y|x)P(x) \\ &\sim \prod_i P(y|x_i)P(x_i) \end{aligned}$$

où

- $P(y|X_i = 1)$ (resp. $P(y|X_i = 0)$) sont des modèles de présence (resp. non présence) de classes
- $P(y|x_i) \sim$ GMM, 64 composantes
- y est un vecteur de 16 $MFCC + \Delta MFCC + \Delta\Delta MFCC + \Delta E + \Delta\Delta E$

cela revient à faire des tests d'hypothèses, les seuils étant déterminés par les probabilités à priori.

Détection du sexe

1. **UBM dépendant du sexe du module SVL**
   - sélection de trames, centrage/réduction global
   - 256 gaussiennes
   - femmes : 140 locutrices, 1h de parole
   - hommes : 430 locuteurs, 3h de parole
2. **GMM avec les paramÉtres du module TRS**
   - centrage/réduction glissant
   - 256 gaussiennes
   - femmes : 220 locutrices, 2h de parole (1h25 + 35m)
   - hommes : 481 locuteurs, 2h de parole (1h15 + 45m)

Détection du sexe (suite)

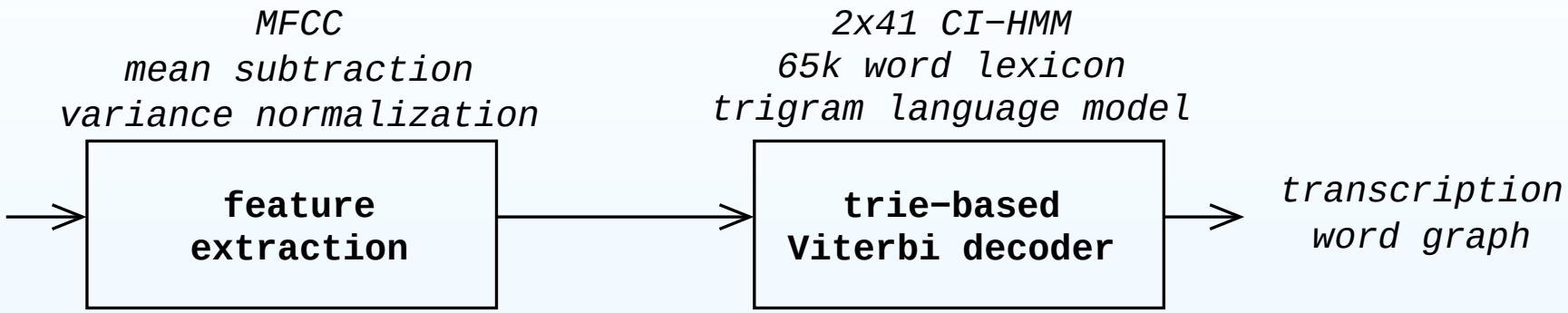
- **Taux d'identi£cation** (1667 segments)

UBM-256	94%
GMM-256	96%

- **Impact sur la transcription**

source	UBM-256	GMM-256	REF	IND
France-Inter	40.4	38.4	36.7	37.6
RFI	51.8	50.4	49.6	50.4
moyenne	45.5	43.8	42.5	43.3
France-Inter	34.8	34.8	34.8	–
RFI	48.3	48.5	48.6	–
moyenne	40.8	40.9	41.0	–

Transcription (architecture)



	50k/100	100k/200	200k/200
x RT (750Mb RAM)	15	30?	60?
France-Inter	34.8	33.6	32.8
RFI	48.3	46.4	45.7

Lexiques et modÉles de langage

- **Lexiques** (65k mots)
  - interpolation optimale entre ’Le Monde’ et le corpus d’apprentissage
  - phonétisation par ILPho (83% des mots) et automatique
  - moyenne de 3.7 variantes par forme graphique
  - 3 variantes
- **ModÉles de langage** (trigramme)
  - lm1 - 2M bigrammes, 1M trigrammes
  - lm2a - 5.3M bigrammes et 2.7 M trigrammes
  - lm2b - 8.5M bigrammes et 4.3 M trigrammes
  - lm3 - 8.6M bigrammes et 4.3 M trigrammes

Modèles acoustiques

- 41 (pseudo-)phones hors-contexte
- MMC gauche-droit, 3 états x 64 gaussiennes
- estimation des paramètres
  1. phonétisation du corpus 'train' ESTER (après sélection des bons segments)
  2. bootstrap à partir de modèles 32 gaussiennes BREF-120
  3. réalignement phonétique : liaisons, silences et schwas
  4. estimation ML modèles 32 puis 64 gaussiennes
- spécialisation des modèles
  - sexe : MAP, MLLR ou ML
  - sexe et bande passante : (MLLR+)MAP

Transcription : résultats

- **Lexiques**

source	lex1	lex2		lex3
	lm1	lm2a	lm2b	lm3
France-Inter	36.7	35.0	34.8	40.7
RFI	49.6	49.0	48.6	52.7
moyenne	42.5	41.2	41.0	46.0

- **Modéles spéci£que**

source	sexe		bande
	MAP	ML	MAP
France-Inter	33.6	32.8	32.4
RFI	46.4	46.3	46.2
moyenne	39.3	38.8	38.6

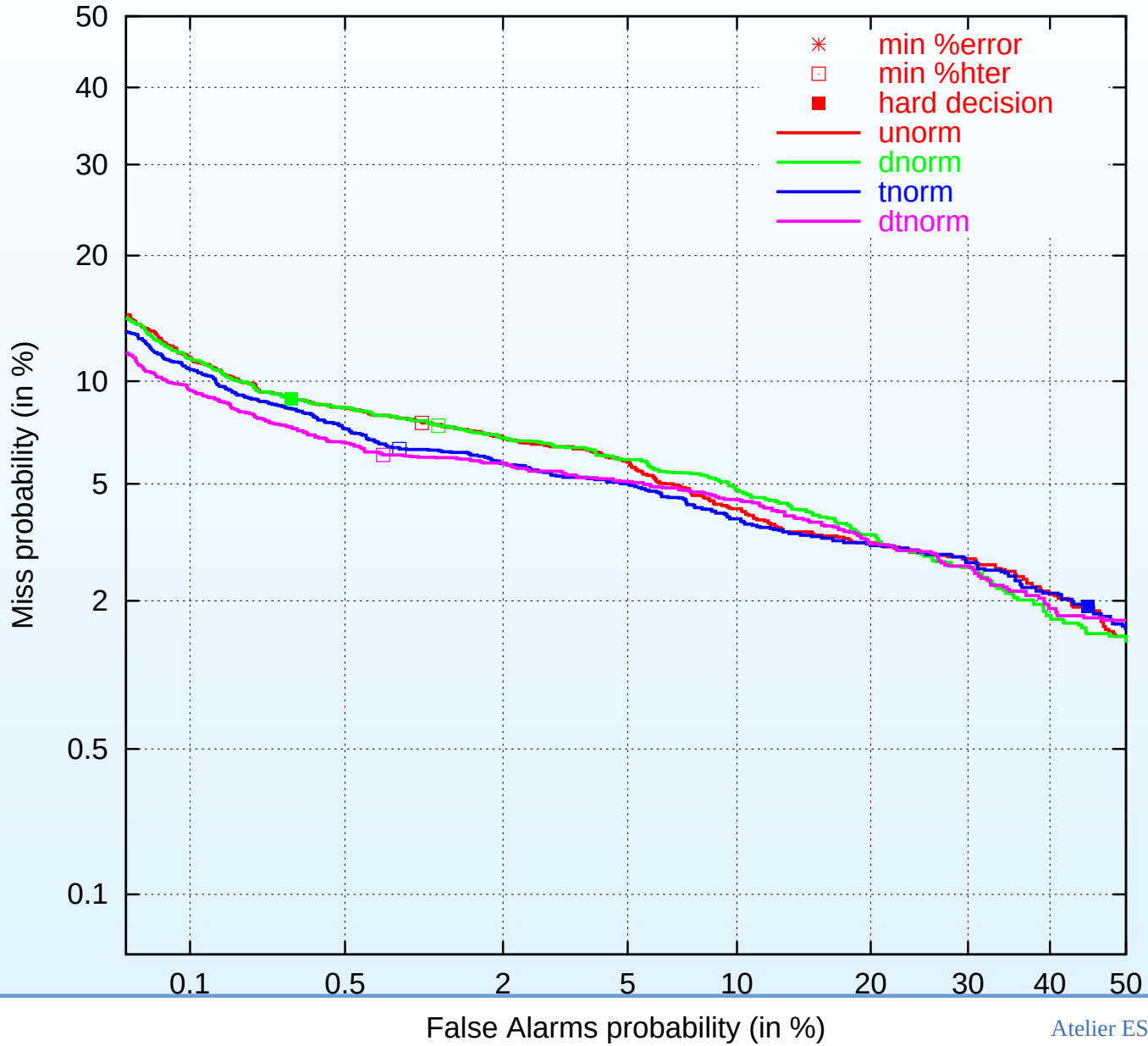
Suivi de locuteur

- **Front-end**
  - 16 LFCC + Deltas + Delta-logE
  - selection de trame : modèle bi-gauss. de l'energie + ML
  - normalisation globale (centrage+réduction)
- **UBM**
  - 512 gauss. indep. du genre (concat. de 2 UBM dép. du genre)
  - (corpus *train*) - (données locuteurs cibles et T-norm)
  - pas + de 3 min. du même locuteur
  - hommes: $\sim$ 430 hommes, 3 h de parole et 140 femmes
  - femmes: $\sim$ 140 femmes, 1h de parole femmes
- **modÉles des locuteurs cibles**
  - adaptation MAP de l'UBM (facteur de con£ance τ = 5)
  - (données cibles du corpus *train*) - (segments multi-locuteurs)

Suivi de locuteur (suite)

- **scoring**
  chaque segment testé contre toutes les identités cibles X
  score brut S = LLR/N-Best (N=10) entre X et UBM
  normalisation de score : DT-norm
  - D-norm : $S_D = \frac{S}{KL_X}$
    KL_X = distance KL entre modèle X et UBM
  - DT-norm : $S_{DT} = \frac{S_D - \mu_{S_D}}{\sigma_{S_D}}$
    (μ_{S_D}, σ_{S_D}) = moy. et ec. type des scores D-norm du segment vs modèles imposteurs
  - données T-norm :
    loc. imposteurs du corpus *train* avec plus d'1 min. de parole
    51 femmes et 51 hommes
- **décison**
  comparaison du score normalisé à un seuil θ
  θ optimisé par rapport à la métrique sur corpus *dev*

Suivi de locuteur (suite de la suite)



Regroupement de locuteur

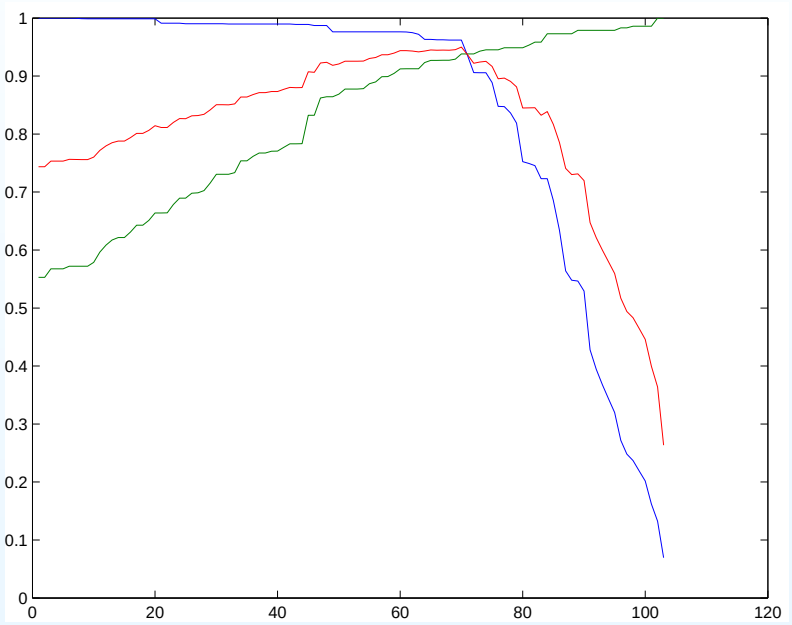
Approche MAP pour un algorithme 'chop & cluster'

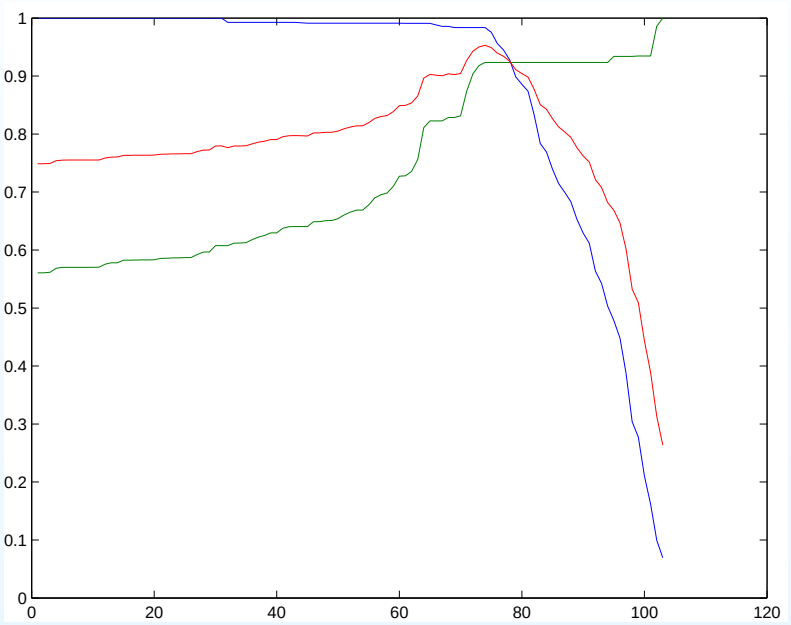
- estimation MAP d'un GMM par cluster
- modèle du monde estimé
  - soit sur l'ensemble du document,
  - soit au préalable sur un ensemble de locuteurs
- distance entre modèles
  - distance euclidienne pondérée

$$(1) \qquad D(\Lambda^{(1)}, \Lambda^{(2)}) = \sqrt{\sum_i w_i \sum_j \frac{(m_{ij}^{(1)} - m_{ij}^{(2)})^2}{\sigma_{ij}^2}}$$

  - distance angulaire
- critère d'arret en cours de ré¤exion...

Regroupement de locuteur (suite)





taux d'erreur (optimiste) pour le système MAP: 14.6%

Mais encore...

- comprendre pourquoi les mots composés ne marchent pas mieux
- améliorer le module de segmentation
- modélisation contextuelle
- utiliser l'information sur le sexe pour SVL et SRL
- intégrer le clustering dans SVL
- adaptation
- et encore tout plein d'autres choses...