Intégration de connaissances par modèles probabilistes pour l'analyse de documents multimédias

Guillaume Gravier

guig@irisa.fr

UMR IRISA

Document multimédia
=
ensemble de sources d'informations complémentaires

plus ou moins synchrones

Exemples

- texte + image
- vidéos – image, son et parfois texte
- analyse de scènes multimodales

⇒ modéliser les documents multimédias en s'appuyant sur l'ensemble des connaissances disponibles

Proposer et étudier des modèles (probabilistes) permettant de modéliser conjointement les informations disponibles

Traitement des documents multimédias professionnels selon deux axes privilégiés

1. **multimodalité** – exploiter conjointement les indices sonores et visuels
   - analyse de vidéos sportives
   - nécessité de meilleurs modèles pour cette intégration
2. **sémantique** – exploiter le sens porté par la composante orale
   - analyse de documents oraux
   - modéliser différents niveaux de connaissances (acoustique, phonétique, syntaxique)

$\Rightarrow$ **convergence entre ces deux axes**

Connaissance
=
quelque chose que l'on sait sur un document

- *a priori* **= connaissances du domaine**

  e.g., règles de montage d'une vidéo, syntaxe en parole, *etc.*

  ⟶ structure des modèles
  ⟶ modèle *a priori*

- *a posteriori* **= observations**

  e.g., les images, le son, *etc.* ⟶ flux d'observations multiples

  - synchronisation
  - corrélation

Difficulté : comment concilier des connaissances hétérogènes plus ou moins synchronisées ?

Difficultés

intégration précoce *vs.* intégration tardive

↓

modélisation conjointe

Ce qui soulève quelques problèmes

- représenter dans un même modèle des informations différentes
- représenter les relations entre ces informations
- contrôler la complexité des modèles

Quelques approches en ce sens : multiflux, couplés, multicouche, asynchrones, modèles de segments, *etc.*

Intégration de connaissances

par modèles probabilistes

pour l'analyse de documents multimédias

1. intégration de connaissances homogènes
2. intégration de connaissances hétérogènes
3. intégration de connaissances sporadiques
4. quelques pistes de recherche

Partie I

Intégration de connaissances homogènes

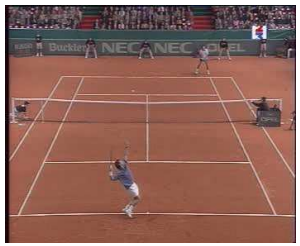
structuration multimodale de vidéos sportives

Thèses de Ewa Kijak, Emmanouil Delakis et Siwar Baghdadi,
en étroite collaboration avec P. Gros et Thomson (L. Oisel, C.-H. Demarty).

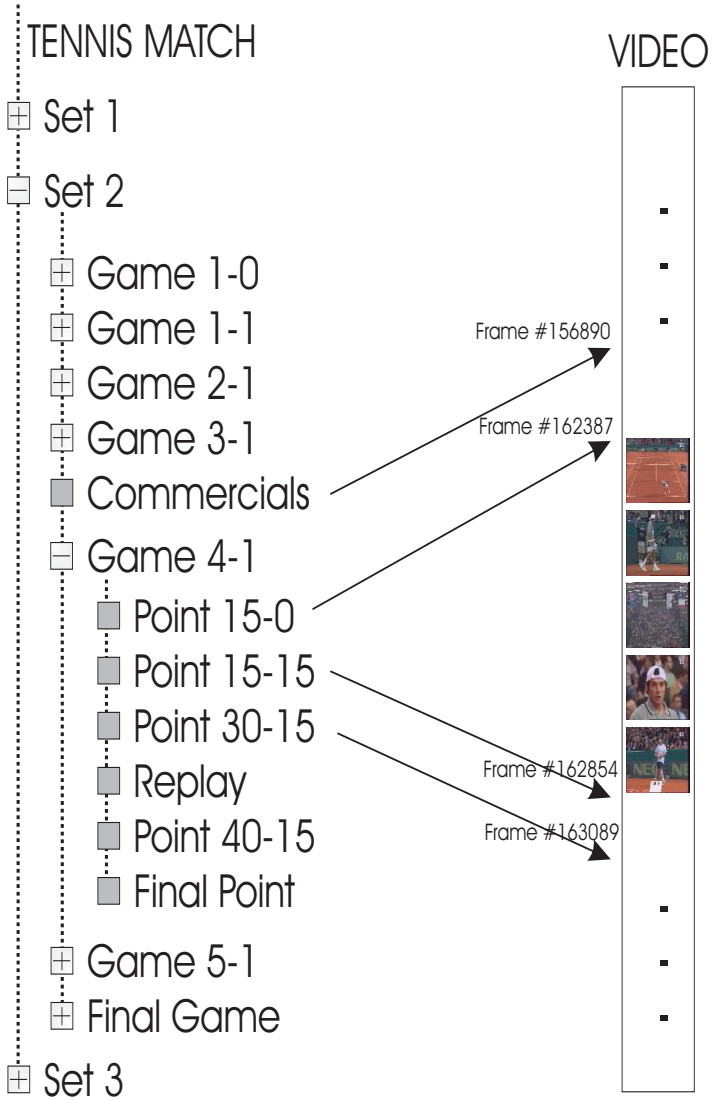
UMR IRISA

Structuration de vidéos de tennis

- quatre scènes génériques
  - ⟶ service + rallye, rallye, rediffusion, temps morts
- connaissances *a priori*
  - ⟶ règles de montage, règles du tennis
- **connaissances** *a posteriori*
  - ⟶ **images & événements sonores**



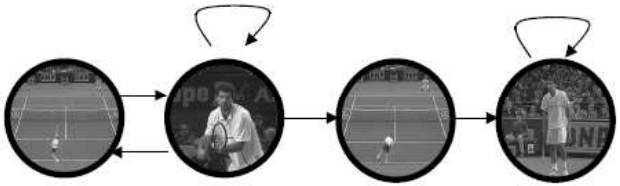


UMR IRISA

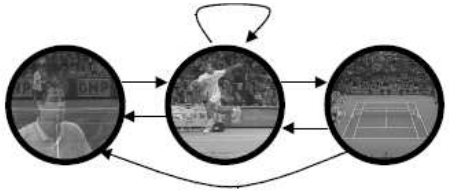
Approche visuelle [Kijak *et al.*, 03]

un état = un plan = une observation

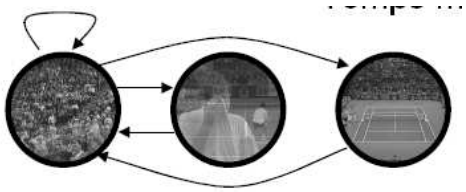
- attributs = distance à une vue globale, présence de fondu enchainé, durée
- modélisation des scènes par modèles de Markov cachés



service + rally



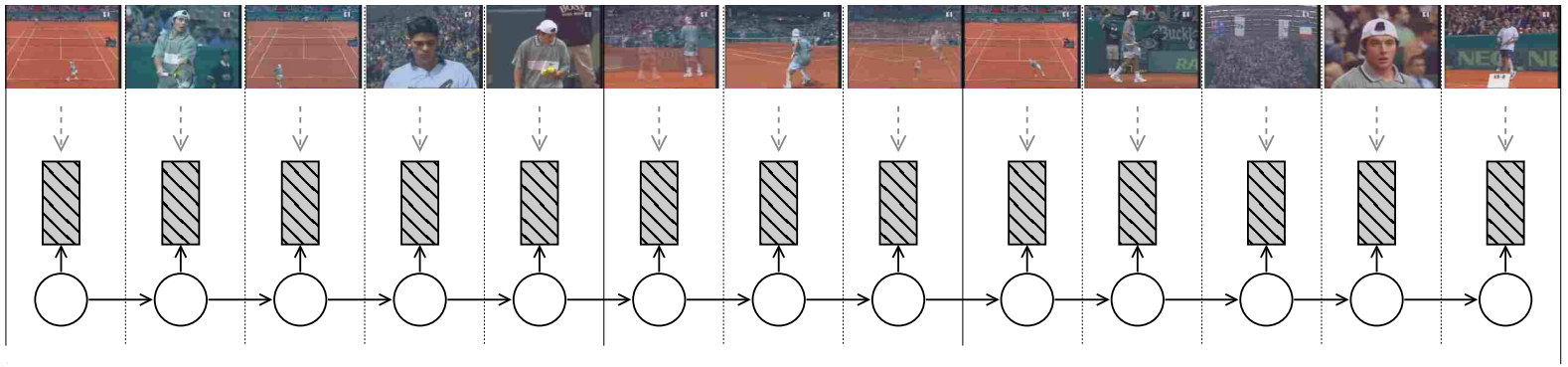
rediffusion



temps mort

- décodage par algorithme de Viterbi

$$\widehat{q}_1^N = \arg\max_{q_1^N} \sum_{i=1}^{N} \ln p(q_i|q_{i-1}) + \ln p(v_i|q_i)$$

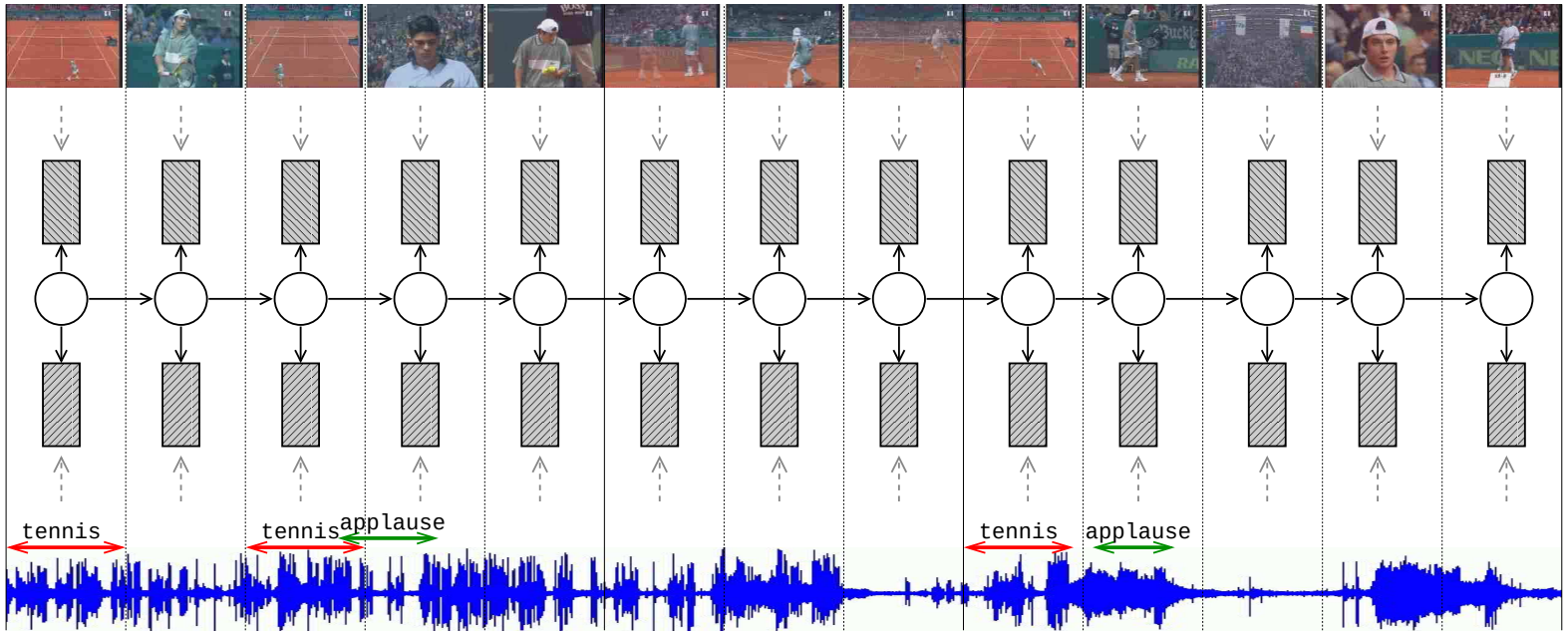


Modèle multiflux synchrone [Kijak *et al.*, 03]

un état = un plan = plusieurs observations

- attributs sonores = frappes de balle, applaudissements, musique(, parole)
- décodage par Viterbi avec pondération des flux d'observations

$$\widehat{q}_1^N = \arg\max_{q_1^N} \sum_{i=1}^{N} \ln p(q_i|q_{i-1}) + \alpha \ln p(v_i|q_i) + \beta \ln p(a_i|q_i)$$



↘ synchronisation des descripteurs au niveau du plan

Modèle de segments multiflux [Delakis *et al.*, 06]

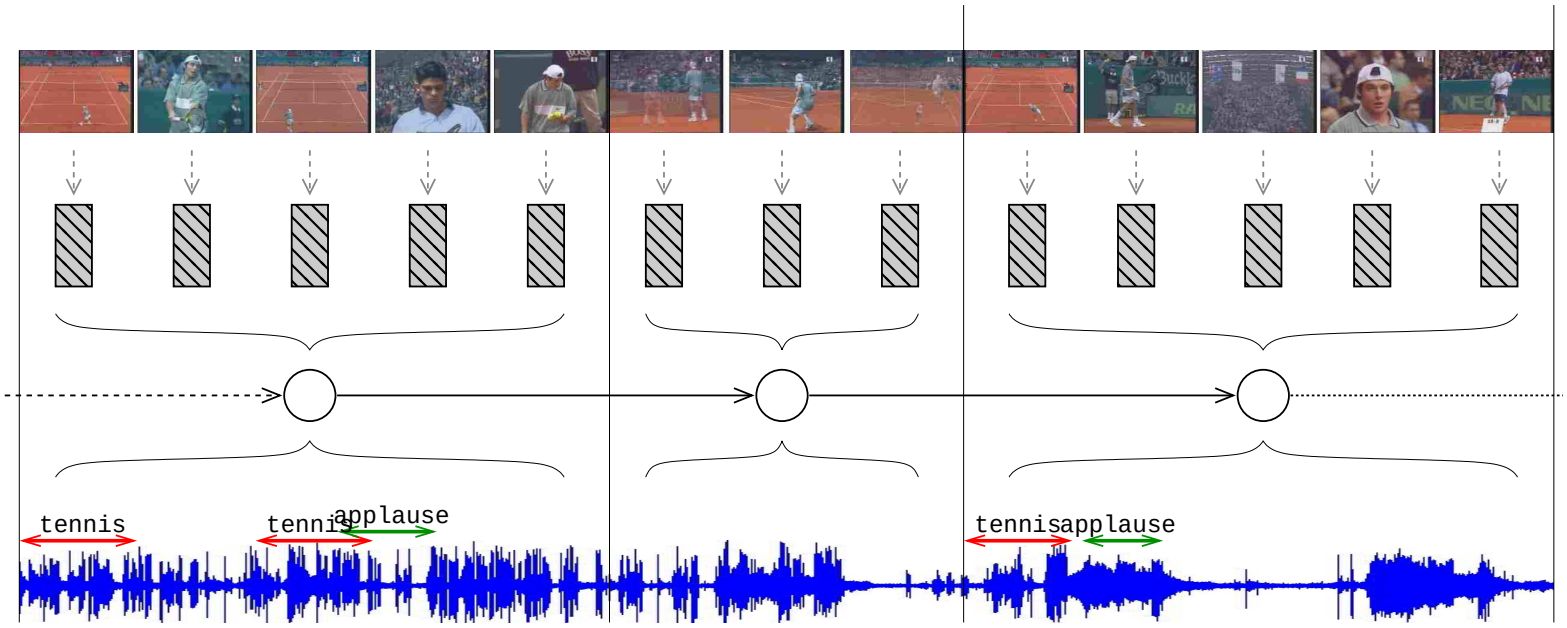
un état = une scène = plusieurs séquences d'observations

- modèle de segments = associer une séquence d'observations à un état

[Ostendorf *et al.*, 96]

- recherche conjointe de la segmentation et de la séquence d'états optimales

$$(\widehat{q}_1^{L^*}, \widehat{l}_1^{L^*}) = \arg\max_{q_1^L, l_1^L} \sum_{i=1}^{L} \ln p(q_i|q_{i-1}) + \ln p(l_i|q_i) + \alpha \ln p(v_{s_i}^{e_i}|q_i) + \beta \ln p(a_{s_i}^{e_i}|q_i)$$

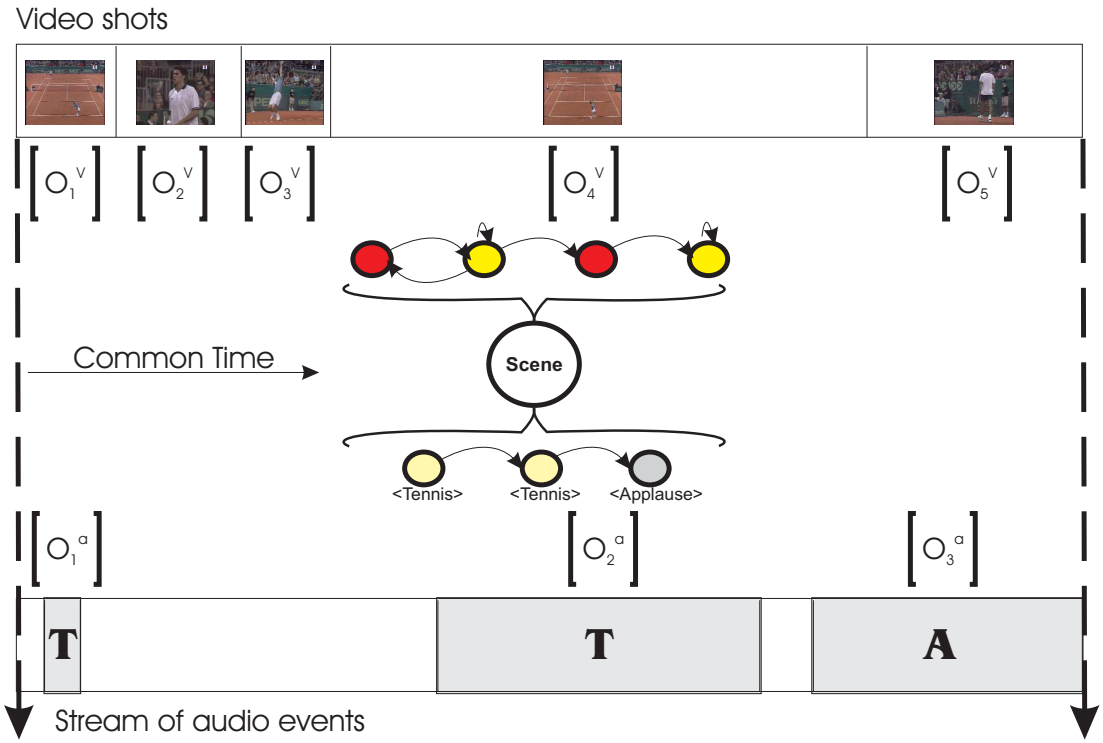


Modèle de segments multiflux (suite)

1. modèle de la probabilité conditionnelle pour le flux visuel

   ⟶ MMC
2. modèles de la probabilité conditionnelle pour le flux sonore

   ⟶ unigramme, bigramme ou encore MFCC/MMC
3. combinaison MMC multiflux + bigramme



Quelques résultats

Classification des plans en scènes (% classification correcte)

	visuels	fusion précoce	multiflux bigramme	précoce & bigramme
modèles de Markov	76,3	80,2	–	–
modèles de segments	79,7	84,4	81,8	**84,7**

+ importance d'intégrer les connaissances sonores
+ modèle de durée dans les modèles de segments
+ souplesse offerte par le modèle de segments multiflux
- hypothèse d'indépendance conditionnelle des flux d'observations

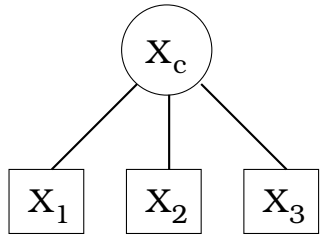
Apprentissage de structure [Baghdadi *et al.*, 09]

- formalisme permettant de décrire des interactions complexes
- MMC & modèle de segments sont des cas particulier [Murphy, 02]

$\Rightarrow$ **apprentissage de la structure**

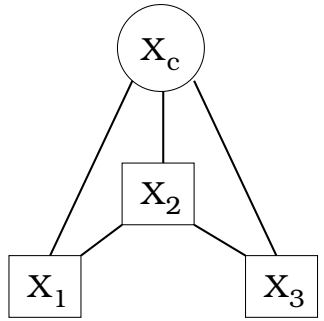
Cadre expérimental

- détection des événements dans les vidéos de football
- apprentissage de la structure statique
- modèle de Markov pour la structure dynamique



Principaux résultats

- possibilité d'apprendre la structure
- nécessité d'un critère adapté pour la classification
- difficulté d'apprentissage de la structure temporelle



Peut-on combiner modèles de segments et réseaux bayésiens ?

Bilan et perspectives

On a mis en évidence

1. l'avantage d'une modélisation conjointe
2. la souplesse offerte par les modèles de segments
3. la capacité des réseaux bayésiens pour modéliser les dépendances entre descripteurs

mais deux problèmes majeurs demeurent

- modéliser les dépendances entre descripteurs
- gérer la synchronisation entre flux de descripteurs

Pour cela, nous proposons de

1. aller plus loin sur l'apprentissage de structure

   $\longrightarrow$ modèle temporel, variables conceptuelles, *etc.*

2. combiner modèle de segments et réseaux bayésiens

Pourquoi une seule séquence d'états ?

Partie II

Intégration de connaissances hétérogènes

analyse de documents contenant de la parole

Thèses de Stéphane Huet, Gwénolé Lecorvé, Camille Guinaudeau,
en étroite collaboration avec P. Sébillot et le groupe RAPTAL.

Connaissances et parole

parole = acoustique, phonétique, morphologique, syntaxique, sémantique, pragmatique

En pratique

- *a posteriori* : observations acoustiques
- *a priori* : lexique de prononciation, modèle de langage (syntaxe faible)

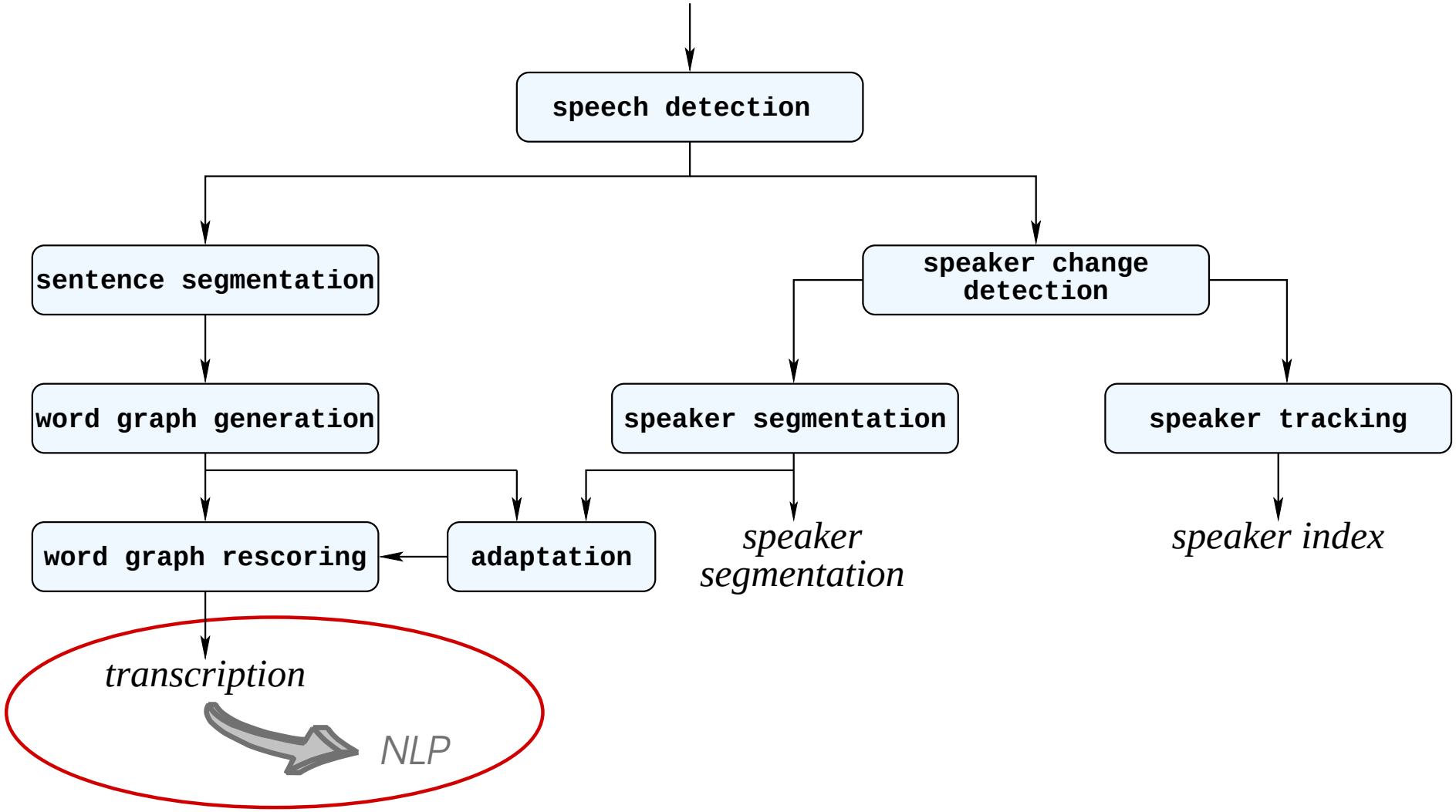
$$\widehat{w}_1^{L^*} = \arg\max_{w_1^L} \ln p(o_1^T | w_1^L) + \ln P[w_1^L]$$

Objectif : introduire de nouvelles sources de connaissances pour

1. améliorer la transcription
2. exploiter la transcription

en se plaçant à la frontière entre traitement automatique de la parole et traitement automatique des langues

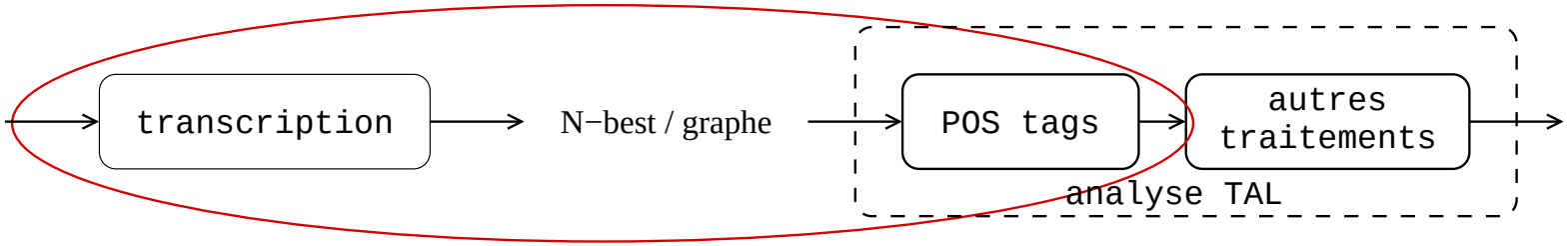
La plateforme Irene



Travaux de thèses de M. Ben et W. X. Teng . Intégration P. Cauchy. Collaboration : F. Yvon.

Validation : ESTER 1 & 2 [Galliano *et al.*, 05 & 09].

Intégration de connaissances morphosyntaxiques [Huet *et al.*, 07]



1. étiquetage morphosyntaxique de chaque hypothèse selon,
   $\widehat{c} = \arg\max_{c} P[c]P[w|c]$, *e.g.*

```
une   date  qui  à   donner  le   vertige  à   une   partie  de   la   france
_une  NCFS  _qui  _à  VINF    _le  NCMS     _à  _une  NCFS    _de  _la  NPFS
```

2. réordonancement des hypothèses sur la base d'un score a linguistico-acoustico-syntaxique

$$\ln p(w;o) = \ln p\left(o|w\right) + \beta \ln\left(P[w]\right) + \alpha \ln P[\widehat{c}]$$

```
une  date  qui    a       donné    le   vertige  à   une   partie  de   la   france
_une NCFS  _qui   AVOIR3S VPARPMS  _le  NCMS     _à  _une  NCFS    _de  _la  NPFS
```

Intégration de connaissances morphosyntaxiques (suite)

- ESTER eva05, LM 4g, POS LM 7g, 100-best

	sans POS	avec POS
réordonnancement	24,7	23,9
consensus	24,0	23,5

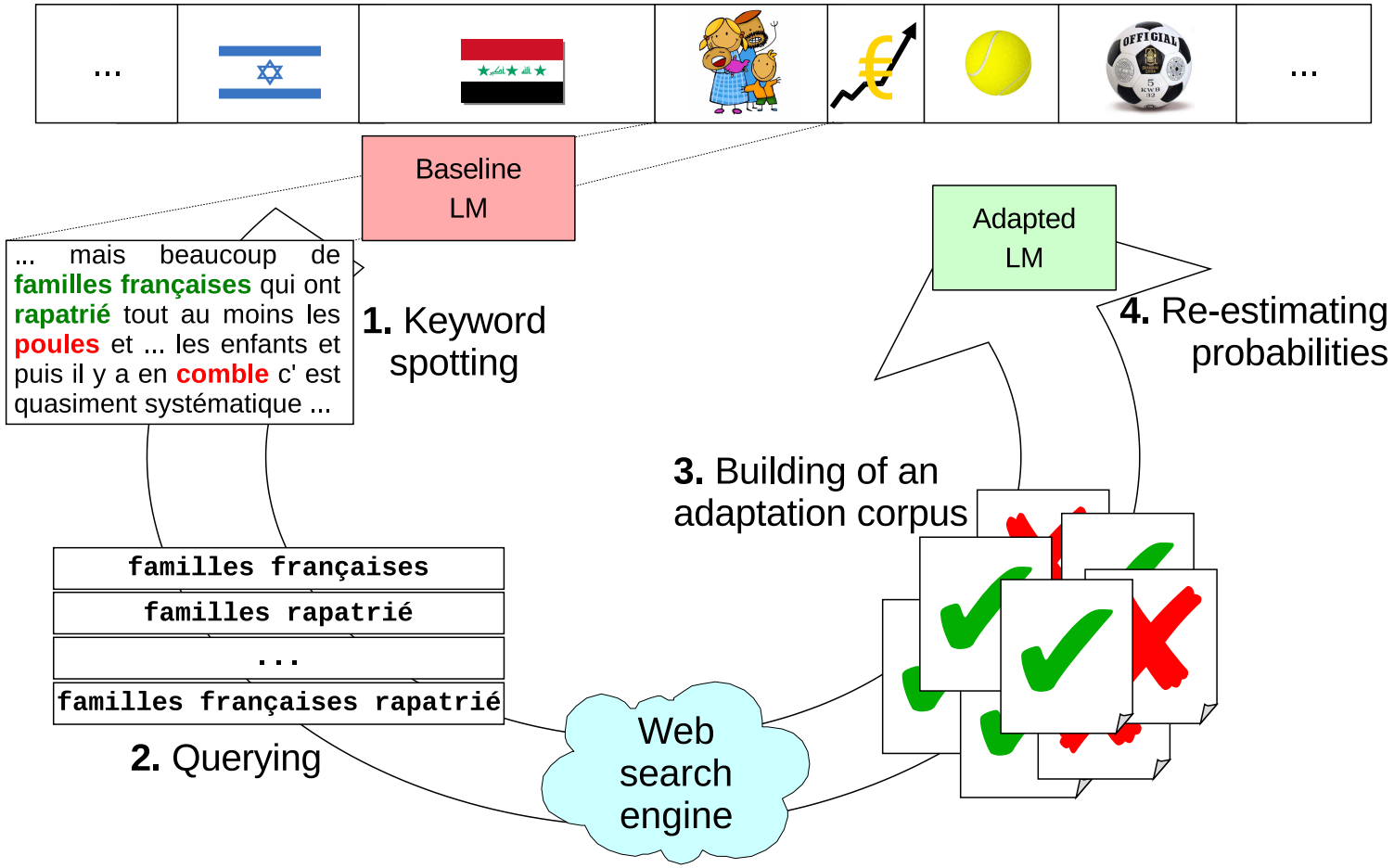
%WER sur ESTER eva05

- apport d'une nouvelle connaissance
- apport d'un modèle d'ordre plus élevé
- amélioration des mesures de confiance

- autres types de connaissances à considérer
  - sémantique
  - syntaxe forte

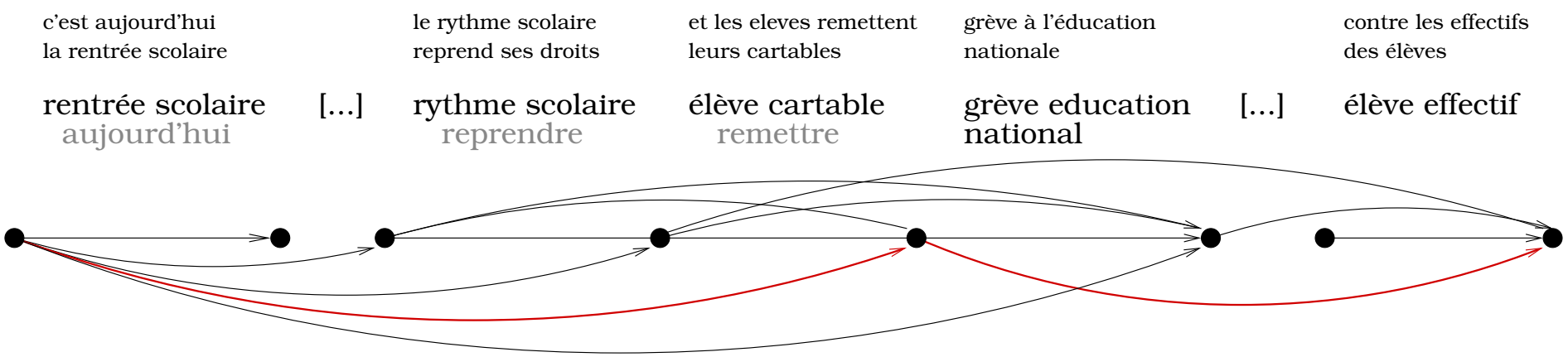
  → projet RAPSODIS

Intégration de connaissances pragmatiques [Lecorvé *et al.*, 08]



- collaboration entre modules (intégration tardive)
- transcriptions => modifications des critères RI/TAL
- connaissances sémantique et linguistique pour adaptation MDI [Lecorvé *et al.*, 09]

Segmentation thématique [Utiyama *et al.*, 02]



valuation des arcs = cohésion lexicale

1. estimation d'un modèle unigramme (avec lissage)
2. calcul de la probabilité de la séquence de mots

Modèle proche d'un modèle de segments

$$\widehat{s}_1^{L^*} = \arg\max_{s_1^L} \ln p(w_1^n | s_1^L) + \ln p(s_1^L)$$

Segmentation thématique par modèle de segments multiflux

- extension du modèle à deux nouvelles sources d'informations

$$\widehat{s}_1^{L^*} = \arg\max_{s_1^L} \ln p(w_1^n|s_1^L) + \underbrace{\beta_a \ln p(a|s_1^L)}_{\text{acoustique}} + \underbrace{\beta_m \ln p(m|s_1^L)}_{\text{morphosyntaxe}} + \gamma \ \ln p(s_1^L)$$

- modèle acoustique
  - attributs = changement de sexe du locuteur, pauses, musique
  - modèle = arbre de décision
- modèle morphosyntaxique
  - attributs = séquence des étiquettes morphosyntaxiques
  - modèle = N-gramme caché
- extension à d'autres connaissances

  images structurantes [Quaero], sémantique [C. Guinaudeau]

$\Rightarrow$ inclure la parole dans un modèle de segments !

Bilan et perspectives

On a mis en évidence

1. l'intérêt d'un meilleur couplage entre RAP et TAL
2. l'adéquation du modèle de segments pour prendre en compte la parole
3. la possibilité de combiner la parole avec d'autres indices

mais il est nécessaire de **mieux compenser les erreurs de transcriptions**

Perspectives

1. vers toujours plus de couplage entre RAP et TAL
   - introduction de relations sémantiques
   - extraction d'informations de la transcription
   - recherche d'information (thèse J. Fayolle)
2. effectivement combiner la transcription avec d'autres indices

exploiter la transcription pour structurer un flux TV
en conjonction avec d'autres indices (visuels)

Partie III

Intégration de connaissances sporadiques

alignement dynamique guidé

Thèses de Xavier Naturel, Emmanouil Delakis ; Post-doc de Daniel Moraru.

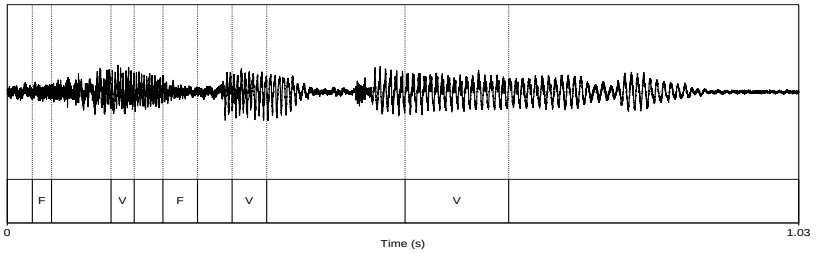
Connaissances sporadiques

Qu'est-ce qu'une connaissance sporadique ?

⟶ connaissance qui porte sur *une partie* de l'objet à reconnaître

Quelques exemples

- images-clés pour l'alignement flux/EPG [Naturel *et al.*, 06]
- scores affichés dans les vidéos de tennis [Delakis *et al.*, 06]
- points d'ancrage phonétiques [Gravier *et al.*, 07]

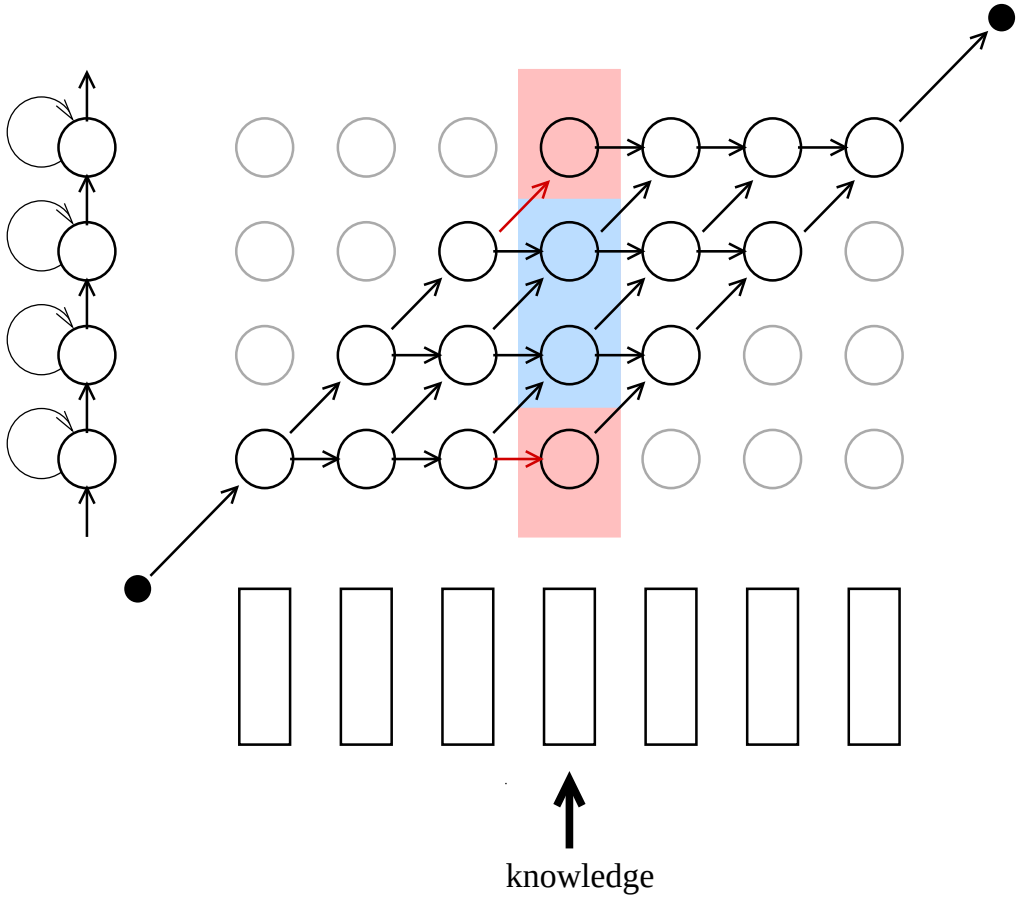


- fusion de système de transcription [Lecouteux *et al.*, 08]

⇒ un même paradigme, l'**alignement dynamique guidé**

UMR IRISA

Principe de l'alignement dynamique guidé



landmarks

=

indices locaux sur le meilleur chemin

⇓

pénaliser les chemins incompatibles

$$S_j(t) = \min_i S_i(t-1) + c_{ij} + d_j(t) + r_j(t)$$

= modèle multiflux ou modèle non stationnaire

Vers une approche collaborative

On a mis en évidence

- un paradigme valide si l'on dispose d'une connaissance sûre
- avec des résultats positifs dans des cas réalistes (EPG, tennis, combinaison de systèmes de transcription)

à valider dans un cadre réaliste

$\longrightarrow$ transcription guidée par des ancres phonétiques

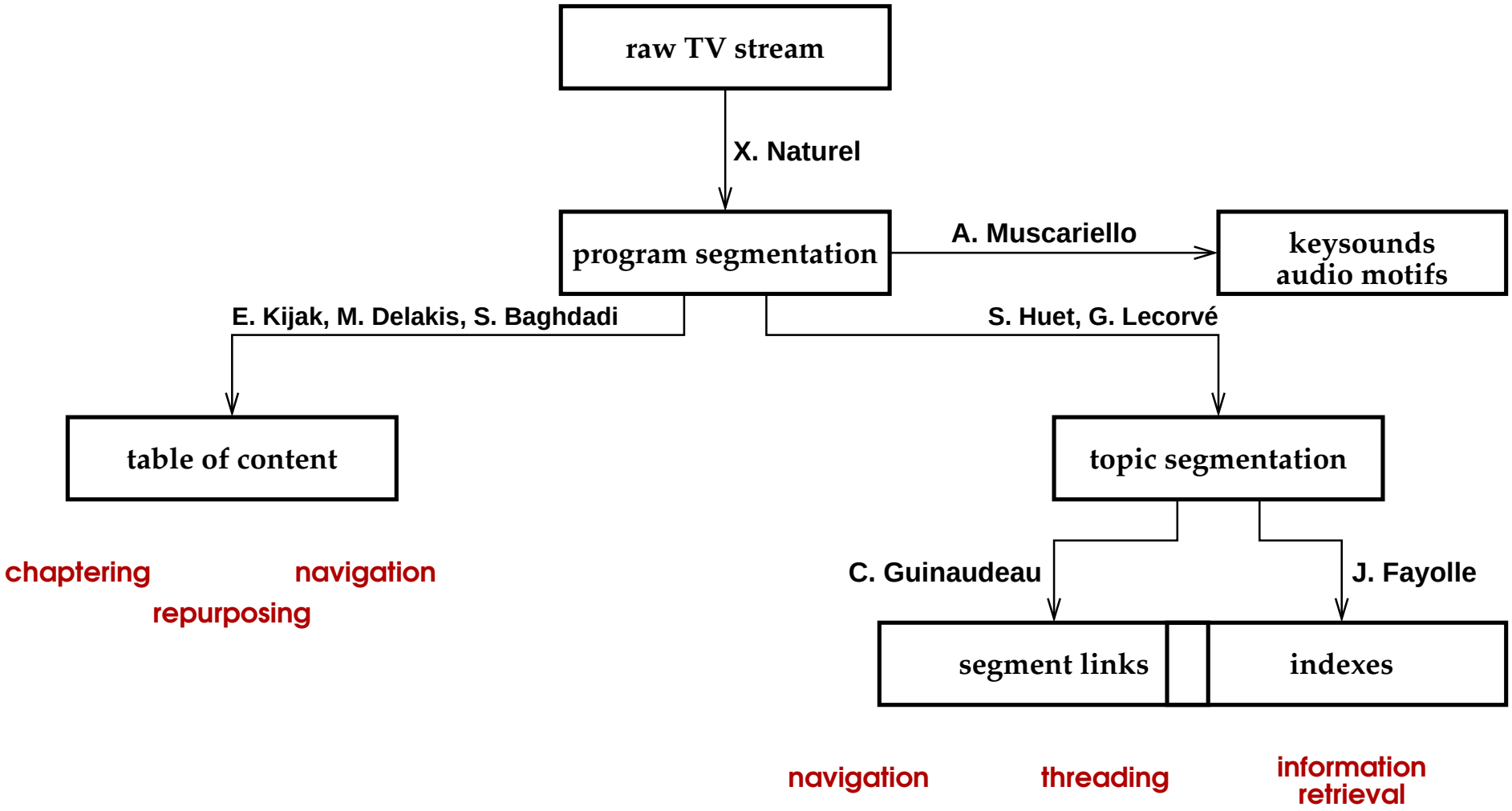
De nombreuses questions sur le plan scientifique

- comment choisir les connaissances à injecter
- coapprentissage des systèmes
- après le coapprentissage, le codécodage !

Partie IV

Quelques pistes de travail

Les éléments pour structurer un flux télévisé



Vers une intégration – génération d'hypervidéos

hypervidéo
=
"flux" vidéo enrichi d'hyperliens pour la navigation

- macro- et micro-segmentation
- liens endogènes et exogènes
- indexation

Pour répondre à plusieurs objectifs (technico-)scientifiques

- Comment combiner les techniques de structuration existantes ?
- Quel impact ont les erreurs pour un utilisateur ?
- Que signifie "naviguer" dans un flux télé ?
- Quels axes de recherche pour les services du futur ?

Retour vers le futur...

modèles conjoints
=
beaucoup de paramètres

+

décodage guidé
=
intégration tardive

⇓

analyse collaborative

Coupler plusieurs algorithmes qui s'échangent en permanence des informations permettant une influence croisée des analyses

- Comment synchroniser les modules d'analyse ?
- Quelles informations échanger ?
- Comment influencer un module par les résultats d'un autre ?

⟶ projet ANR Attelage de Systèmes Hétérogènes (ASH)

Vers un monde sans contraintes

Annotation et *design* = $$$$$$$$

⇒ **limiter, voire supprimer, la supervision !**

- apprentissage faiblement supervisé (*bootstrap*)
- sélection et apprentissage de descripteurs
- apprentissage de structure dans les réseaux bayésiens
- découverte de motifs [Muscariello *et al.*, 09]

⇒ apprendre ce qui est détectable

plutôt que d'apprendre à détecter

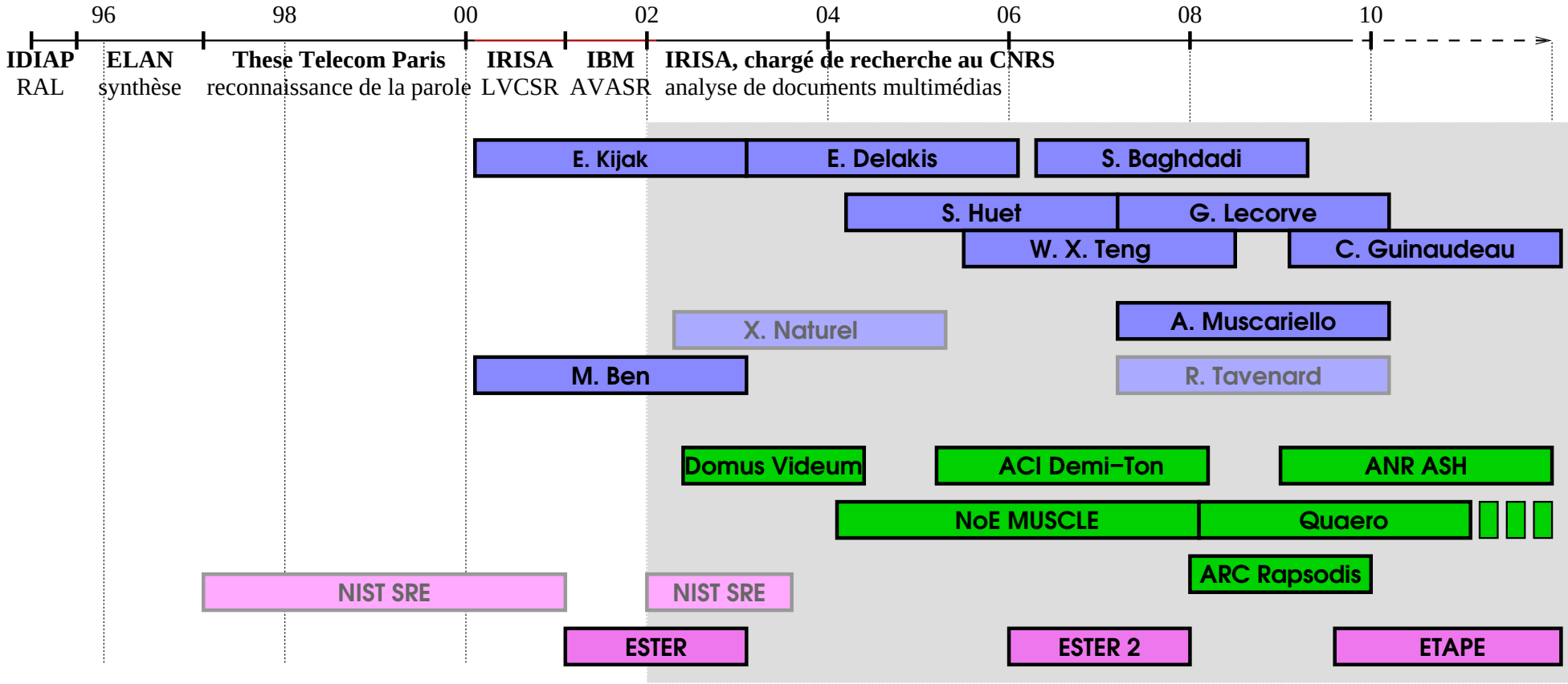
Merci aux gens qui ont fait le travail avec/pour moi (par ordre alphabétique) : S. Baghdadi, M. Betser, E. Delakis, C. Guinaudeau, S. Huet, E. Kijak, G. Lecorvé, A. Mohan, D. Moraru, A. Muscariello, A. Ozerov, R. Tavenard, W. X. Teng et S. Ziegler.

Mes travaux ont largement bénéficié de la collaboration de (par ordre alphabétique) : L. Amsaleg, R. André-Obrecht, M. Ben, F. Bimbot, J.-F. Bonastre, S. Campion, P. Deléglise, C.-H. Demarty, Y. Estève, G. Linarès, S. Galliano, E. Geoffrois, R. Gribonval, P. Gros, S. Meigner, F. Moreau, L. Oisel, P. Sébillot, F. Soufflet, L. Villaseñor Pineda et F. Yvon.

Sans compter tous ceux que j'ai oubliés dans ce transparent !

ANNEXES TECHNIQUES

Mon parcours



D’autres travaux

Mathieu Ben (2004) – reconnaissance du locuteur

- comparaison de modèles de mélanges
- adaptation MAP hiérarchique

Wen Xuan Teng (2008) – adaptation rapide au locuteur

- interpolation de modèles de références
- sous-espace de référence variable

Romain Tavenard (2010) – comparaison de séquences

- modélisation de séquences par SVM
- approximations rapides de la DTW

Armando Muscariello (2010) – découverte de motifs sonores

- algorithmes de reherche approximative
- découverte de motifs par alignement dynamique

Projets, campagnes d'évaluation

Projets

- internationaux : Pelops, NoE MUSCLE, Quaero
- nationaux : Domus Videum, Demi-Ton, ESTER, ETAPE, Rapsodis, ASH

Évaluations

- NIST SRE
- ESTER 1 & 2
- The Star Challenge (avec NII)

⟶ travail collaboratif dans un cadre réaliste
⟶ comparaison objective des méthodes
⟶ animation de la communauté

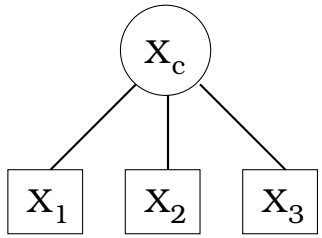
Réseaux bayésiens et apprentissage de structure

- un formalisme permettant de décrire des interactions complexes
- MMC & modèle de segments sont des cas particulier [Murphy, 02]

$\Rightarrow$ **apprentissage de la structure**

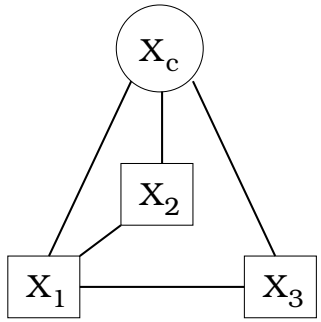
- proposition de plusieurs critères d'apprentissage
  - K2 (Bayesian Information Criterion) [Cooper *et al.*, 92]

$$\mathcal{L}(X_i) = \ln p(X_i | \{X_j \quad \forall j \in \mathcal{P}_i\}) - \frac{\lambda}{2} C(X_i, \mathcal{G})\ N$$



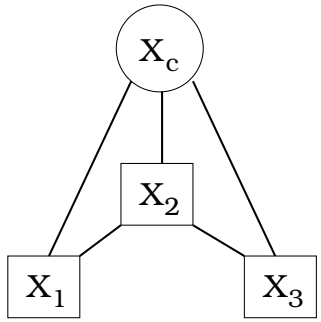
  - Tree Augmented Network [Friedman *et al.*, 97]
  - K2 Augmented Network [Baghdadi *et al.*, 09]

$$\mathcal{L}(X_i) = \ln p(X_i | \{X_j \quad \forall j \in \mathcal{P}_i\}, X_c) - \frac{\lambda}{2} C(X_i, \mathcal{G})\ N$$

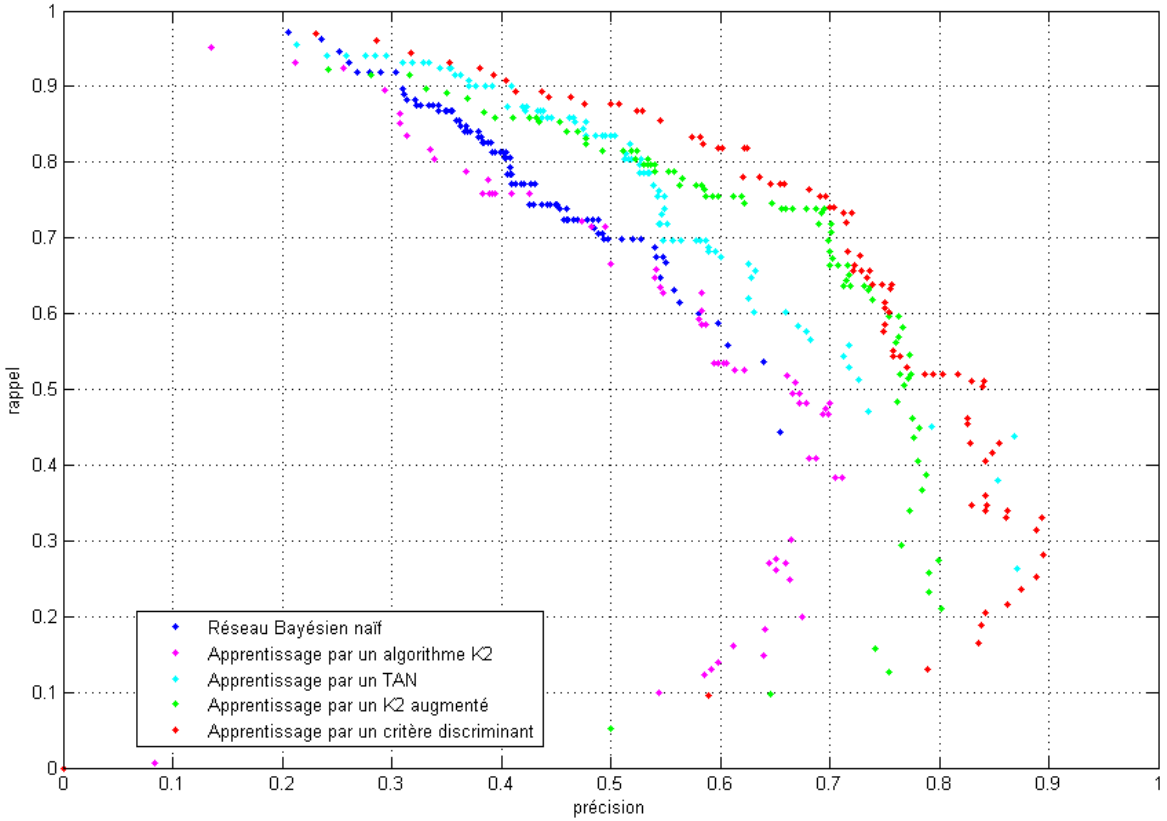


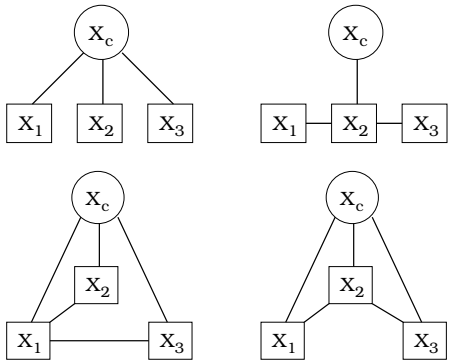
  - critère discriminant [Baghdadi *et al.*, 09]

$$\text{CLL}(\mathcal{G}|\mathcal{D}) = \sum \ln p_{\mathcal{G}}(X_c | X_1^N)$$



Réseaux bayésiens et apprentissage de structure (suite)





+ apprendre la structure, c'est possible et c'est mieux !
+ avantage du critère discriminant pour la classification
- requiert une sélection des descripteurs pertinents

Segmentation thématique par modèle de segments multiflux

$$\widehat{s}_1^{L^*} = \arg\max_{s_1^L} \ln p(w_1^n|s_1^L) + \underbrace{\beta_a \ln p(a|s_1^L)}_{\text{acoustique}} + \underbrace{\beta_m \ln p(m|s_1^L)}_{\text{morphosyntaxique}} + \gamma \ \ln p(s_1^L)$$

Inclusion d'un score acoustique et d'un score morphosyntaxique au poids d'un arc

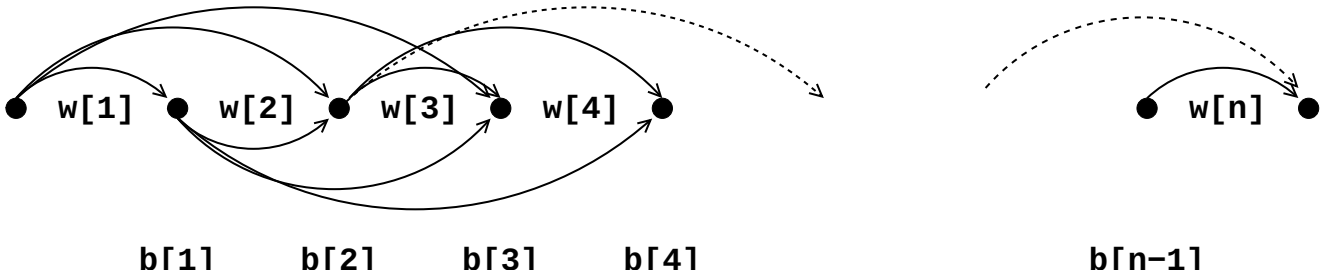
$$C(S_i^j) = C(S_i^j|w) + \sum_{x\in\{a,m\}} \sum_{k=i}^{j-1} lnP(B_k = 0|x) + \ln P(B_j = 1|x)$$

* arbre de décision pour le calcul de $P(B_k|a)$

  attributs = changement de sexe du locuteur, pauses, musique

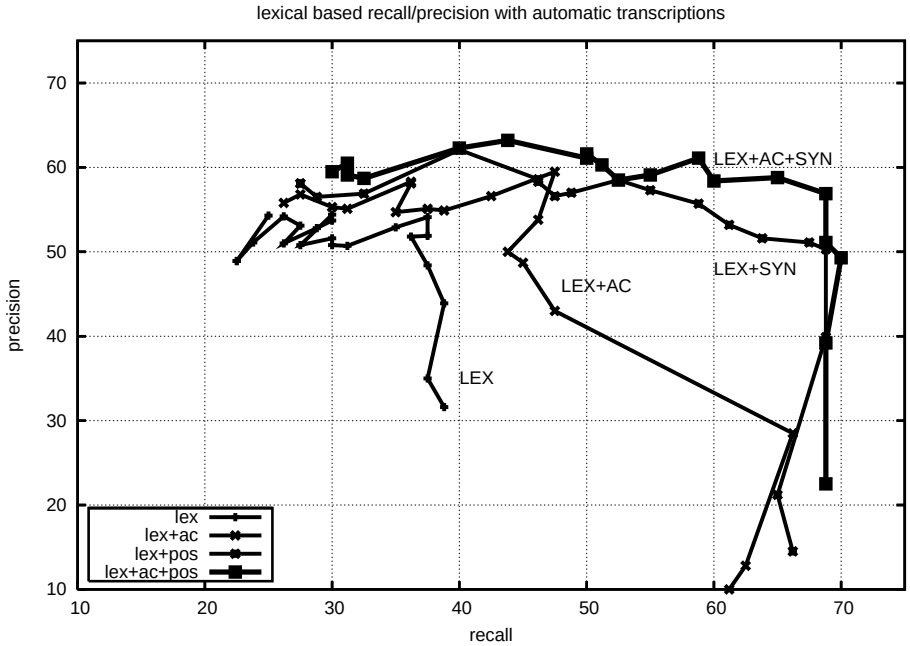
* N-gramme caché pour le calcul de $P(B_k|m)$

  attributs = mots ou etiquettes morphosyntaxiques



Résultats [Huet *et al.*, 08]

- segmentation d'émissions radiophoniques (ESTER 1, dev05b)



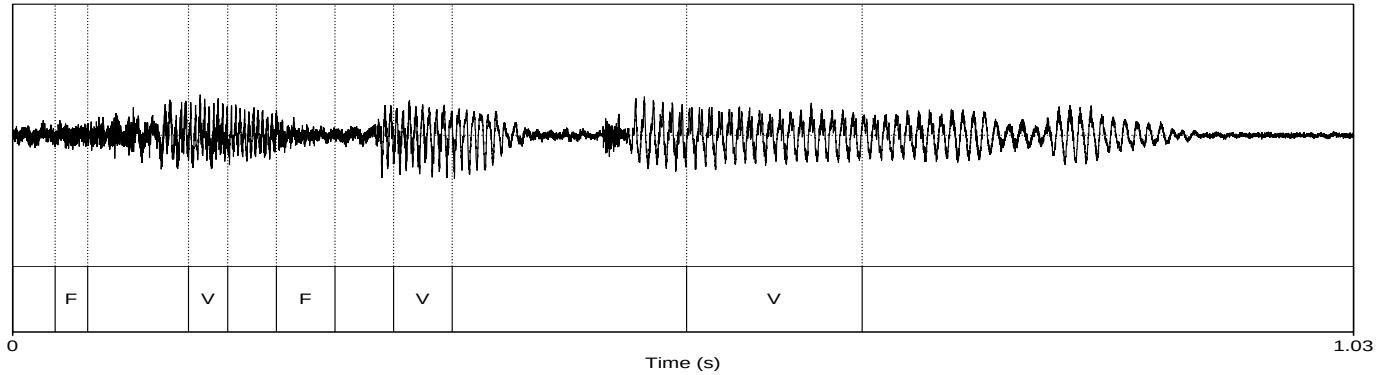
+ intégration des connaissances au niveau du segment
+ apport de la syntaxe à la radio
- apprentissage des modèles

- autres sources d'information segmentale = image, prosodie
- sémantique au niveau de l'estimation du LM [travail de C. Guinaudeau]

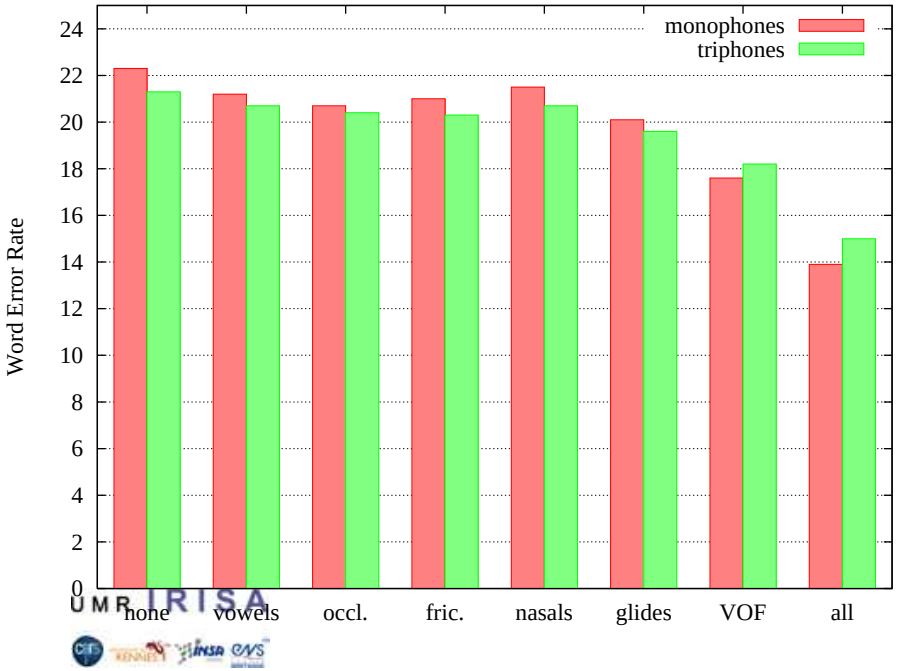
⇒ intégration de connaissances hétérogènes dans un modèle de segments

Cas des ancres phonétiques [Gravier *et al.*, 07]

- utilisation d'informations locales sur le contenu phonétique



- étude préliminaire sur l'apport d'ancres macrophonétiques (oracle)



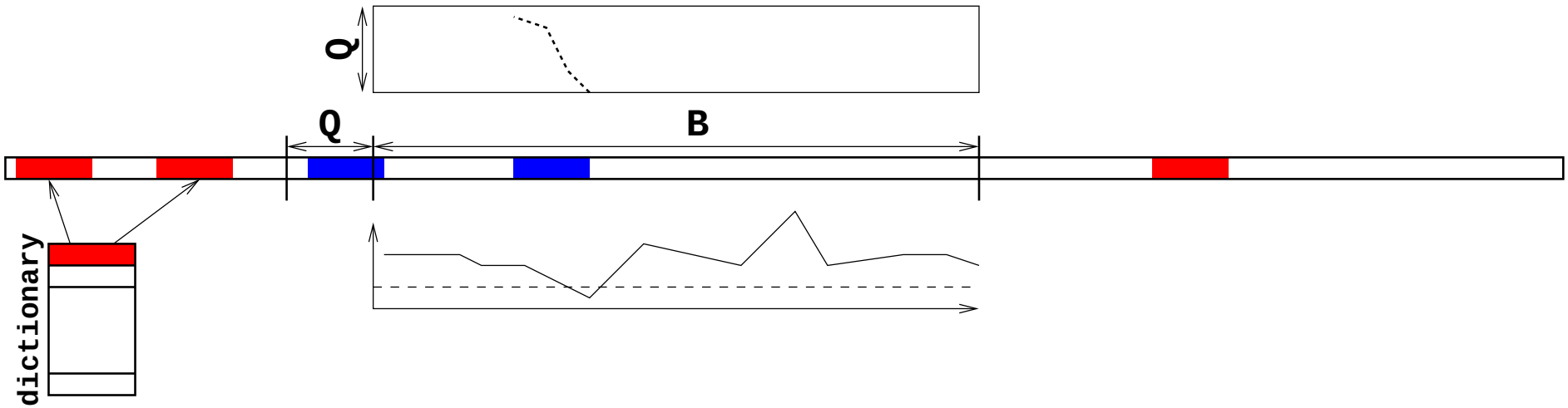
- chaque classe apporte un léger gain
- peu sensible aux frontières des ancres
- peu sensible aux omissions (linéaire)
- et en pratique ????

Découverte de motifs sonores

- Finding repetitions in audio streams
  - not as “simple” as with images (no equivalent of SIFT descriptors)
  - DTW based comparison might be slow

  ⇒ summarize data sequences with a (SVM) model and compare/index models directly

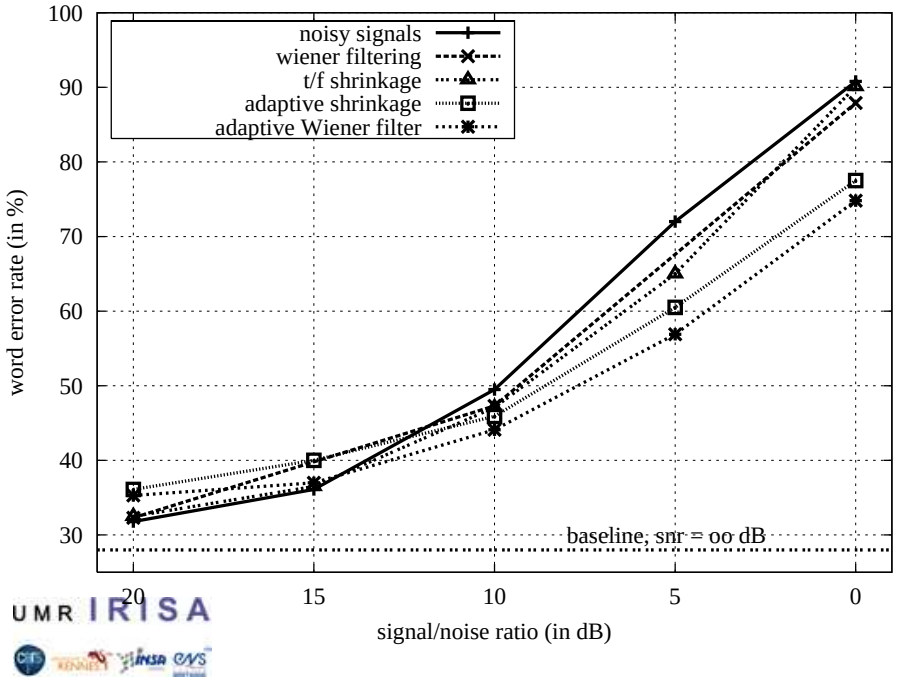
- Word discovery



Motif 1: #1, #2, #3, #4, #5, #6; Motif 2: #1, #2, #3, #4; Motif 3: #1, #2

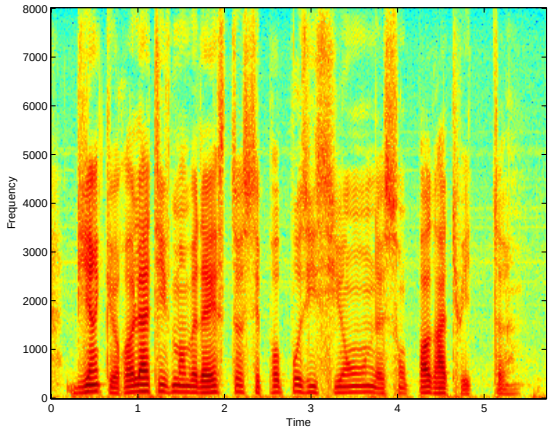
Source separation and localization

- Applications
  - source separation: over and under determined cases
  - source localization and numbering (under-determined stereo case)
  - multisensor audio scene analysis
- Techniques
  - sparse decomposition (matching pursuit): decompose a signal on a small set of *atoms*, localized in time and frequency
  - beam-forming (to a lesser extent)

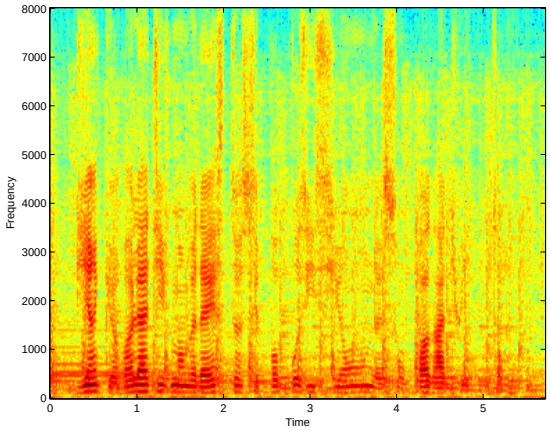


An example of application: speech signal denoising using single sensor source separation techniques for noise robust speech recognition.

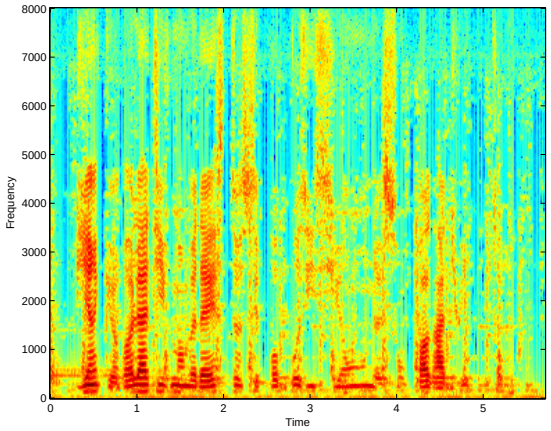
Source separation for denoising



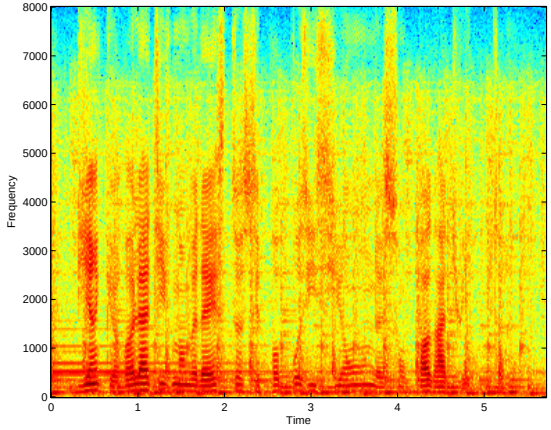
initial



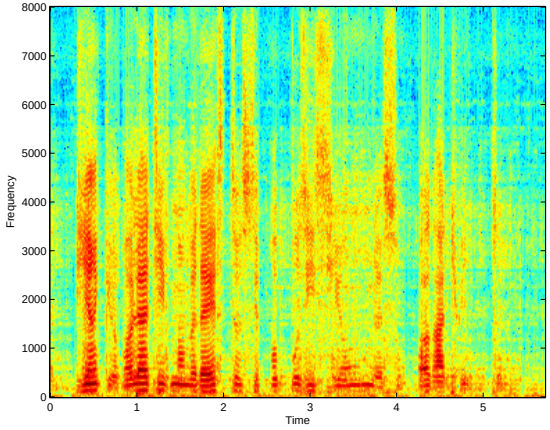
wiener



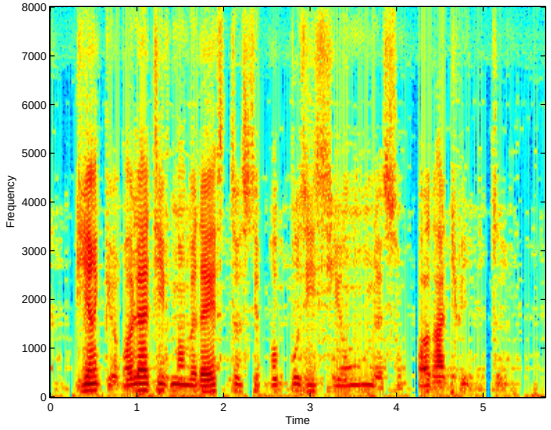
seuillage



bruit



wiener adaptatif



seuillage adaptatif

Speaker characterization

- Fast and compact models for speaker recognition
  - Frame selection and quantization, fast Gaussian selection, sub-Gaussian quantization
  - Decision trees to approximate the normalized log-likelihood ratio function
  - Model space approaches to replace the log-likelihood function by a distance between model parameters, easier to compute
- SVM approaches in model space
- applied to speaker verification (NIST evaluations)
- applied to speaker segmentation and tracking

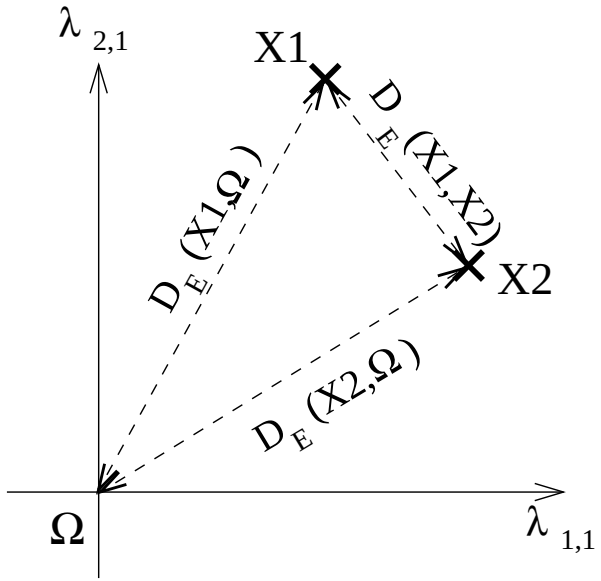
Speaker characterization (cont'd)

Fast and compact approaches for speaker verification and segmentation.

- Decision trees [Gonon *et al.*, 2005]
  - approximate the score function with a decision tree
  - polynomial interpolation in the leaves
  - only a slight degradation

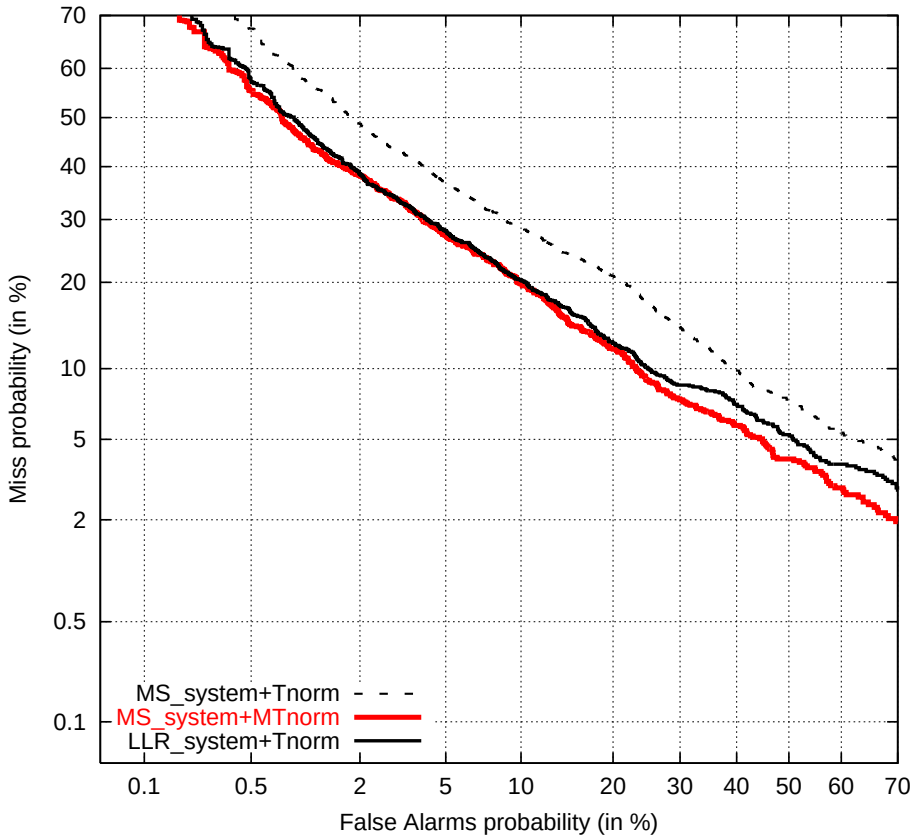
- Model space distance [Ben *et al.*, 2005]
  - distance between adapted models
  - approximation of the KL divergence
  - model normalization
  - used for verification and segmentation

- Fast Gaussian likelihood approximations
  - frame quantization
  - Gaussian selection
  - sub-Gaussian quantization



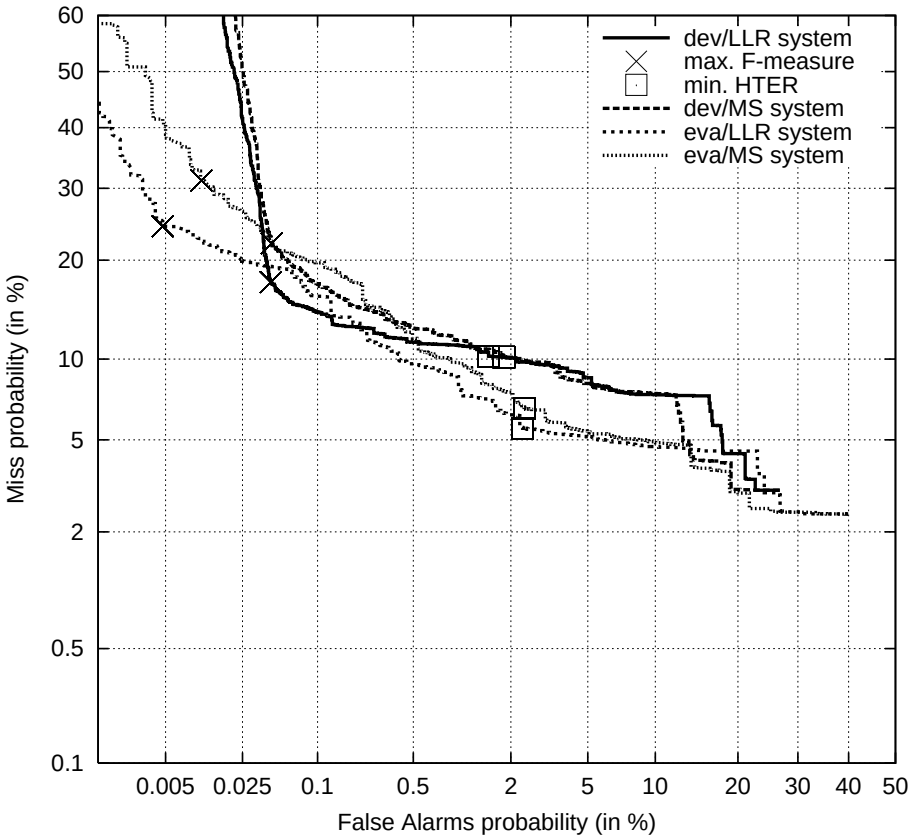
Speaker characterization (cont'd)

Model space distance applied to speaker verification (NIST SRE 04) and tracking (ESTER 2005).



NIST SRE 04

reduction of 75%



SVL, Ester Phase 2

reduction of 60%

UMR IRISA