Eléments de processus stochasiques

Projet: Modèles de Markov cachés pour la reconnaissance vocale

IC Majeures en biomédical, électricité, informatique, physique

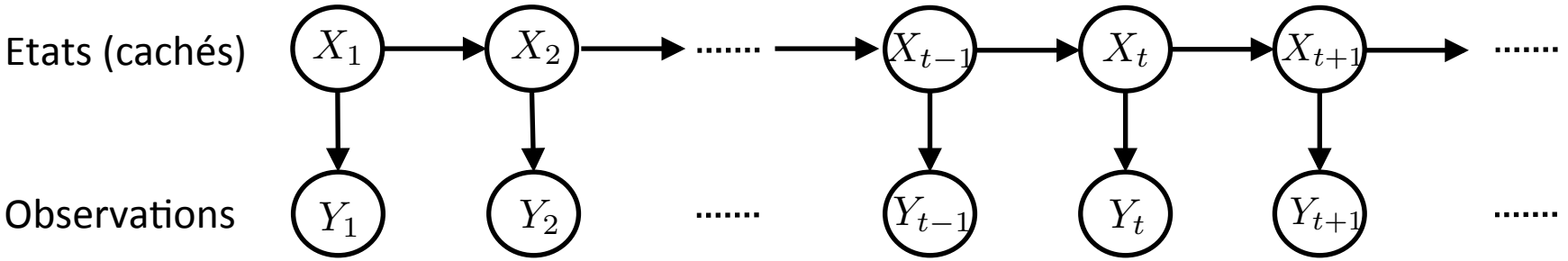
Profs: Louis Wehenkel, Pierre Geurts
Assistante: Laurine Duchesne

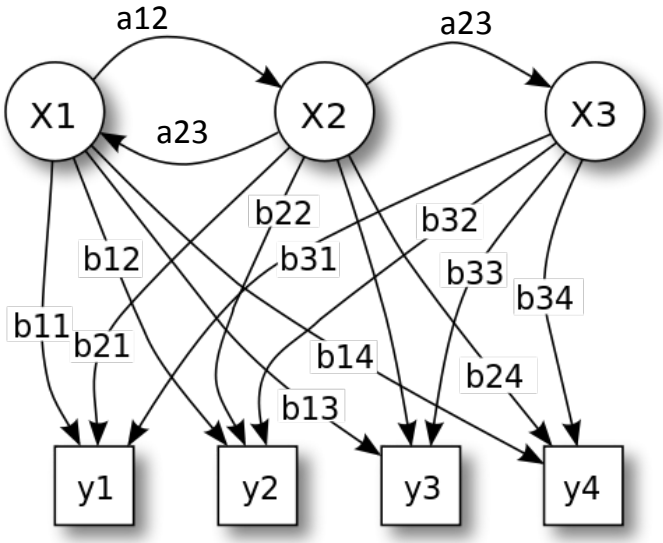
{l.wehenkel, p.geurts, l.duchesne}@ulg.ac.be

Site web: http://www.montefiore.ulg.ac.be/~lduchesne/stocha/

Modèles de Markov cachés

Une chaîne de Markov qui émet des observations à chaque pas de temps, dépendant uniquement de l'état à ce pas de temps.





$$P(X_1, \ldots, X_t, Y_1, \ldots, Y_t) = P(X_1) \prod_{i=2}^{t} P(X_i|X_{i-1}) \prod_{i=1}^{t} P(Y_i|X_i)$$

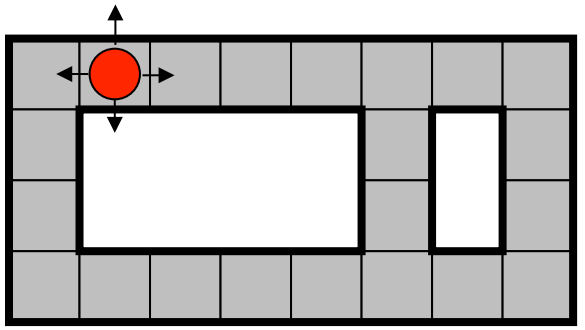
$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{pmatrix}$$

Exemple d'applications

Localisation d'un robot

X_t = position exacte du robot

Y_t = information partielle sur la position (p.ex., obtenue de capteurs)



Annotation de génome

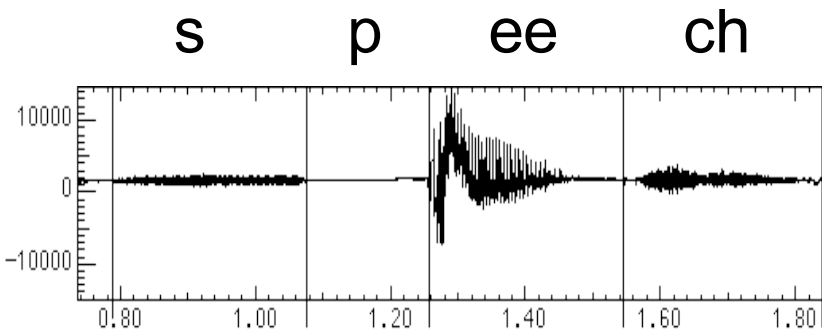
X_t = label associé à la position t

Y_t = nucléotide à la position t

Reconnaissance vocale

X_t = phonème (=son)

Y_t = signal acoustique



Utilisations

Prédictions des états cachés à partir des observations

$$\arg\max_{x_t} P(X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t)$$

Quelle est la position actuelle la plus probable du robot étant donné les observations déjà collectées ?

$$\arg\max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t)$$

Quelle est l'annotation la plus probable du génome étant donnés la séquence de nucléotides ?

Vraisemblance des observations étant donné un modèle

$$P(Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

Est-ce que le mot prononcé est "speech" ?

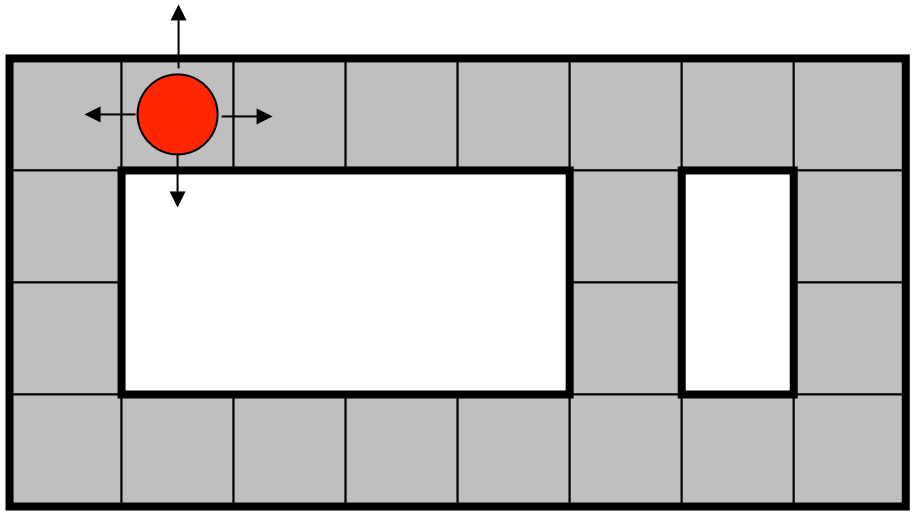
Apprentissage des paramètres à partir de données

$$\arg\max_{\lambda} P(Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

A partir d'instances du mot "speech", déterminer les paramètres du modèle maximisant la vraisemblance de ces données.

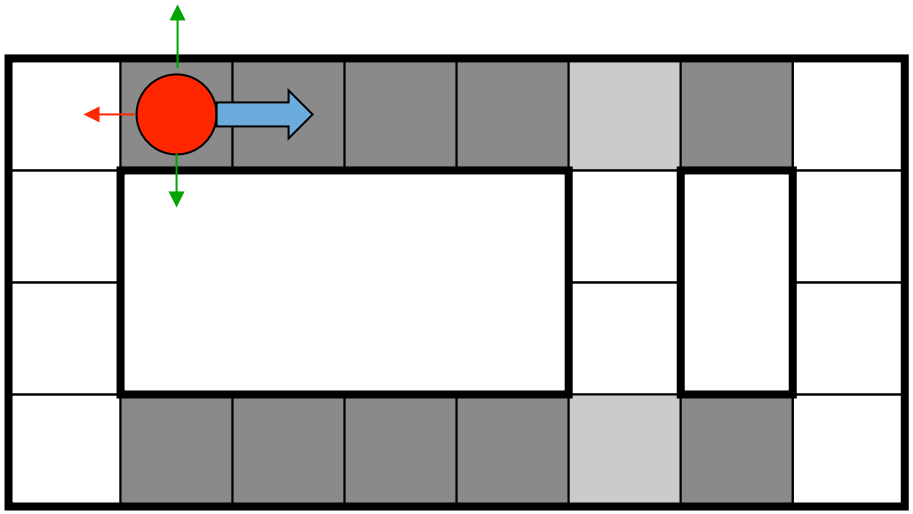
Example: Robot Localization

Example from Michael Pfeiffer



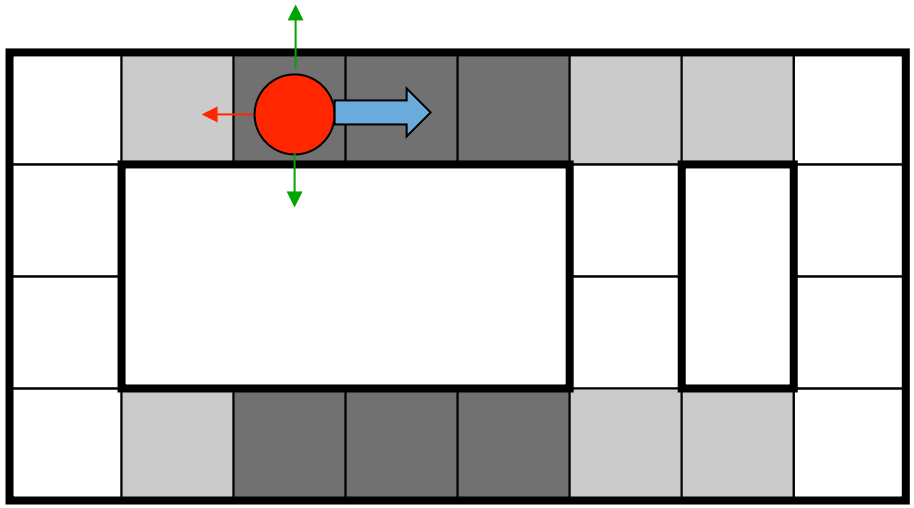
Prob 0 1

Example: Robot Localization



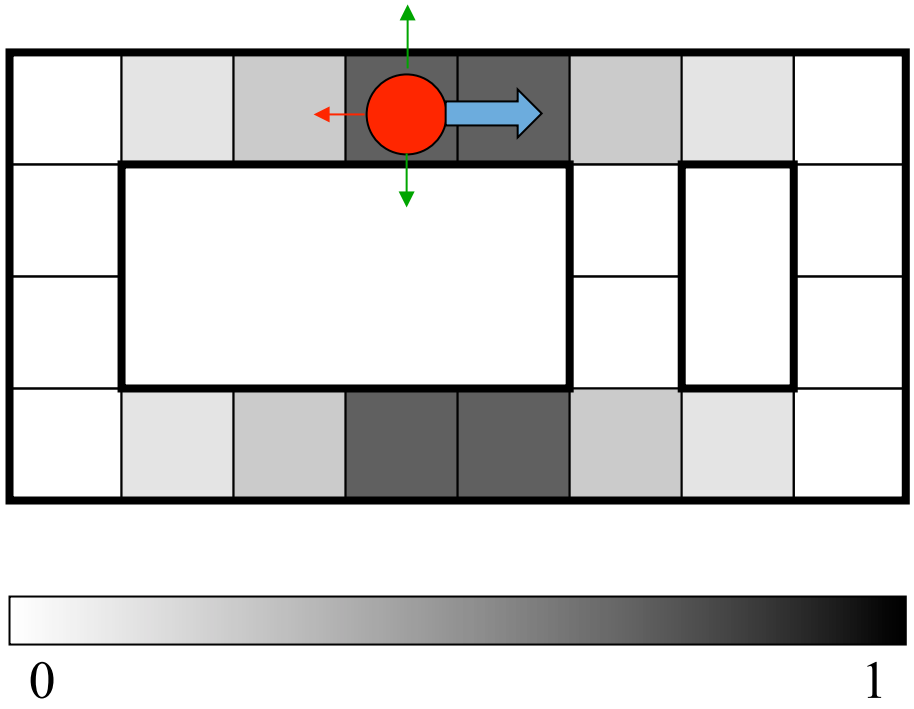
t=1

Example: Robot Localization



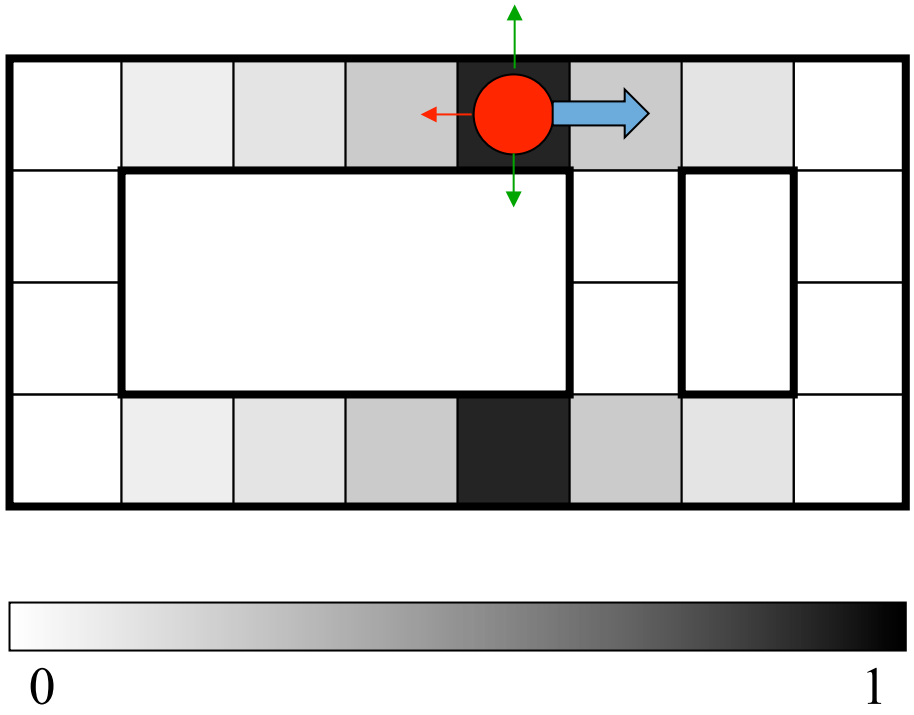
t=2

Example: Robot Localization



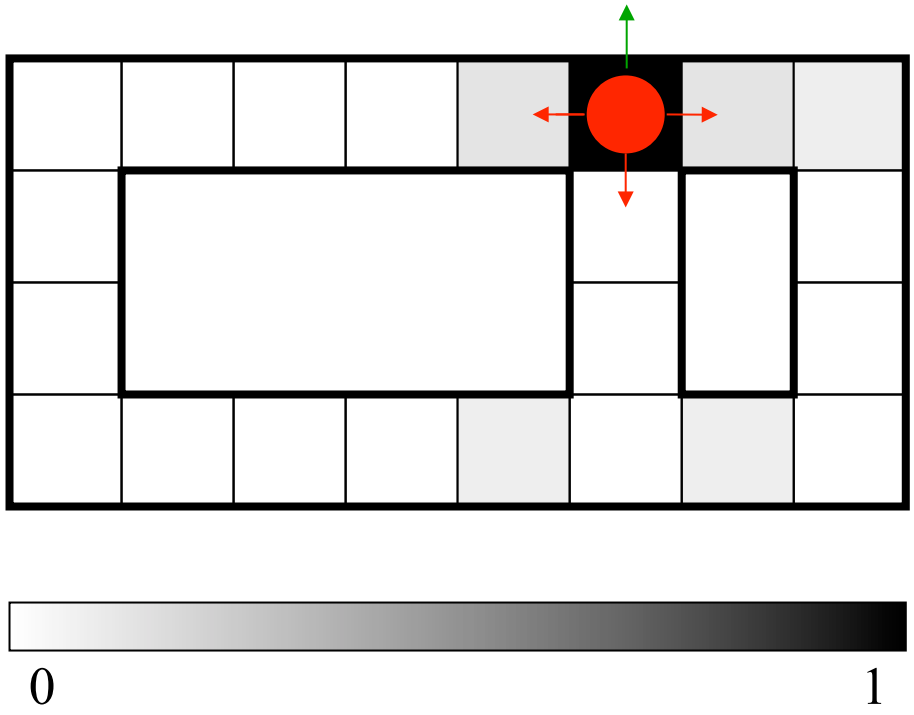
t=3

Example: Robot Localization



t=4

Example: Robot Localization



t=5

Utilisations

Prédictions des états cachés à partir des observations

$$\arg\max_{x_t} P(X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t)$$

Algorithme Forward

$$\arg\max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t)$$

Algorithme de Viterbi

Vraisemblance des observations étant donné un modèle

$$P(Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

Algorithme Forward

Apprentissage des paramètres à partir de données

$$\arg\max_{\lambda} P(Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

Algorithme Baum-Welch ou Algorithme de Viterbi

Algorithme de Viterbi

Permet de calculer efficacement la séquence cachée la plus probable étant donné les observations

$$\arg \max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t, \lambda)$$

$$= \arg \max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t, Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

Calculons d'abord la probabilité maximale:

$$\max_{x_1,\ldots,x_t} P(X_1 = x_1, \ldots, X_t = x_t | Y_1 = y_1, \ldots, Y_t = y_t, \lambda)$$

$$= \max_{1 \le i \le N} \max_{x_1,\ldots,x_{t-1}} P(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = i, Y_1 = y_1, \ldots, Y_t = y_t | \lambda)$$

$$\triangleq \max_{1 \le i \le N} \delta_t(i)$$

(avec $\{1, \ldots, N\}$ l'ensemble des états cachés)

Algorithme de Viterbi

$$\delta_t(i) = \max_{x_1,\ldots,x_{t-1}} P(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = i, Y_1 = y_1, \ldots, Y_t = y_t|\lambda)$$

Les $\delta_t(i)$ peuvent se calculer récursivement (en se basant sur la factorisation de la distribution conjointe)

$$\begin{aligned}\delta_1(i) &= P(X_1 = i|\lambda)P(Y_1 = y_1|X_1 = i, \lambda) \\ &= \pi_i B_{i,y_1}\end{aligned}$$

$$\begin{aligned}\delta_k(i) &= \max_{1\leq j\leq N} \delta_{k-1}(j)P(X_t = i|X_{t-1} = j, \lambda)P(Y_t = y_t|X_t = i, \lambda) \\ &= \max_{1\leq j\leq N} \delta_{k-1}(j)A_{j,i}B_{i,y_t} \qquad 1 < k \leq t\end{aligned}$$

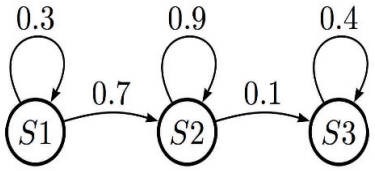
Pour récupérer la solution:

$$\psi_k(i) = \arg\max_{1\leq j\leq N} \delta_{k-1}(j)A_{j,i}B_{i,y_t} \qquad 1 < k \leq t$$

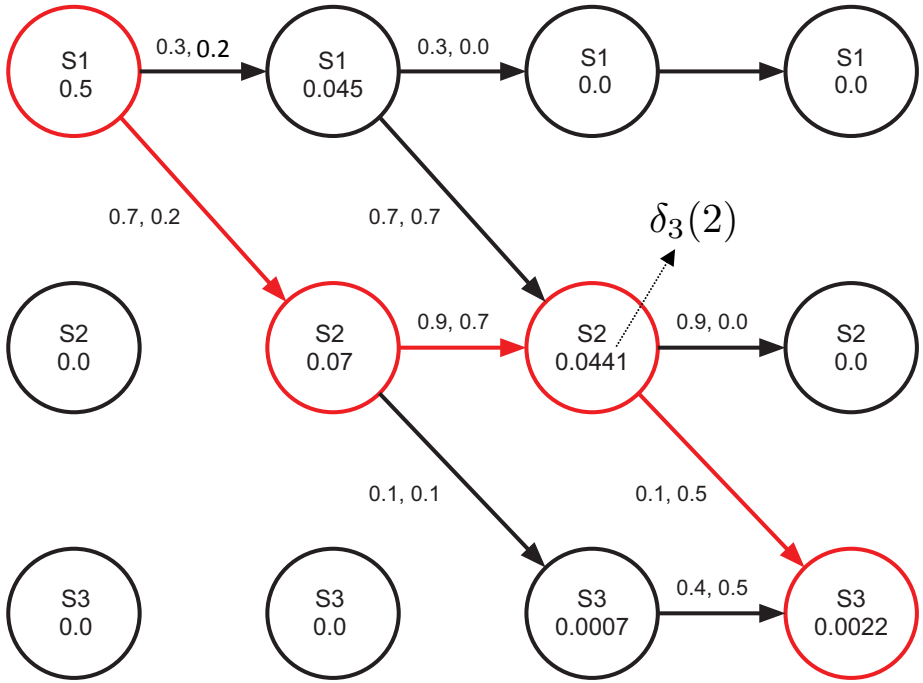
$$x_t^* = \arg\max_{1\leq i\leq N} \delta_t(i) \qquad x_{k-1}^* = \arg\max_{1\leq i\leq N} \psi_k(i) \qquad k = t, t-1, \ldots, 2$$

Algorithme de Viterbi: exemple

$P(X_1 = S1) = 1$



C1	0.5	0	0
C2	0.3	0	0
C3	0.2	0.2	0
C4	0	0.7	0.1
C5	0	0.1	0
C6	0	0	0.5
C7	0	0	0.4



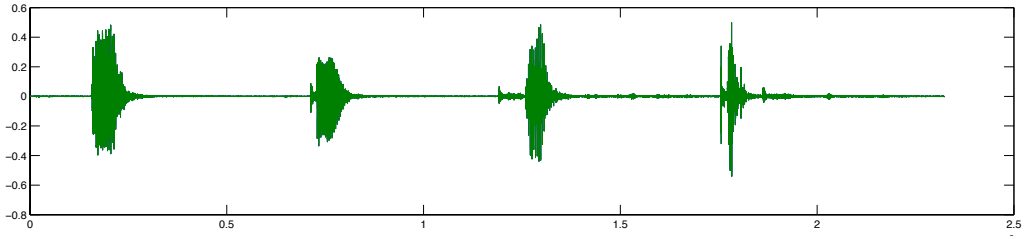
Séquence observée: C1 C3 C4 C6

(Murphy, 2012)

Le projet

Objectif général:

Mettre au point un système de reconnaissance vocale de chiffres, prononcés d'abord isolément puis en séquence.

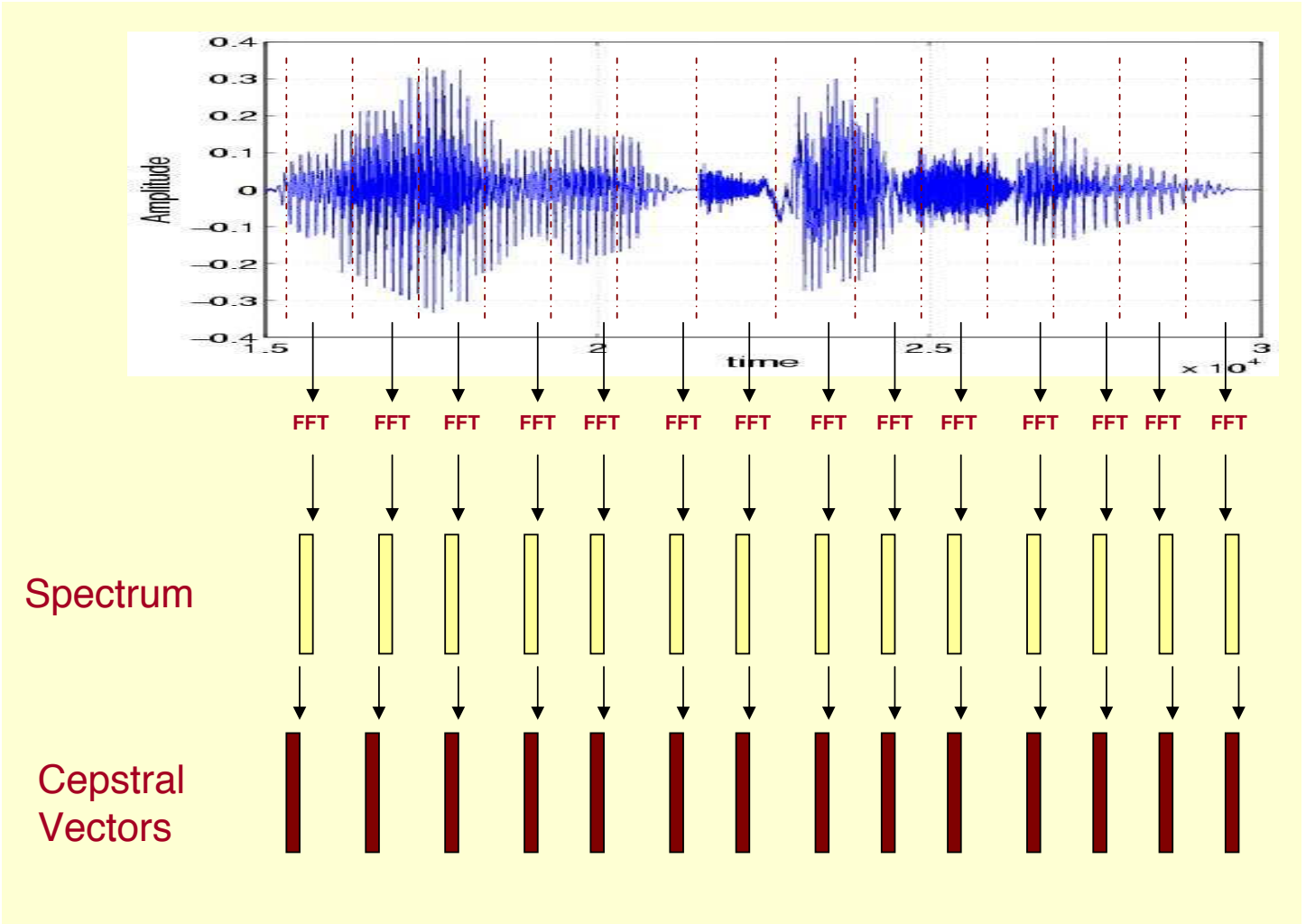


1,2,3,4

Idée générale de la solution à mettre en place:

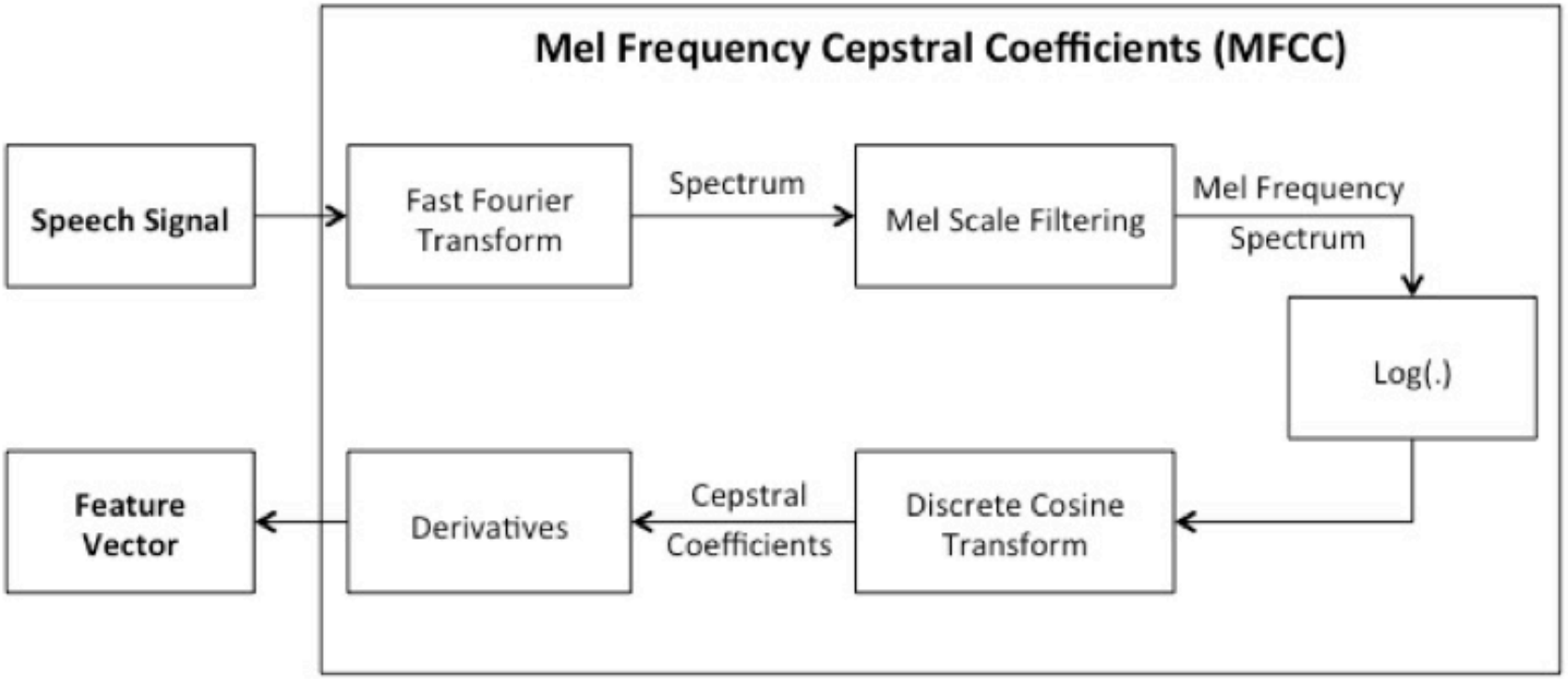
- Collecter des données (fichiers wav)
- Les mettre dans une forme acceptable pour leur modélisation à l'aide de modèles de Markov cachés
- Choisir la structure des modèles et apprendre leurs paramètres à partir des données
- Tester les performances du système. Si nécessaire itérer.

MFCC feature extraction



http://recognize-speech.com/feature-extraction/mfcc

MFCC feature extraction



http://recognize-speech.com/feature-extraction/mfcc

MFCC feature extraction