Présentation de travaux en Reconnaissance de la Parole

Bruno Jacob

Maître de Conférence stagiaire au LIUM

Points abordés

- Introduction à la RAP et aux MMC
- Vérification du locuteur (PICASSO/IRISA)
- Fusion de données (AMIBE/IRIT)
- Dictée Vocale (SIROCCO/IRISA)

Introduction à la RAP

Deux grandes tendances en Reconnaissance Automatique de la Parole :

1. Répondre à la question: “Qui parle ?”
   - Reconnaissance d’un locuteur parmi N
   - Vérification du locuteur (est-ce bien lui ?)
2. Répondre à la question: “Que dit-on ?”
   - Reconnaissance de mots isolés (nb mots $\searrow$ nb locuteurs $\nearrow$)
   - Reconnaissance grand vocabulaire (nb mots $\nearrow$ nb locuteurs ≈ 1)

   Système grand vocabulaire indépendant du locuteur
   $\Rightarrow$ pas encore pour tout de suite

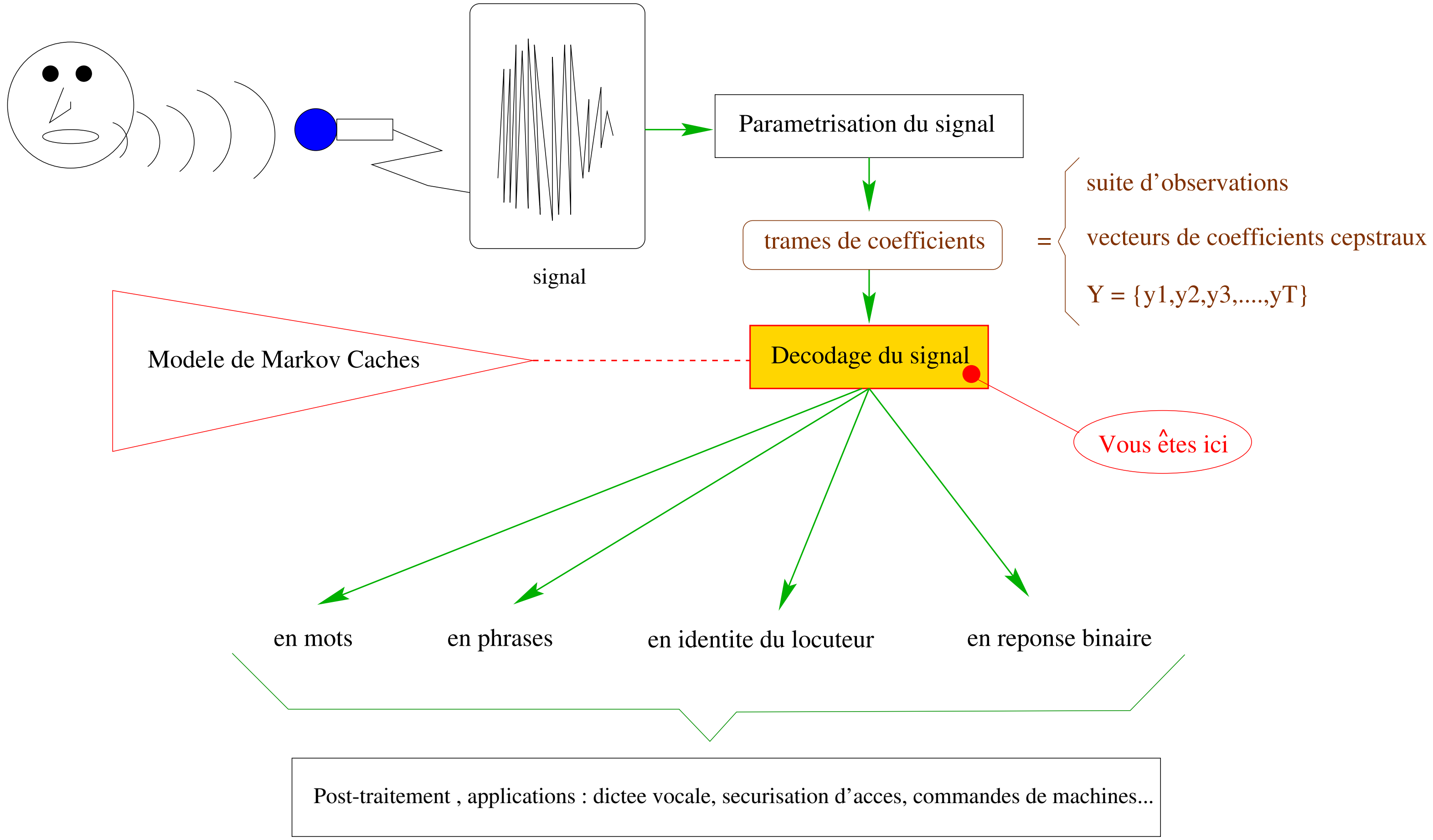
Difficultés de la RAP

Mot clé : très grande **variabilité** de la voix .

- entre les personnes
  - caractéristiques physiologiques
  - origine géographique
  - contexte socio-culturel ...
- pour une personne
  - non reproductibilité (état de santé, nervosité ...)
  - bruit ambiants, distortion du canal ...
  - modification intentionnelle

$\Rightarrow$ on a une *signature* vocale et non une *empreinte* vocale

Là où on intervient dans un système de RAP

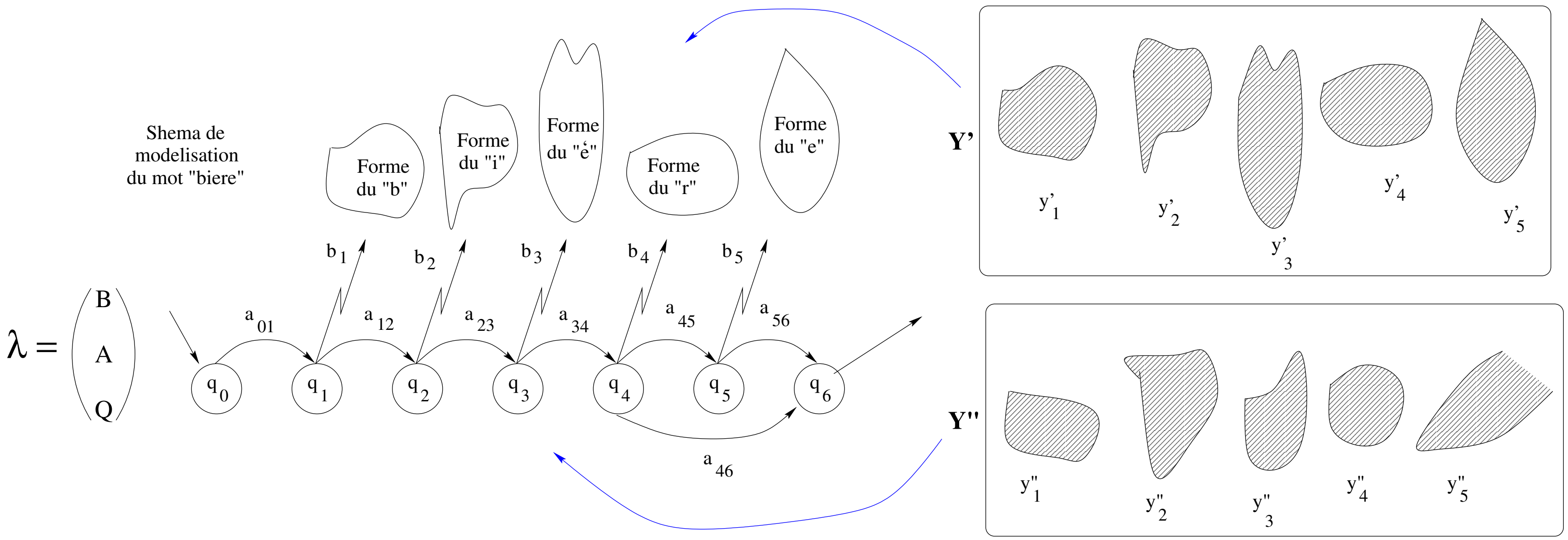


Les modèles de Markov Cachés (*ou HMM*)

Que sont les MMC ?

- ce sont des outils mathématiques pour la reconnaissance des formes.
- ils sont utilisés en Parole pour reconnaitre les formes acoustiques du signal
- on peut les voir comme des graphes orientés et probabilisés

Quelle tête ça a ?



$$P(Y'/\lambda) \gg P(Y''/\lambda)$$

2 phases :

1. Apprentissage à partir d'observations connues
2. Reconnaissance d'une suite d'observations inconnues

À la base, dans les MMC, tout est simple

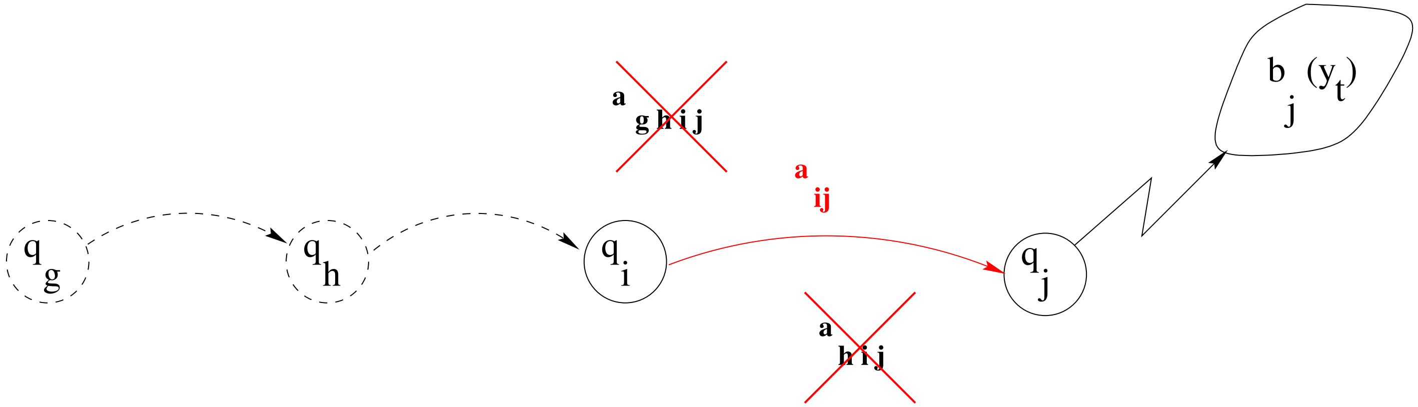
L'objet modélisé :

1 MMC ⇔ reco d'1 forme acoustique

La topologie

Parole : on utilise en général des MMC d'ordre 1

⇒ la proba d'aller dans un état ne dépend que de l'état précédent

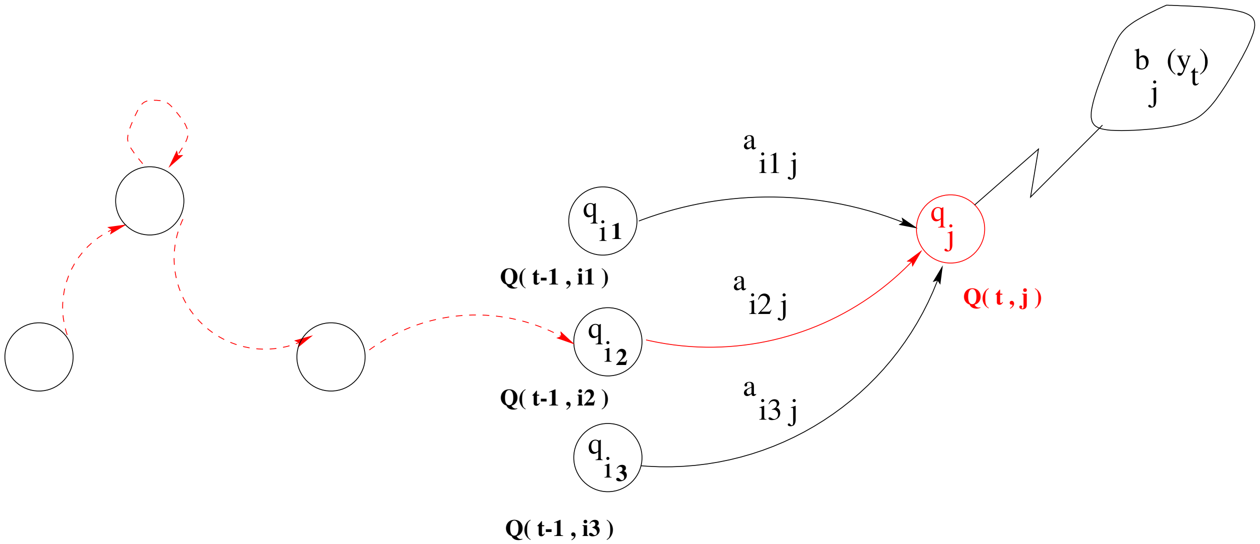


Exploitation simple des MMC

En général on utilise l'algo de (***Viterbi***)

→ Calcul de $P(Y/\lambda)$ = Recherche du meilleur chemin dans le MMC λ ayant généré une suite d'observations Y

$$Q(t, q_j) = \max_i [a_{ij}\ b_j(y_t)\ Q(t-1, q_i)]$$



Mais la mode est de faire des modèles et/ou des traitements de plus en plus complexes

Introduction aux cadres d'études

- Dans le thème "Qui Parle ?"

  Vérification du *qui*

  → *Les MMC modélisent les caractéristiques du locuteur*

  → Vérification Automatique du Locuteur par ses caractéristiques acoustiques

- Dans le thème "Que dit-on ?"

  Reconnaître mieux

  → *Topologie complexe des MMC*

  → Fusion de plusieurs sources d'informations pour rendre plus robuste la reconnaissance

  Reconnaître plus

  → *Algorithme d'exploitation complexe des MMC*

  → Système de Dictée Vocale pour reconnaître de grands vocabulaires en parole continue

Généralités sur la Vérification du Locuteur

$\boxed{\text{Vérification du Locuteur}}$

=

Processus de décision consistant à utiliser des caractéristiques du signal de parole pour déterminer si un individu donné est ou non le locuteur d'un énoncé particulier

Un locuteur X proclame son identité. Est-ce bien le locuteur X ?
⇒Réponse : oui/non

→ Représentation et modélisation des caractéristiques d'un locuteur
→ Élaboration de mesures de ressemblances et de stratégies de décisions

Types d'erreurs :

- **Fausse acceptation** (non-détection)
  un imposteur est accepté par le système
- **Faux rejet** (fausse alarme)
  un utilisateur authentique (client) est rejeté

$\Rightarrow$ **SEUIL de DÉCISION**

Approche statistique

- Mise en compétition de 2 modèles probabilistes :
  - Client (X) → énoncé prononcé par le client
  - Monde (Ω) → énoncé prononcé par l'ensemble de la population
- 2 phases :

  Apprentissage : estimer les paramètres des modèles Client et Monde

  Reconnaissance : Pour 1 énoncé donné (observation Y) → 2 scores de vraisemblance
  - Client → $\widehat{P}(Y \mid X)$
  - Monde → $\widehat{P}(Y \mid \Omega)$

- Décision :
  - rapport des scores de vraisemblance

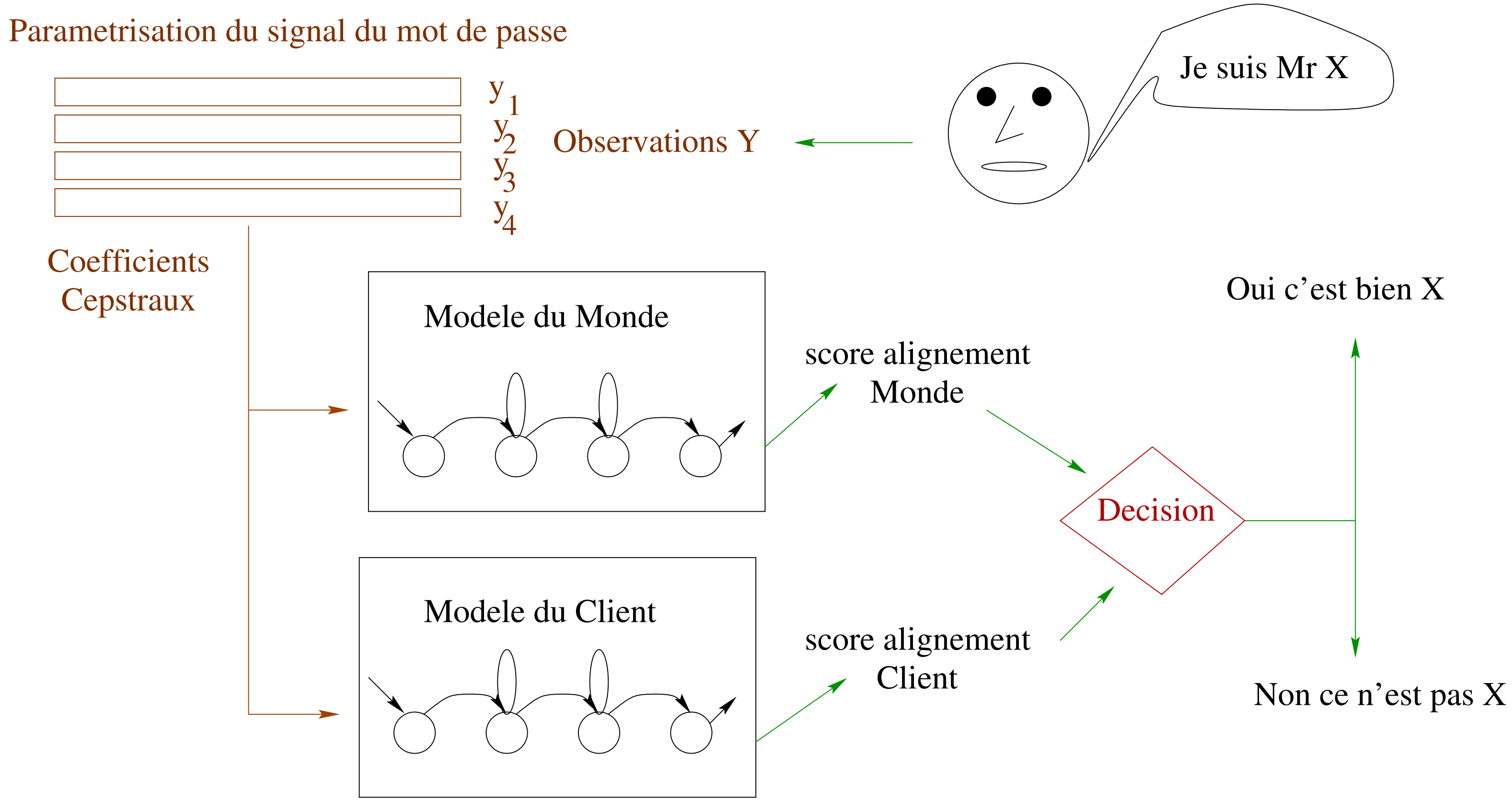
$$llr_X(Y) = log\Big(\frac{\widehat{P}(Y \mid X)}{\widehat{P}(Y \mid \Omega)}\Big)$$

  - comparaison par rapport à un seuil dépendant ou non du Client

$$llr_X(Y) \quad \underset{\text{rejet}}{\overset{\text{acceptation}}{\gtrless}} \quad \theta_X(R, Y)$$

$$\text{avec } R = \frac{P_{imp} C_{FA}}{P_{cli} C_{FR}}$$

Schéma d'un système de vérification du locuteur



Vérification du locuteur dans le cadre du projet PICASSO

Plan

- Présentation du projet PICASSO
- Travaux sur le mot de passe personnalisé

Présentation du projet PICASSO

- Projet de la CE LE-Telematics Programme (LE4-8369) + OFES (projet 97.0494-2)
- Début en Janvier 1998 (suite du projet CAVE)
- Durée : 30 mois
- Partenaires : KPN-Telecom (NL), ENST (F), Fortis (NL), IDIAP (CH), IRISA (F), KPN-Research (NL), KTH (S), KUN (NL), Swisscom (CH), UBS-Ubilab (CH), Vocalis (UK)

Difficultés

Améliorer la ROBUSTESSE du système

- aux conditions de prise du son
- au manque de données d'apprentissage
- au manque de données de test
- à la variabilité de la voix dans le temps
- aux impostures intentionnelles

Robustesse aux fraudes

- Étude des techniques d'impostures par
  - Mots ou chiffres concaténés
  - *Sous-mots* concaténés + synthèse

  Lindberg et al. : Vulnerability in Speaker Verification. A study of technical impostor techniques, Eurospeech'99, S6.PO3.7, Vol 3, p 1211

- Résultats : avec un système vocabulaire figé → imposture efficace par simple concaténation de mots synthétiques
- Prévention
  - détection des voix de synthèse
  - utiliser des mots rares
  - utiliser des mots de passe personnalisés

Mot de passe personnalisé

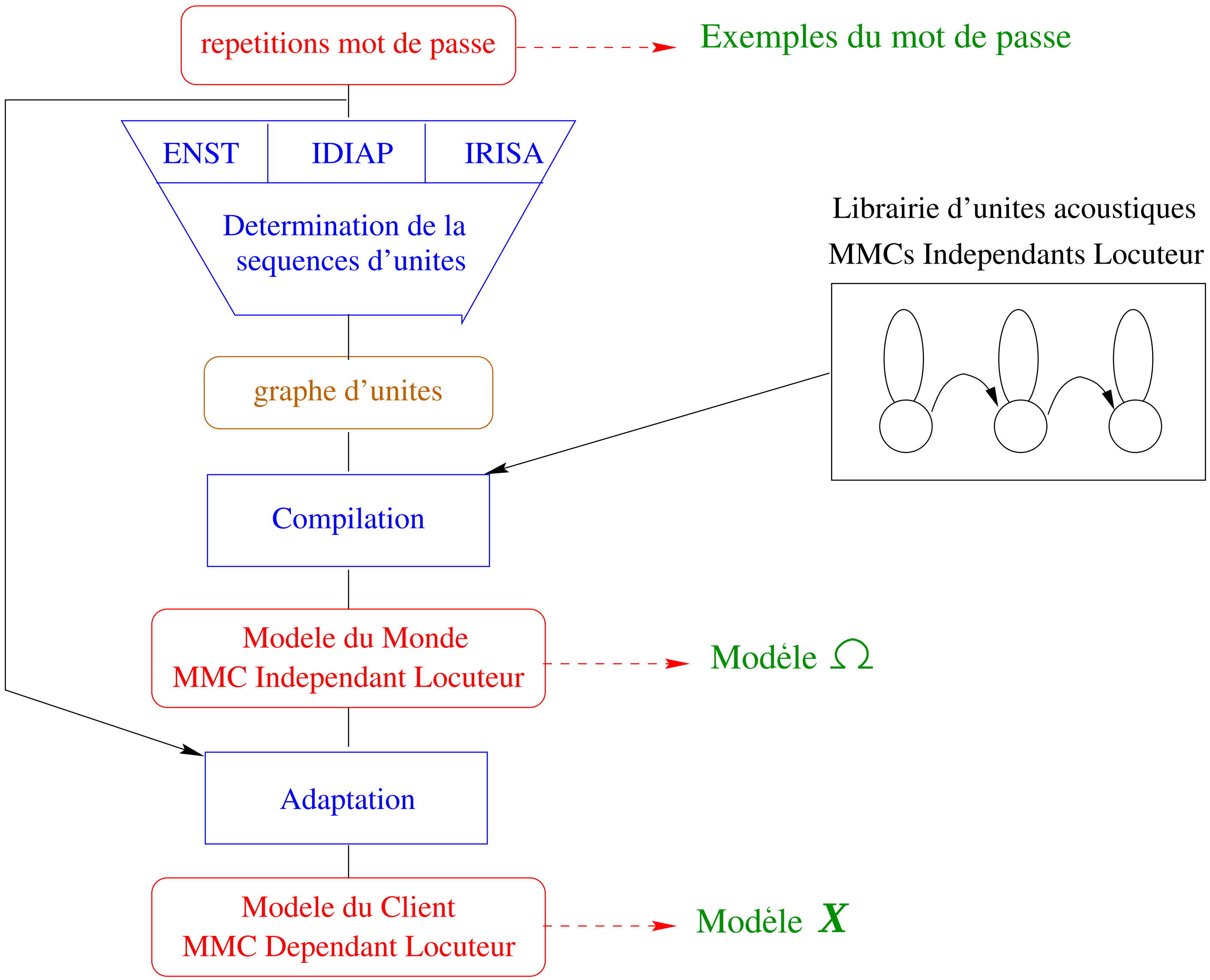
Buts :

- offrir un service plus ergonomique
- renforcer le système contre les impostures intentionnelles ($\neq$ vocabulaire figé)

Difficulté supplémentaire :

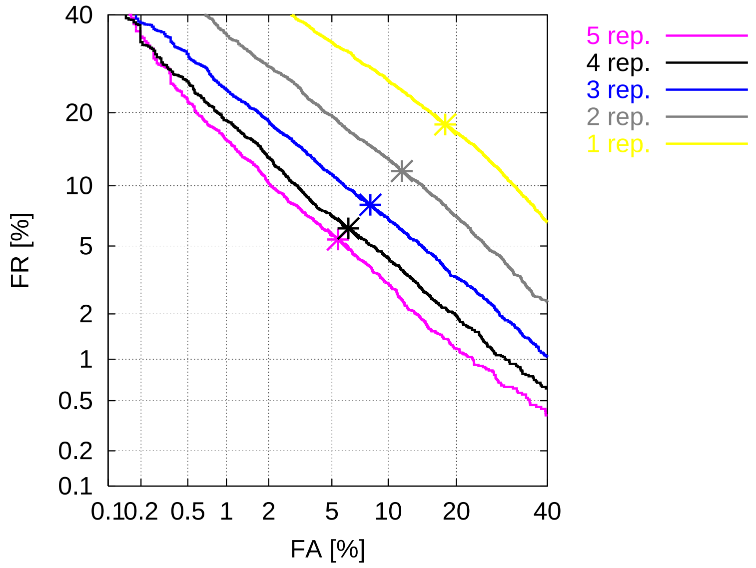
- Construire un modèle Ω du mot de passe du client
  $\rightarrow$ mot de passe tel qu'il serait prononcé par l'ensemble de la population
  $\Rightarrow$ construire un modèle Ω Indépendant du Locuteur à partir des répétitions du mot de passe **d'un seul locuteur client donné**

Synopsis de la création du mot de passe personnalisé

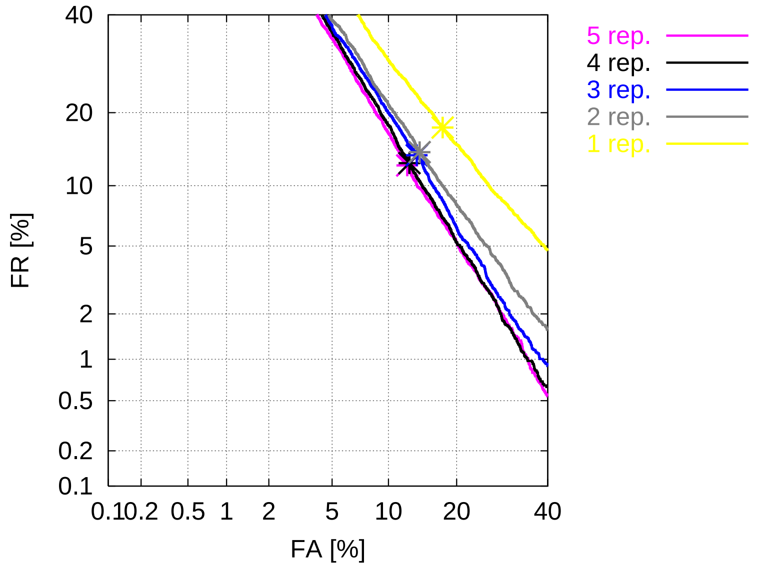


Résultats en fonction du nombre de répétitions

Référence (mot de passe connu)



Mot de passe personnalisé

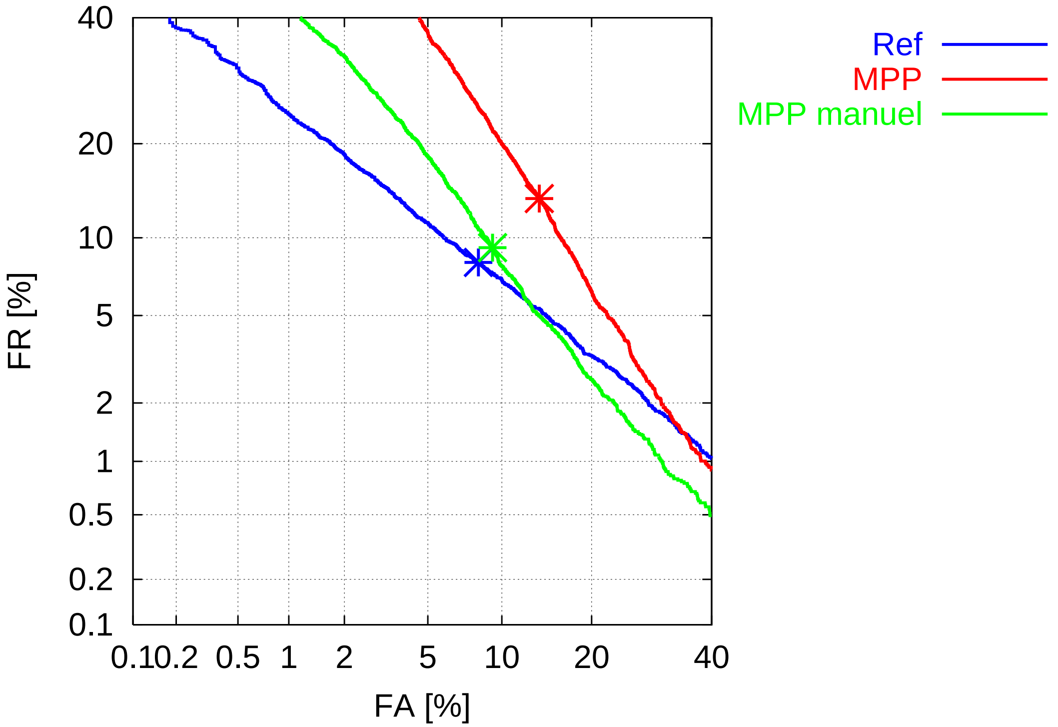


Répétitions	1	2	3	4	5
% EER Ref	18.1	11.6	8.1	6.2	5.4
% EER MPP	17.6	14	13.6	12.6	12.3

$\Rightarrow \approx 3$ répétitions du mot de passe

Résultats avec 3 répétitions du mot de passe

MPP_{Manuel} = MPP avec transcription manuelle du mot de passe



Méthode	Référence	MPP	MPP_{Manual}
EER %	8.1	13.6	9.2

Bilan :

- pour 1 répétition du mot de passe (cas le plus ergonomique) : MPP $\approx$ Référence
- faisabilité du mot de passe personnalisé : 3 répétitions sont suffisantes
- si l'on possède une bonne transcription $\rightarrow$ résultats équivalents

Perspectives

- recherche d'une meilleure transcription automatique du mot de passe

2^{ieme} thème "Que dit-on ?"

. $\hookrightarrow$ Robustesse

. $\hookrightarrow$ Fusion d'informations

Cadre : projet AMIBE

- projet soutenu par les PRC Informatiques 93-96
- Étude de la reconnaissance audiovisuelle

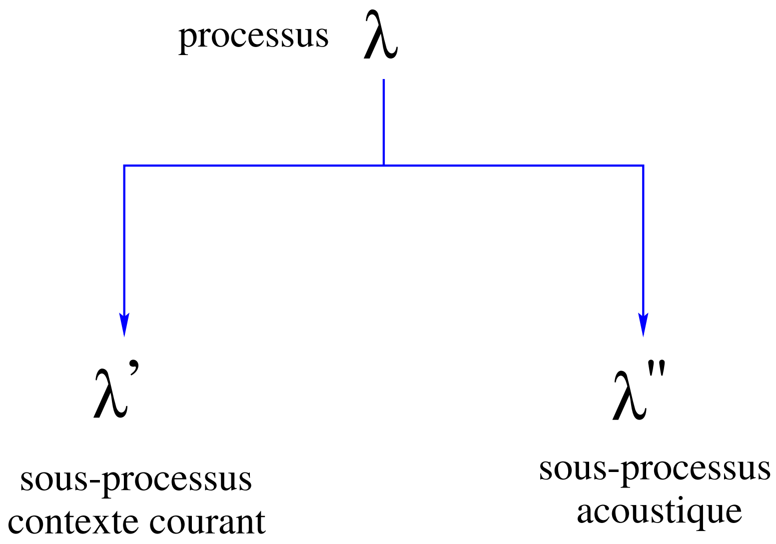
Participants : `IRIT Toulouse, LIA Avignon, LIUM Le Mans`

Problème : comment fusionner des informations auditives et visuelles ?

But : $\nearrow$ la reconnaissance surtout en milieu bruité

Méthode

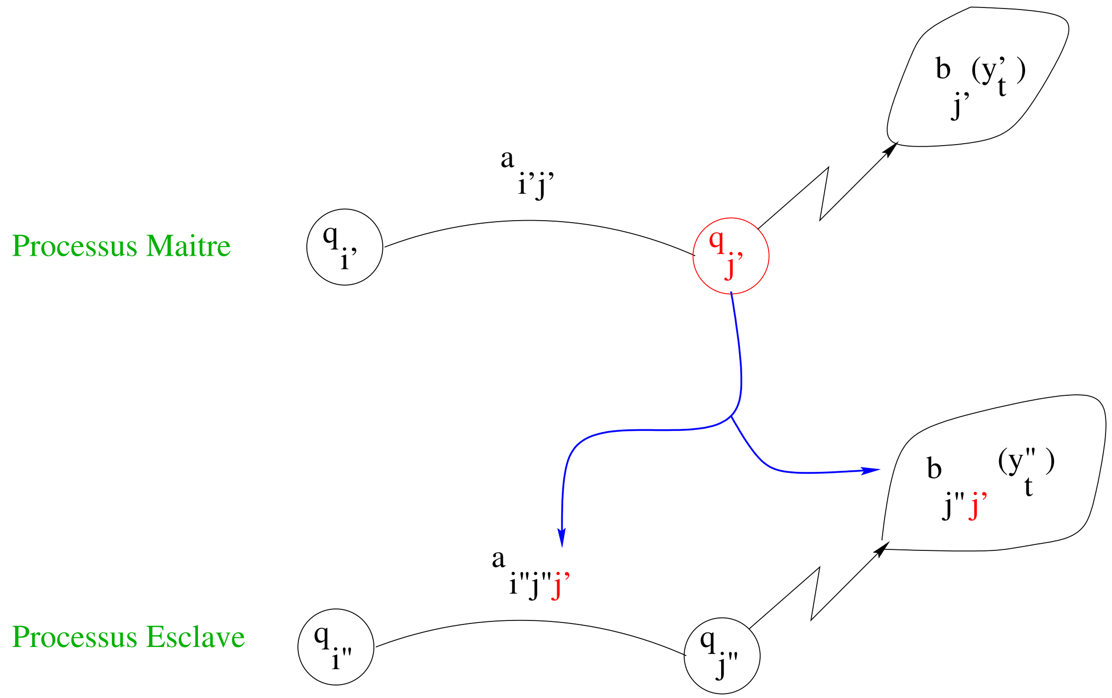
Idée de F. Brugnara et R. De Mori `[ICASSP,92]` :
Adaptation dynamique des lois de probabilité d'un Modèle de Markov Caché en fonction du contexte courant



`Contexte` : voisement, nasalisation, articulation, labialité ...

Hypothèses :

1. λ' MMC "classique" d'ordre 1
2. λ" est un MMC d'ordre 1 dont les paramètres dépendent du processus λ'



Intérêt : Modélisation non supervisée de la désynchronisation labial/acoustique

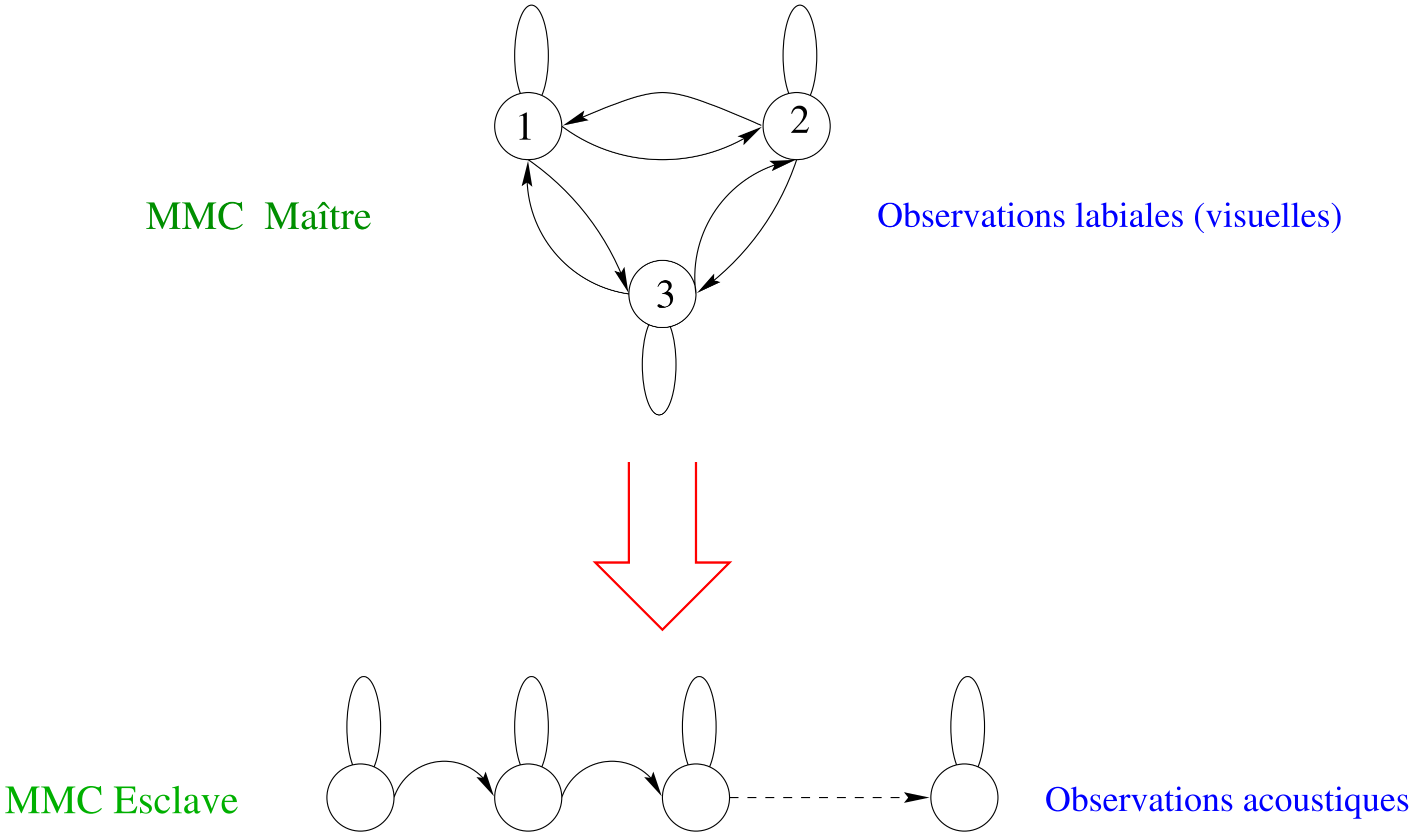
Application :

R&D d'un gestionnaire de MMC (thèse)

⇒ construction d'un MMC Maître-Esclave.

- Processus Maître
  - ergodique
  - 3 états (visèmes `[Benoît,92]`)
  - modélise les positions des lèvres
- Processus Esclave
  - Gauche-droite
  - phrases possibles (petites phrases de 4 mots)
  - modélise le signal acoustique

Visualisation du modèle



Le Modèle Maître (*données visuelles*) pilote
le Modèle Esclave (*données acoustiques*)

Expérience

Validation : comparaison du ME avec un MMC global concaténant les 2 types d'informations

- Apprentissage : 632 mots
- Test : 192 mots

Résultats, taux de reconnaissance :

	Calme	Bruité 10dB
labiale	40%	40%
acoustique	90,1%	55%
Acoustique+labiale (global)	96,5%	78,7%
Acoustique+labiale (ME)	96,5%	77,6 %

Conclusion

Résultats des 2 approches :

$\approx$ 96% calme

$\approx$ 78% bruité

Mais :

- Nombre de paramètres du ME $\nearrow$
- Étude de la désynchronisation entre labial et acoustique paraît plus abordable par cette approche.

Perspectives

Savoir faire du LIUM :

- en image
- en fusion de données

Collaboration pour poursuivre de cet axe de recherche

- Acquisition de nouveaux paramètres visuels ?
- Nouvelle méthode de fusion ?

2^{ieme} thème “Que dit-on ?”

. $\hookrightarrow$ Grand Vocabulaire

. $\hookrightarrow$ Dictée Vocale

But : Décoder un signal acoustique en séquences de mots (phrases)

Problème : On cherche une structure qui modélise toutes les phrases possibles

MAIS 1 seul modèle global impossible à réaliser

Donc, on “découpe” cette structure en niveaux de modélisation.

Généralement, on ne sait modéliser que 3 niveaux :

acoustique : modélisation des unités sonores

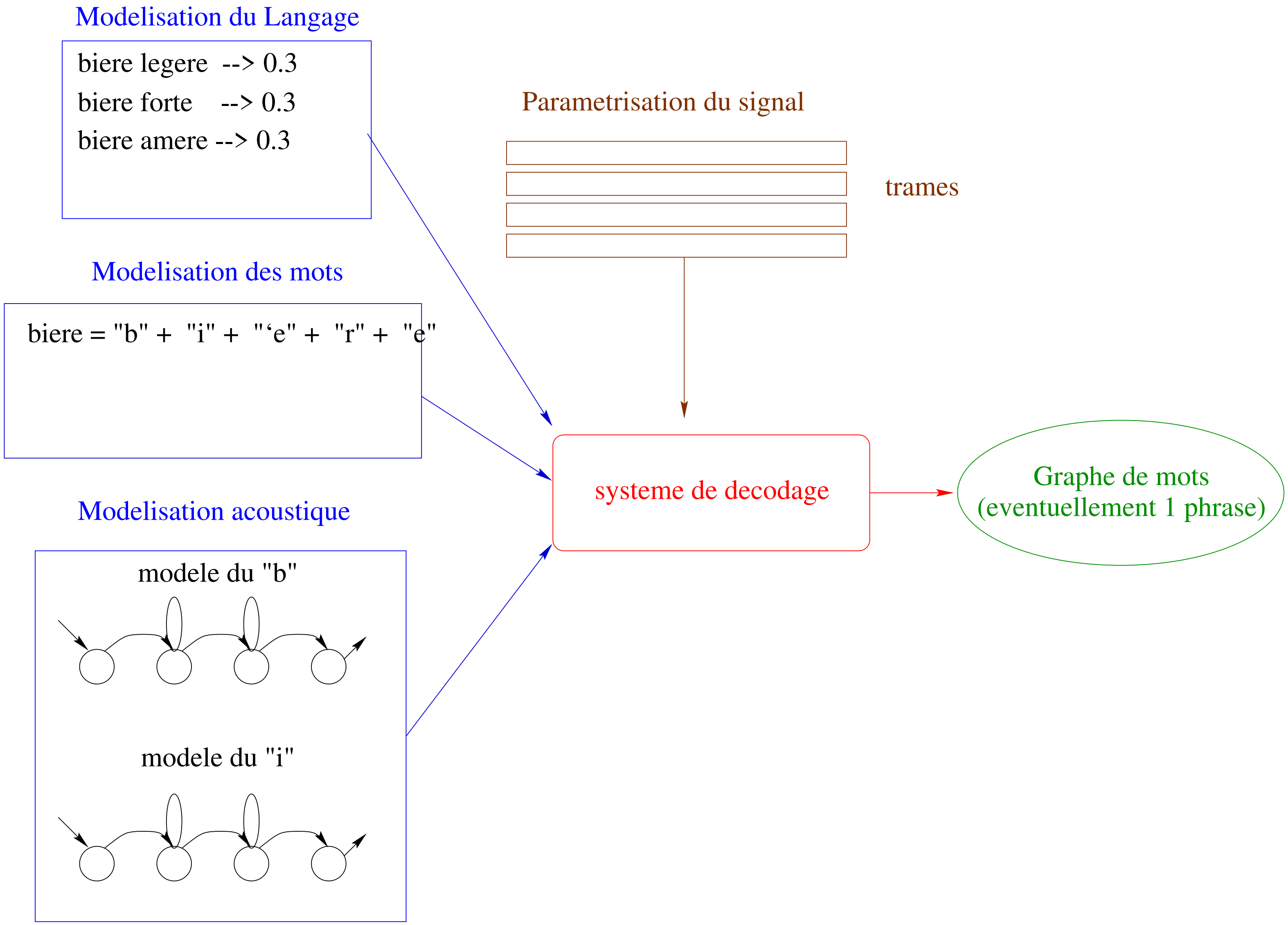
$\updownarrow$ articulation des sons en mots

lexical : modélisation des mots

$\updownarrow$ articulation des mots en phrases

syntaxique : modélisation des phrases

Modules intervenant dans le décodage



Cadre de travail pour le décodage

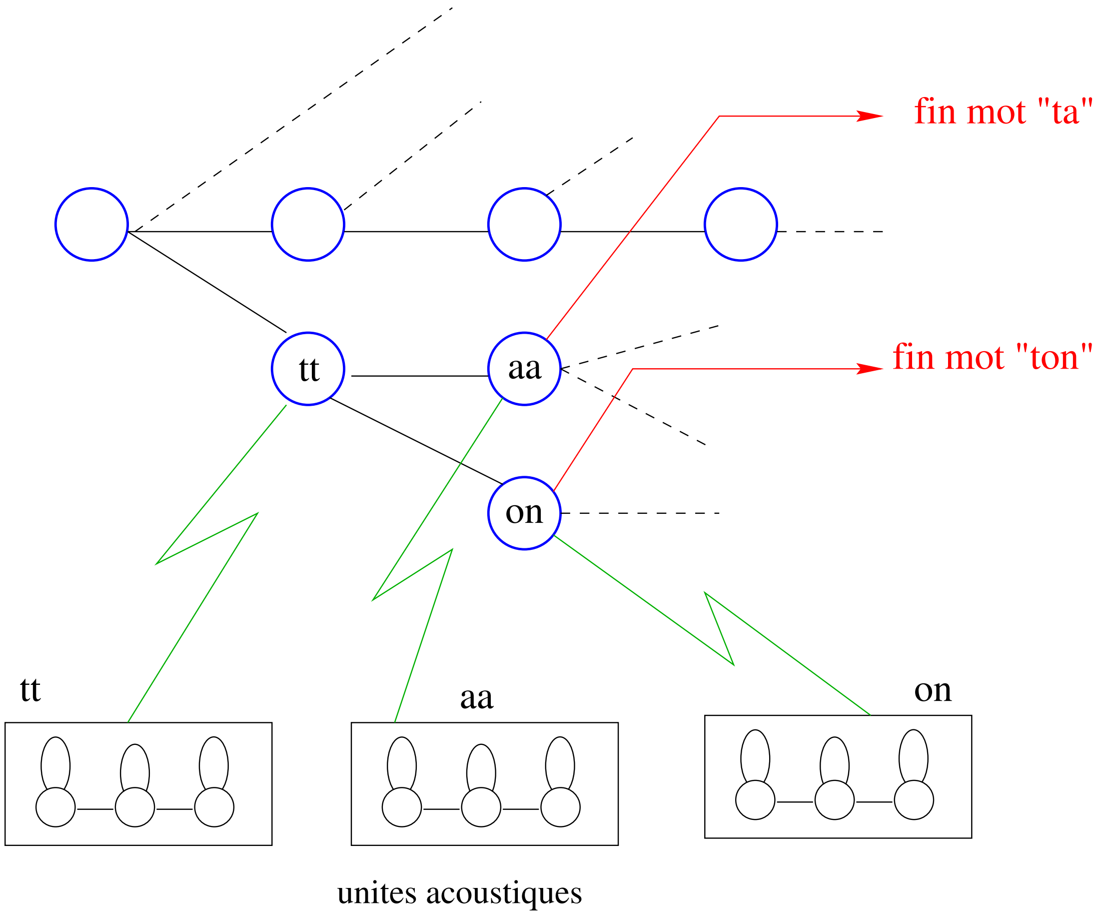
Dans les systèmes de reconnaissance performants le décodage se fait **simultanément** dans les 3 niveaux de modélisation ($\neq$ avant $\rightarrow$ successivement dans les niveaux $\Rightarrow$ propagation d'erreurs)

Le module de décodage

- pilote conjointement:
  - un *Graphe de Décodage* $\rightarrow$ informations aux niveaux lexical et acoustique
  - un *Modèle de Langage* $\rightarrow$ informations au niveau syntaxique
- organise l'espace de recherche des phrases possibles.

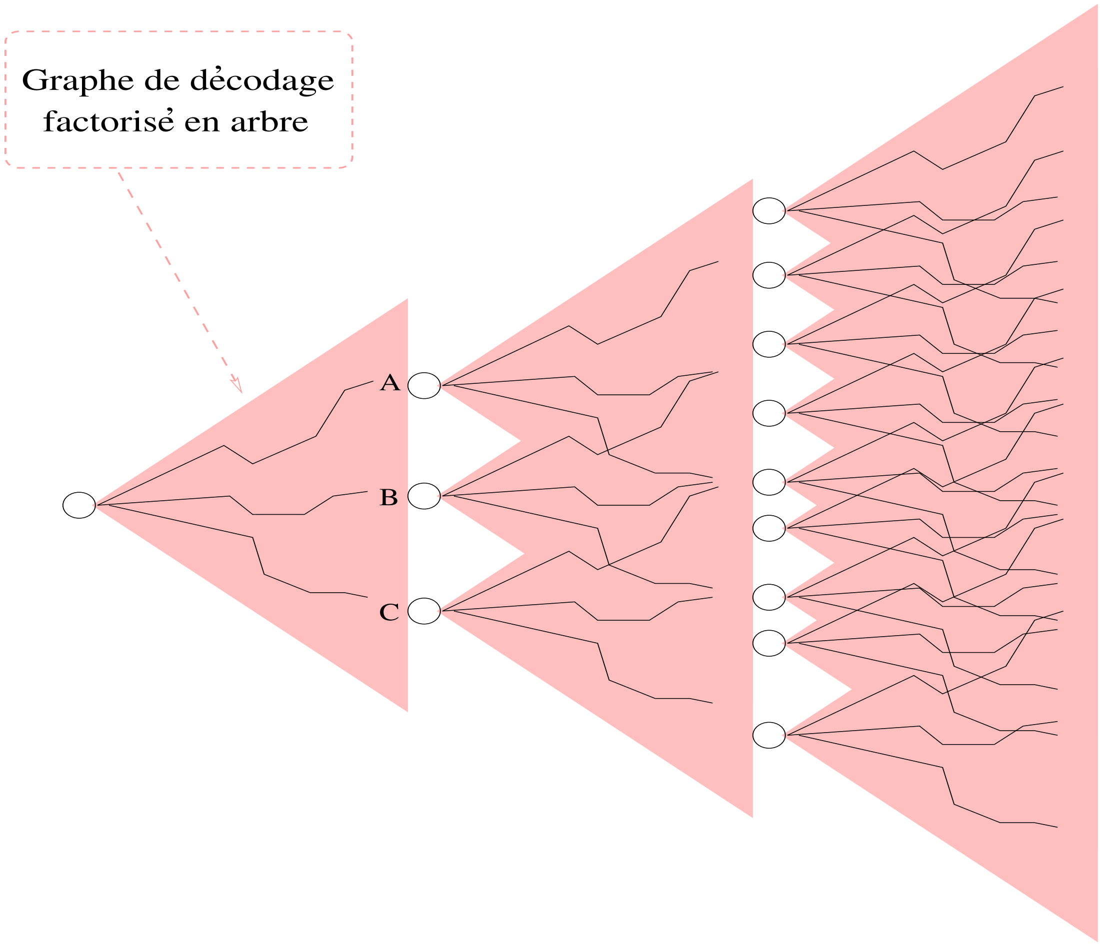
Graphe de décodage

⇒ Modèles acoustiques des mots du vocabulaire basé sur un lexique en arbre



Espace de recherche global

Exemple avec un vocabulaire $\{A, B, C\}$



Présentation du projet SIROCCO

- Projet soutenu par les actions coopératives de l'INRIA 1999-2000
- Début en Janvier 1998
- Durée : 3 ans
- Partenaires : ENST Paris, IRISA Rennes, IRIT Toulouse, LIA Avignon, LORIA Nancy

Spécifications du décodeur acoustico-phonétique

- taille du lexique > 100.000 mots ($\Rightarrow$ Énorme graphe)
- Modèle de Langage bigramme (proba d'apparition d'un mot sachant uniquement le mot précédant)

$$P(mot_1, mot_2) = proba$$

Rappel du pb : Prise en compte de la totalité de l'espace de recherche impossible

Méthode : Recherche en **faisceaux** par l'algorithme de **Viterbi** (*Beam-Search*)
Permet le contrôle de la largeur des faisceaux $\Rightarrow$ taille espace de recherche

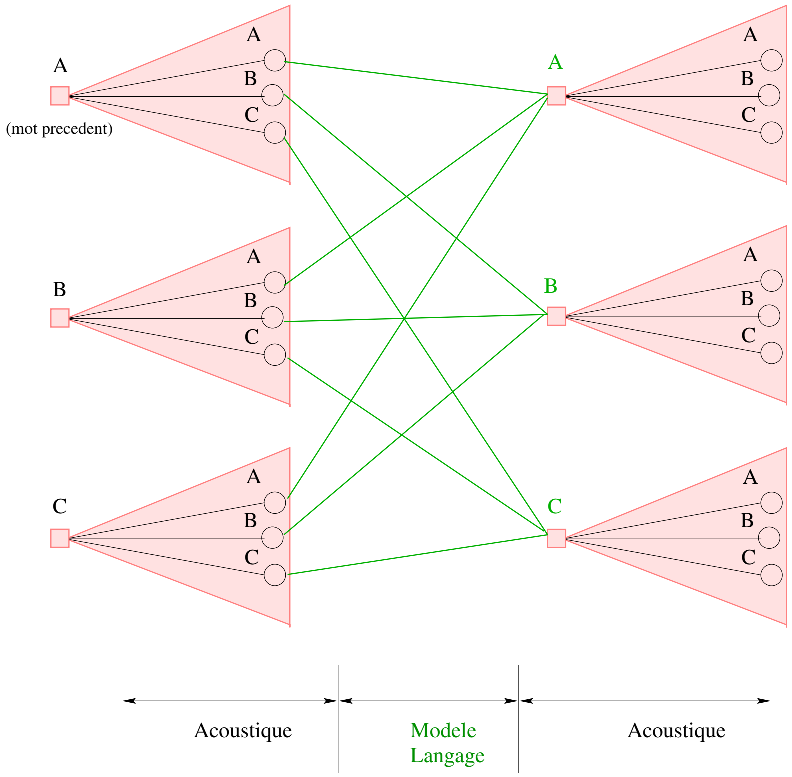
Principe

Remarques :

Organisation en arbre ⇒ mot courant décodé sur un noeud terminal

Pour appliquer le ML bigramme ⇒ il faut décoder 2 mots

Solution : concaténation de 2 arbres de décodage



Principe

⇒ Décorrélation du calcul des hypothèses intra-mots et inter-mots

- propagation dynamique intra-mots (intra-copie) :

$$Q_v(t+1, j) = b_j(y_{t+1}) \max_i Q_v(t, i)$$

- propagation inter-mot (inter-copie) :
  - optimisation :

$$H(w, t) = \max_v Q_v(t, Etat\_Final_w) P(v, w)^\beta$$

  - démarrage d'une nouvelle copie d'arbre avec w comme prédécesseur

Contrôle de la taille de l'espace de recherche

Au niveau intra-mots : Élagage à chaque observation y_t

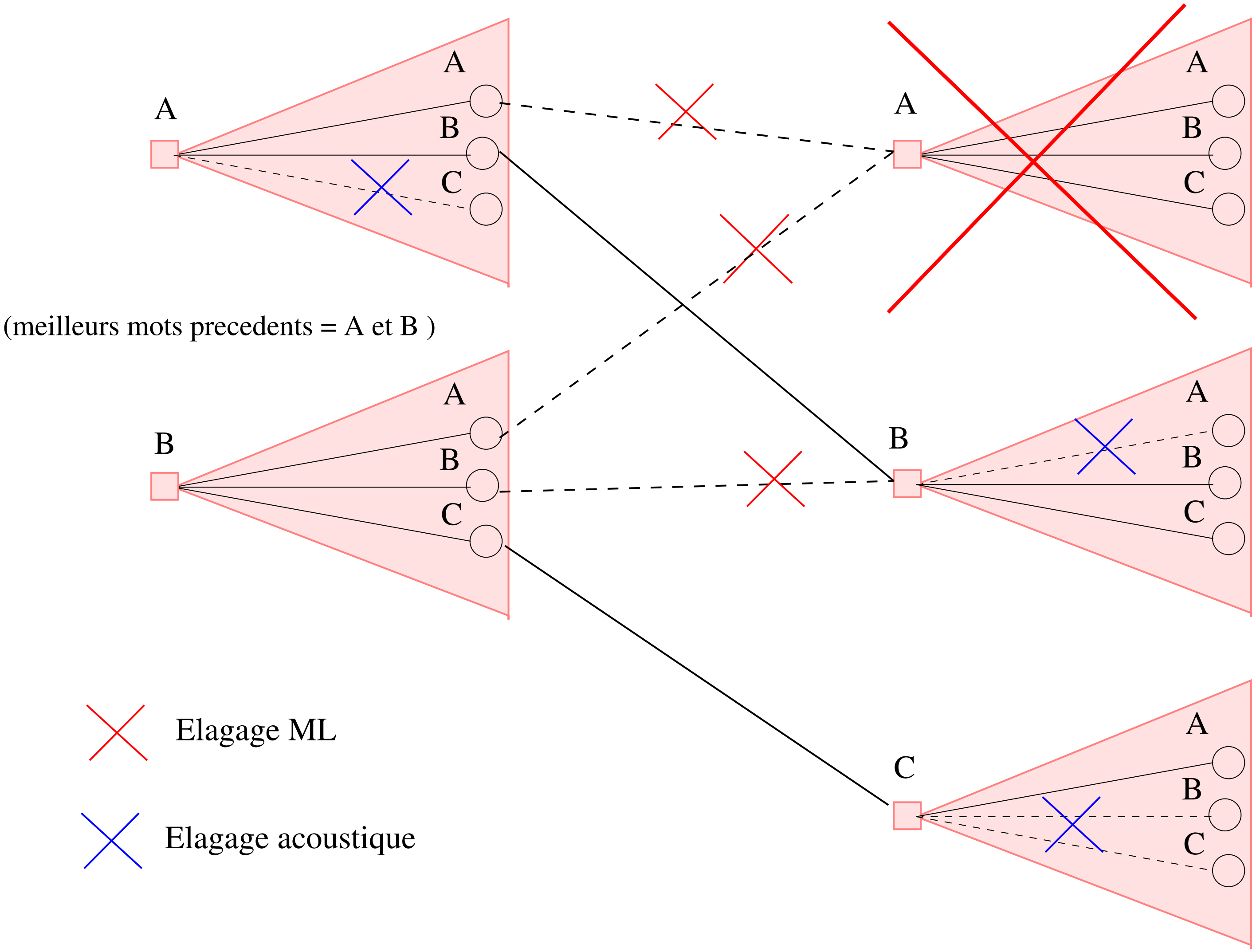
- *acoustique* : % du score acoustique max sur $Q_v(t, s)$
- *histogramme* : nombre max d'hypothèses
- *look-ahead* : recherche “en avant” dans le début de décodage des mots

Au niveau inter-mots : Élagage des mots décodés au temps t

- *ML* : % du score linguistique max sur $H(w, t)$
- *histogramme* : nombre max de mots
- *purge* des chemins perdus

$\Rightarrow$ détermine la largeur des *faisceaux* de recherche (*Beam-Search*)

Faisceaux dans l'espace de recherche acoustique pour un ML bigramme



Faisabilité de la méthode

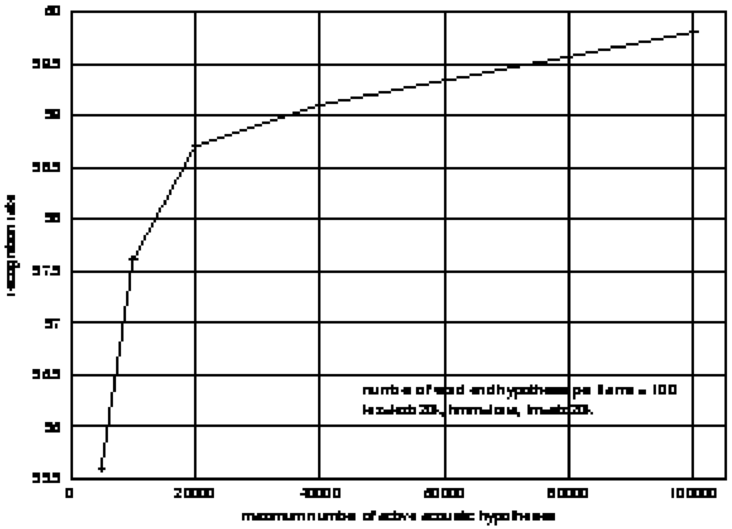
- Vocabulaire de 40 mots
- ML 40×40
- 37 modèles de phonèmes à 5 états acoustiques

Test sur un fichier de 1069 trames = environ 11 secondes

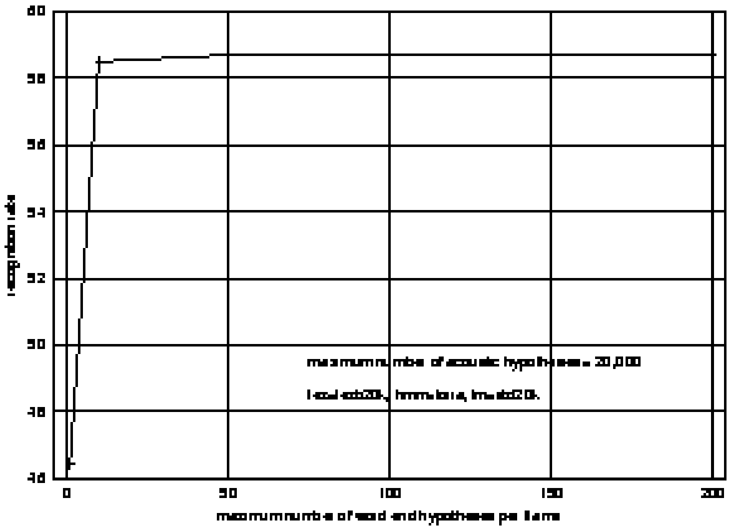
- recherche dans l'espace complet: $>$ 24 heures
- recherche en faisceaux :

Max hypothèses acoustiques à chaque trame	Temps d'exécution (environ)	Temps Réel
1000	30"	$\times$ 2,5
5000	2'	$\times$ 12
10000	5'	$\times$ 27
50000	30'	$\times$ 150

Résultats pour ≠ taille de faisceaux



Nb hyp. **acoustiques** / trames



Nb hyp. **mots** / trames

Perspectives

Collaboration : équipe Dialogue

Idée : Amélioration du Modèle de Langage

Conclusion

Le CPER TEMEDI sur les TICs en EIAD $\rightarrow$ cadre de recherche

- Fusion de données pour augmenter la robustesse et la richesse de la communication avec un étudiant à distance
- Dictée Vocale pour dialoguer sans clavier
- Adaptation au locuteur par techniques proches de Vérification du Locuteur
- toutes les idées sont les bienvenues ...

... *That all folks*