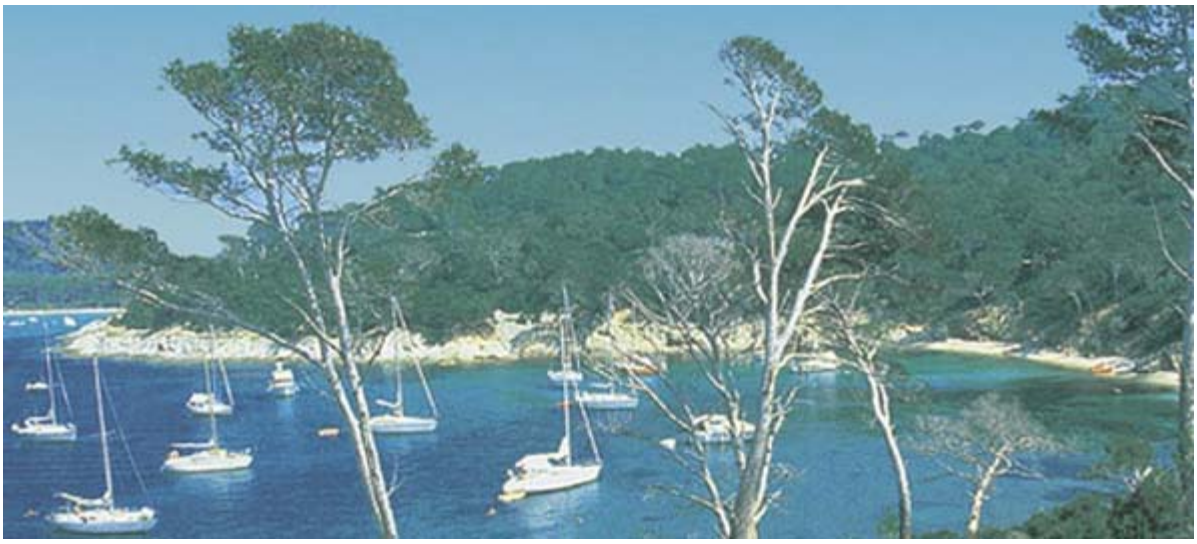


ERMITES

Ecole Recherche Multimodale d'Information – Techniques et Sciences

Giens

4 – 6 septembre 2006

Recueil des présentations

réunies par

Hervé Glotin et Jacques Le Maitre

Avant-propos

La recherche d'information, sur le web notamment, est de plus en plus complexe et hasardeuse compte tenu du volume sans cesse croissant des informations disponibles et de leur nature multimodale (textes, images, sons, vidéos…). C'est afin :

- d'analyser les dernières avancées, théoriques et pratiques, des Systèmes Robustes de Recherche d'Information Multimodale (SRIM), couplant textes, images, sons ou vidéos.
- de sensibiliser les jeunes scientifiques au vaste champ scientifique nécessaire à l'élaboration de SRIM et au problème de leur fiabilité.

que nous avons décidé d'organiser à la presqu'île de Giens dans le Var du 4 au 6 septembre, l'Ecole Recherche Multimodale d'Information: Techniques & Sciences (ERMITES) avec le concours de l'Association Francophone de la Communication Parlée (AFCP), du Laboratoire des Sciences de l'Information et des Systèmes (LSIS) et de l'Université du Sud Toulon-Var.

Ce recueil réunit les présentations faites à ERMITES qui portent sur les thèmes suivants : les théories de l'information, du signal, des processus aléatoires et de l'apprentissage automatique ; l'analyse de scène (audio et vidéo) ; l'intelligence artificielle ; le traitement automatique du langage et de la parole ; les sciences cognitives et la neurophysiologie de la perception ; la recherche d'information textuelle. Un des objectifs d'ERMITES est que ses participants ne voient plus d'antagonismes entre ces diverses disciplines qui se lieront de plus en plus pour générer des SRIM efficaces.

Nous tenons à remercier tous ceux qui ont contribué à la tenue d'ERMITES : les intervenants, les participants ainsi que le LSIS, l'AFCP, et l'USTV pour leur soutien matériel et financier.

Les organisateurs : Hervé Glotin et Jacques Le Maitre

Sommaire

Programme

Lundi 4 septembre

14h : Accueil, salle 'Tour Fondue'
Ouverture & présentation des participants

14h45 : Café & boissons & biscuits

15h15-16h15 : S. Bengio (IDIAP), *Apprentissage pour le Traitement de Séquences Multimodales*

17h-18h30 : P. Gallinari (LIP6), *Méthodes statistiques pour l'apprentissage de structures*

Rafraîchissements
Pause & Diner

20h30-21h30 : Discussions sur les thèmes 'Apprentissage et RI', Torch, 'Une approche discriminante pour la recherche d'images à partir de requêtes',…

Mardi 5 septembre

9h-10h : S. Bengio (IDIAP), *Recherche d'Information dans les Séquences Multimodales*

10h-11h : P. Giraudet (USTV), *Neurophysiologie de la Vision : des Traits au Sens*

Café, rafraîchissements

11h30-12h30 : P. Mulhem (CLIPS), *Indexation & Recherche Sémantique d'Images*

Repas

13h30-14h30 : H. Glotin (LSIS), *Recherche Robuste d'Information dans des Scènes Audiovisuelles*

14h30-15h30 : J. Farinas (IRIT), *Identification & Classification Automatique de Langues*

15h30-16h30 : J.-F. Bonastre (LIA), *Reconnaissance du Locuteur & Indexation de Documents Audio*

Café, rafraîchissements

17h-18h : G. Gravier (IRISA), *Reconnaissance Automatique de la Parole*

Pause et Diner

20h30-21h30 : Discussions sur le thème Transcription Enrichie d'Emissions Radiophoniques (détection parole/musique, suivi et regroupement de locuteur, transcription),…

Mercredi 6 septembre

9h-10h : P. Joly (IRIT), *Segmentation & Thématisation de Séquences Vidéo, Similarité de contenu visuel*

10h-11h : G. Gravier (& P. Gros) (IRISA), *Structuration Multimodale de Vidéos de Sports*

Café, rafraîchissements

11h30-12h30 : J. Le Maitre (LSIS), *Recherche d'Information Textuelle dans des Documents XML*

Repas

14h30-15h30 : Table ronde

16h : Clôture

Café, rafraîchissements

Machine Learning Approches for Multi Channel Sequence Processing

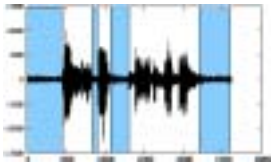
Samy Bengio

IDIAP Research Institute
Martigny, Switzerland
bengio@idiap.ch
http://www.idiap.ch/~bengio

September 4, 2006

Sequence Processing

- One microphone:
  speech and speaker recognition, music processing, ...
- One video camera:
  person tracking, object tracking, gesture recognition, ...
- One sensor:
  DNA analysis, bio-signal control, ...



Multi Channel Sequence Processing

- Many applications involve multiple streams of data.
- The same event is represented by more than one sequence.
- Multiple cameras, microphones, EEG signals, bio-signals, etc.

Applications Involving Multiple Channels

- Audio-visual speech recognition.
- Audio-visual person authentication.
- Multimedia document retrieval.
- Multi-rate and multi-stream speech recognition.
- Multimodal tracking of objects/humans (with several cameras/microphones).
- Multimodal human-computer interfaces.
- Multimedia analysis (news, sports, home videos, scenes, meetings...).
- Wearable computers.

How to Analyze Such Data?

- We should use all the available information:
  - Prior knowledge of what we expect (and do not expect!) to see
  - Data, data, data...
- A principled approach: statistical machine learning
- Roadmap of this presentation:
  - The statistical learning theory
  - Models for single sequence processing: HMMs
  - Extension to the case of multi channel processing
  - Applications to the meeting scenario
  - Torch: a statistical machine learning library

The Data

Available training data

- Let $Z_1, Z_2, \cdots, Z_n$ be an n-tuple random sample of an unknown distribution of density $p(z)$.
- All Z_i are independently and identically distributed (iid).
- Let D_n be a particular instance $= \{z_1, z_2, \cdots, z_n\}$.

Various forms of the data

- Classification: $Z = (X, Y) \in \mathbb{R}^d \times \{-1, 1\}$
  objective: given a new x, estimate $P(Y|X = x)$
- Regression: $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$
  objective: given a new x, estimate $E[Y|X = x]$
- Density estimation: $Z \in \mathbb{R}^d$
  objective: given a new z, estimate $p(z)$

The Function Space

Learning: search for a good function in a function space $\mathcal{F}$

Examples of functions $f(\cdot; \theta) \in \mathcal{F}$:

- Regression:
$$\hat{y} = f(x; a, b) = a \cdot x + b$$
- Classification:
$$\hat{y} = f(x; a, b) = \text{sign}(a \cdot x + b)$$
- Density estimation
$$\hat{p}(z) = f(z; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{|z|}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z - \mu)^T \Sigma^{-1} (z - \mu)\right)$$

The Loss Function

Learning: search for a good function in a function space $\mathcal{F}$

Examples of loss functions $L : \mathcal{Z} \times \mathcal{F}$

- Regression:
$$L(z, f) = L((x, y), f) = (f(x) - y)^2$$
- Classification:
$$L(z, f) = L((x, y), f) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise} \end{cases}$$
- Density estimation:
$$L(z, f) = -\log p(z)$$

The Risk and the Empirical Risk

Learning: search for a good function in a function space $\mathcal{F}$

- Minimize the Expected Risk on $\mathcal{F}$, defined for a given f as
$$R(f) = E_Z[L(z, f)] = \int_Z L(z, f)p(z)dz$$
- Induction Principle:
  - select $f^* = \arg\min_{f \in \mathcal{F}} R(f)$
  - problems: $p(z)$ is unknown, and we don't have access to all $L(z, f)$!!!
- Empirical Risk:
$$\hat{R}(f, D_n) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f)$$

The Empirical Risk

- The empirical risk:

$$\hat{R}(f, D_n) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f)$$

- The (expected) risk:

$$R(f) = E_Z[L(z, f)] = \int_Z L(z, f) p(z) dz$$

- The empirical risk is an unbiased estimate of the risk

The principle of empirical risk minimization (ERM):

$$f^*(D_n) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

The Risk and the Training Error

- Training error:

$$\hat{R}(f^*(D_n), D_n) = \min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

- Is the training error a biased estimate of the risk? YES.

$$E[R(f^*(D_n)) - \hat{R}(f^*(D_n), D_n)] \geq 0$$

- The solution $f^*(D_n)$ found by minimizing the training error is better on D_n than on any other set D'_n drawn from $p(z)$.

Bounding the Risk

Can we bound the difference between the training error and the generalization error?

$$|R(f^*(D_n)) - \hat{R}(f^*(D_n), D_n)| \leq ?$$

- Answer: under certain conditions on $\mathcal{F}$, yes.
- These conditions depend on the notion of capacity h of $\mathcal{F}$.

The Capacity

- The capacity $h(\mathcal{F})$ is a measure of its size, or complexity.
- Classification:

  The capacity $h(\mathcal{F})$ is the largest n such that there exist a set of examples D_n such that one can always find an $f \in \mathcal{F}$ which gives the correct answer for all examples in D_n, for any possible labeling.

- Example: for the set of linear functions ($y = w \cdot x + b$) in d dimensions, the capacity is $d + 1$.
- Regression and density estimation: capacity exists also, but more complex to derive (for instance, we can always reduce a regression problem to a classification problem).

Bounding the Risk

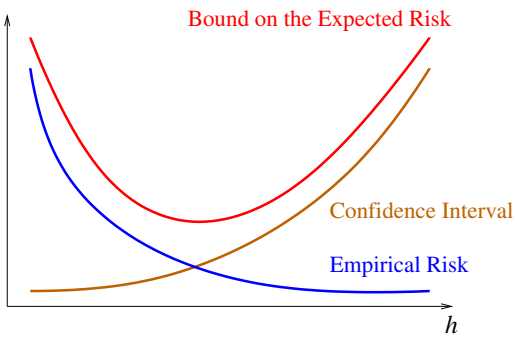
Bound on the expected risk:

- let $\tau = \sup L - \inf L$.
- $\forall \eta$ we have

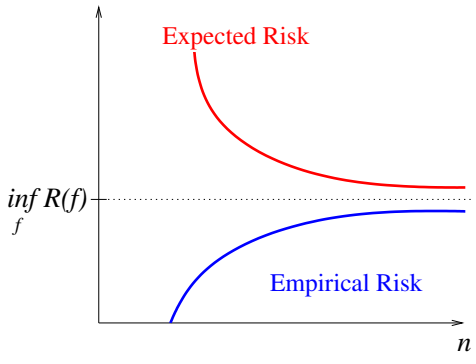
$$P\left(\sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f, D_n)| \leq 2\tau \sqrt{\frac{h\left(\ln\frac{2n}{h}+1\right) - \ln\frac{\eta}{9}}{n}}\right) \geq 1-\eta$$

- with h the capacity of $\mathcal{F}$ and n the number of training examples in D_n

Structural Risk Minimization - Fixed n



Consistency - Fixed h



Methodology

- First: identify the goal! It could be
  1. to give the best model you can obtain given a training set?
  2. to give the expected performance of a model obtained by empirical risk minimization given a training set?
  3. to give the best model and its expected performance that you can obtain given a training set?
- If the goal is (1): use need to do model selection
- If the goal is (2), you need to estimate the risk
- If the goal is (3): use need to do both!
- There are various methods that can be used for either risk estimation or model selection:
  - simple validation
  - cross validation (k-fold, leave-one-out)

Model Selection - Validation

- Select a family of functions with hyper-parameter θ
- Divide your training set D_n into two parts
  - $D^{tr} = \{z_1, z_2, \cdots, z_{tr}\}$
  - $D^{va} = \{z_{tr+1}, z_{tr+2}, \cdots, z_{tr+va}\}$
  - $tr + va = n$
- For each value θ_m of the hyper-parameter θ
  - select $f^\star_{\theta_m}(D^{tr}) = \arg\min_{f \in \mathcal{F}_{\theta_m}} \hat{R}(f, D^{tr})$
  - estimate $R(f^\star_{\theta_m})$ with $\hat{R}(f^\star_{\theta_m}, D^{va}) = \frac{1}{va} \sum_{z_i \in D^{va}} L(z_i, f^\star_{\theta_m}(D^{tr}))$
- select $\theta^\star_m = \arg\min_{\theta_m} R(f^\star_{\theta_m})$
- return $f^\star(D_n) = \arg\min_{f \in \mathcal{F}_{\theta^\star_m}} \hat{R}(f, D_n)$

Model Selection - Cross-validation

- Select a family of functions with hyper-parameter θ
- Divide your training set D_n into K distinct and equal parts $D^1, \cdots, D^k, \cdots, D^K$
- For each value θ_m of the hyper-parameter θ
  - For each part D^k (and its counterpart $\bar{D}^k$)
    - ⋆ select $f^\star_{\theta_m}(\bar{D}^k) = \arg\min_{f \in \mathcal{F}_{\theta_m}} \hat{R}(f, \bar{D}^k)$
    - ⋆ estimate $R(f^\star_{\theta_m}(\bar{D}^k))$ with $\hat{R}(f^\star_{\theta_m}(\bar{D}^k), D^k) = \frac{1}{|D^k|} \sum_{z_i \in D^k} L(z_i, f^\star_{\theta_m}(\bar{D}^k))$
  - estimate $R(f^\star_{\theta_m}(D_n))$ with $\frac{1}{K} \sum_k R(f^\star_{\theta_m}(\bar{D}^k))$
- select $\theta^\star_m = \arg\min_{\theta_m} R(f^\star_{\theta_m}(D))$
- return $f^\star(D_n) = \arg\min_{f \in \mathcal{F}_{\theta^\star_m}} \hat{R}(f, D_n)$

Estimation of the Risk - Validation

- Divide your training set D_n into two parts
  - $D^{tr} = \{z_1, z_2, \cdots, z_{tr}\}$
  - $D^{te} = \{z_{tr+1}, z_{tr+2}, \cdots, z_{tr+te}\}$
  - $tr + te = n$
- select $f^*(D^{tr}) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, D^{tr})$

  (this optimization process could include model selection)
- estimate $R(f^*(D^{tr}))$ with $\hat{R}(f^*(D^{tr}), D^{te}) = \frac{1}{te} \sum_{z_i \in D^{te}} L(z_i, f^*(D^{tr}))$

Estimation of the Risk - Cross-validation

- Divide your training set D_n into K distinct and equal parts $D^1, \cdots, D^k, \cdots, D^K$
- For each part D^k
  - let $\bar{D}^k$ be the set of examples that are in D_n but not in D^k
  - select $f^*(\bar{D}^k) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, \bar{D}^k)$

    (this process could include model selection)
  - estimate $R(f^*(\bar{D}^k))$ with $\hat{R}(f^*(\bar{D}^k), D^k) = \frac{1}{|D^k|} \sum_{z_i \in D^k} L(z_i, f^*(\bar{D}^k))$
- estimate $R(f^*(D_n))$ with $\frac{1}{K} \sum_k R(f^*(\bar{D}^k))$
- When $k = n$: leave-one-out cross-validation

Estimation of the Risk and Model Selection

- When you want both the best model and its expected risk.
- You then need to merge the methods already presented.
  For instance:
  - train-validation-test: 3 separate data sets are necessary
  - cross-validation + test: cross-validate on train set, then test on separate set
  - double-cross-validation: for each subset, need to do a second cross-validation with the $K-1$ other subsets
- Other important methodological aspects:
  - compare your results with other methods!!!!
  - use statistical tests to verify significance
  - verify your model on more than one datasets

Train - Validation - Test

- Select a family of functions with hyper-parameter θ
- Divide your training set D_n into three parts D^{tr}, D^{va}, and D^{te}
- For each value θ_m of the hyper-parameter θ
  - select $f^*_{\theta_m}(D^{tr}) = \arg\min_{f \in \mathcal{F}_{\theta_m}} \hat{R}(f, D^{tr})$
  - let $\hat{R}(f^*_{\theta_m}(D^{tr}), D^{va}) = \frac{1}{va} \sum_{z_i \in D^{va}} L(z_i, f^*_{\theta_m}(D^{tr}))$
- select $\theta^*_m = \arg\min_{\theta_m} \hat{R}(f^*_{\theta_m}(D^{tr}), D^{va})$
- select $f^*(D^{tr} \cup D^{va}) = \arg\min_{f \in \mathcal{F}_{\theta^*_m}} \hat{R}(f, D^{tr} \cup D^{va})$
- estimate $R(f^*(D^{tr} \cup D^{va}))$ with $\frac{1}{te} \sum_{z_i \in D^{te}} L(z_i, f^*(D^{tr} \cup D^{va}))$

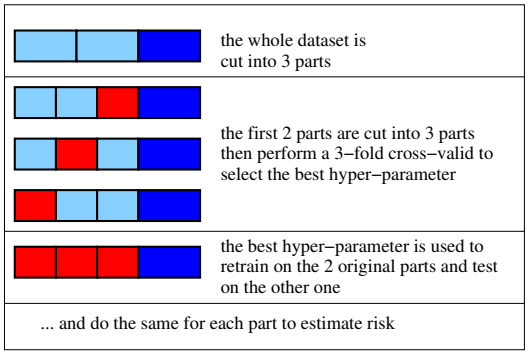
Cross-validation + Test

- Select a family of functions with hyper-parameter θ
- Divide you dataset D_n into two parts:
  a training set D^{tr} *and a test set* D^{te}
- For each value θ_m of the hyper-parameter θ
  estimate $R(f^*_{\theta_m}(D^{tr}))$ *with* D^{tr} *using cross-validation*
- select $\theta^*_m = \arg\min_{\theta_m} R(f^*_{\theta_m}(D^{tr}))$
- retrain $f^*(D^{tr}) = \arg\min_{f \in \mathcal{F}_{\theta^*_m}} \hat{R}(f, D^{tr})$
- estimate $R(f^*(D^{tr}))$ with $\frac{1}{te} \sum_{z_i \in D^{te}} L(z_i, f^*(D^{tr}))$

Double Cross-validation

- Select a family of functions with hyper-parameter θ
- Divide your training set D_n into K distinct and equal parts $D^1, \cdots, D^k, \cdots, D^K$
- For each part D^k
  - select the best model $f^*(\bar{D}^k)$ by cross-validation on $\bar{D}^k$
  - estimate $R(f^*(\bar{D}^k))$ with $\hat{R}(f^*(\bar{D}^k), D^k) = \frac{1}{|D^k|} \sum_{z_i \in D^k} L(z_i, f^*(\bar{D}^k))$
- estimate $R(f^*(D))$ with $\frac{1}{K} \sum_k R(f^*(\bar{D}^k))$
- Note: this process only gives you an estimate of the risk, but not a model. If you need the model as well, you have to perform a separate model selection process!

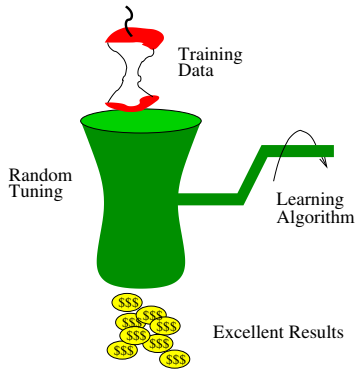
Double Cross-validation



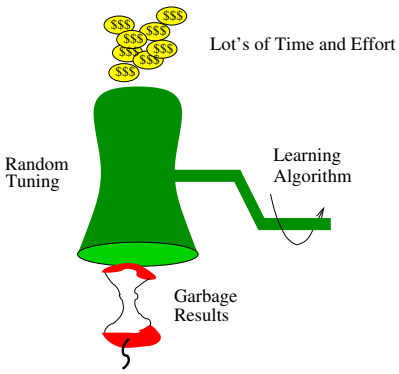
Examples of Known Models

- Multi-Layer Perceptrons (regression, classification)
- Radial Basis Functions (regression, classification)
- Support Vector Machines (classification, regression)
- Gaussian Mixture Models (density estimation, classification)
- Hidden Markov Models (density estimation, classification)
- Graphical Models (density estimation, classification)
- AdaBoost and Bagging (classification, regression, density estimation)
- Decision Trees (classification, regression)

Beware of the Machine Learning Magic



Beware of the Machine Learning Magic (con't)



Markov Models

- Stochastic process of a temporal sequence: the probability distribution of the variable q at time t depends on the variable q at times $t-1$ to 1.

$$P(q_1, q_2, \ldots, q_T) = P(q_1^T) = P(q_1) \prod_{t=2}^{T} P(q_t | q_1^{t-1})$$

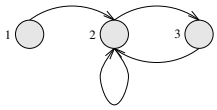
- First Order Markov Process:

$$P(q_t | q_1^{t-1}) = P(q_t | q_{t-1})$$

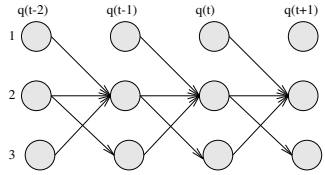
- Markov Model: model of a Markovian process with discrete states.

Markov Models (Graphical View)

- A Markov model:



- A Markov model unfolded in time:



Training Markov Models

- A Markov model is represented by all its transition probabilities:

$$P(q_t = i | q_{t-1} = j) \quad \forall i, j$$

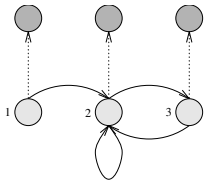
- Given a training set of sequences X, training means re-estimating these probabilities.
- Simply count them to obtain the maximum likelihood solution:

$$P(q_t = i | q_{t-1} = j) = \frac{\#(q_t = i \text{ and } q_{t-1} = j | X)}{\#(q_{t-1} = j | X)}$$

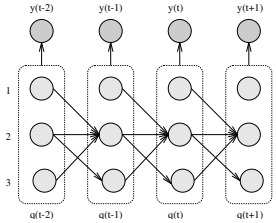
- Example: observe the weather today assuming it depends on the previous day.

Hidden Markov Models

- A hidden Markov model:



- A hidden Markov model unfolded in time:



Elements of an HMM

Hidden Markov Model: Markov Model whose state is not observed, but of which one can observe a manifestation (a variable x_t which depends only on q_t).

- A finite number of states N.
- Transition probabilities between states, which depend only on previous state: $P(q_t{=}i|q_{t-1}{=}j, \theta)$.
- Emission probabilities, which depend only on the current state: $p(x_t|q_t{=}i, \theta)$ (where x_t is observed).
- Initial state probabilities: $P(q_0 = i|\theta)$.
- Each of these 3 sets of probabilities have parameters θ to estimate.

Derivation of the Forward Variable α

the probability of having generated the sequence x_1^t and being in state i at time t:

$$\begin{aligned}
\alpha(i,t) &\stackrel{\text{def}}{=} p(x_1^t, q_t = i) \\
&= p(x_t|x_1^{t-1}, q_t = i)p(x_1^{t-1}, q_t = i) \\
&= p(x_t|q_t = i)\sum_j p(x_1^{t-1}, q_t = i, q_{t-1} = j) \\
&= p(x_t|q_t = i)\sum_j P(q_t = i|x_1^{t-1}, q_{t-1} = j)p(x_1^{t-1}, q_{t-1} = j) \\
&= p(x_t|q_t = i)\sum_j P(q_t = i|q_{t-1} = j)p(x_1^{t-1}, q_{t-1} = j) \\
&= p(x_t|q_t = i)\sum_j P(q_t = i|q_{t-1} = j)\alpha(j, t-1)
\end{aligned}$$

From α to the Likelihood

- Reminder: $\alpha(i,t) \stackrel{\text{def}}{=} p(x_1^t, q_t = i)$
- Initial condition:

$$\alpha(i,0) = P(q_0 = i) \rightarrow \text{ prior probabilities of each state } i$$

- Then let us compute $\alpha(i,t)$ for each state i and each time t of a given sequence x_1^T
- Afterward, we can compute the likelihood as follows:

$$\begin{aligned}
p(x_1^T) &= \sum_i p(x_1^T, q_T = i) \\
&= \sum_i \alpha(i,T)
\end{aligned}$$

- Hence, to compute the likelihood $p(x_1^T)$, we need $\mathcal{O}(N^2 \cdot T)$ operations, where N is the number of states

Basics of Training with Expectation-Maximization

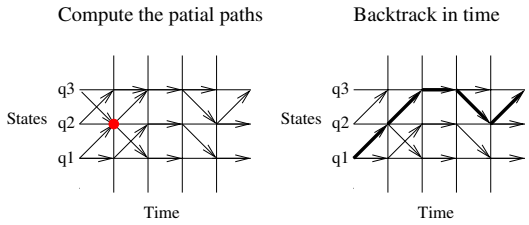
- Objective: maximize the likelihood $p(X;\theta)$ of the data X drawn from an unknown distribution, given the model parameterized by θ:

$$\theta^* = \arg\max_\theta p(X|\theta) = \arg\max_\theta \prod_{p=1}^{n} p(x_p|\theta)$$

- Basic ideas of EM:
  - Introduce a hidden variable such that *its knowledge would simplify the maximization of* $p(X;\theta)$
  - At each iteration of the algorithm:
    - ⋆ E-Step: estimate the distribution of the hidden variable given the data and the current value of the parameters
    - ⋆ M-Step: modify the parameters in order to maximize the joint distribution of the data and the hidden variable
  - For HMM, hidden variables: $p(q_t = i|X), p(q_t = i, q_{t-1} = j|X)$

The Most Likely Path (Graphical View)

- The Viterbi algorithm finds the best state sequence.



The Viterbi Algorithm for HMMs

The Viterbi algorithm finds the best state sequence.

$$
\begin{aligned}
V(i,t) &\stackrel{\text{def}}{=} \max_{q_1^{t-1}} p(x_1^t, q_1^{t-1}, q_t{=}i) \\
&= \max_{q_1^{t-1}} p(x_t|x_1^{t-1}, q_1^{t-1}, q_t{=}i) p(x_1^{t-1}, q_1^{t-1}, q_t{=}i) \\
&= p(x_t|q_t{=}i) \max_{q_1^{t-2}} \max_j p(x_1^{t-1}, q_1^{t-2}, q_t{=}i, q_{t-1}{=}j) \\
&= p(x_t|q_t{=}i) \max_{q_1^{t-2}} \max_j p(q_t{=}i|q_{t-1}{=}j) p(x_1^{t-1}, q_1^{t-2}, q_{t-1}{=}j) \\
&= p(x_t|q_t{=}i) \max_j p(q_t{=}i|q_{t-1}{=}j) \max_{q_1^{t-2}} p(x_1^{t-1}, q_1^{t-2}, q_{t-1}{=}j) \\
&= p(x_t|q_t{=}i) \max_j p(q_t{=}i|q_{t-1}{=}j) V(j, t-1)
\end{aligned}
$$

From Viterbi to the State Sequence

- Reminder: $V(i,t) = \max_{q_1^{t-1}} p(x_1^t, q_1^{t-1}, q_t{=}i)$
- Let us compute $V(i,t)$ for each state i and each time t of a given sequence x_1^T
- Moreover, let us also keep for each $V(i,t)$ the associated argmax previous state j
- Then, starting from the state $i = \arg\max_j V(j,T)$ backtrack to decode the most probable state sequence.
- Hence, to compute all the $V(i,t)$ variables, we need $\mathcal{O}(N^2 \cdot T)$ operations, where N is the number of states

Applications of HMMs

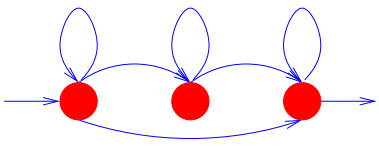
- Classifying sequences such as...
  - DNA sequences (which family)
  - gesture sequences
  - video sequences
  - phoneme sequences
  - etc.
- Decoding sequences such as...
  - continuous speech recognition
  - handwriting recognition
  - sequence of events (meeting, surveilance, games, etc)

HMMs for Speech Recognition

- Application: continuous speech recognition:

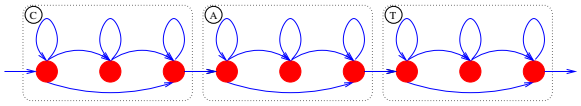
  Find a sequence of phonemes (or words) given an acoustic sequence

- Idea: use a phoneme model



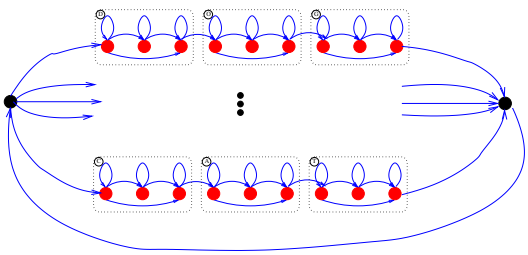
Embbeded Training of HMMs

- For each acoustic sequence in the training set, create a new HMM as the concatenation of the HMMs representing the underlying sequence of phonemes.
- Maximize the likelihood of the training sentences.



HMMs: Decoding a Sentence

- Decide what is the accepted vocabulary.
- Optionally add a language model: P(word sequence)
- Efficient algorithm to find the optimal path in the decoding HMM:



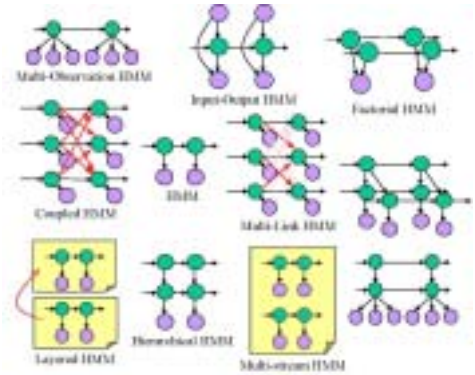
Measuring Error

- How do we measure the quality of a speech recognizer?
- Problem: the target solution is a sentence, the obtained solution is also a sentence, but they might have different size!
- Proposed solution: the Edit Distance:
  - assume you have access to the operators insert, delete, and substitute,
  - what is the smallest number of such operators we need to go from the obtained to the desired sentence?
  - An efficient algorithm exists to compute this.
- At the end, we measure the error as follows:

$$\text{WER} = \frac{\#\text{ins} + \#\text{del} + \#\text{subst}}{\#\text{words}}$$

- Note that the word error rate (WER) can be greater than 1...

How to Handle Multiple Streams?



Some Well Known Models

Early Integration HMMs

- Concatenate frames of all streams, frame by frame.
- Training maximizes the joint likelihood of streams.
- Assumption: streams are frame-synchronous.

Multi-Stream HMMs

- Train one HMM per stream, independently.
- Training does not maximize the joint likelihood of streams.
- Likelihoods are merged during decoding, generally at each state.
- Assumption: streams are state-synchronous.

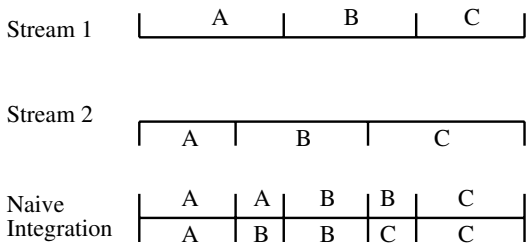
Challenges in Multi Channel Integration

- How to handle more than two streams: computational complexity?
  - Most solutions that model the joint probability of streams need exponential resources with respect to the number of streams.
  - Heuristics to limit the search space.
- How to handle learning in high dimensional spaces?
  - The observation space grows with the number of streams.
  - Often, the number of parameters follows linearly or worse...
- How to handle long term temporal dependencies?
  - This is already a problem with one stream!

Challenges in Multi Channel Integration

- Joint feature extraction and heterogeneity of sources
  - For the moment features of each stream are extracted independently...
  - Extracting jointly should increase the robustness.
  - What about streams of different nature (such as text and videos)?
- How to handle different levels of *a priori* knowledge constraints? a scene can be described by pixels, persons, actions, language...
- Available benchmark datasets for evaluation
  - This is a key point for progress in this field.
  - Audio-visual speech recognition.
  - Meeting scenario.
  - Any others? (surveillance, games, ...)

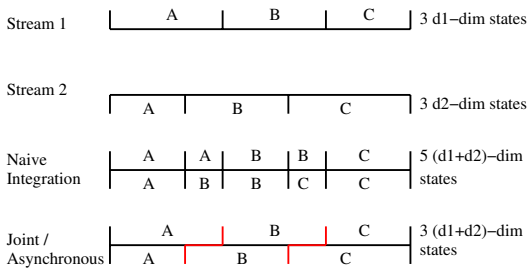
Challenges in Multi Channel Integration: Asynchrony



Some Evidence of Stream Asynchrony

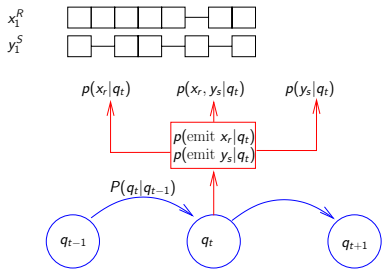
- Audio-visual speech recognition with lip movements: lips do not move at the same time as we hear the corresponding sound.
- Speaking and pointing: pointing to a map and saying "I want to go there".
- Gesticulating, looking at, and talking to someone during a conversation.
- In a news video, the delay between the moment when the newscaster says "Bush" and the moment when Bush's picture appears.
- ...

Asynchrony Revisited

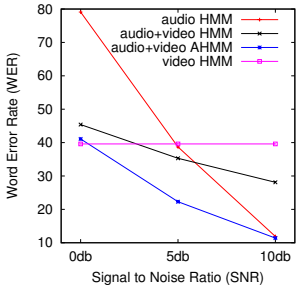


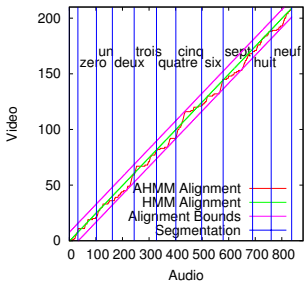
Asynchronous HMMs

- Enables re-synchronization of streams.
- One HMM: maximizes the likelihood of all streams jointly.



Alignments with Asynchronous HMMs





Training AHMMs and Results

- EM algorithm maximizes the joint probability $p(\mathbf{x}_1^R, \mathbf{y}_1^S, \cdots)$.
- Complexity grows with number of streams, can be controlled.
- Easily adaptable to complex tasks such as speech recognition.
- Significant performance improvements in
  - audio-visual speech recognition,
  - audio-visual speaker verification,
  - meeting analysis.

- S. Bengio. Multimodal Speech Processing Using Asynchronous Hidden Markov Models. *Information Fusion*, 5(2):81–89, 2004.
- S. Bengio. An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition. *NIPS 15*, 2002.

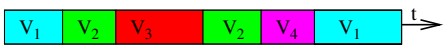
Complexity of the Meeting Scenario

- Modeling multimodal group-level human interactions in meetings.
- Multimodal nature: data collected from multiple sensors (cameras, microphones, projectors, white-board, etc).
- Group nature: involves multiple interacting persons at the same time.

Meeting Analysis

- Structure a meeting as a sequence of group actions taken from an exhaustive set V of N possible actions:

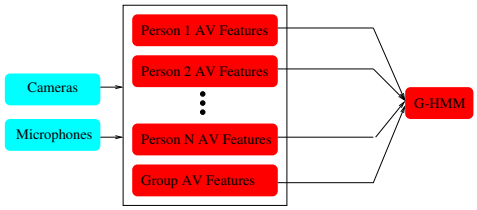
$$V = \{v_1, v_2, v_3, \cdots, v_N\}$$



- Recorded and annotated 30 training a 30 test meetings.
- Extract high level audio and visual features.
- Try to recover the target action sequence of unseen meetings.

I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic Analysis of Multimodal Group Actions in Meetings. *IEEE Transactions on PAMI*, 27(3), 2005.

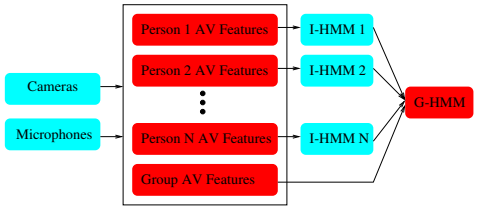
A One-Layer Approach



Classical Approach

- A large vector of audio-visual features from each participant and group-level features are concatenated to define the observation space.
- A general HMM is trained using a set of labeled meetings.

A Two-Layer Approach



Advantages

- Smaller observation space.
- I-HMMs share parameters, person independent, trained on simple task (write, talk, passive).
- Last layer less sensitive to variations of low-level data.

Experiments

Group Actions: Turn-Taking

- Discussion
- Monologue
- Monologue/Note-Taking
- Presentation
- Presentation/Note-Taking
- White-board
- White-board/Note-Taking

Individual Actions

Speaking - Writing - Passive

Results

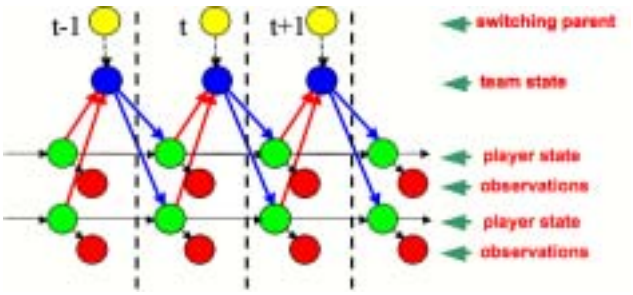
Method	Features	AER
One-Layer	Visual Only	48.20
	Audio Only	36.70
	Audio Visual	23.74
Two-Layer	Visual Only	42.45
	Audio Only	32.37
	Audio Visual	16.55
	Async HMM	15.11

D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling Individual and Group Actions in Meetings: a Two-Layer HMM Framework. In *CVPR*, 2004.

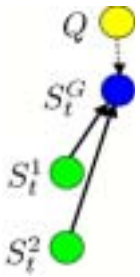
Learning Influence in Human Interactions

What is the influence of each person in a meeting? who is the dominant one? who drives the decisions taken in the meeting?

The Team-Player Influence Model



Defining Influence



$$P(S_t^G|S_t^1 \cdots S_t^N) =$$

$$= \sum_{i=1}^{N} P(S_t^G, Q = i|S_t^1 \cdots S_t^N)$$

$$= \sum_{i=1}^{N} P(Q = i|S_t^1 \cdots S_t^N) P(S_t^G|S_t^i \cdots S_t^N, Q = i)$$

$$= \sum_{i=1}^{N} P(Q = i) P(S_t^G|S_t^i) = \sum_{i=1}^{N} \alpha_i P(S_t^G|S_t^i).$$

α_i describes the influence of player i over the team.

Training the Team-Player Influence Model

Advantages of the Model

- The switching parent makes the system tractable.
- Complexity grows linearly with the number of individual chains.
- The actual (trained) value of the switching parent provides the influence (posterior probability) of using a player state to determine the team outcome.

Experiments on Meetings

Meeting Corpus

- 30 meetings (5-min, 4-participants)
- 3 annotators, good agreement between them (Kappa = 0.68)

Audio Features

- Speech activity (SRP-PHAT)
- Speech pitch
- Speech energy
- Speaking rate

Results on Meetings

Models:

Method	KL divergence
model + Audio	0.135
Speaking length	0.226
Random guess	0.863

Annotators:

Human Annotation	KL divergence
A_i vs. A_j	0.09
A_i vs. $A_{\bar{i}}$	0.053
A_i vs. GT	0.037

D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning Influence Among Interacting Markov Chains. In *NIPS 18*, 2005.

Introduction

- Context: large research collaboration
  - Having a common software platform is important
  - Modularity, portability, and efficiency, would be important
- At IDIAP, we are working on various but related problems:
  - Speech processing
  - Image and video processing
  - Multimodal processing
  - Text retrieval
  - Machine Learning algorithms

What is Torch?

- Machine Learning library written in C++ under FreeBSD license.
- Modular Design and efficiency (space and time).
- Objective: it should contain most of the state-of-the-art Machine Learning algorithms.
- Examples already available:
  - all kinds of neural networks including multi-layer perceptrons, mixtures of experts, radial basis functions, LeNet, recurrent NNs.
  - support vector machines, for classification and regression,
  - kmeans, Gaussian mixture models, hidden Markov models,
  - non-parametric models such as k-nearest neighbors or Parzen,
  - ensemble models such as bagging and adaboost.

Other Interesting Features of Torch

- Easy-to-use Command Line manager.
- General purpose matrix package (inversion, eigenvectors, etc).
- General output functions (warning, error, print, message, etc).
- Some random generator functions (uniform, gaussian, etc).
- Automatic documentation generator using DOC++.
- Easy-to-create any kind of machine or neural networks.
- Easy-to-create any kind of criterion to optimize.

A Word about Performance

- Torch has been designed to be essentially time efficient.
- Some comparisons have been performed with other packages
  - MLP is most often faster than the other tested models,
  - SVM with Gaussian Kernels was the fastest implementation (not anymore true now),
  - GMM for speaker verification is the fastest implementation,
  - HMM for speech recognition using embedded training is as fast as HTK on small vocabulary tasks.
- Torch has also been designed to be modular, as it is a research oriented library: it is easy to add new classes and test new research ideas.

How to Obtain Torch

- Go to www.Torch.ch
- Download linux or windows source and doc packages
- Download and use the xmake package
- (you need gcc and python, of course!)
- compile everything!!!!

An Example of a Main.cc

The Command Line

```
// Construct the command line with help
CmdLine cmd;
cmd.info(help);
// Ask for arguments
cmd.addText("\nArguments:");
cmd.addSCmdArg("file", &file, "the train or test file");
cmd.addICmdArg("n_inputs", &n_inputs, "input dimension");
cmd.addICmdArg("n_targets", &n_targets, "target dimension");
// Propose some options
cmd.addText("\nModel Options:");
cmd.addICmdOption("-nhu", &nhu, 25, "# of hidden units");
cmd.addBCmdOption("-norm", &norm, true, "normalize inputs");

cmd.read(argc, argv);
```

An Example of a Main.cc

Creating the Objects

```
MLP mlp(3,n_inputs, "linear", nhu,
                    "tanh",   nhu,
                    "linear", n_targets);

MatDataSet data(train_file, n_inputs, n_targets);
MeanVarNorm mv_norm(&data)
data.preProcess(mv_norm);

MseCriterion mse(n_targets);

StochasticGradient trainer(&mlp,&mse);
trainer.setIOption("max iter", max_iter);
trainer.setROption("learning rate", learning_rate);
```


An Example of a Main.cc

Measure and Train by Cross-Validation

```
MeasurerList measurers;
MseMeasurer mse_m(mlp.outputs, &data, "valid_mse");
measurers.addNode(&mse_m);

KFold k(&trainer,k_fold);
k.crossValidate(data, NULL, measurers);
mlp.save(model_file);
```

Example of Methods

GaussianKernel

```
real GaussianKernel::eval(real *x, real *y)
{
  real sum = 0.;
  for(int i = 0; i < frame_size; i++)
  {
    real z = x[i] - y[i];
    sum -= z*z;
  }

  return exp(g*sum);
}
```

Concluding Remarks

Multi Channel Sequence Processing is Challenging!

- It is a relatively new research domain.
- Multiple applications require such a framework.
- It gives rise to several interesting research challenges.

Some Attempts at Current Multi Channel Challenges:

Asynchrony: Asynchronous HMMs.

Complexity/Levels: Layered Approach.

Complexity/Numbers of channels: Influence Model.

Benchmark Datasets: The Meeting datasets, available at `http://mmm.idiap.ch`.

Machine Learning: The Torch library, available at `http://www.torch.ch`. $>$ 15000 downloads.

A Discriminative Approach for the Retrieval of Images from Text Queries

David Grangier, Florent Monay and Samy Bengio

IDIAP Research Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
{grangier, monay, bengio}@idiap.ch

September 2006

Outline

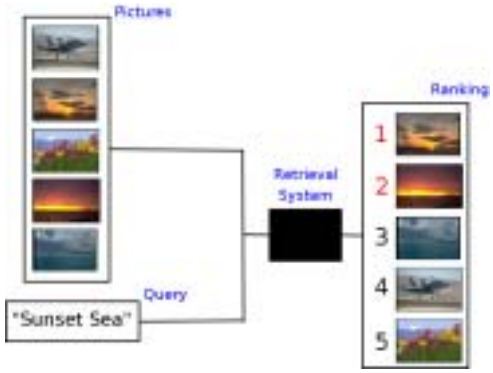
- **Introduction**
  What is Image Retrieval from Text Queries ?
- **State-of-the-Art**
  Manual Annotation / Generative Auto-Annotation Models
- **Proposed Approach**
  Model Parameterization, Training Criterion and Learning Procedure.
- **Experiments and Results**
  Dataset, Features and Results.
- **Conclusions**

Image Retrieval from Text Queries ?

- **Input:**
  a set of pictures P and a text query q.
- **Output:**
  a picture ranking in which the pictures relevant to q appear above the others.

- Ranking task
- Need to generalize to unseen pictures and queries

Image Retrieval from Text Queries ?



State-of-the-Art Approaches

Mainly two types of approaches to this retrieval problem:

- Manual Annotation Approaches
- Generative Models for Auto-Annotation

Manual Annotation

- Each picture is annotated with a manually-produced caption,
- Text retrieval techniques are applied over the captions.

e.g. Corbis Stock Photography, Google Images...

Manual Annotation

- Each picture is annotated with a manually-produced caption,
- Text retrieval techniques are applied over the captions.

e.g. Corbis Stock Photography, Google Images...

\+ effectiveness
\- annotation cost, incomplete/biased annotations

Generative Models for Automatic Annotation

Objective:

- Avoid the manual annotation step.

Generative Models for Automatic Annotation

Approach:

- a model generating picture/caption pairs $P(p, c|\theta)$ is trained over a training set of pictures with captions.
- for any test picture without caption, we can infer the most likely caption or a distribution over the vocabulary.

Generative Models for Automatic Annotation

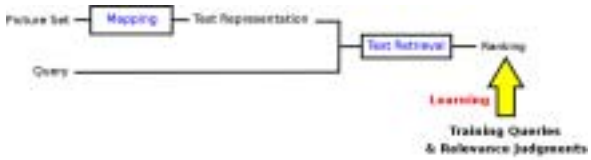
Such Models include:

- Cross Media Relevance Model (CMRM),
  → [Jeon et al., SIGIR'03]
- Cross Media Translation Table (CMTT),
  → [Pan et al., ICME'04]
- Probabilistic Latent Semantic Analysis (PLSA)...
  → [Hofmann, ML '01], [Monay et al., ACM–MM '04]

Generative Models for Automatic Annotation

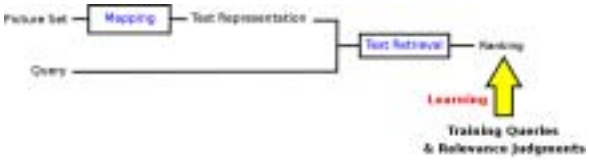
\+ no need for manual annotation
\- focus on an intermediate task,
do not optimize retrieval performance

The Proposed Approach: a Discriminative Model



Goal: learn a model to optimize the retrieval performance.

The Proposed Approach: a Discriminative Model



Passive-Aggressive Model for Image Retrieval (PAMIR)

- a parametric function for image ranking,
  → inspired from text retrieval.
- a loss related to the retrieval performance,
  → adapted from Joachims's ranking SVM (KDD'02).
- a training procedure.
  → adapted from Crammer et al.'s Passive-Aggressive (JMLR'06).

The Proposed Approach: Parameterization

What is our parametric ranking model ?

Ranking Approach:

Given a query q, our model assign a score $F_W(q, p)$ to each picture p and then rank the pictures by decreasing scores.

The Proposed Approach: Parameterization

What is our parametric ranking model ?

Ranking Approach:

> Given a query q, our model assign a score $F_W(q, p)$ to each picture p and then rank the pictures by decreasing scores.

Scoring Function: $F_W(q, p)$ is computed as follows,

1. map the picture p into the text space, with a linear mapping:

$$p \rightarrow f_W(p) = Wp,$$

   where W is the parameter matrix.

2. compute the dot product between the text vectors $f_W(p)$ and q:

$$F_W(q, p) = q \cdot f_W(p).$$

The Proposed Approach: Learning Objective

How to select the parameters W of the scoring function $F_W(q, p)$?

Recall that:

> For any query, the relevant pictures should appear above the others.
> The pictures are ranked by decreasing scores $F_W(q, p)$.

Therefore, $F_w(\cdot, \cdot)$ would achieve perfect ranking iif

> for any query q,
> for any picture p^+ relevant to q, $F_W(q, p^+) > F_W(q, p^-)$.
> for any picture p^- non-relevant to q,

The Proposed Approach: Learning Objective

Therefore, $F_w(\cdot,\cdot)$ would achieve perfect ranking iif

for any query q,
for any picture p^+ relevant to q, $\quad F_W(q,p^+) > F_W(q,p^-)$.
for any picture p^- non-relevant to q,

Given a training set D_{train}, we want to minimize the number of non-satisfied constraints.

We rely on the loss,

$$L(F_W; D_{train}) = \sum_{(q,p^+,p^-)\in D_{train}} \max(0, 1 - F_W(q,p^+) + F_W(q,p^-)),$$

which is a bound over the number of non-satisfied constraints.

The Proposed Approach: Learning Procedure

How to minimize $W \rightarrow L(F_W; D_{train})$ **?**

Passive-Aggressive Training

- Training constraints are examined sequentially.
- For each constraint, W is updated as a trade-off between
  - remaining close to the current weight,
  - satisfying the current constraint

The Proposed Approach: Learning Procedure

How to minimize $W \rightarrow L(F_W; D_{train})$ **?**

Passive-Aggressive Training

- Training constraints are examined sequentially.
- For each constraint, W is updated as a trade-off between
  - remaining close to the current weight,
  - satisfying the current constraint

i.e. for the t^{th} training constraint (q_t, p_t^+, p_t^-),

$$\begin{aligned} W_t &= \arg\min_W \|W - W_{t-1}\|^2 + C \overbrace{\max(0, 1 - F_W(q_t, p_t^+) + F_W(q_t, p_t^-))}^{l_t(W)} \\ &= W_{t-1} + \alpha_t v_t \end{aligned}$$

where
- C is a hyperparameter controlling the trade-off,
- v_t is the gradient of the loss term $l_t(W)$, i.e. $v_t = q_t(p_t^- - p_t^+)^{\mathrm{T}}$,
- $\alpha_t = \max(C, \frac{l_t(W)}{\|v_t\|^2})$

Introduction State-of-the-Art Proposed Approach Experiments & Results Conclusions

The Proposed Approach: Learning Procedure

How to minimize $W \rightarrow L(F_W; D_{train})$ **?**

Passive-Aggressive Training

- Training constraints are examined sequentially.
- For each constraint, W is updated as a trade-off between
  - remaining close to the current weight,
  - satisfying the current constraint

Advantages

- online training $\rightarrow$ efficient, scalable
- minimal update rule $\Leftrightarrow$ margin maximization $\rightarrow$ good generalization

The Proposed Approach

Passive-Aggressive Model for Image Retrieval (PAMIR)

- a parametric function for image ranking,
- a loss related to the retrieval performance,
- a training procedure.

Experimental Setup

Data

- Corel Dataset, standard split (Dyugulu ECCV'02):
  - train: 4,500 images and 7,200 queries,
  - test: 500 images and 2,200 queries,
- Query Representation: ▸ Def.
  - bag-of-words, normalized *idf* weighting.
- Picture Representation: ▸ Def.
  - visterms over blobs (Dyugulu ECCV'02) and SIFTs (Lowe IJCV'04).

Evaluation

- Average Precision evaluation ▸ Def.
- Comparison with alternative models:
  - CMRM, CMTT, PLSA

Results

Average Precision (%)

	CMRM	CMTT	PLSA	PAMIR
all queries	14.7	11.5	16.7	**21.6** (+29%)
1-word queries	19.2	19.1	24.5	**30.7** (+25%)

- Compare favorably with all alternatives.
- Consistent improvement over query set (Wilcoxon test, 95% confidence).

▸ Examples

Conclusions & Future Work

PAMIR model to retrieve images from text queries

- discriminative approach, → learning focus on the final retrieval task.
- efficient training, → online minimization procedure.
- good generalization properties → margin maximization.

Positive Experimental Results

- PAMIR outperforms state-of-the-art models.

Future Work

- tackle larger datasets,
- non-linear model with kernel-based parameterization.

That's all folks !

Do you have any comment, question ?

Results: Examples

Top 5 pictures for the queries 'Pillar' and 'Landscape'

'Pillar' PLSA

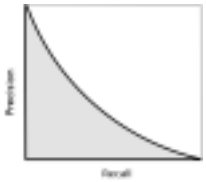
PAMIR

'Landscape' PLSA

PAMIR

Back to Results

Average Precision



Precision at position n

$$P(n) = \frac{\#\text{ of relevant pic. ranked above } n}{n}$$

Recall at position n

$$R(n) = \frac{\#\text{ of relevant pic. ranked above } n}{\text{total }\#\text{ of relevant pic.}}$$

For each query,

- Precision VS Recall $\Leftrightarrow$ $P(n)$ and $R(n)$ at each pos. n of the ranking.
- Average Precision corresponds to the Area Under Curve (AUC).

For a set of queries,

- Average Precision is averaged over the set.

Back to experimental setup

Bag-of-Words vectors

- each query vector q is vocabulary sized,
- the i^{th} component q_i is called the weight of the term i in query q

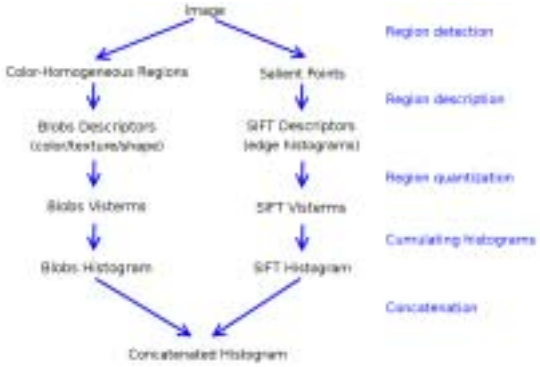
$$q_i = \frac{b_i \cdot idf_i}{\sqrt{\sum_j (b_j idf_j)^2}}$$

where b_j is 1 if term i appears in q and 0 otherwise,
idf_i is the inverse document frequency, $idf_i = -log(r_i)$,
r_i is the fraction of training captions containing term i.

$\rightarrow$ more weight is given to the discriminant, rare terms of the query.

Back to experimental setup

Visterm Histograms



Back to experimental setup

Methodes statistiques pour l'apprentissage de structures

P. Gallinari

Patrick.gallinari@lip6.fr

www-connex.lip6.fr

LIP6

University Pierre et Marie Curie – Paris - Fr

Plan

- Contexte et motivation
- Modèles pour l'apprentissage structuré
  - Modèles génératifs
  - Champs de Markov Conditionnels
  - Modèles à larges marges
  - Modèles à variables latentes
- XML Document Mining Challenge
- Webspam challenge

Apprentissage automatique et données structurées

- Modéliser, Classifier, regrouper des données structurées
  - Domaines: Chimie, biologie, XML, etc
  - Modèles: discriminant e.g. noyaux, génératifs e.g. densités d'arbres
- Predire des sorties structurées
  - Domaines: langage naturel, taxonomies, etc
  - Modèles: apprentissage relationnel, extension des méthodes à large marge
- Apprendre à associer des représentations structurées
  - Domaines: bases de données, recherche d'information
  - Modèles: génératifs, larges marges

Exemple 1
Document semi-structuré

Director Ang Lee Takes Risks with Mean Green 'Hulk'

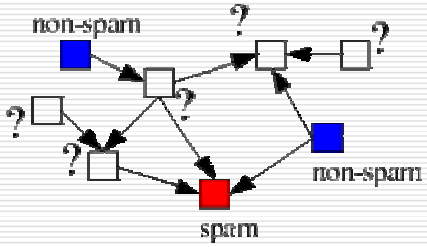
LOS ANGELES (Reuters) - Taiwan-born director Ang Lee, perhaps best known for his Oscar-winning "Crouching Tiger, Hidden Dragon," is taking a big risk with the splashy summer popcorn flick .....

FAMILY DRAMA, BIG ACTION

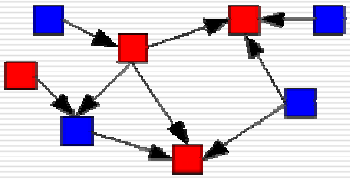
For loyal comic book fans who may think Lee's "Hulk" will be too touchy-feely, think again. " This is a drama, a family drama," said Lee, "but with big action." His slumping shoulders twitch and he laughs.....

Exemple 2
Adversial computing : web-spam

entrée



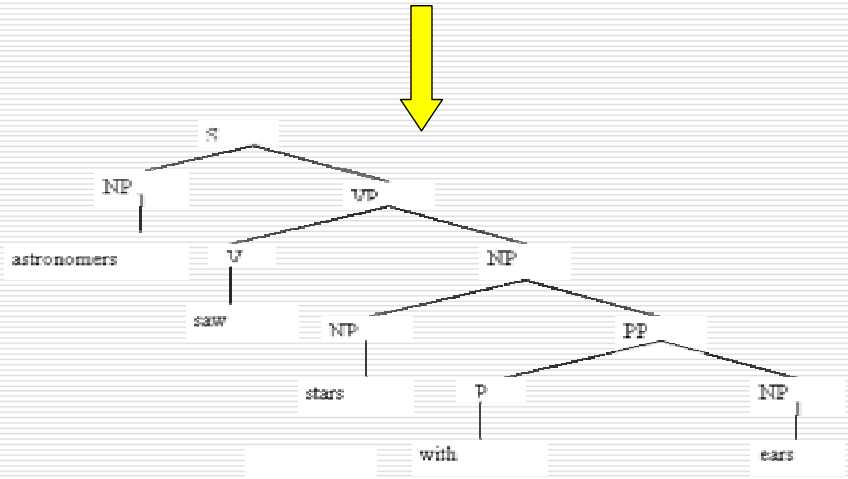
sortie



- Considérer contenu et structure externe
- Peu d'exemples en apprentissage

Exemple 3
Analyse syntaxique

- Astronomers saw stars with ears



Example 4
Semi-structured documents: Structural heterogeneity

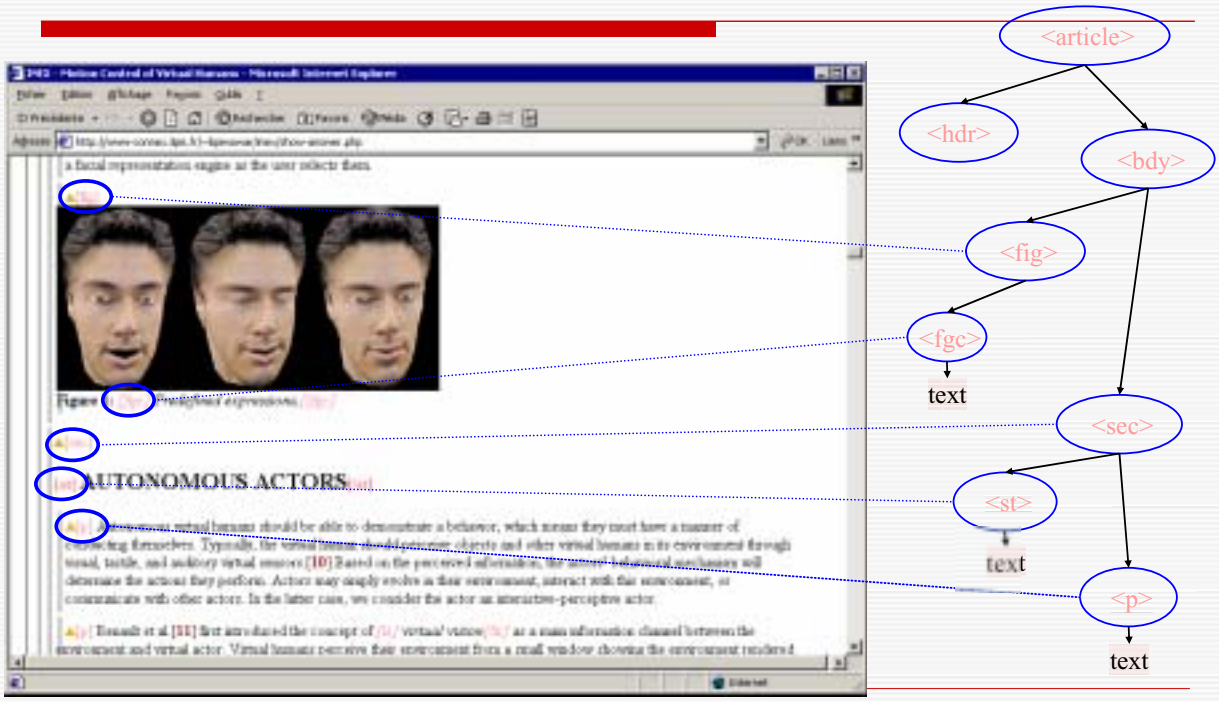
<**Restaurant**> <**Nom**>Tokyo Bar</**Nom**> <**Adresse**> <**Ville**>Paris</**Ville**> <**Arrd**>19</**Arrd**> <**Rue**>Bolivar</**Rue**> <**Num**>127</**Num**> </**Adresse**> <**Plat**>Sushi</**Plat**> <**Plat**>Sashimi</**Plat**> </**Restaurant**>	<**Restaurant**> <**Nom**>La cantine</**Nom**> <**Adresse**> 65 rue des pyrénées, Paris, 19ème, FRANCE </**Adresse**> <**Spécialités**> Canard à l'orange, Lapin au miel </**Spécialités**> </**Restaurant**>	<**Restaurant**> <**Nom**>L'olivier</**Nom**> <**Description**> Ce joli restaurant localisé près du métro Jaurès, au 19 du boulevard de la vilette, perdu dans le 19ème arrondissement de Paris propose une cuisine italienne, notamment des pâtes fraîches au 3 fromages. </**Description**> </**Restaurant**>

- Problem: Query heterogeneous XML databases or collections, Storage, etc
- Needs to know the correspondence between the structured representations

Modèles génératifs d'arbres

Classification / clustering de documents semi-structurés (Denoyer et al. 2003)

Context-XML semi-structured documents



Document model

$$d \mid (s^d, t^d)$$

$$P(D \mid d/\mathrm{N}) \mid P(S \mid s^d, T \mid t^d/\mathrm{N})$$

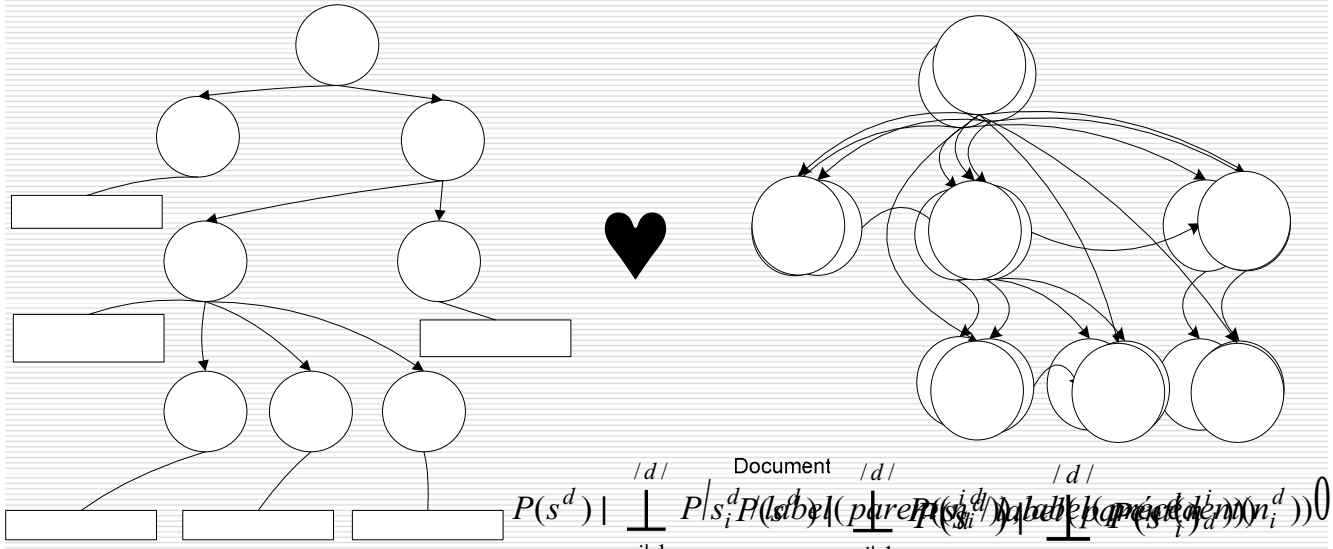
$$\mid P(S \mid s^d/\mathrm{N}) P(T \mid t^d/S \mid s^d, \mathrm{N})$$

Structural probability

Content probability

Document Model: Structure

- Belief Networks



$$P(s^d) \mid \prod_{i \mid 1}^{\mid d \mid} P(s_i^d \mid label(parent(s_i^d))) \quad P(s_i^d \mid label(parent(s_i^d))) \mid P(s_i^d \mid label(previous(s_i^d)))$$

Titre

Corps

Document Model: Content

- Model for each node

$$t_d \mid (t_d^1, \ldots, t_d^{\mid d \mid})$$

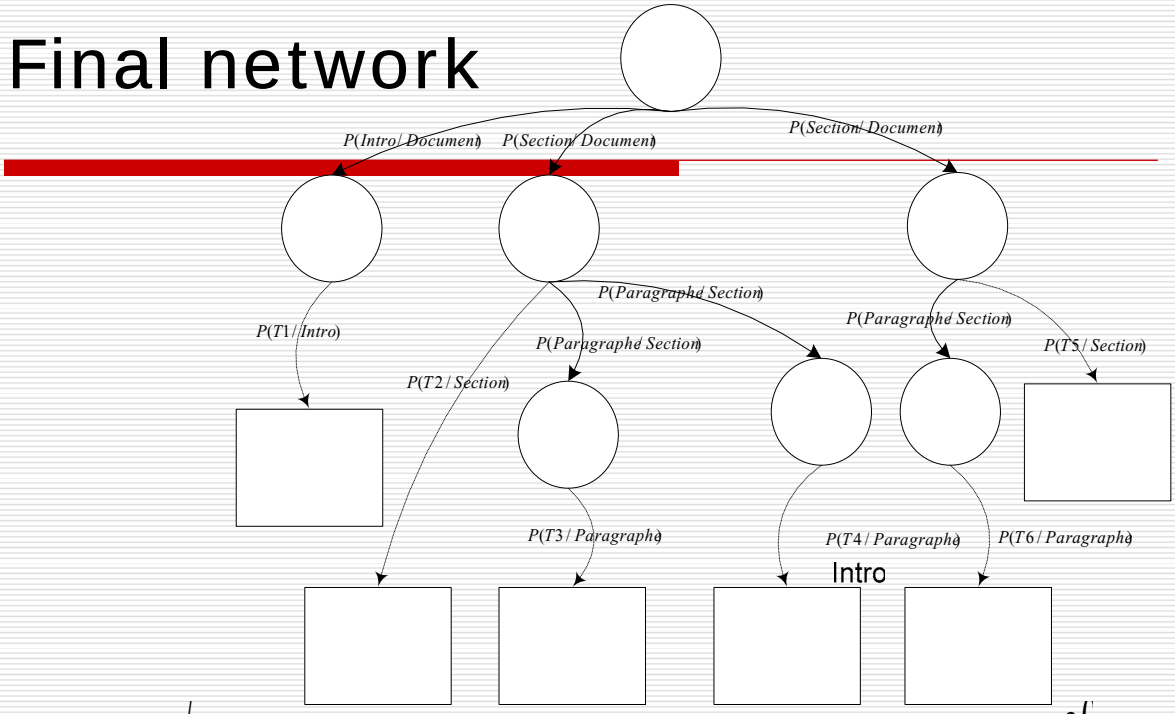
- 1st order dependency

$$P(t_d / s_d, \chi) \mid \prod_{i \mid 1}^{\mid d \mid} P(t_d^i / s_d^i, \chi)$$

- Use of a local generative model for each label

$$P(t_d^i / s_d^i, \chi) \mid P(t_d^i / \chi_{s_d^i})$$

Final network



$$P(d) = P(Intro/Document)P(Section/Document)^2 P(Paragraphe/Section)^3 \left(\right.$$
$$* P(T1/Intro)P(T2/Section)P(T3/Paragraphe)$$
$$* P(T4/Paragraphe)P(T5/Section)P(T6/Paragraphe)$$

Different learning techniques

- Likelihood maximization

$$L = \sum_{d \subset D_{TRAIN}} \log P(d/N)$$

$$= \left[\sum_{d \subset D_{TRAIN}} \log P(s^d/N^s) \right] + \left[\sum_{d \subset D_{TRAIN}} \sum_{i=1}^{|d|} \log P(t_i^d / s_i^d, N_{s_i^d}^t) \right]$$

$$= L_{structure} + L_{contenu}$$

- Discriminant learning

$$P(c/x) = \frac{1}{1 + e^{-\log \frac{P(x/c)}{P(x/\bar{c})}}}$$

$$= \frac{1}{1 + e^{-\sum_{i=1}^{n} \log \frac{\chi^c_{x_i, pa(x_i)}}{\chi^{\bar{c}}_{x_i, pa(x_i)}}}}$$

- Logistic function
  - Error minimization
- Fisher Kernel

Example: Multimedia model

Director Ang Lee Takes Risks with Mean Green 'Hulk'

LOS ANGELES (Reuters) - Taiwan-born director Ang Lee, perhaps best known for his Oscar-winning "Crouching Tiger, Hidden Dragon," is taking a big risk with the splashy summer popcorn flick …...

FAMILY DRAMA, BIG ACTION

For loyal comic book fans who may think Lee's "Hulk" will be too touchy-feely, think again. " This is a drama, a family drama," said Lee, "but with big action." His slumping shoulders twitch and he laughs…..

	Macroaverage recall	Microaverage recall
NB	89.9 [89.2 ;90.4]	88.4 [87.7 ;89]
Structure model with text	92.5 [91.9 ;93]	92.9 [92.3 ;93.3]
Structure model with pictures	83 [82.2 ;83.7]	82.7 [81.9 ;83.4]
Structure model text and pictures	**93.6 [93.1 ;94]**	**94.7 [94.2 ;95.1]**

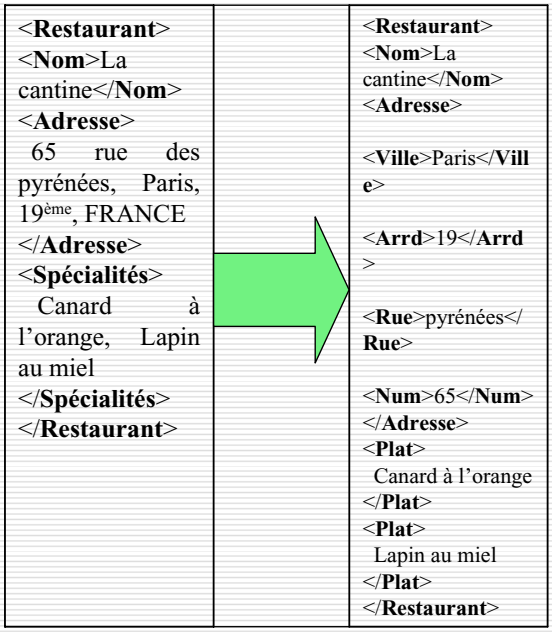
Modèles génératifs d'arbres (2)

Mapping d'arbres (Denoyer et al. 2005)

Document mapping problem

- Problem
  - Learn from examples how to map heterogeneous sources onto a predefined target schema
  - Preserve the document semantic
  - Sources: semistructured, HTML, PDF, flat text, etc

Labeled tree mapping problem



Document mapping problem

- Central issue: Complexity
  - Large collections
  - Large feature space: 10^3 to 10^6
  - Large search space (exponential)
- Approach
  - Learn generative models of XML target documents from a training set
  - Decoding of unknown sources according to the learned model

Problem formulation

Given

S_T a target format

$d_{sin(d)}$ an input document

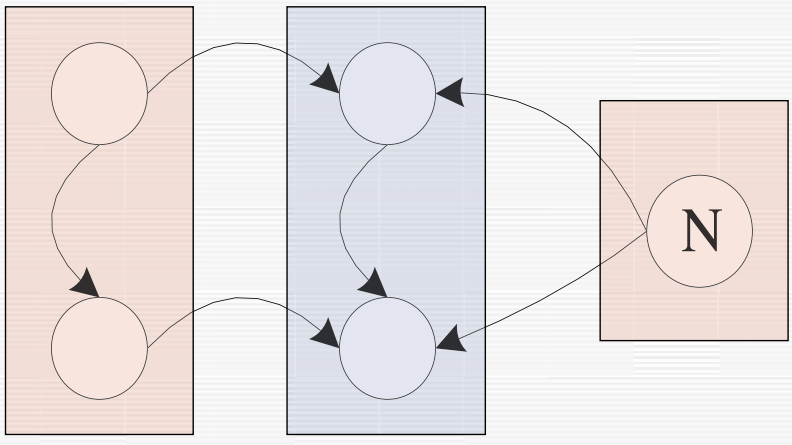
Find the most probable target document

$$d_{S_T} \mid \arg\max_{d' \subset S_T} P(d' \mid d_{S_{in(d)}})$$

Decoding

Learned transformation model

General restructuration model

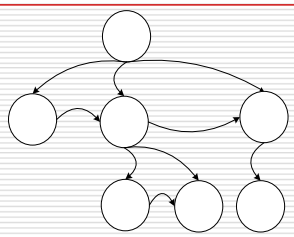


$$d_1 \mid \arg\max_{d'} P(s^{d'} / s^{d}, \mathrm{N}) P(t^{d'} / s^{d}, t^{d}, \mathrm{N})$$

Example : HTML to XML

- Hypothesis
  - Input document
    - HTML tags mostly for visualization
    - Remove tags
    - Keep only the segmentation (leaves)
  - Transformation
    - Leaves are the same in the HTML and XML document
    - Target document model: node label depends only on its local context
      - Context = content, left sibling, father

Model and training



- Probability of target tree

$$P(d_T \mid d_{Sin(d)}) \mid P(d_T \mid d_1,...,d_{|d|})$$

$$P(d_T \mid d_1,...,d_{|d|}) \mid \perp_{n_i} P(n_i \mid c_i, sib(n_i), father(n_i))$$

- Document model : max-entropy conditional model learned from a training set of target docs

$$P(n_i|c_i, sib(n_i), father(n_i)) = \frac{1}{Z_{c_i,sib(n_i),father(n_i)}} exp\left(< W_{n_i}, F_{c_i,sib(n_i),father(n_i)} >\right)$$

Decoding

- Solve $$d_{S_T} \mid \arg\max_{d' \subset S_T} P(d' \mid d_{S_{in(d)}})$$

$$d_{s_{FINAL}} = \underset{\substack{d_T \text{ such as} \\ (d^1,\ldots,d^{|d|})=(c_1,\ldots,c_{|d|})}}{\operatorname{argmax}} \prod_{n_i \in N_{d_T}} \frac{exp\left(< W_{n_i}, F_{c_i,sib(n_i),father(n_i)} >\right)}{Z_{c_i,sib(n_i),father(n_i)}}$$

- Exact Dynamic Programming decoding
  - O(|Leaf nodes|^3.|tags|)
- Approximate solution with LASO (Hal Daume ICML 2005)
  - O(|Leaf nodes|.|tags||tree nodes|)

Experiments : HTML to XML

- IEEE collection / INEX corpus
  - 12 K documents,
    - Average: 500 leaf nodes, 200 int nodes, 139 tags
- Movie DB
  - 10 K movie descriptions (IMDB)
    - Average: 100 leaf nodes, 35 int. nodes, 28 tags
- Shakespeare 39 plays
  - Few doc, but:
    - Average: 4100 leaf nodes, 850 int nodes, 21 tags
- Mini-Shakespeare
  - Randomly chosen 60 scenes from the plays
    - 85 leaf nodes, 20 int. nodes, 7 tags

Performances

Collection	Method	Micro	Macro	Internal	Full	Learning time	Testing time
INEX	DP	**79.6%**	**47.5%**	51.5%	**70.5%**	30 min	$\simeq$ 4 days
	LaSO	75.8%	42.9%	**53.1%**	67.5%	$>$ 1 week	**3h20min**
Movie	DP	**95.3%**	**91.2%**	77.1%	**90.4%**	20 min	$\simeq$ 2 days
	LaSO	90.5%	88.6%	**86.8%**	89.6%	$>$ 1 week	**1h15min**
Shakespeare	LaSO	95.3%	78.0%	77.0%	92.2%	$\simeq$ 5 days	**30 min**
Mini-shakespeare	DP	**98.7%**	**95.7%**	**94.7%**	**97.9%**	2 min	$\simeq$ 1 hour
	LaSO	89.4%	83.9%	63.2%	84.4%	20 min	**1 min**

Champs conditionnels de Markov (CRF) (Lafferty et al 2001)

- X, Y variables aléatoires, Y prend ses valeurs dans un alphabet fini
- X, Y sont structurés
  - e.g. séquences, arbres, etc
- CRF : modélise P(Y/X)
- Dans la plupart des travaux actuels
  - X : séquence d'observations
  - Y : séquence d'étiquettes

- Modèle discriminant
- Le conditionnement est fait sur X entier
  - Pas d'hypothèse d'indépendance sur les composantes de X
  - Permet de prendre en compte des composantes globales de la séquence X et donc (en principe) de modéliser des dépendances quelconques

CRF : définition

Ä G = (V,E) graphe non orienté défini sur Y (Y = $(Y_v)_{v \subset V}$).

(X,Y) est un CRF si on a la propriété Markovienne suivante :

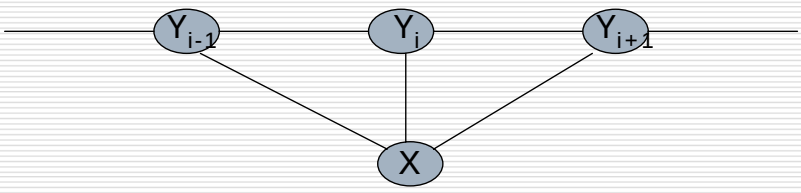
$$P(Y_i | X, Y_{V \setminus i}) \propto P(Y_i | X, Y_{\text{voisins de i dans G}})$$

Cas des séquences

X et Y deux séquences aléatoires

Â X = (X_1, …, X_n), Y = (Y_1, …, Y_n)

Â Si Y est une séquence, V = {1,…,card(V)} et E = {(y_i, y_{i+1})}



Forme des probabilités conditionnelles : théorème de Hammersley - Clifford

- Pour un CRF la probabilité conditionnelle suit un modèle log-linéaire

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{c \in C} \sum_k \varsigma_k f_k(c, y(c), x)\right)$$

$$Z(x) = \sum_y \exp\left(\sum_{c \in C} \sum_k \varsigma_k f_k(c, y(c), x)\right)$$

c : clique maximale du graphe G

C : ensemble des cliques maximales de G

x, y : séquences

Y(c) : points de G dans la clique max de y

Les f_k sont des fonctions réelles, elles sont fixées à l'avance

Les ς_k sont des paramètres à apprendre

- Pour le cas des séquences on a :

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_k \varsigma_k f_k(y_{i-1}, y_i, x) + \sum_i \sum_k \varsigma_k g_k(y_i, x)\right)$$

Apprentissage et inférence dans les CRF

- Critère d'apprentissage
  - Cas du MV : $L(\chi) = \sum_{x,y} p(x,y) \log p_\chi(y|x)$

    p(x,y) distribution empirique des exemples

    $p_\chi(y|x)$ modèle conditionnel
- Algorithmes : méthodes de gradients du 2nd ordre
- Inférence
  - Programmation dynamique

$$\hat{y} = \arg\max_y (p(y|x))$$

$$= \arg\max_y \left(\sum_i \sum_k \varsigma_k f_k(y_{i-1}, y_i, x) + \sum_i \sum_k \varsigma_k g_k(y_i, x)\right)$$

Exemple 1
Etiquetage morpho-syntaxique

- Caractéristiques
  - $F_{y,y'}$ = 1 si (y,y') est dans D (ens. d'apprentissage), 0 sinon
  - $G_{x,y}$ = 1 si (x,y) est dans D, 0 sinon
  - Autres caractéristiques
    - 1 si mot commençant par une majuscule, un chiffre
    - 1 si terminaison –ing, -ed, -ly, -ies, etc
    - On peut ajouter toutes les caractéristiques que l'on veut qui sont des fonctions des x et y.
- Vecteur global de caractéristiques
  - Vecteur binaire de grande taille et creux
    - chaque composante correspond à une des caractéristiques précédentes

Exemple 2
Extraction d'information

Titre Ecole Recherche Multimédia d'Information

Techniques & Sciences

ERMITES

-avec les soutiens du LSIS, du département d'informatique de l'UFRST USTV, et de l'Association Francophone de la Commmunication Parlée (AFCP)-

Date 4 - 6 septembre 2006

Lieu Presqu'île de Giens - Var

24 participants maximum plus les 11 tuteurs

- Il reste des places -

Web http://glotin.univ-tln.fr/ERMITES06

ermites@univ-tln.fr

Objectifs :

La recherche d'information, sur le web notamment, est de plus en plus complexe et hasardeuse compte tenu du volume sans cesse croissant des informations disponibles et de leur nature multimédia (textes, images, sons, vidéos...). L'Ecole Recherche Multimédia d'Information: Techniques & Sciences (ERMITES), organisée par l'Université du Sud Toulon-Var, le Laboratoire des Sciences de l'Information et des Systèmes (LSIS) et l'Association Française de la Communication Parlée (AFCP), regroupera dans un cadre convivial, ...

Organisateurs :

Hervé Glotin & Jacques Le Maitre

Prédiction de sorties structurées

Extension des classifieurs à marges

SVM ISO (Tsochantaridis et al. 2004)

SVM iso

Extension des méthodes à marge pour prédire des sorties structurées

- Recherche d'une fonction

  $f : X \Downarrow Y$ avec Y un espace discret structuré

- Approche

Apprentissage d'une fonction discriminante

$$F : X * Y \Downarrow R$$

Inférence : la solution pour une entrée donnée est obtenue

en maximisant F sur l'espace de sortie

$$f(x;w) \mid \arg\max_{Y} F(x,y;w)$$

Hypothèse : F linéaire pour une représentation conjointe des entrées - sorties

$$F(x,y;w) \mid \{ w, \Phi(x,y) \}$$

Apprentissage

Coût empirique

$$C(w) = \frac{1}{N}\sum_{i=1}^{N} c(y_i, f(x_i))$$

Cas du coût 0/1 en classification et pb linéairement séparable

N inégalités non linéaires :

$$\forall i = 1,..N \quad \max_{y \in Y \setminus y_i} \langle w, \Phi(x_i, y) \rangle \leq \langle w, \Phi(x_i, y_i) \rangle$$

N * card(Y) 4 N contraintes linéaires :

$$\forall i = 1,..N, \forall y \in Y \setminus y_i : \langle w, \Phi(x_i\ 4\ y) \rangle 4 \langle w, \Phi(x_i, y_i) \rangle \geq 0$$

Formulation SVM : problème d'optimisation quadratique convexe

$$QP \quad \begin{cases} \min \|w\|^2 \\ \text{sous contraintes} \quad \langle w, \iota \Phi_i(y) \rangle = \langle w, \Phi(x_i, y) \rangle 4 \langle w, \Phi(x_i, y_i) \rangle \geq 1 \end{cases}$$

Ä Principe de l'algorithme

Â Maintenir et résoudre itérativement un faible nombre de contraintes actives

$S_i = 0$ (ensemble de contraintes actives)

pour i = 1 à N

calculer $\hat{y} = \underset{Y}{\text{argmax}}(1 - \langle w, \iota \Phi_i(y) \rangle)$

Si $\hat{y}$ viole la contrainte locale, ajouter cette contrainte à S_i

résoudre le problème QP avec S_i

Condition arrêt

- Caractéristiques du problème
  - Nombre potentiellement exponentiel de contraintes linéaires
    - Ex : tous les arbres possibles en sortie
    - Problème d'optimisation complexe
    - Calcul de argmax( ) : programmation dynamique – dont la complexité limite les applications

- Extensions
  - Discrimination: cas non linéairement séparable
  - Regression
  - Fonctions de coût plus complexes mieux adaptées au cas structuré

Exemple 1: Multiclasse

- A(x) caractéristiques de x
- k classes $y_1, ..y_k$
- Θ(y) indicatrice de classe
- Coût 0/1

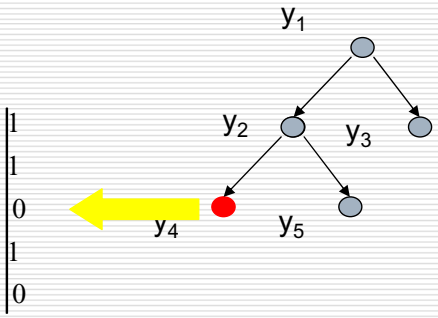
$$\Phi(x,y) \mid A(x) \cup \Theta(y)$$

$$\begin{vmatrix} x_1 \\ x_2 \end{vmatrix} \cup \begin{vmatrix} 1 \\ 0 \end{vmatrix} \mid \begin{vmatrix} x_1 * 1 \\ x_2 * 1 \\ x_1 * 0 \\ x_2 * 0 \end{vmatrix}$$

Exemple 2: Hiérarchie de classes

$$\Phi(x,y) \mid A(x) \cup \Theta(y)$$

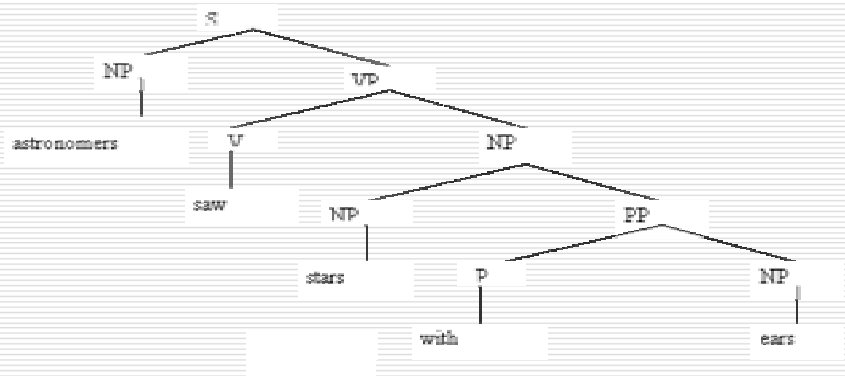
$$\Theta_z(y) \mid \begin{cases} 1 & \text{si } y \,\Omega z \\ 0 & \text{sinon} \end{cases}$$



Exemple 3 : pcfg

- PCFG définition
- quadruplet <W, N, N_1, R>
  - W = {w_1,…,w_v} est un ensemble de terminaux
  - N = {N_1,…,N_n} est un ensemble de non terminaux, N_1 est le symbole initial
  - R est un ensemble de règles de la forme $N_i \Downarrow z_j$,
    - z_j une chaîne de non terminaux et de terminaux.
    - On considère uniquement des pcfg telles que $z_j = w_j$, ou $z_j = N_k\ N_m$
  - A chaque règle est associée une probabilité $P(N_i \Downarrow z_j)$
  - Pour un ensemble R donné, on a $\sum_j P(N^i \Downarrow z^j) = 1$

- Astronomers saw stars with ears



PCFG

S ⇓ NP VP 1.0	NP ⇓ NP PP 0.4
PP ⇓ P NP 1.0	NP ⇓ astronomer 0.1
VP ⇓ V NP 0.7	NP ⇓ ears 0.18
VP ⇓ VP PP 0.3	NP ⇓ saw 0.04
P ⇓ with 1.0	NP ⇓ stars 0.18
V ⇓ saw 1.0	NP ⇓ telescopes 0.1

$$P(N^j_{k,l} \Downarrow z^m \,/\, \text{n'importe quoi hors de k},l) \mid P(N^j_{k,l} \Downarrow z^m)$$

$$P(N^j_{k,l} \Downarrow z^m \,/\, \text{n'importe quoi au dessus de } \mathrm{N}^{\mathrm{j}}_{\mathrm{k,l}} \text{ dans l'arbre}) \mid P(N^j_{k,l} \Downarrow z^m)$$

$$\sum_j P(N^i \Downarrow z^j) \mid 1$$

Codage PCFG

$\Phi =$

S ⇓ NP VP	1
PP ⇓ P NP	1
VP ⇓ V NP	1
VP ⇓ VP PP	0
P ⇓ with	1
V ⇓ saw	0
NP ⇓ NP PP	1
NP ⇓ astronomer	1
NP ⇓ ears	1
NP ⇓ saw	1
NP ⇓ stars	1
NP ⇓ telescopes	0

Modèles à variables latentes

Découverte de relations cachées entre variables d'un problème

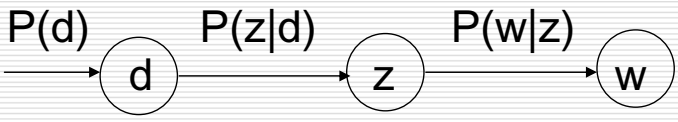
Modèles à variables latentes

- Une ou plusieurs variables latentes, inconnues, conditionnent la génération des données
- Elles représentent une « relation » entre ces données
- Exemples
  - Mélanges de densités (var.latente : classe)
  - HMM (var. latente : état)
  - Nombreux modèles latents utilisés en texte / image pour représenter et inférer des relations cachées entre variables
- Détails sur un exemple populaire : « PLSA »

PLSA (Hofmann 99)

- Modélisation stochastique de LSA -
  - Modèle à variable latente
  - Une variable latente est associée à chaque occurrence d'un mot dans un document
  - Processus génératif
    - Choisir un document *d* ,P(*d*)
    - Choisir une classe latente *z*, P(*z*|*d*)
    - Choisir un mot *w* suivant P(*w*| *z*)

Modèle PLSA



$$\begin{cases} P(d,w) = P(d) * P(w|d) \\ P(w|d) = \sum_z P(w|z)P(z|d) \end{cases}$$

- Hypothèses
  - # valeurs de z est fixé
  - Indépendence des observations (d, w), i.e. sac de mots
  - Connaissant z, w ne dépend pas de d
- Apprentissage
  - MV et EM

Applications

- Extraction de concepts
  - Z_k : concept
  - $P(w_i|z_k)$ représentation du concept z_k
  - $P(z_k|d_i)$ importance du concept dans le document
  - Un concept sera commun à plusieurs mots
  - Un même mot peut être associé à différents concepts

Applications (autres)

- Segmentation thématique
- Construction de hiérarchies de documents (# modèles plsa hiérarchiques)
- Recherche d'information
- Annotation d'images
  - Pour une image inconnue : P(w|image)

Bibliographie

- Denoyer L., Gallinari P., 2005, Bayesian Network Model For Semi-Structured Document Classification, Information Processing and Management. Volume 40 , Issue 5 (September 2004)- 807 - 827
- Denoyer L., Gallinari P., Vittaut JN, Brunessaux S., Brunessaux S., (2003) Structured multimedia document classification DOCENG 2003
- Denoyer (Ludovic), Gallinari (Patrick), 2004, Document Structure Matching for heterogeneous corpora, In XML and Information Retrieval workshop of SIGIR 2004
- Hofmann T., Probabilistic latent semantic indexing, SIGIR 1999
- Lafferty J., McCallum A., Pereira F., 2001, Conditional random fields : probabilistic models for segmenting and labeling sequence data, ICML 2001
- Tsochantaridis, I, Hofmann T., Joachims T., Altun Y., Support vector machine learning for independent and structured output spaces, ICML 2004
- Wisniewski (Guillaume), Denoyer (Ludovic), Gallinari (Patrick), 2005, Restructuration automatique de documents dans les corpus semi structurés hétérogènes, In EGC'2005

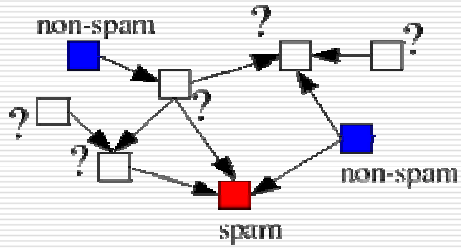
XML Document Mining Challenge 2006

- Challenge
  - INEX-Delos and Pascal networks of excellence
- Three tasks
  - Classification
  - Clustering
  - Document mapping
- 3 XML corpora
  - IEEE collection
  - IMDB (Movie descriptions)
  - Wikipedia in 4 languages
- Web site : http://xmlmining.lip6.fr
- Email : xmlmining@lip6.fr

Graph Labelling and Web Spam Challenge (to be announced)

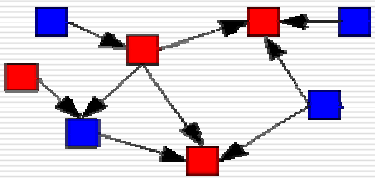
- 2 collections
- LIP6 corpus: connected graph of 5,000 Web pages labelled at the page level.

entrée



- WEBSPAM-UK-2006 compiled by University of Rome ``La Sapienza'' and University of Milan, and hosted by Yahoo! Research Barcelona. 18 million pages from 12,000 hosts, annotated at the level of hosts.

sortie



Notes

Notes

Neurophysiologie de la vision: des traits au sens

Pascale Giraudet
Univ. du Sud Toulon Var
giraudet@univ-tln.fr

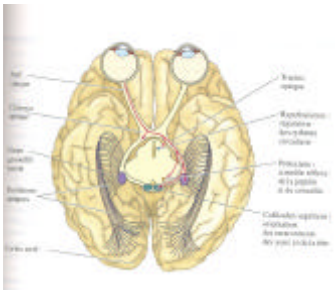
Neurophysiologie de la vision: des traits au sens

- Vision = sens le plus développé chez l'homme
- Le capteur (œil) est sensible à des intensités lumineuses très variables
- Le système nerveux central obtient une information très rapide de position, taille, couleur, texture des objets, ainsi que de direction et de vitesse des mouvements
- Comment l'information visuelle est-elle décomposée en traits ?
- Comment le SNC regroupe-t-il ces traits pour donner du sens à l'information perceptive ?

Neurophysiologie de la vision: des traits au sens

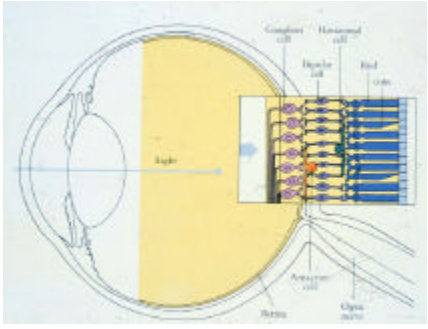
Plan de l'exposé

- L'œil: réception et prétraitement de l'information visuelle
- Le cortex visuel: extraction de traits
- Le cortex associatif: émergence du sens
- Conclusion



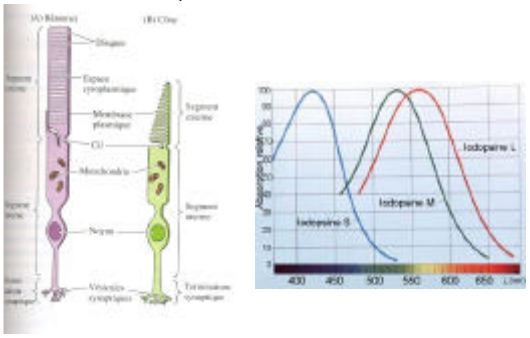
1. L'œil: réception et prétraitement de l'information visuelle

1.1. Anatomie de l' œil



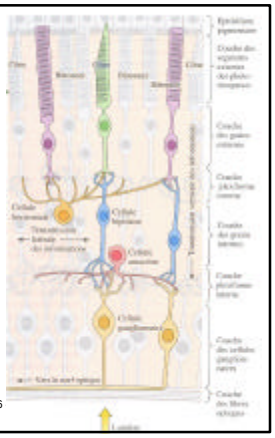
1. L'œil: réception et prétraitement de l'information visuelle

1.2. Réception de l'information visuelle

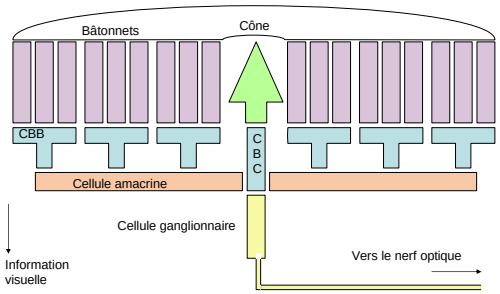


1. L'œil: réception et prétraitement de l'information visuelle

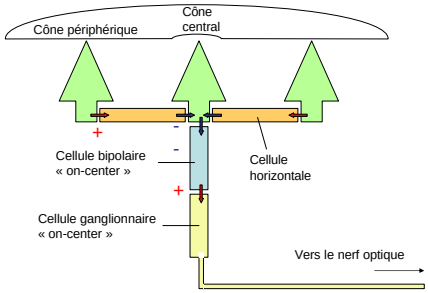
1.3. Connectique de la rétine



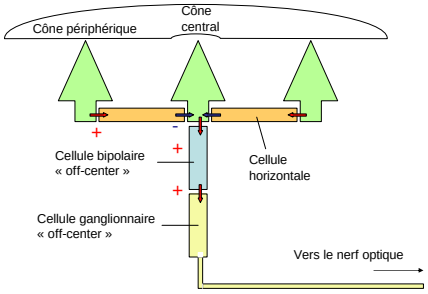
1. L'œil: réception et prétraitement de l'information visuelle
1.4. Sensibilité vs acuité



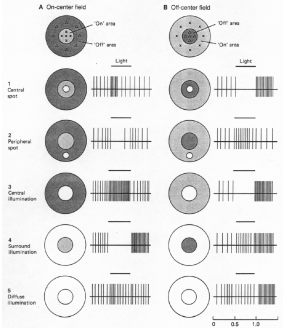
1. L'œil: réception et prétraitement de l'information visuelle
1.5. Champs récepteurs des cellules ganglionnaires



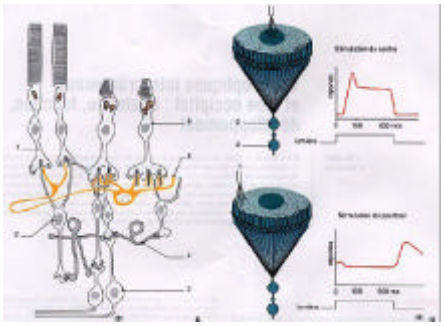
1. L'œil: réception et prétraitement de l'information visuelle
1.5. Champs récepteurs des cellules ganglionnaires



1. L'œil: réception et prétraitement de l'information visuelle
1.6. Extraction des contrastes spatiaux



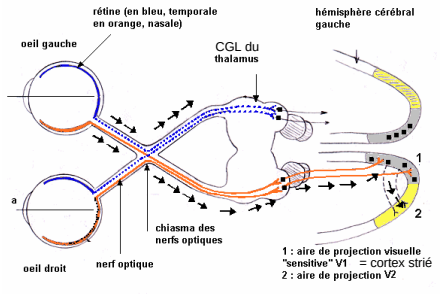
1. L'œil: réception et prétraitement de l'information visuelle
1.7. Extraction des contrastes temporels



1. L'œil: réception et prétraitement de l'information visuelle
1.8. Bilan: l'information visuelle à la sortie de la rétine

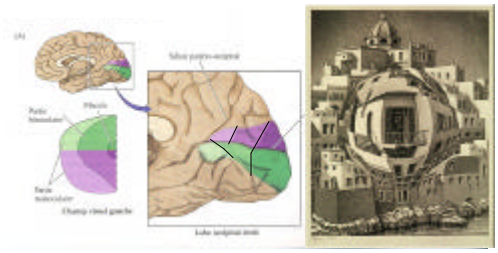
- Type d'information captée : luminosité ou couleur RGB pour chaque zone du champ visuel (± grande selon position et luminosité)
- Mode de codage : impulsionnel, fréquence fonction du contraste pour chaque cellule
- Type de connectique : inhibitions latérales locales
- Prétraitement : amplification des contrastes spatiaux et temporels ? accentuation des aspects les plus informatifs

2. Le cortex visuel: extraction de traits
2.1. La voie visuelle rétino-géniculo-striée



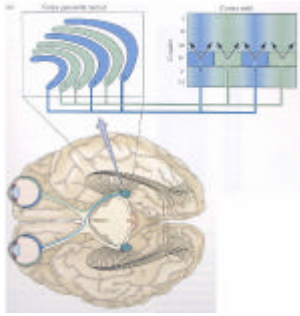
2. Le cortex visuel: extraction de traits
2.2. La rétinotopie et l'information de position

- Rétinotopie conservée à tous les niveaux du système visuel
- Magnification de la région fovéale

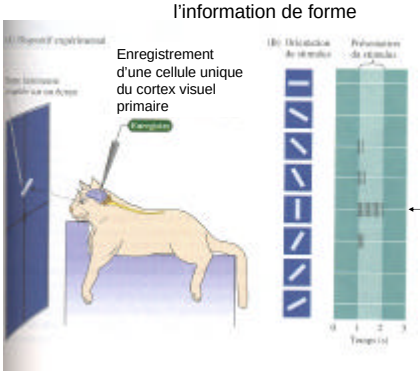


2. Le cortex visuel: extraction de traits
2.3. La stéréopsie et l'information de profondeur

- Rétine : information monoculaire
- Chiasma optique: croisement de l'information de la rétine nasale de chaque oeil
- CGL: neurones monoculaires mais alternance œil droit / œil gauche dans chaque hémisphère
- Cortex visuel primaire (entre les colonnes de dominance oculaire) et surtout secondaire: neurones binoculaires dont les champs récepteurs monoculaires sont ± décalés ? cellules "près" ou "loin" ? information de profondeur



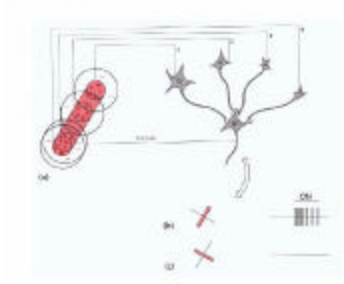
2. Le cortex visuel: extraction de traits
2.4. L'organisation en colonnes du cortex strié et l'information de forme



- Cellules gangl. et du CGL : sensibilité concentrique on/off
- Cellules du cortex strié : réponses à des bords orientés

Orientation préférentielle du neurone enregistré

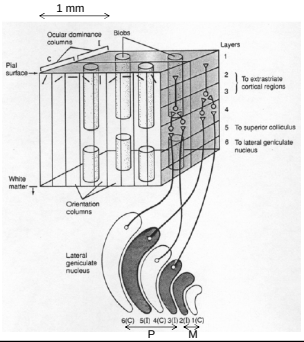
2. Le cortex visuel: extraction de traits
2.4. L'organisation en colonnes du cortex strié et l'information de forme



- Convergence de plusieurs champs récepteurs concentriques on/off ? réponse maximale pour :
- une orientation,
- une longueur donnée
? trait de forme

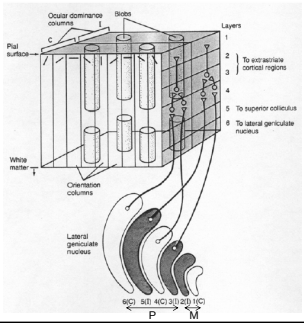
2. Le cortex visuel: extraction de traits
2.4. L'organisation en colonnes du cortex strié et l'information de forme

- Rétine et CGL: cellules aux champs récepteurs concentriques on/off
- Couches 3 à 6 (voie P): précision spatiale ? info utile aux traits de forme
- Cortex visuel primaire: neurones sensibles aux orientations et longueurs des stimuli, avec organisation en colonnes
- Cortex visuels secondaires: neurones sensibles à des formes + complexes



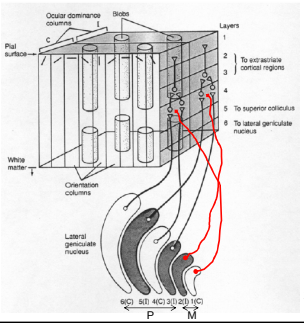
2. Le cortex visuel: extraction de traits
2.5. L'organisation en colonnes du cortex strié et l'information de couleur

- Couches 3 à 6 (voie P): cellules ganglionnaires connectées à des cônes centraux et périphériques sensibles à des longueurs d'onde différentes (? voie M) ? information de couleur
- Cortex visuel primaire (blobs): neurones sensibles aux couleurs
- Cortex visuel secondaire: neurones de V4 sensibles à la couleur d'un stimulus quel que soit son déplacement



2. Le cortex visuel: extraction de traits
2.6. La voie magno-cellulaire et l'information de mouvement

- Couches 1 à 2 (voie M): précision temporelle ? information de mouvement
- Cortex visuel primaire: neurones sensibles aux mouvements
- Cortex visuel secondaire: neurones de V5 sensibles à une direction donnée de déplacement d'un stimulus quelle que soit sa couleur



2. Le cortex visuel: extraction de traits
2.7. Bilan: l'organisation fonctionnelle du cortex visuel

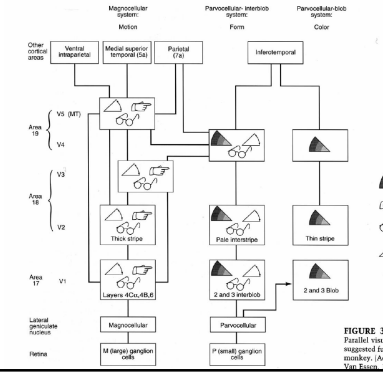
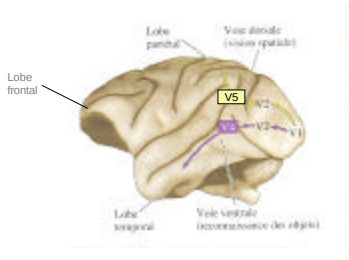


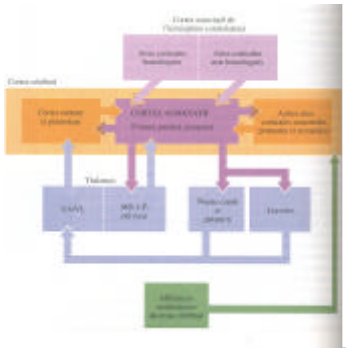
FIGURE 30–9
Parallel visual pathways and their suggested functions in the macaque monkey. (Adapted from DeYoe and Van Essen, 1988.)

Lésions cortex extrastrié?
Perte fonctionnelle spécifique à la perception d'un trait
Ex : mvts,couleurs

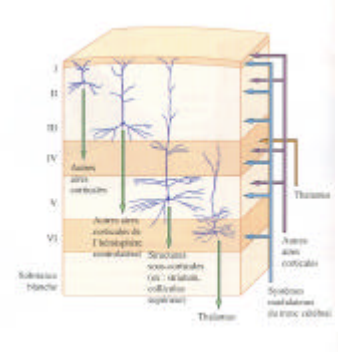
3. Le cortex associatif: émergence du sens
3.1. Présentation du cortex associatif



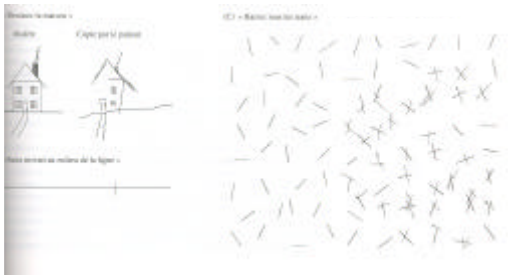
3. Le cortex associatif: émergence du sens
3.1. Présentation du cortex associatif



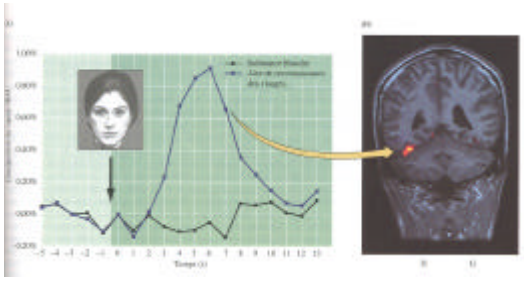
3. Le cortex associatif: émergence du sens
3.2. Connectique du cortex associatif



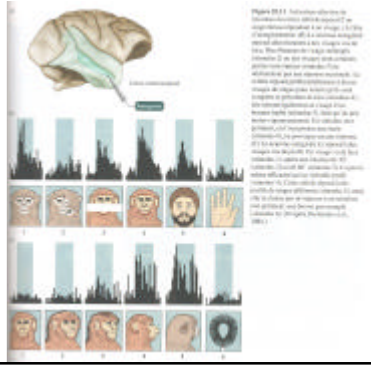
3. Le cortex associatif: émergence du sens
3.3. Le lobe pariétal, l'attention et la représentation spatiale



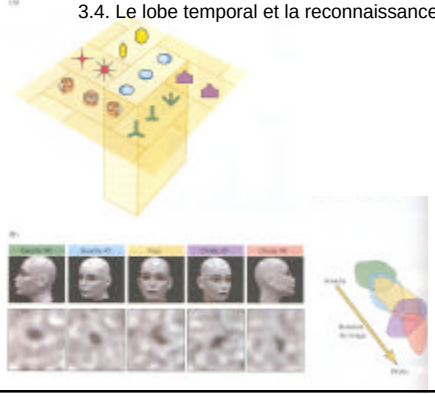
3. Le cortex associatif: émergence du sens
3.4. Le lobe temporal et la reconnaissance



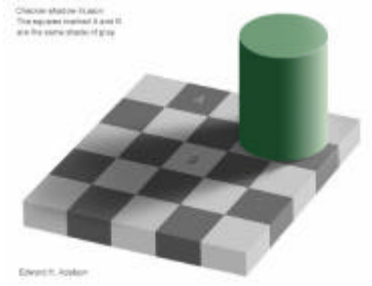
3. Le cortex associatif: émergence du sens
3.4. Le lobe temporal et la reconnaissance



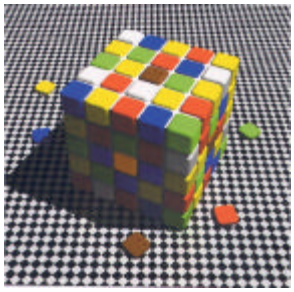
3. Le cortex associatif: émergence du sens
3.4. Le lobe temporal et la reconnaissance



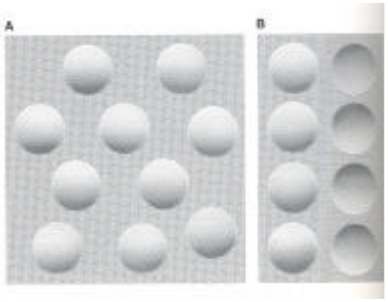
3. Le cortex associatif: émergence du sens
3.5. L'utilisation du contexte dans l'interprétation



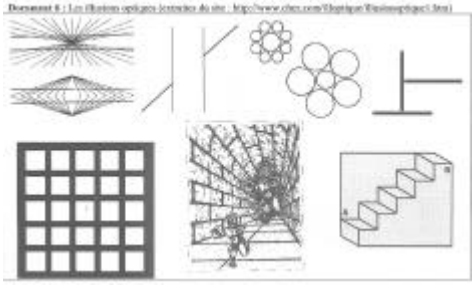
3. Le cortex associatif: émergence du sens
3.5. L'utilisation du contexte dans l'interprétation



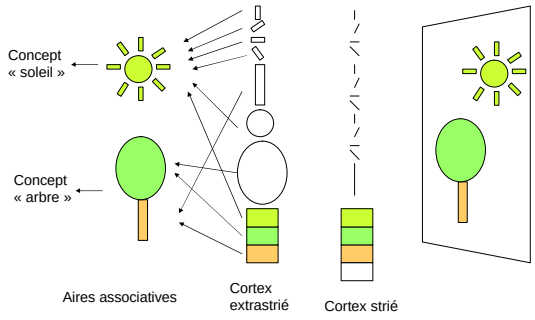
3. Le cortex associatif: émergence du sens
3.5. L'utilisation de la connaissance dans l'interprétation



3. Le cortex associatif: émergence du sens
3.5. Autres exemples

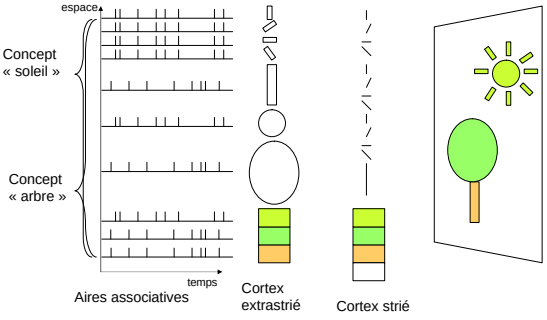


3. Le cortex associatif: émergence du sens
3.6. Comment un stimulus visuel fait-il sens ?



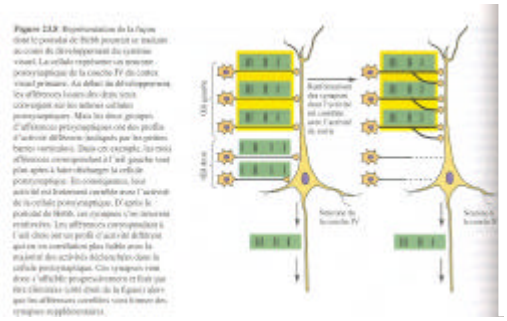
Théorie du neurone « grand-mère »

3. Le cortex associatif: émergence du sens
3.6. Comment un stimulus visuel fait-il sens ?



Sens = activation d'un réseau synchro.

3. Le cortex associatif: émergence du sens
3.7. Apprendre le sens



Conclusion

• Stimulus visuel décomposé et transmis au SNC sous forme de traits (forme, couleur, mouvement en particulier) de plus en plus complexes, la position restant majoritairement rétinotopique

• Recomposition du sens au niveau des aires corticales associatives du SNC peut-être par activation de réseaux de neurones synchronisés

• Activation corticale qui « fait sens » s'étendant sans doute à :
- différentes aires sensitives qui permettent de se représenter le concept,
- une zone du cortex associatif pariétal qui permet une représentation spatiale de l'objet,
- une zone du cortex associatif inféro-temporal qui permet une reconnaissance consciente,
- ainsi qu'aux aires du langage qui permettent de nommer et sans doute manipuler les concepts.

• Transmission de l'information vers le cortex frontal pour réflexion et prise de décision et/ou les cortex prémoteur et moteur pour réponse à la stimulation

Notes

Notes

Indexation et Recherche Sémantique d'images

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

M r i M

ERMITES

septembre 2006

P. Mulhem

Philippe.Mulhem@imag.fr

But

- Sensibiliser sur les problèmes complexes liés à la Recherche d'Information images en utilisant des descriptions symboliques.
- Faire un survol des approches et travaux dans ce domaine.
- Esquisser le futur de cette problématique.

Plan

1. Introduction
2. Spécificités par rapport à la RI textuelle
3. RI d'images web par le contexte
4. RI d'images par le contenu « brut »
5. RI d'images par le contenu sémantique
6. Pondération
7. Conclusion
8. Bibliographie

1. Introduction

- Qu'est qu'une image?
  - TLFI (atilf.atilf.fr/tlf.htm):
    - « Représentation (ou réplique) perceptible d'un être ou d'une chose.
      - La relation entre l'objet et son image est de nature physique plus précisément optique ou physico-chimique, notamment dans les techniques de photographie

      …»

1. Introduction

- Pourquoi de la RI d'images fixes?
  - L'exemple des photographies[1]
    - Getty Images : 70 millions de photographies
    - Corbis Images : 65 millions de photographies
    - Photos personnelles, estimation : 750 milliards
  - Le besoin de retrouver des photographies est donc évident.

1. www.sims.berkeley.edu/research/projects/how-much-info/index.html

1. Introduction

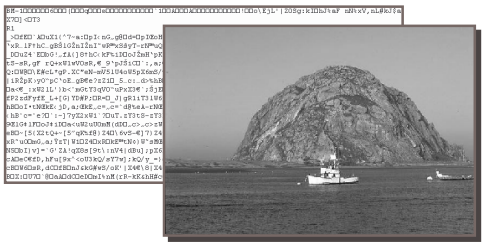
- Multiples sortes d'images pour multiples usages
  - Diagnostic médical
  - Analyse scientifique
  - Stockage pour vente
  - Se remémorer des souvenirs (photos personnelles) et les montrer

1. Introduction

- L'utilisation de systèmes d'indexation et de recherche d'images fixes est donc un besoin crucial pour pratiquement tout le monde : du contexte professionnel au contexte personnel.
- … mais comment faire???

2. Spécificités / RI textuelle

- Mettons-nous à la place d'un ordinateur, voilà une image :

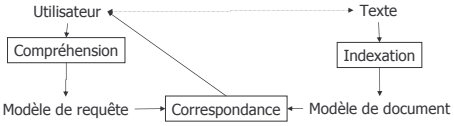


2. Spécificités / RI textuelle

- Le média image (photographies par ex.)
  - N'est pas basé sur un langage
  - Interprétation compliquée pour un humain
    - différents niveaux : émotions, objets, actions
  - Interprétation encore plus compliquée pour un ordinateur…

2. Spécificités / RI textuelle

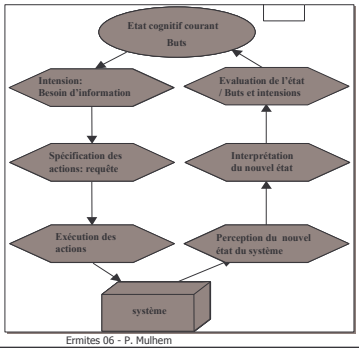
- Rappel d'un système de recherche d'information (SRI) textuel



- Problèmes d'un SRI image
  - Comment indexer
  - Comment rechercher

2. Spécificités / RI textuelle

- Cycle de Norman sur une interaction de recherche



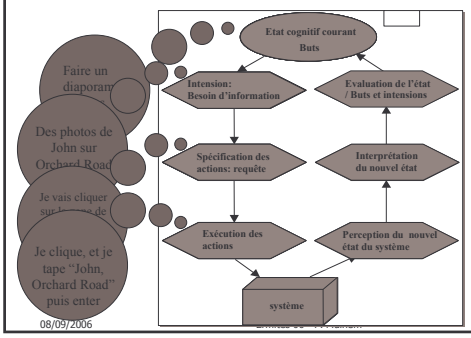
[D. Norman, Cognitive Engineering, Chapter 3. User Centered System Design, New Perspectives on Human Computer Interqction, Hillsdqle, pp. 31-61, 1986]

2. Spécificités / RI textuelle

- Par rapport à cette interaction de recherche, le niveau de représentation de l'index des images a un impact important
  - Index non-symbolique
    - Gouffre entre le niveau de représentation des deux interacteurs
  - Index symbolique
    - Le symbole est manipulable par les deux interacteurs

2. Spécificités / RI textuelle

- Indexation Symbolique



2. Spécificités / RI textuelle

- L'indexation symbolique est donc ce qui semble la « bonne » voie, mais
  - Elle est compliquée
  - Elle n'est pas générale
  - Elle n'est pas totalement automatique
- Ce qui a amené à proposer des approches non-symboliques (voir partie 4)

2. Spécificités / RI textuelle

- [Smeulder et al. 2000]
  - Paramètres à considérer

	contexte réduit	contexte large
Variance	faible	grande
Connaissance	spécifique	générique
Base établie	probable	peu probable
Application	professionnelle	personnelle

3. RI image WEB par le contexte

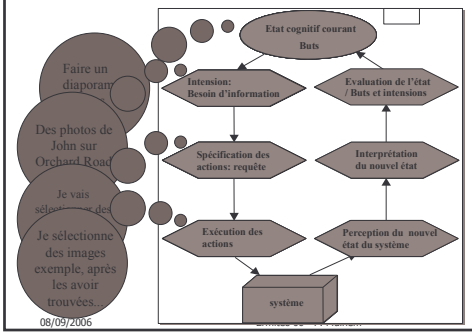
- Ce cas est lié à l'utilisation du contexte d'apparition des images pour les indexer
  - Google (www.google.com)
    - « Pour déterminer le contenu graphique d'une image, Google analyse le texte de la page qui entoure l'image, le titre de l'image et de nombreux autres critères »
  - Altavista (www.altavista.com)
    - « Lorsque vous saisissez une requête et que vous cliquez sur le bouton Chercher, AltaVista récupère les images contenant les mots cherchés dans le nom du fichier, le texte alternatif ALT, le texte proche et/ou les métatags de la page. »
  - C'est de la recherche de texte "déguisée"
    - Avantages
      - Transparent
      - Général
    - Inconvénients
      - Très sensible à la qualité du contexte
      - Pas d'utilisation du contenu des images

4. RI d'images par le contenu « brut »

- Idée : se servir du contenu des images pour les indexer et les retrouver
  - Similaire aux approches sur le texte
- Problème
  - Quel niveau de description utiliser?
- Une solution
  - Niveau de description proche du signal
    - Couleurs, textures, formes, positions

4. RI d'images par le contenu « brut »

- Indexation non-symbolique



L'utilisateur se met au niveau du système

4. RI d'images par le contenu « brut »

- Représenter les couleurs, textures, formes présentes dans des images ou des régions en gardant en tête les deux facettes indissociables
  - Indexation : compact et fidèle
  - Recherche : bonne précision et bon rappel
  - Passage à l'échelle?
    - Oui en terme d'applications
    - Difficulté en terme d'espace de caractéristiques

4. RI d'images par le contenu « brut » - Couleurs -

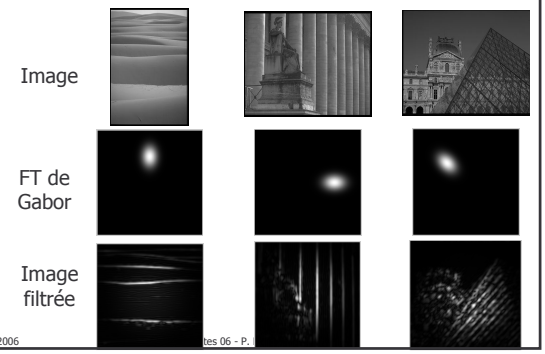
- Une fois un espace de couleurs choisi (RGB, HSV, …), comment représenter les couleurs d'une région ou d'une image?
  - Le principe le plus utilisé est l'histogramme (vecteur)
  - Mais combien de dimensions, et correspondant à quelles couleurs?
- Fonctions de correspondance
  - Normes L_1 ou L_2 ne donnent pas de très bons résultats
  - Intersection d'histogrammes
  - Distance Euclidienne pondérée (QBIC)
  - …

4. RI d'images par le contenu « brut » - Textures -

- Texture :
  - "Disposition et entrelacement des fibres, des éléments constitutifs du tissu organique"
  - Joue un rôle dans la perception de l'orientation et de la profondeur spatiale entre des objets se recouvrant.

4. RI d'images par le contenu « brut » - Textures -

- Filtres de Gabor – Illustration (de Anne Guérin-Dugué)

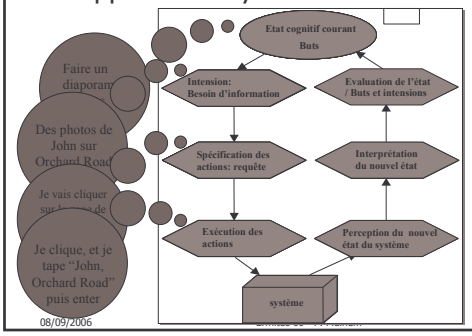


4. RI d'images par le contenu « brut » - Textures -

- Filtres de Gabor
  - On calcule ensuite l'énergie (somme des carrés) de chaque image filtrée
  - On utilise une banque de filtres pour détecter différentes textures
    - E échelles (exemple 4)
    - O orientations (exemple 6)
  - Histogrammes d'énergies

5. RI d'im. par contenu sémantique

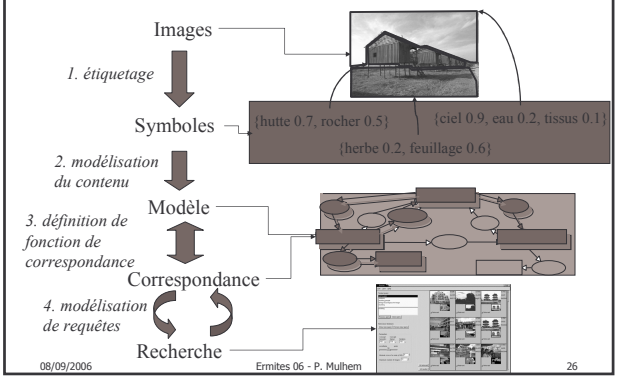
- Rappel sur le cycle de Norman



5. RI d'images par contenu sémantique

- En fait, on est actuellement plus sur de la recherche d'images symbolique que sémantique, car on ne fait pas vraiment appel à de la sémantique…

5. RI d'im. par contenu sémantique



5. RI d'im. par contenu sémantique

- Le système de recherche d'information doit se mettre au niveau symbolique de représentation de l'utilisateur pour la recherche.
- Niveau Processus de Recherche ?
  - oui
- Niveau Processus d'Indexation ?
  - oui

5. RI d'im. par contenu sémantique

- Indexation/Recherche
  - Images globales
    - Ladret & Guérin 2001: KNN sur histogrammes d'orientation à partir de coefficents DCT d'images JPEG
    - Wang & Li 2002 : Modèles de Markov Multi-échelles
    - Bradshaw 2000 : Etiquetage probabiliste multi-échelle
    - Jeon, Lavrenko et Manmatha 2003 : Etiquetage probabiliste
    - La Cascia et al. 1998 : Approche LSA
    - Tollari 2005 : classification ascendante hiérarchique
    - Glotin et al. 2006 : sélection de caractéristiques pertinentes par symbole
  - Régions
    - Rowe & Frew 1997 : réseaux de neurones
    - Town & Sinclair 2000 : réseaux de neurones
    - Lim 2000: réseaux de neurones, Machines à support de vecteurs

5. RI d'im. par contenu sémantique

- Etiqueter des images globales
  - Jeon, Lavrenko et Manmatha 2003
    - Idée : Modéliser par des probabilité les associations entre termes et « blobs » (régions) d'images pour annoter automatiquement de nouvelles images
    - Principe :
      - Pour une image I, on évalue la probabilité que I soit indexée par w_t en utilisant la formule

$$P(w_t / I) \approx P(w_t / b_1 b_2 ... b_m)$$

avec

$$P(w_t / b_1 b_2 ... b_m) = \frac{P(w_t, b_1 b_2 ... b_m)}{\sum_{v \in V} P(v, b_1 b_2 ... b_m)}$$

5. RI d'im. par contenu sémantique

- Etiqueter des images globales
  - Jeon, Lavrenko et Manmatha 2003 (2)
    - Principe
      - Utiliser une collection d'apprentissage T composés de |T| images, telles que
        - $J=\{b_1, ..., b_m; w_1, ..., w_n\}$
        - Avec

          $\{b_1, ..., b_m\}$ ensemble de blobs extraits

          $\{w_1, ..., w_n\}$ ensemble de termes qui annotent l'image
      - Pour un mot w_t, on calcule la distribution conjointe :

$$P(w_t, b_1, b_2, ..., b_m) = \sum_{J \in T} P(J)\, P(w_t, b_1, b_2, ..., b_m / J)$$

5. RI d'im. par contenu sémantique

- Etiqueter des images globales
  - Jeon, Lavrenko et Manmatha 2003 (3)
    - Principe
      - Si on suppose que les événements w et b_i sont mutuellement indépendants pour une image J

$$P(w,b_1,b_2,...,b_m)=\sum_{J\in T}(P(J)P(w/J)\prod_{i=1}^{m}P(b_i/J))$$

      - P(J) est calculé simplement par 1/|T|
      - P(w/J) et P(b_i/J) par maximisation de vraisemblance lissée:

$$P(w/J)=(1-\alpha_J)\frac{|(w,J)|}{|J|}+\alpha_J\frac{|(w,T)|}{|T|}\qquad P(b_i/J)=(1-\beta_J)\frac{|(b_i,J)|}{|J|}+\beta_J\frac{|(b_i,T)|}{|T|}$$

5. RI d'im. par contenu sémantique

- Etiqueter des images globales
  - Jeon, Lavrenko et Manmatha 2003 (4)
    - Recherche :
      - Pour une requête symbolique Q = {w_1,..,w_k}

$$P(Q/I)=\prod_{j=1}^{k}P(w_j/I)$$

    - Expérimentations
      - Sur 5000 images, caractéristiques (couleur, orientation, taille, convexité, …), 371 termes, avec α_J=0.1 et β_J=0.9

Q. Length (in words)	1	2	3	4
Av. Prec.	0.1501	0.1419	0.1730	0.2364

5. RI d'im. par contenu sémantique

- Etiqueter des images globales [Tollari 2005]
  - Idée : Définition de classes visuelles (hyper-rectangle dans l'espace de caractéristiques) associées à un ou plusieurs symboles.
  - Modéliser des associations entre blobs et symboles avec classification ascendante hiérarchique : apprentissage
    - Pour un symbole, on prend tous les blobs associés
    - On part de singletons (clusters avec un blob)
    - On regroupe itérativement les clusters les plus proches (distance min)
    - Chaque classe est décrite par son cdg et son vecteur d'écarts-type par dimension.

5. RI d'im. par contenu sémantique

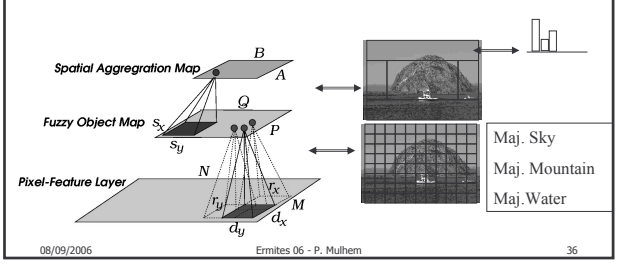
- Etiqueter des images globales [Tollari 2005]
  - Choix du meilleur regroupement par évaluation du clustering hérarchique
    - On définit l'appartenance d'un blob à une classe si pour chaque dimension de l'espace les caractéristiques du blob sont à moins de 2 fois l'écart type pour toutes les dimensions.
    - Évaluation par Normalized Score :
      - NS = right/n – wrong / (N-n)
  - Expérimentations
    - 1000 images (apprentissage 7000, test 2900)
    - Blobs avec 40 dimensions (6 forme, 18 couleur, 16 texture)
    - 31 mots avec plus de 150 images
    - NS entre 0.05 (people) et 0.4 (cat) par symbole
    - Test sur le filtrage de symboles préliminaire.

5. RI d'im. par contenu sémantique

- Etiqueter des régions - Visual Keywords [Lim 2000]
  - Apprentissage sur des bloc exemple
    - couleurs : espace YIQ
      - Moyenne et écart-type des caractéristiques
      - $x^c=(\mu_Y, \sigma_Y, \mu_I, \sigma_I, \mu_Q, \sigma_Q)$
    - Textures filtres de Gabor 5 échelles x 6 directions
      - Moyenne et écart-type des caractéristiques
      - $x^t=(\mu_{1,1}, \sigma_{1,1}, \mu_{1,2}, \sigma_{1,2}, \ldots \mu_{5,6}, \sigma_{5,6})$
    - Utilisation de techniques d'apprentissage pour caractériser les étiquettes
      - Machines à Support de Vecteurs

5. RI d'im. par contenu sémantique

- Etiqueter des régions - Visual keywords
  - Indexation sur des blocs
  - Étiquetage par SVM et softmax ($e^{svm\_i(X)}/\sum_k e^{svm\_k(X)}$).



5. RI d'im. par contenu sémantique

- Etiqueter des régions - Visual keywords
  - Termes

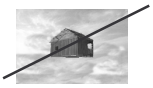
  - Recherche
    - Regroupement pondéré avec poids ω

    - Correspondance entre requête (image x) et une image y (distance "city-block")

$$Match_{\omega}(x,y)=\frac{\sum_{(a,b)}\omega(a,b)\lambda(a,b)}{\sum_{(a,b)}\omega(a,b)} \qquad \lambda(a,b)=1-\frac{1}{2}\left|SAM_x(a,b)-SAM_y(a,b)\right|$$

5. RI d'im. par contenu sémantique

- Étiquetage avec contexte [Mulhem 2002]
  - Constat
    - [A. Clark, T. Trscianko, N. Campbell, B. Thomas : A comparison between human and machine labelling of i,mage regions, Perception, Vol. 29, 2000, pp. 1127-1138] : les humains utilisent le contexte pour reconnaître les objets dans des photographies :



5. RI d'im. par contenu sémantique

- Étiquetage avec contexte
  - Utiliser le contexte "statistique" des photographies personnelles pour l'étiquetage.
    - "ciel entre deux eaux et en bas de l'image peu probable"
    - "ciel en haut de l'image très probable"
  - 2 étapes

1. Étiquetage initial

Image segmentée

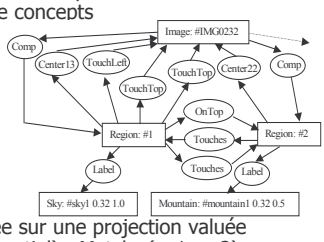
ciel 0.49, eau 0.44
0.84 0.44

ciel 0.59, eau 0.57
0.59 0.72

2. Mise-à-jour

5. RI d'im. par contenu sémantique

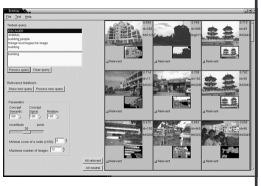
- Indexation/Recherche par graphes conceptuels étendus
  - Utilise le meilleur label de VK par bloc + relations + hiérarchie de types de concepts



  - Correspondance basée sur une projection valuée (subgraph matching partiel) : $Match_{cg}(cg1, cg2)$

5. RI d'im. par contenu sémantique

- Graphes conceptuels étendus
  - Recherche
  - 2 modes :
    - Requête par texte
      - "Je voudrais une image avec un arbre à droite d'un bâtiment, si possible avec du ciel"
      - Transformation en graphe requête
    - Requêtes par l'exemple (QBE)
      - Sélection de N images => fusion des N images
      - Génération du "Plus Petit Graphe Composé Généralisant"
        - Abstraction des graphes des N images
        - Types de concepts et relations complexes
        - Calcul statistique d'importance des concepts et relations



6. Pondération

6. Pondération

- Les SRI d'images basés sur le signal ne prennent pas en compte ce qui est important pour les personnes qui recherchent des images
  - Distances entre histogrammes (QBIC, NETRA)
  - Question : est-ce que les distances entre les petits "bins" d'histogrammes sont aussi importantes que les distances sur les "bins" avec de grandes valeurs?

6. Pondération

- Hypothèses sur les objets images [Martinet et al. 2005]
  - Objet Image (IO):
    - Projection 2D d'un objet physique réel
    - Relié à seulement un label d'un langage d'indexation (plusieurs IO peuvent être liés au même label).
  - Importance d'un IO
    - L' importance d'un IO io d'une image I étiqueté par un terme t est directement relié au fait que l'image est au sujet de t.

6. Pondération - hypothèses

- Hypothèse de taille
  - L'importance d'un IO varie de la même manière que sa taille S.

- Hypothèse de position
  - L'importance d'un IO est maximale quand sa position P est au centre de l'image et décroit quand sa distance du centre augmente.

6. Pondération - hypothèses

- Hypothèse de fragmentation
  - L'importance d'un IO est maximale quand il n'est pas fragmenté et décroit quand sa fragmentation F augmente.

- Hypothèse d'homogénéité
  - L'importance d'un IO varie de la même manière que l'homogénéité H de l'image le contenant.

6. Pondération

- Conclusion
  - Les hypothèse de taille, de position et d'homogénéité sont validées expérimentalement
  - L'hypothèse d'agrégation n'est pas valide
  - Renforcement des critères
  - D'autres critères sont à évaluer :
    - Visibilité, utilisation de régions d'intérêt

7. Conclusion

- Pas d'approches globales validées
  - Pb de passage à l'échelle
    - Termes
    - Généralisation
    - Taille de collections
    - Recherche
- Choix du vocabulaire
  - Aller vers des représentations sémantiques
- Indexation d'image en contexte
  - Contexte multimédia
  - Contexte de connaissance (sémantique)
    - "Common sense knowledge" (openmind initiative)…

Bibliographie

- Général
  - A. Smeulder, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence (2000), vol 22, N. 12, 1349-1380.
  - A. del Bimbo, Visual Infrmation retrueval, Morgan Kaufman, 1999.
- QBIC
  - C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, D. Petkovic, R. Barber Efficient and Effective Querying by Image Content, Journal of Intelligent Information Systems (1994),3, 231-262.
- NETRA
  - Ma, W. Y., & Manjunath, B. S. (1997). *NETRA: A Toolbox for navigating large image databases*. IEEE International Conference on Image Processing (ICIP'97), USA, Vol. I, 568-571
- VisualSeek
  - VisualSeek: a fully auotmated content-based image query system, ACM Multmedia, Boston, USA, 87-98, 1998.
- Blobworld
  - Chad Carson and Megan Thomas and Serge Belongie and Joseph M. Hellerstein and Jitendra Malik, Blobworld: A system for region-based image indexing and retrieval, Third International Conference on Visual Information Systems, 1999.

Bibliographie

- Textures
  - H. Tamura, S. Mori, and T. Yamawaki, Textural features corresponding to visual perception, IEEE Trans. on Sys., Man and Cyb., SMC - 8, no. 6, pp. 460 - 472, 1978.
  - Ma, W. Y., & Manjunath, B. S. (1996). Texture Features and Learning Similarity. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, 425-430.
- Contours
  - F. Mokhtarian, S. Abbasi and J. Kittler, Efficient and Robust retrieval by Shape Content through Curvature Scale Space, Proc. International Workshop on Image Databases and MultiMedia Search, pages 35--42, Amsterdam, The Netherlands, 1996
- Relations
  - 4-intersection model
    - http://www.spatial.maine.edu/~max/pointset.pdf
  - Distances entre relations
    - http://www.spatial.maine.edu/~max/pisa.pdf
  - 2D-Strings
    - S.-K Chang, Q.Y. Shi, and C.Y. Yan. Iconic indexing by 2-d strings. IEEE Transactions on Pattern Analysis and Machine Intelligence, 9(3):413-428, May 1987.

Bibliographie

- Symboles
  - Town, C., & Sinclair, D. (2000). Content-based image retrieval using semantic visual categories. Technical report 2000.14, AT&T Laboratories Cambridge, UK
  - Ladret & Guérin, Categorisation and retrieval of Scene Photographs from a JPEG Compressed Database, Pattern Analysis and Applications (2001)4 : 185-199
  - J. Wang & J. Li, Learning-based Indexing of Pictures with 2D-MHMMs, ACM Multimedia, Juan les Pins; France, 2002, pp. 436-445.
  - B. Bradshaw, Semantic Based Image Retriveval: A Probabilistic Approach, ACM Multimedia, Los Angeles, USA, 2000, pp. 167-176.
  - N. Rowe & B. Frew, Automatic Classification of Objects in Captioned Depictice Photographs for Retrieval, Chapter 4 of Itelligent Multimedia Information Retrieval, AAAI Press/The MIT Press, 1997.
  - Lim, J.H., Visual keywords: From text IR to multimedia IR. In F.Crestani & G.Pasi (ed.), Soft Computing in Information Retrieval: Techniques and Applications, Physica-Verlag, Springer Verlag, Germany, 2000, pp. 77-101.
  - P. Mulhem and J.-H. Lim, Symbolic Photograph Content-Based Retrieval, ACM CIKM 2002, MC Lean, USA, 92-101.
  - P. Mulhem, Indexation par apprentissage local et global pour la recherche symbolique d'images fixes, Revue Ingénierie des Systèmes d'Information, Numéro spécial Recherche et Filtrage d'Information, éditions Hermès, Vol. 7, No. 1-2, 2002, pp. 183-205.

Bibliographie

- Symboles
  - J. Jeon, V. Lavrenko, R. Manmatha Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, ACM SIGIR 2003, pp. 119-126.
  - S. Tollari, Filtrage de l'indexation textuelle d'une image au moyen du contenu visuel pour un moteur de recherche d'images sur le web, CORIA'05, pages 261-275, Grenoble, mars 2005
  - H. Glotin, S. Tollari, P. Giraudet, Approximation of Linear Discriminant Analysis for Word Dependent Visual Features Selection, ACIVS2005, LNCS 3708, Springer, pages 170-177, Antwerp, Belgium, September 2005
  - M. La Cascia, S. Sethi, and S. Sclaroff, "Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web", Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, June 1998, Santa Barbara, CA, USA.

Notes

Recherche ‘robuste’ d’informations dans des scènes audiovisuelles

ou

“ cibler les informations pertinentes dans du bruit ”

Hervé Glotin

Laboratoire Sciences de l’Information & des Systèmes (LSIS)
Université du Sud Toulon Var

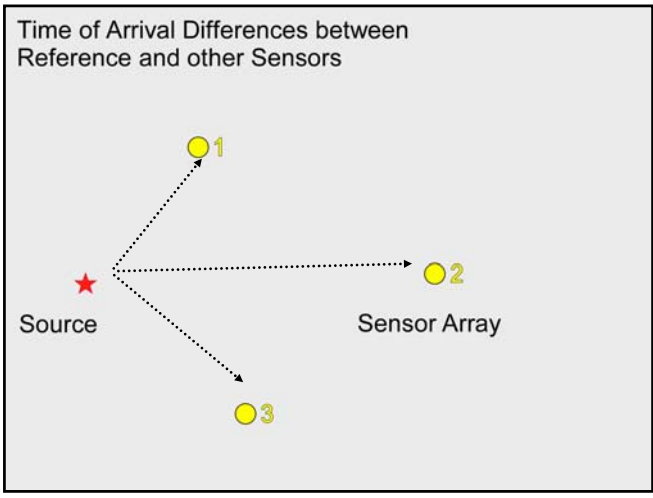
glotin @ univ-tln.fr
http://glotin.univ-tln.fr

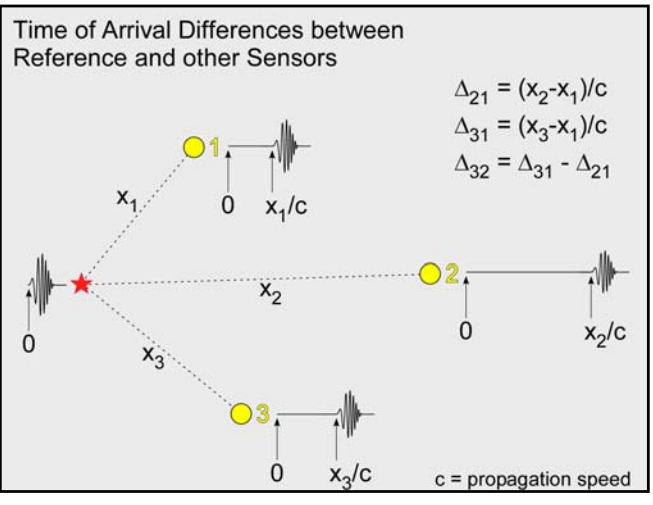
LSIS UMR CNRS

Objectifs

- Présenter différents modes de l’information
- Montrer que des calculs simples permettent un filtrage relativement efficace du bruit
- Montrer que l’information se traite de façons assez similaires sous différents modes
- Donner quelques les ordres de grandeurs des gains de systèmes sur des bases de données connues des différentes communautés pour appréhender leur relative robustesse.
- Discuter d’une orientation possible de ces systèmes vers des architectures ‘pro-actives’

PLAN





Définition d’une intercorrélation ou ‘CrossCorrelation’

Le produit de convolution $h = f * g$ de deux fonctions f et g est défini par : $h(x) = \int_{-\infty}^{\infty} f(t) \cdot g(x-t)dt$

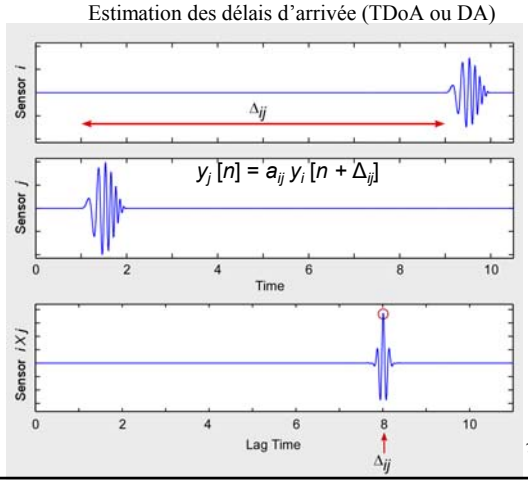
Lorsque les signaux sont à valeurs continues, le CC $s_1 x s_2$ est calculé de la façon suivante :

$$CC(s_1 x s_2) = s_1' * s_2 \text{ où } \forall t, s_1'(t) = s_1(-t)$$

Et l’AutoCorrélation :

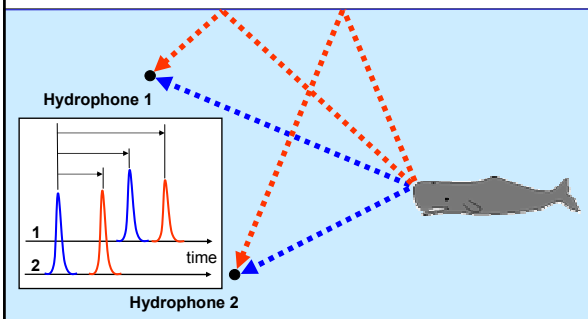
$$AC(s_1) = CC(s_1 x s_1)$$

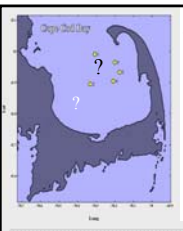
Estimation des délais d'arrivée (TDoA ou DA)



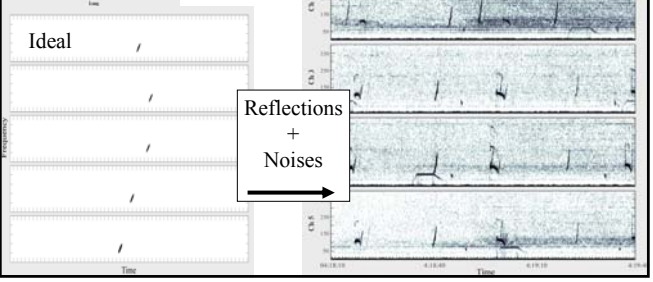
Cross Corr (i x j)

Echoes effects

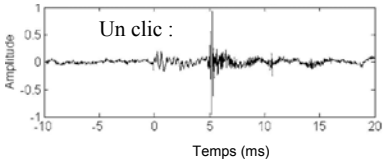


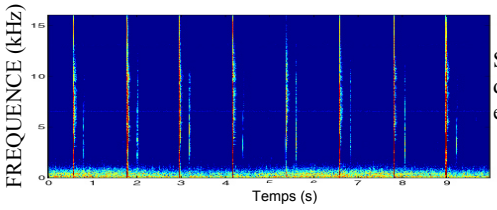


- Real data recorded on 5 channels (Cape Cod)



Le cas du cachalot





Spectrogramme de 8 clics et leurs échos.

Example of a real time acoustic tracking method

1. Click extraction (rectification & decimation)
2. TDOA estimation
3. Echo detection & TDOA selection (echo/noise cancelation)
4. Quadruplet selection
5. Source localization

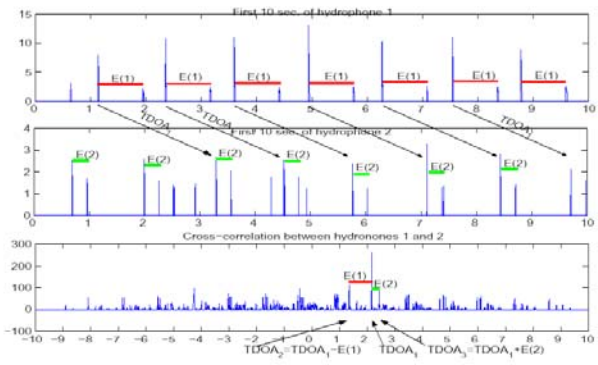
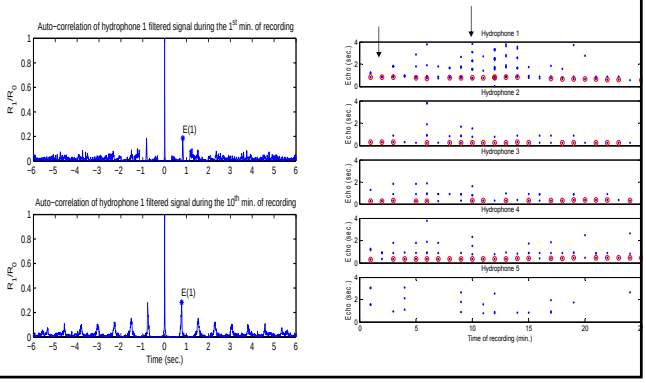


Fig.2: Rough TDOA estimation, including echoes (reflections of the click train off the ocean surface, or bottom, or different water layers).

3. Echo detection by AutoCorrelation



3. TDOAs selection

To eliminate TDOAs between direct clics and echoes, for each pair of hydrophones (i, j), $TDOA_x(i, j)$ satisfying one of the following equations are removed (Fig.2):

$$TDOA_x(i,j) - TDOA_1(i,j) = k * E(i) \pm 0.002, k \in \{1..4\}, x \in \{2..5\}$$
$$TDOA_x(i,j) - TDOA_1(i,j) = -k * E(j) \pm 0.002, k \in \{1..4\}, x \in \{2..5\}$$

4. Hydrophone quadruplets selection

The remaining TDOAs are combined every 10 s to select all quadruplets of hydrophones (i, j, k, h) whose TDOAs (u, v, w, x, y, z) correspond to the same source, therefore if:

$$\begin{aligned} TDOA_u(i,j) + TDOA_v(j,k) &= TDOA_w(i,k) \pm 0.004, \\ TDOA_u(i,j) + TDOA_x(j,h) &= TDOA_y(i,h) \pm 0.004, \\ TDOA_w(i,k) + TDOA_z(k,h) &= TDOA_y(i,h) \pm 0.004, \\ TDOA_v(j,k) + TDOA_z(k,h) &= TDOA_x(j,h) \pm 0.004. \end{aligned}$$

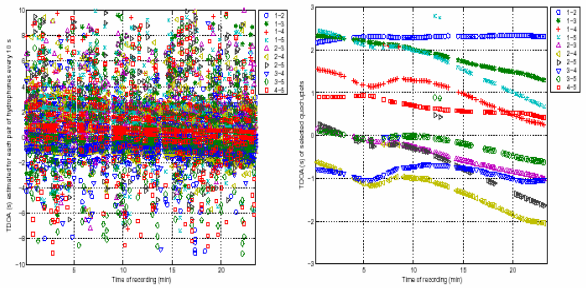


Fig.3: The initial 6630 TDOAs (Left). The final 1065 ones (Right) of the pairs of hydrophones belonging to the selected quadruplets, after isolated TDOAs and echoes elimination.

5. Source localization

for each 10-sec window
  for each selected quadruplets (i,j,k,h) with TDOA (u,v,w,x,y,z)
    localization of source S by solving by hyperbolic least-squares minimization :

$$\begin{cases} \text{dist(S,i)-dist(S,j)} = \text{TDOA}_u\text{(i,j)} * C, \\ \text{dist(S,i)-dist(S,k)} = \text{TDOA}_w\text{(i,k)} * C, \\ \text{dist(S,i)-dist(S,h)} = \text{TDOA}_y\text{(i,h)} * C. \end{cases}$$

  end
end

C = fixed sound speed = 1500 m/s

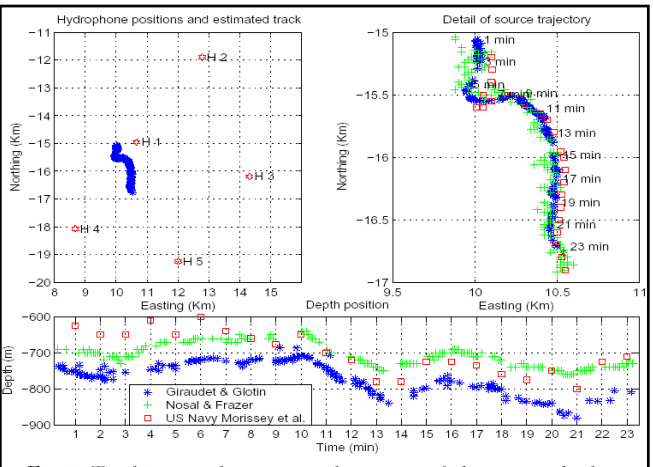


Fig.4: Tracking results compared to state-of-the-art methods.

Bilan

- En pratique 5 flux sont nécessaires pour situer dans l'espace la source (données incomplètes, bruitées).
- Réflexions et bruits très perturbants, problème difficile.
- Les simples calculs d'Inter- et Auto-corrélations des flux permettent de filtrer une grande partie du bruit et d'estimer rapidement (temps réel) une trajectoire vraisemblable.

PLAN

Problématique

Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux et paroles simultanées

Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

4/ RAP audiovisuelle

Pondération des flux audio et visuel dans le bruit

Sélection des flux par Analyse Linéaire Discriminante

La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles

Sélection automatique concept dépendante des traits visuels

Conclusion

Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

Annexes et pistes bibliographiques

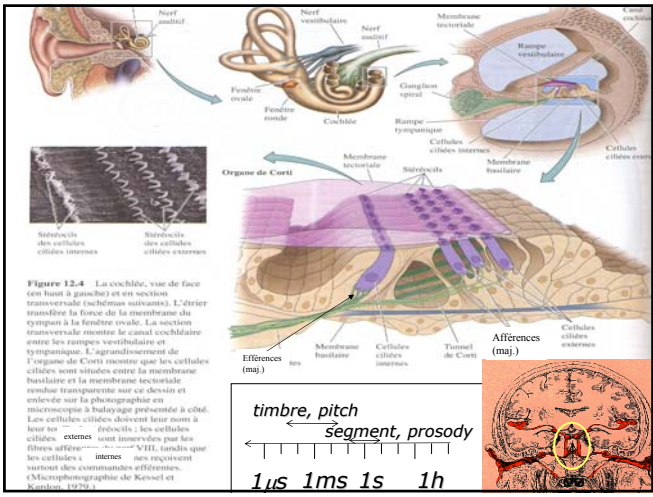


Figure 12.4 La cochlée, vue de face (en haut à gauche) et en section transversale (schémas suivants). L'étrier transfère la force de la membrane du tympan à la fenêtre ovale. La section transversale montre le canal cochléaire entre les rampes vestibulaire et tympanique. L'agrandissement de l'organe de Corti montre que les cellules ciliées sont situées entre la membrane basilaire et la membrane tectoriale rendue transparente sur ce dessin et enlevée sur la photographie en microscopie à balayage présentée à côté. Les cellules ciliées doivent leur nom à leur to… stéréocils ; les cellules ciliées externes sont innervées par les fibres affére… VIII, tandis que les cellules c… internes nes reçoivent surtout des commandes efférentes. (Microphotographie de Kessel et Kardon, 1979.)

Un modèle de CrossCorrélation naturelle

Hypothèse de Jeffress (1948)
(Observée chez la chauve souris)

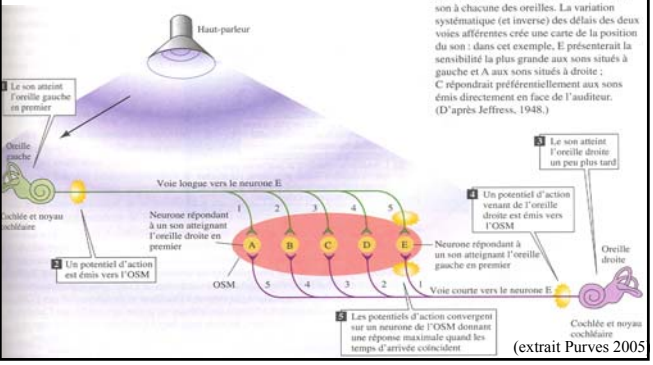
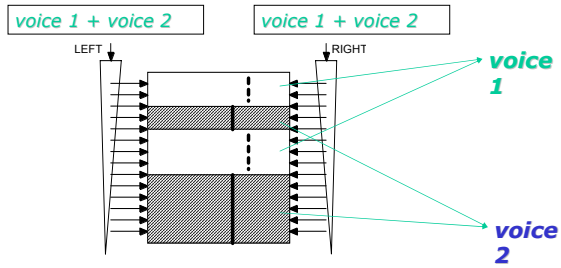


Figure 12.13 Schéma illustrant la façon dont l'OSM calcule la position d'un son au moyen des écarts temporels entre oreilles. Un neurone donné de l'OMS présente une réponse maximale quand deux messages afférents arrivent en même temps, comme c'est le cas lorsque les voies afférentes ipsi- et controlatérales compensent exactement (par leur différence de longueur) les différences entre les moments d'arrivée du son à chacune des oreilles. La variation systématique (et inverse) des délais des deux voies afférentes crée une carte de la position du son : dans cet exemple, E présenterait la sensibilité la plus grande aux sons situés à gauche et A aux sons situés à droite ; C répondrait préférentiellement aux sons émis directement en face de l'auditeur. (D'après Jeffress, 1948.)

(extrait Purves 2005

Application : binaural segregation

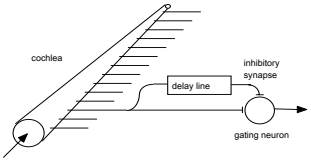
(Lyon 1983)



22

F0-guided segregation (Licklider 1951)

Unité de base :



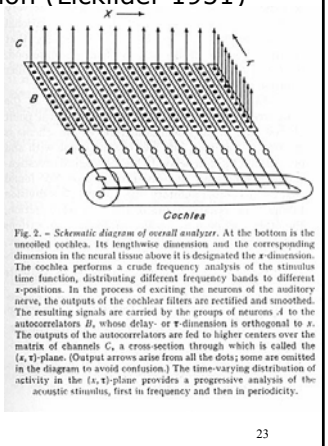
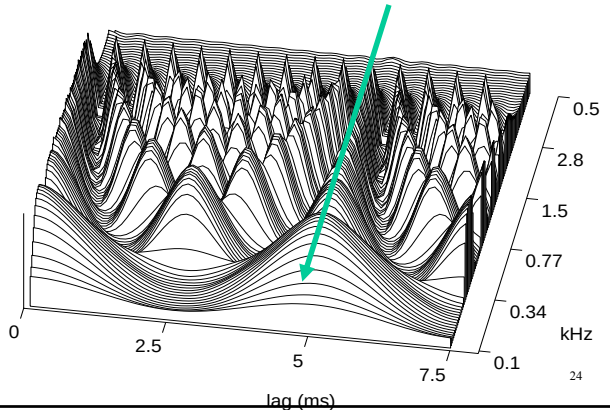


Fig. 2. – *Schematic diagram of overall analyzer.* At the bottom is the uncoiled cochlea. Its lengthwise dimension and the corresponding dimension in the neural tissue above it is designated the x-dimension. The cochlea performs a crude frequency analysis of the stimulus time function, distributing different frequency bands to different x-positions. In the process of exciting the neurons of the auditory nerve, the outputs of the cochlear filters are rectified and smoothed. The resulting signals are carried by the groups of neurons A to the autocorrelators B, whose delay- or τ-dimension is orthogonal to x. The outputs of the autocorrelators are fed to higher centers over the matrix of channels C, a cross-section through which is called the (x, τ)-plane. (Output arrows arise from all the dots; some are omitted in the diagram to avoid confusion.) The time-varying distribution of activity in the (x, τ)-plane provides a progressive analysis of the acoustic stimulus, first in frequency and then in periodicity.

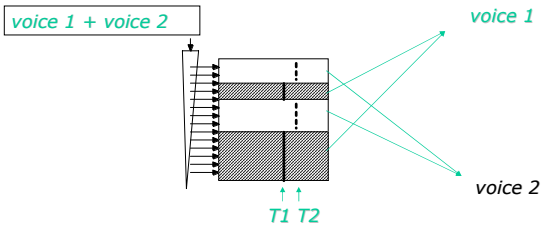
23

period --> timbre du locuteur

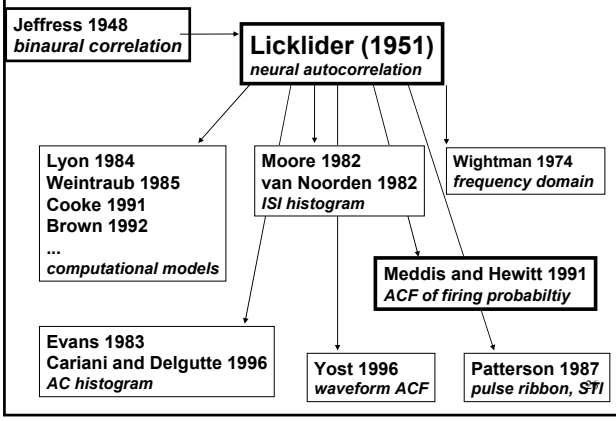


24

Application : F0-guided segregation (Meddis & Hewitt 1991)



Très nombreuses applications en analyses de scènes acoustiques



PLAN

Problématique

Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux

paroles simultanées Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

4/ RAP audiovisuelle

Pondération des flux audio et visuel dans le bruit
Sélection des flux par Analyse Linéaire Discriminante
La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles

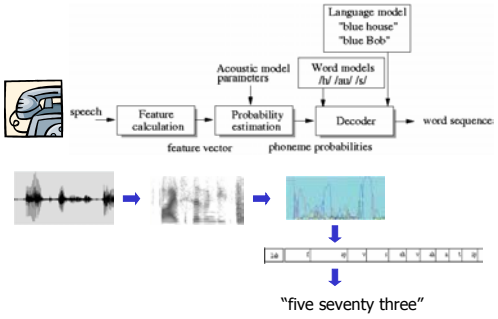
Sélection automatique concept dépendante des traits visuels

Conclusion

Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

Annexes et pistes bibliographiques

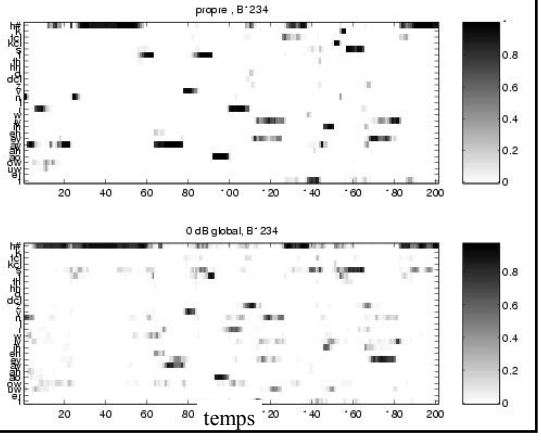
Automatic speech recognition



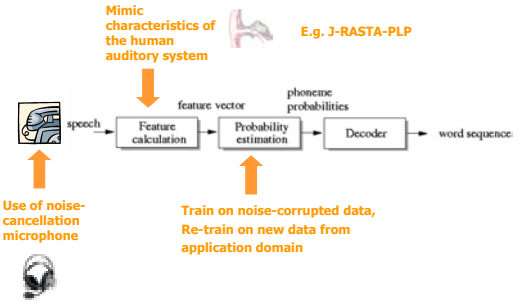
Démo : suivi des P(qk | X)

Parole propre

+ bruit blanc 0dB



Methods to enhance noise-robustness



Method for noise robustness:

Missing data processing

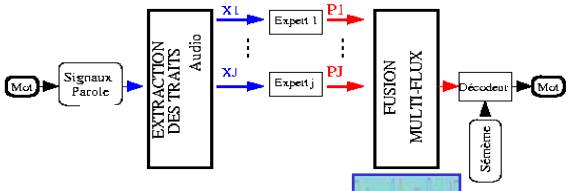
Clean speech

Speech with added helicopter noise (0 dB)

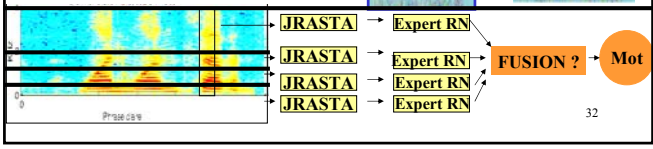
The reliable speech regions form a **mask** over the spectral data

➡ Problems: How to find reliable regions?
How to carry out recognition on partial data?

Modèle multi-bande / flux



(Fletcher 1920)



Fusion en Identification Séparée : Modèle "Toutes Combinaisons" (TC = Loi des Probabilités Totales)

•Chaque probabilité a posteriori est décomposée suivant le TC par la somme pondérée :

$$P(qk \mid X)=\Sigma_j \, Wj \, .P(qk \mid Xj)$$

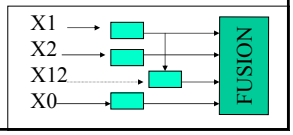
•qk = 'La classe phonétique k est reconnue'

•Cj : événements qui doivent former une partition sur l'ensemble des flux Xj

•Wj = P(Cj | X)

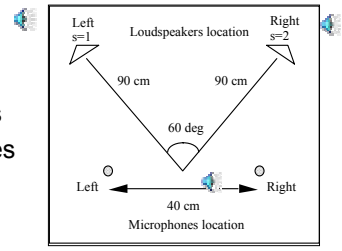
•Avantage : pas d'hypoth. d'Indép.

(Glotin Bourlard 1999)



Application à la double reconnaissance de parole (en coll. avec Tessier)

- Mélanges binaires à 0 db de 2 sources simultanées fixes.



Recouvrement de 75 % des mots des deux locuteurs

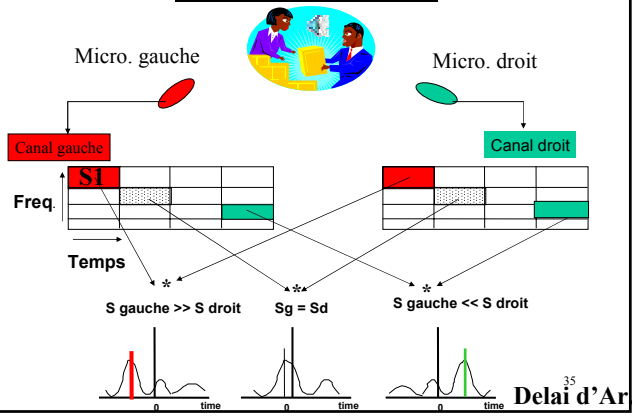
On conserve les niveaux de références : 3 enregistrements

Idée : étiqueter les pavés TF suivant le délais d'arrivée des sources

Application : RAP double par Indice de localisation de 2 locuteurs

intercorrélation Temps-Fréquence

(d'après le principe de Lyon 1983)



Delai d'Ar

Niveau relatif des deux sources

$$\mathbf{NR}_{1,i} = 10 \log \frac{\sigma^2_{1,i,left} + \sigma^2_{1,i,right}}{\sigma^2_{2,i,left} + \sigma^2_{2,i,right}}$$

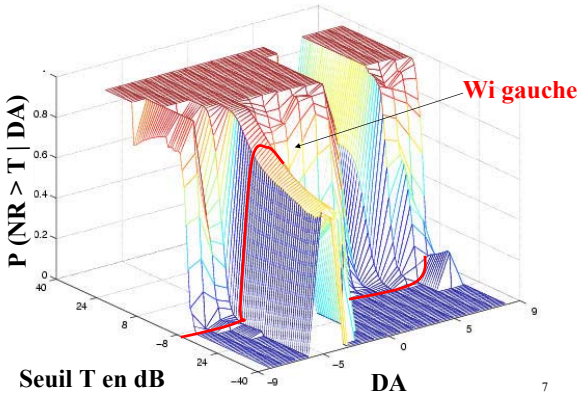
$$NR_{2,i} = - NR_{1,i}$$

Objectif : associer NR et délais d'arrivée (DA)

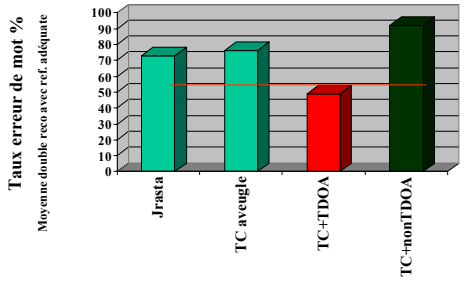
$$W1,_i = P(NR1,_i > Ti \mid DA_i)$$

Dans le train set on dispose du NR de reference on modélise alors la relation entre DA et NR

Carte (DA, niveau relatif)



RAP double reconnaissance multiflux, pondération par loi des probabilités totales Résultats de double reconnaissance



PLAN

Problématique

Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux et paroles simultanées
Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

4/ RAP audiovisuelle
Pondération des flux audio et visuel dans le bruit
Sélection des flux par Analyse Linéaire Discriminante
La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles
Sélection automatique concept dépendante des traits visuels

Conclusion

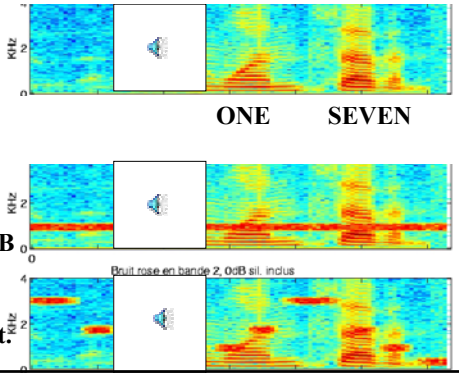
Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

Annexes et pistes bibliographiques

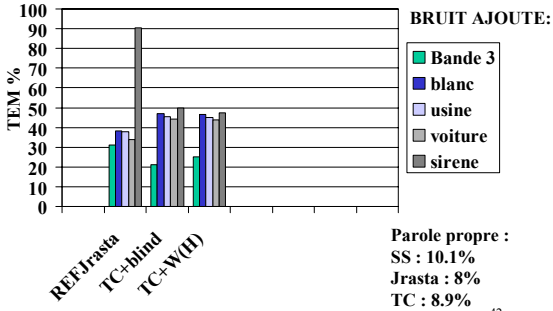
On présente trois types de pondérateurs :

- Pondération "aveugle" : Wi = cste
- Intrinsèque : qualité des estimations phonétiques
  ex : entropie
  Wi = P ('H(P(qk | Xi)) > seuil')
- Extrinsèque : sur critère qualité signal
  Wi = P('RSB > seuil')

Résistance aux bruits focalisés …



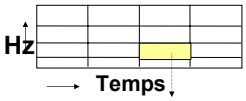
Taux Erreur de Mot %



Moyennes sur 6 niveaux RSB : [-12,-6,0,6,12,18] dB, 1200 phrases (6*200)

Le taux de voisement issu de l'AutoCorréation comme indice de qualité de la parole

Analyse **Temps-Fréquence**

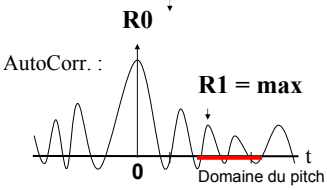


- 4 sous-bandes
- 128ms * 700 Hz de large

Indice de voisement :

$$R = \frac{R1}{R0}$$

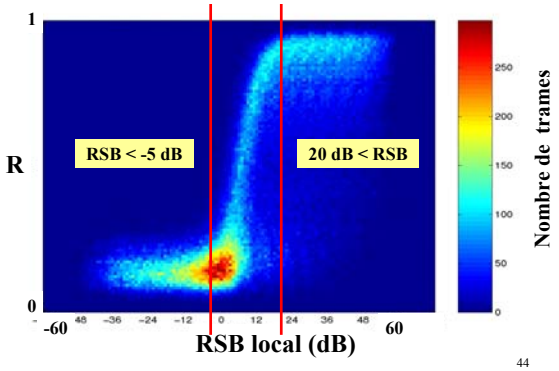
DEMODULATION



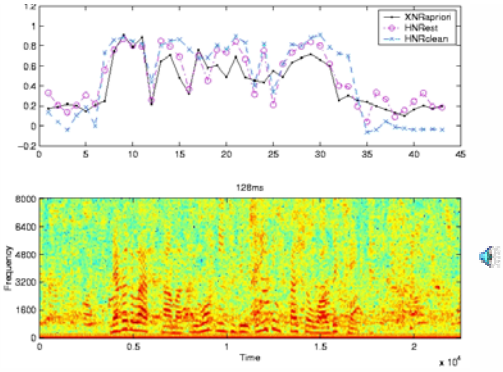
Domaine du pitch = [1/350 , 1/90]

Corrélation entre R et RSB local

(sur 1000 paves = [0 600] Hz x 128ms)



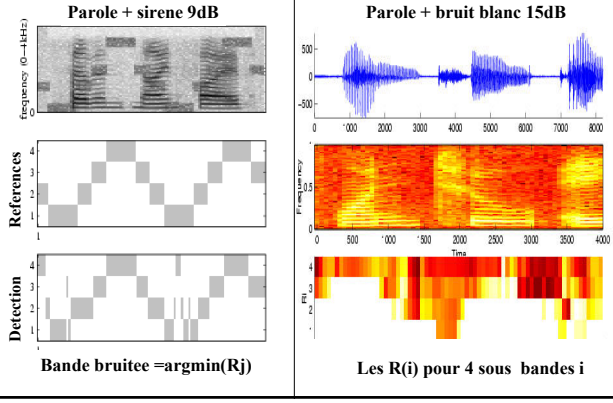
Corrélation entre RSB et R (=0.84) avec du bruit de cafétéria 8.5 dB RSB



Usage de R pour détecter du bruit ou de la parole

Parole + sirene 9dB

Parole + bruit blanc 15dB



Bande bruitee =argmin(Rj)

Les R(i) pour 4 sous bandes i

Pondération externe du FC via Mapping (R,RSB) in band 2

- **P(RSB > T | R)**

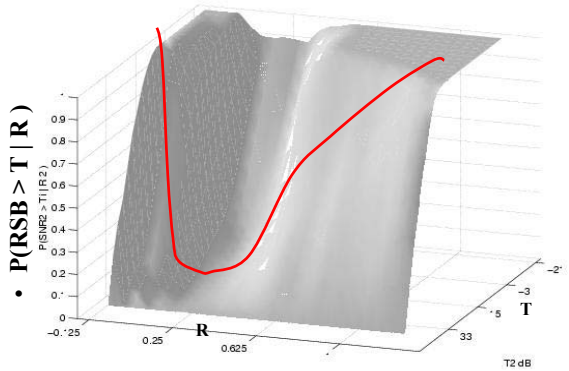


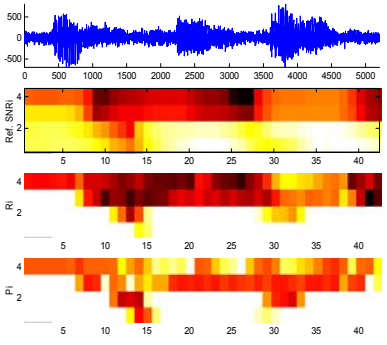
Illustration des Wi = P(RSB i > Ti | Ri)

Signal + bruit Blanc 5 dB

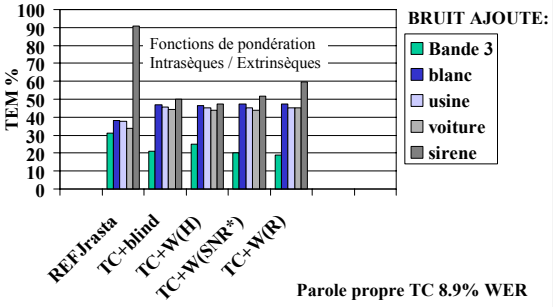
RSB(i) des 4 Sous-bandes

R(i) des 4 Sous-bandes

W(i) =P(RSB > T | R)



Taux Erreur de Mot %



Parole propre TC 8.9% WER

Moyennes sur 6 niveaux RSB : [-12,-6,0,6,12,18] dB, 1200 phrases (6*200)

PLAN

Problématique
Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux et paroles simultanées
Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

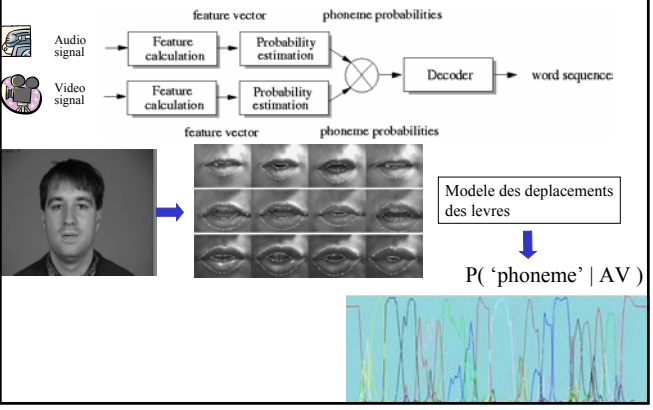
4/ RAP audiovisuelle
Pondération des flux audio et visuel dans le bruit
Sélection des flux par Analyse Linéaire Discriminante
La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles
Sélection automatique concept dépendante des traits visuels

Conclusion
Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

Annexes et pistes bibliographiques

Reconnaissance Audio-visuelle



THE VISUAL FRONT END - INTRODUCTION

- Three groups of visual front ends:
  - Lip contour based features.
    Height, width, area (Benoît);
    Moments, Fourier descriptors (Potamianos);
    Template, B-spline parameters (Silsbee, Blake).
  - Video pixel based features.
    PCA (Bregler); whole *ROI* (Waibel); *DCT* (Duchnowski);
    DWT (Potamianos); *LDA* (Duchnowski).
  - Lip-contour driven, pixel based features.
    Active appearance models - AAMs (Matthews);
    Active shape models (Luettin).
- In WS00-AVSR, two visual front ends have been considered:
  - A 3-stage cascade image transform visual front end.
    - *Image transform* for data compression (DCT).
    - *Linear discriminant analysis* data projection (LDA).
    - *Maximum likelihood linear transform* data rotation (MLLT).
  - Active appearance models (AAMs).
  - Comparison caveats: ROI, static feature, temporal window sizes.

AVSR WS'00

The Recognition Problem

$$M^* = \arg\max_M P\left(M \middle| O^A, O^V\right)$$

$$P\left(M \middle| O^A, O^V\right) = \frac{P\left(O^A, O^V \middle| M\right) P(M)}{P\left(O^A, O^V\right)}$$

M: word (phoneme) sequence
M^*: most likely word sequence
O^A: acoustic observation sequence
O^V: visual observation sequence

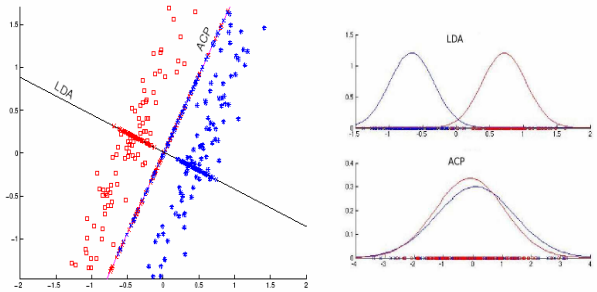
Integration at the Feature Level

Assumption:
- conditional dependence between modalities
- integration at the feature level :

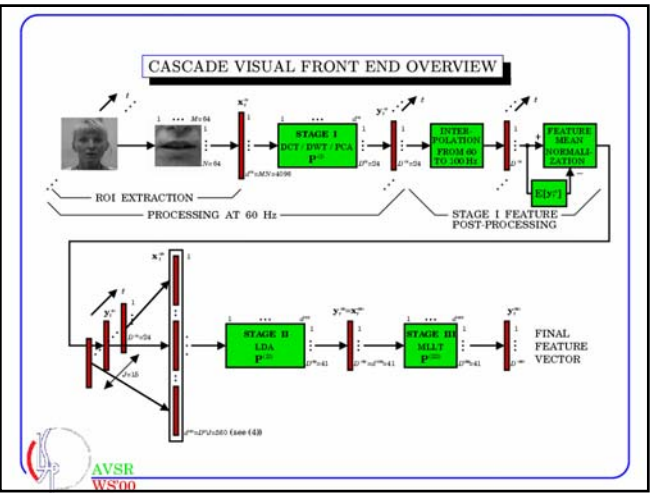
$$P\left(O^A, O^V \middle| M\right) = P\left(O^{AV} \middle| M\right)$$

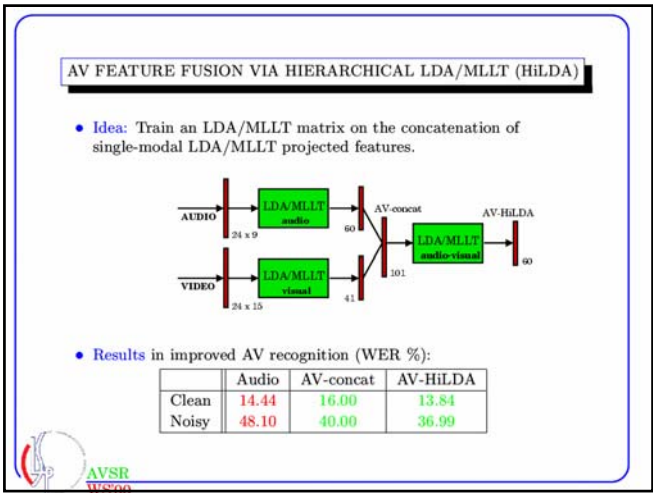
$$where \quad o^{AV}(t) = \left[o_1^A(t), o_2^A(t), ..., o_N^A(t), o_1^V(t), o_2^V(t), ..., o_M^V(t)\right]^T$$

2 Principes de réduction de dimensions : PCA et LDA



(a) Un ensemble de données séparées en deux classes distinctes. Les axes fournis par la LDA et l'ACP. (b) La projection des données sur les axes montre que la LDA sépare les classes (en haut), tandis que l'ACP les confond (en bas).





Integration at the Decision Level

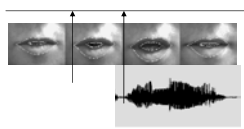
Assumption:

- conditional independence between modalities
- integration at the unit level
- Weithing scheme : alpha, from voicing level

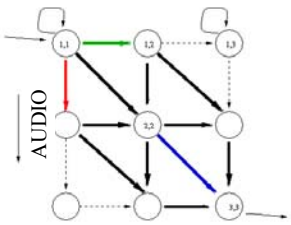
$$P(O^A, O^V | M) = P(O^A | M)^{Alpha} P(O^V | M)^{(1-Alpha)}$$

Exemple : Alpha = P(RSB > seuil | taux de voisement)
(Glotin et al. ICASSP 2001)

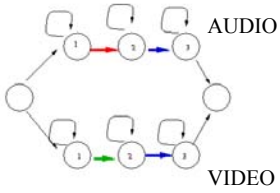
FUSION audio-visuelle asynchrone ou synchrone MMC



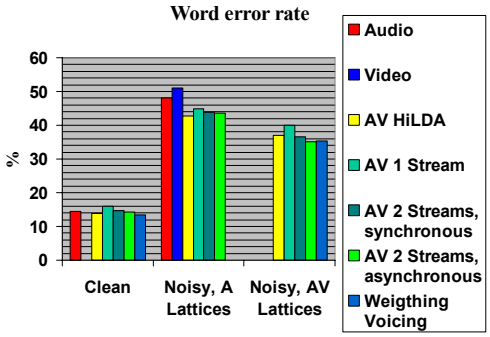
Fusion asynchrone



Fusion synchrone



Large Vocabulary Audio Visual ASR WER Results (Via Voice HTK- CLSP wkp 2000)

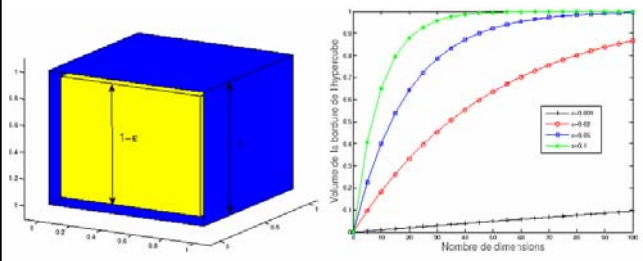


Bilan

- Gains similaires par
  - pondération suivant la fiabilité des flux ou
  - la réduction du nombre de dimension des traits par sélection LDA
- Les deux stratégies sont intéressantes pour des systèmes ‘robustes’, mais insuffisantes…

Le problème des grandes dimensions

- Beyer et al., 1999 démontrent que plus le nombre de dimensions est grand, plus le rapport entre la distance minimale et la distance maximale entre les données tend vers 1,
- Autrement dit plus la dimension des données augmente, plus la notion de plus proche voisin n'existe plus



(a) La bordure d'un hypercube est la zone située entre l'hypercube de coté 1 unité et l'hypercube de coté $1 - \epsilon$ unités

(b) Volume de la bordure d'un hypercube en fonction du nombre de dimensions pour différentes valeurs de ϵ

PLAN

Problématique

Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux et paroles simultanées

Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

4/ RAP audiovisuelle

Pondération des flux audio et visuel dans le bruit
Sélection des flux par Analyse Linéaire Discriminante
La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles

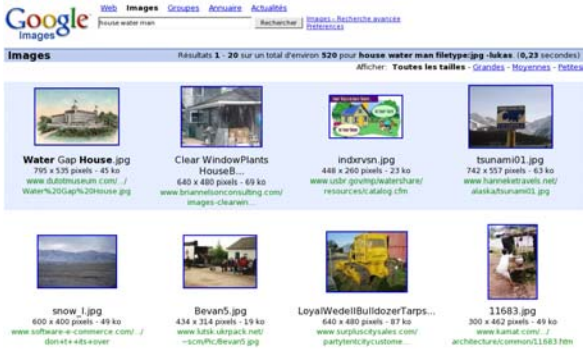
Sélection automatique concept dépendante des traits visuels

Conclusion

Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

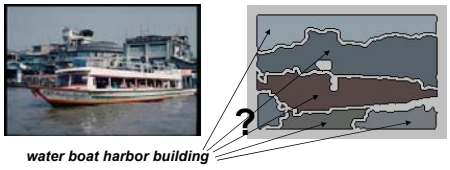
Annexes et pistes bibliographiques

Exemple de problématique : Recherche d'images (ex web) par mots clés



Problématique

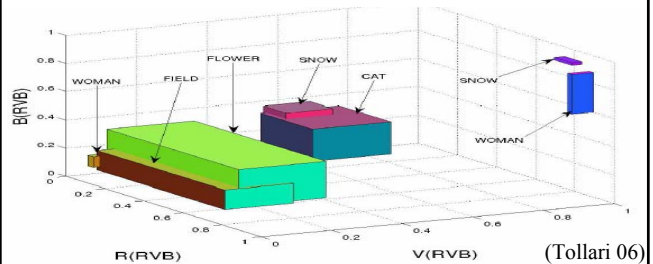
- **Les bases d'images réelles (par exemple les images du web) ne sont pas étiquetées par région d'images.**
- **Comment apprendre les liens sémantiques entre mot-clé et région d'images en connaissant seulement les mots-clés par image ?**



water boat harbor building

Par contre les régions autour de buildings seront très changeantes…

Apprentissage de clusters visuels par Classification Hiérarchique ascendante Exemple dans le sous espace RVB



(Tollari 06)

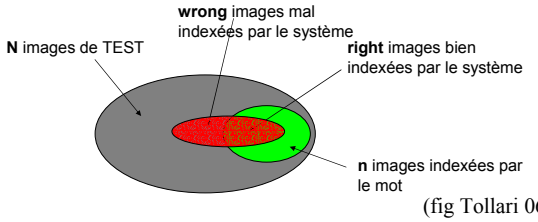
(a) Exemples de distributions dans l'espace RVB des *clusters* visuels des mots *cat, flower, field, snow, woman*. Attention les *clusters* visuels sont en fait en 40 dimensions, mais afin de pouvoir mieux les visualiser, nous ne montrons que les dimensions RVB. Les *clusters* visuels des mots *cat* et *snow*, et des mots *field* et *flower*, ont une partie de leur *cluster* visuel en commun.

Évaluation de l'association : Calcul du score de la classification

Pour chaque mot, on peut calculer le score **« Normalized Score » :**

Score NS = right/n - wrong/(N-n)

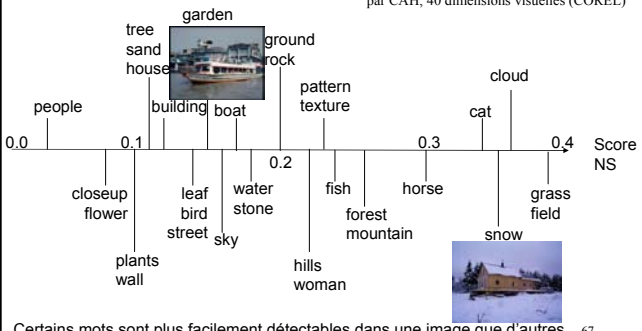
sensibilité 1-spécificité



(fig Tollari 06)

Estimation de la « consistance visuelle » d'un concept

= Echelle des Normalized Score
par CAH, 40 dimensions visuelles (COREL)



Certains mots sont plus facilement détectables dans une image que d'autres.

Construction de nouveaux traits visuels

- Inspiré des travaux en psychovision sur l'air des régions (cf Martinet & Mulhem)
  - Le cerveau humain interprète en contexte
- Extension à tous les traits : la valeur de l'hétérogénéité pour le trait visuel p de l'image d est l'entropie :

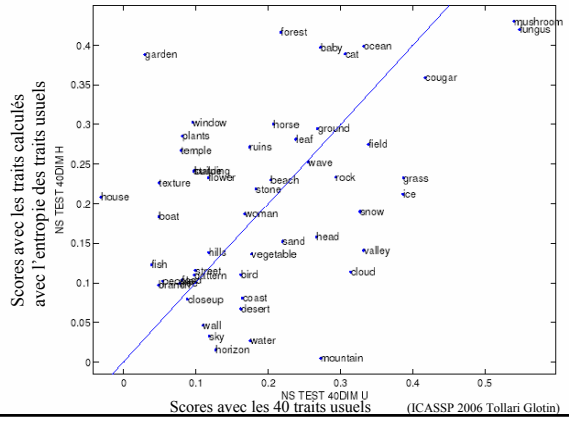
$$H_p = -\sum_{b_j \in d} b_{j,p} \times \log_2(b_{j,p})$$

Homogène en blanc

Hétérogène en blanc

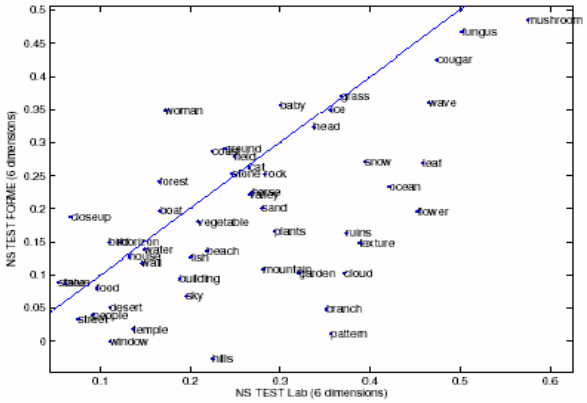
(Inspiré de Martinet PHD)

Des dimensions plus ou moins pertinentes : Usuels vs H(U)



(ICASSP 2006 Tollari Glotin)

Couleur vs Forme



(c) Lab versus F6

Approximation de l'Analyse Linéaire Discriminante (ALDA) pour estimer les traits visuels les plus discriminants

- Pour déterminer les traits visuels les plus discriminants en fonction du mot clé
  - Séparation des images (et pas des régions) en 2 classes :
    - Celles qui sont indexées par ce mot
    - Celles qui ne sont pas indexées par ce mot
  - Calcul des variances interclasses (B) et intraclasses (W)
  - Pour chaque trait visuel v_j et pour chaque mot w_i, on calcule le pouvoir discriminant F:

$$F(v_j; w_i) = \frac{B(v_j; w_i)}{B(v_j; w_i) + W(v_j; w_i)}$$

Approximation de l'Analyse Linéaire Discriminante (ALDA) pour estimer les traits visuels les plus discriminants

- Pour déterminer le nombre **N** de traits visuels nécessaires pour bien discriminer, on peut prendre les N traits visuels qui cumulent 50% de la somme des pouvoirs discriminant de tous les traits.
  - Si les traits visuels v_j sont ordonnés dans l'ordre décroissant des pouvoirs discriminant, N est tel que :

$$\sum_{j=1}^{N} F(v_j; w_i) = \frac{1}{2} \sum_{j=1}^{40} F(v_j; w_i)$$

Principe de l'approximation LDA pour les images globalement annotées

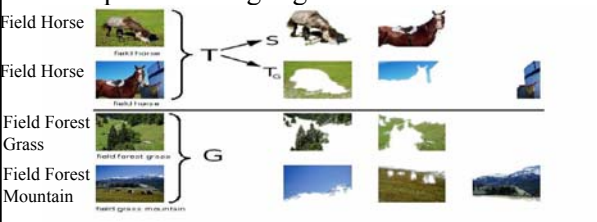


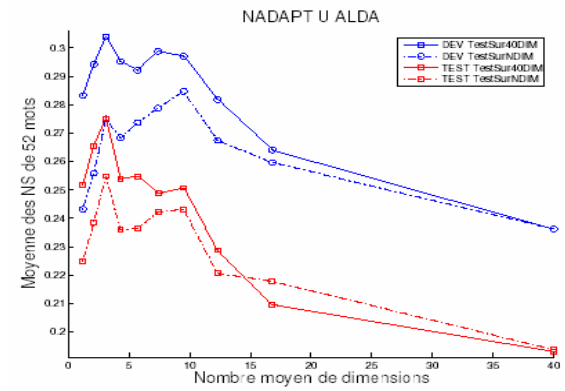
FIG. 7.1 — Principe de la démonstration de l'ALDA. En haut, les images annotées par le mot *horse* (ensemble T) contiennent des blobs qui représentent exactement un *cheval* (ensemble S) et des blobs qui ne le représentent pas (ensemble T_G). En bas, quelques images non-annotées par le mot *horse* (ensemble G).

On montre que l'ordre des pouvoirs discriminants des traits est approximé:

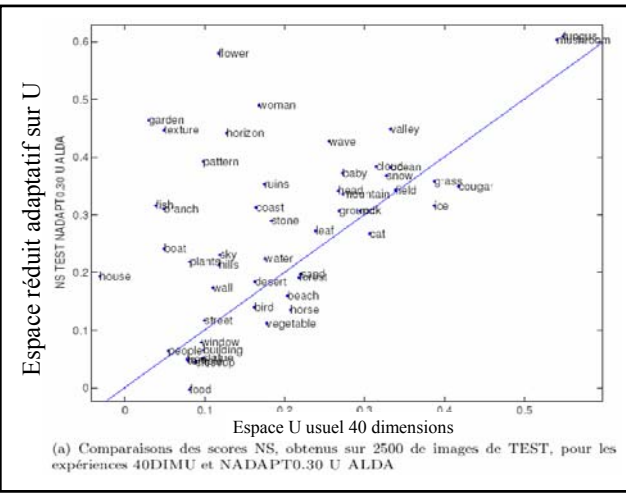
$$\hat{V}_{LDA}(\vec{\omega}; w) = A(w) \cdot V_{LDA}(\vec{\omega}; w) + B(w) \cdot \left(1 - C(\vec{\omega}; w)\right)$$

où V = 1/F

(Glotin,Tollari & Giraudet in Computer Graphics 2006)



(a) Moyennes des scores NS des 52 mots les plus fréquents pour τ variant de 0.10 à 1 par pas de 0.10, sur DEV ou TEST, (Tollari PHD 2006, Tollari et al ICASSP 06)



(a) Comparaisons des scores NS, obtenus sur 2500 de images de TEST, pour les expériences 40DIMU et NADAPT0.30 U ALDA

Exemples montrant la difficulté de l'auto-annotation automatique

Experiments are done on COREL where a maximum of 5 'manual' words are globally labelling each image. DIMATEX is running segmenting each image in 10 regions with Normalized Cuts algorithm [Barnard2003]. We give below images with 1: COREL manual annotation, 2: DIMATEX Auto-Annotations, 3: PLSA, 4: PLSA, 5: DIRECT, and 6: LSA. (3,..,6) are from [Monay2003 & http://www.idiap.ch/monay/acmm04]:

PLAN

Problématique

Exemple d'une recherche acoustique avec présence de réflexions multiples: suivi de baleine

1/ Une recherche efficace d'informations : l'analyse temps-fréquence auditive humaine

2/ Reconnaissance Automatique de la Parole (RAP) multiflux et paroles simultanées
- Analyse temps-fréquence stéréophonique

3/ RAP multibande robuste par analyse temps-fréquence monophonique

4/ RAP audiovisuelle
- Pondération des flux audio et visuel dans le bruit
- Sélection des flux par Analyse Linéaire Discriminante
- La malédiction des grandes dimensions

5/ Analyse robuste de scènes visuelles
- Sélection automatique concept dépendante des traits visuels

Conclusion

Vers une recherche robuste par ciblage ' pro-actif ' des flux informatifs ?

Annexes et pistes bibliographiques

- Faibles performances avec pondérateurs intrasèques,

Bilan

- Contre le principe de Massaro (1987) ("la sortie de chaque processus de decision permet de guider efficacement la fusion des decisions sans qu'il soit besoin d'introduire une information contextuelle externe"),

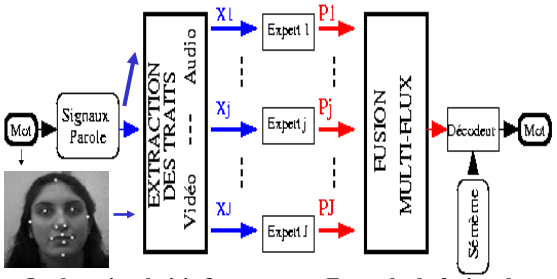
- Vers une analyse en contexte …

=>

- "La reconnaissance est un sous problème de la compréhension. Les résultats obtenus à l'issue de la reconnaissance conditionnent la compréhension (et inversement). Cela exige un auto-contrôle de la reconnaissance sur elle-même, c'est à dire des capacités de reflexivité…il faut tenir compte de la situation pour produire une réponse 'intelligente".

J. Caelen , La Recherche 285, Mars 1996

STRATEGIE : DECOMPOSER → RECOMPOSER



Quels traits choisir ?

Formule de fusion des reconnaissances partielles ?

Techniques :

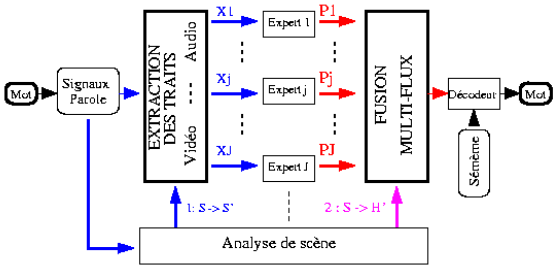
Traitement du signal et Analyse de scène auditive

Fusion de données et théorie de l'information.

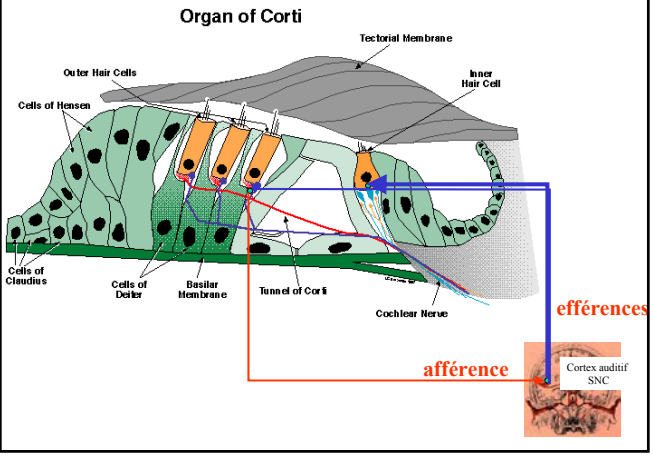
Couplage

rehaussements | pondérations

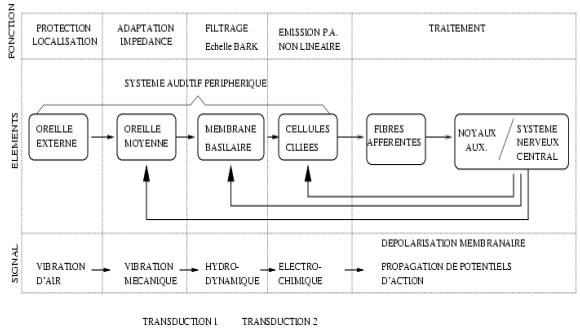
signaux | estimations



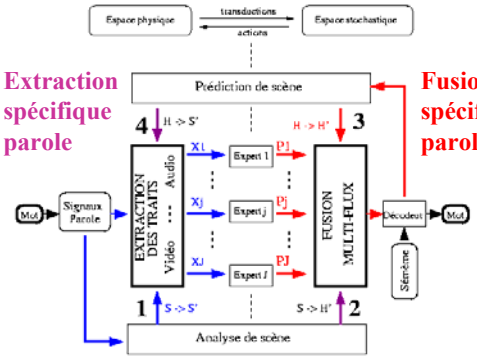
Des efférences du Système Nerveux Central modulent sûrement la perception



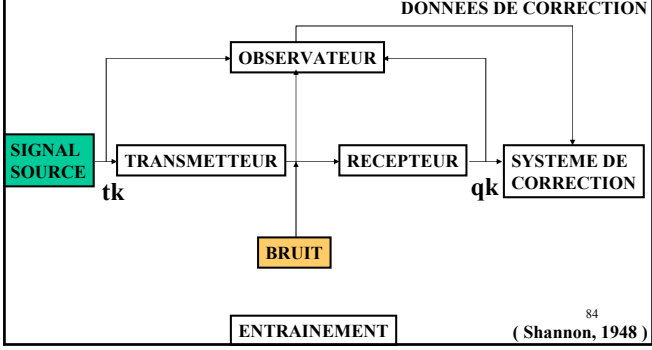
Efferences du Système auditif



Analyses ASCENDANTE & DESCENDANTE



SYSTEME DE PREDICTION DES BIAIS DES POSTERIORS
(PBP Glotin 2001)



(Shannon, 1948)

Le modèle Prédiction des Bias des Posteriors (PBP)

Indices de Fiabilités des Estimations

- Positives = **P(tk| qk)**
- Négatives = **P(¬tk|¬qk)**

Avec :

tk = 'le phonème k est émis'

qk ='le phonème k est reconnu '

On construit ces fonctions à partir des matrices de confusion par tranches de RSB ou d'indice R, …

Modèle Prédiction des Biais des Posteriors (PBP)

On reprend les mêmes P(qk | Xj), Wj que dans le modèle LPT, et on a :

$$P(tk \mid X)=\Sigma j\ Wj.\ (A\ .\ P(qk \mid Xj) + B\ )$$

A = FEP + FEN -1

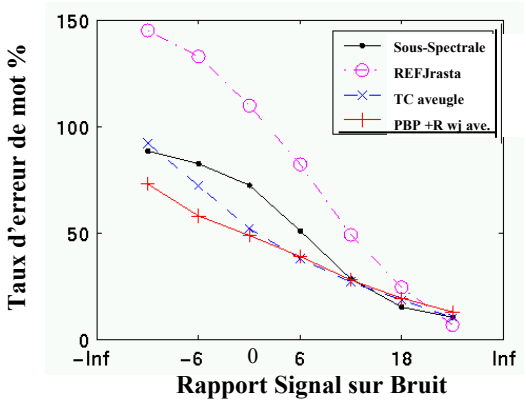
B = 1- FEN

Wj = fonction du RSB, ou intrinsèque …

REM : pour FEP = 1 et FEN = 1 , PBP = le modèle toute combinaison.

(Glotin 2001)

Résultats PBP sur bruit non stationnaire (Numbers95)



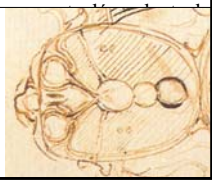
- Pour que le résultat d'une recherche automatique d'information corresponde à notre attente, pourrait-il être nécessaire que le système de recherche s'inspire/simule nos représentations et traitements neurobiologiques de l'information ?

- La neurophysiologie progresse rapidement depuis peu (IRM…), pourrions nous concevoir des Systèmes de Recherche d'Information Multimodaux bio-inspirés ?

- On en observe les prémices : voir les prétraitements acoustiques (PLP) ou visuels (Lab) qui se rapprochent de nos systèmes sensoriels. Mais [illegible] mple modéliser les traitements [illegible] exposé Giraudet) ?

[illegible] Vinci. Les décussations visuelles et auditives sont déjà observés.

• Peut être que ces observations sont aussi obsolètes que ne le sera la neurobiologie actuelle en 2020 ?



Ces travaux ont été effectués de 1997 à 2006 dans différentes équipes :

Institut d'Intelligence Artificielle Perceptive (IDIAP), EPFL – CH,

Institut Communication Parlée (ICP), UMR CNRS - INP Grenoble,

CSLP Johns Hopkins avec IBM – Baltimore USA,

Equipe de Recherche en Syntaxe & Sémantique (ERSS) UMR CNRS – Toulouse,

LSIS UMR CNRS – Toulon.

Mes remerciements particuliers à :
S. Tollari, P. Giraudet, E. Tessier, J. Luettin, M. Potamianos,
C. Neti, S. Bengio, AdC, et anciens IDIAPiens, ICPiens…

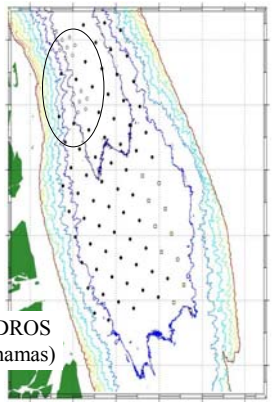
Suivent
Annexes sur les corpus présentés
&
Pistes Bibliographiques

Hydrophones position

Hydrophone	X (m)	Y (m)	Z (m)
1	10 658	-14 953	-1 530
2	12 788	-11 897	-1 556
3	14 318	-16 189	-1 553
4	8 672	-18 064	-1 361
5	12 007	-19 238	-1 522

• 25 audio files
• Total of 25 min
• Recorded on 5 hydrophones
• FS = 48 KHz
• Data set 2 of the workshop

Bathymetry of AUTEC Range



[map from AUTEC]

Test de parole continue téléphonique

BASE :
- Numbers 93/95, multi-locuteur
- 32 nombres / 27 phonèmes
- 2000 'phrases' de cinq à dix nombres

PARAMETRES :
- Prétraitement JRASTA sur chaque flux
- Dérivées et dérivées secondes des coefficients
- Réseaux de neurones type perceptron multicouche (300 000 poids)
- HMM monophone
- Pas de modèle de langage

CALCUL DU TAUX d'ERREUR DE MOT :
- TEM = 100.(1 - (N - omis -remplacés - insérés) / N) %

THE AUDIO-VISUAL DATABASE

- The IBM ViaVoice audio-visual database:
  - 290 subjects.
  - 50 hours, large vocabulary (10.5 K words), continuous speeech.
  - Frontal face color video, 704 × 480 pixels, 30 Hz, MPEG2.
  - 16 KHz audio (clean).

- Experimental setting:
  - *Speaker independent* (SI) data partitioning: 35 hrs *training*, 5 hrs *held-out*, 2 hrs *test* (in addition: 1 hr SI *adaptation* set, and *multi-speaker held-out* and *test* sets).
  - Audio: (a) *clean*; (b) *matched noisy* ("babble", 10 db SNR).
  - Transcriptions, dictionary, language model scores are provided.
  - Clean and matched noisy audio lattices are provided.
  - Lattice audio-only WER: 14.44 % clean, 48.10 % noisy.

AVSR WS'00

REGION OF INTEREST EXTRACTION

- Face detection and mouth location estimation:
  - Statistical face detection and 26 facial feature localization algorithm (A.W. Senior, IBM Research).
  - Multi-scale (image pyramid) search for faces.
  - Uses Fisher discriminant, prior facial part collocation statistics.
  - Requires training (on about 2000 annotated frames).
  - Face detection acc. = 99.7 %; facial feature acc. ≈ 90 %.
- ROI extraction:
  - Obtain smoothed mouth center and size estimates.
  - Extract a 64 × 64 pixel, size normalized ROI.

  - ROI failure example:

AVSR WS'00

Base d'images COREL

- 10 000 images
- 250 mot-clés environs en anglais
- Mots / traits visuels U :
  - De 1 à 5 mot-clés choisis manuellement
  - De 2 à 10 « blobs », des blobs de l'image
  - Chaque blob de l'image possède un vecteur visuel de 40 composantes avec leur moyenne et variance, extraits par Kobus Barnard :

  aire, convexité, inertie, périmètre, pos x, y, RGB, RGS, LAB, 12 coefficients de texture (filtres gaussiens),…)

Kobus Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, *« Matching Words and Pictures »* , Journal of Machine Learning Research, Vol 3, pp 1107-1135, 2003.

http://vision.cs.arizona.edu/kobus/research/data/jmlr_2003/index.html

http://wang.ist.psu.edu/docs/home.shtml

Principe de la CAH

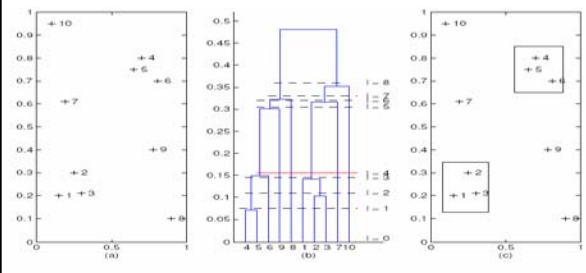


FIG. 6.2 – Exemple d'application de l'algorithme de la page 137 sur des données synthétiques. (a) Les vecteurs visuels correspondant aux images annotées par un mot donné sont répartis dans l'espace visuel. Certains de ces vecteurs appartenant à plusieurs images sont regroupés, car ils correspondent à des propriétés communes à plusieurs images, et donc caractéristiques du mot. (b) Une CAH est effectuée à partir de tous ces vecteurs. Pour chaque niveau l, la qualité des *clusters* courants est évaluée. Les vecteurs moyens et d'écarts-types des *clusters* du niveau l' qui ont obtenu le meilleur score sont conservés (sur l'exemple $l' = 4$). (c) Nous obtenons ainsi les zones visuelles (hyperrectangles) qui caractérisent le mieux un mot.

(extrait PHD Tollari 2006)

Pistes Bibliographiques

Analyse de scènes acoustiques bruitées, localisation :
- Tessier PHD 2003
- Giraudet Glotin, 'Real time robust 3D whales tracking', Applied Acousitics journal Elsevier nov. 2006
- Detection and localization of marine mammals using passive acoustics, 2nd inter. Workshop, Monaco Oceanographics Museum, 2005

RAP 'robuste' multibande ou audiovisuelle :
- Dupont, Bourlard 1996 et après
- Neti Potamianos Luettin Matthews Glotin, 'Large Vocabulary Audiovisual Speech Recognition', Johns Hopkins RR 80p, 2000
- Glotin PHD 2001 et publications (http://glotin.univ-tln.fr)
- Hagen PHD 2001 et publications
- Morris Hagen Glotin Bourlard, in Speech Communication journal Elsevier 2001
- Travaux de Bengio, Hermansky, Gravier, Rogozan, Luettin,…

Analyse de scènes visuelles :
- Barnard, Duygulu, de Freitas, Forsyth, Blei, and Jordan, '*Matching Words and Pictures'* , Jour. of Machine Learning Research, 2003
- Martinet PHD 2004
- Tollari PHD 2006
- Monay PHD 2005
- Berrani PHD 2005
- Tollari Glotin, 'LDA vs MMD approximation on mislabeled images for keyword dependant selection…', ICASSP 2006 – voir aussiTollari, Glotin, revue ISI Hermès, nov 2006.
- Glotin Tollari Giraudet, 'Shape reasonning on missegmented and mislabeled …', Computer Graphics inter. jour. Elsevier 2006
- Et bien sur les travaux de Bengio, Jeon, Joly, Mulhem, Vittaniemi…

Notes

Notes

Identification et classification automatique de langues

ERMITES 2006

Jérôme Farinas
Équipe SAMOVA
(Structuration, Analyse et Modélisation de la Vidéo et de l'Audio)

Organisation de la présentation

Partie I

Analyse du média

Problématique IAL

Identification Automatique de la Langue
(Language Identification)

Définition : détecter la langue parlée à partir de quelques secondes d'un échantillon sonore

Objectif : aiguiller vers un système de reconnaissance de la parole multilingue, aiguiller vers standardiste parlant la langue pour un numéro urgence (ex : 911), central téléphonique hôtelier, bornes interactives multilingues, indexation multimédia, renseignement militaire, etc.

Contraintes : nombre limité de langue connues ou bien pas de limite (rejet), décision rapide (dès les premières secondes)

Sources d'information

Différentes sources d'informations sont exploitables pour l'IAL :

Acoustiques : les sons et leur fréquences d'apparition varient d'une langue à l'autre

Phonotactiques : les enchaînements entre les sons et leur fréquence d'apparition caractérisent les langues

Lexicales : les mots sont souvent propres aux langues. Source d'information peu intéressante si l'on veut pouvoir rajouter une langue au système sans connaissances *a priori*.

Prosodiques : le rythme et l'intonation varient d'une langue à l'autre.

Sources d'information : acoustique

L'inventaire des sons varient d'une langue à l'autre [Vallée 94]

Même si une langue partage les mêmes sons avec une autre, il est fort peu probable que leur fréquence d'apparition soit identique.

Nécessite des décodeurs acoustico-phonétiques ou bien une segmentation au niveau phonétique ou infra phonétique

Sources d'information : phonotactique

L'enchaînement des sons est particulier aux langues.

Certains enchaînements ne se retrouvent pas dans d'autres langues.

Leur fréquence d'apparition est également unique.

Sources d'information : lexicale

Chaque langue possède son propre lexique.

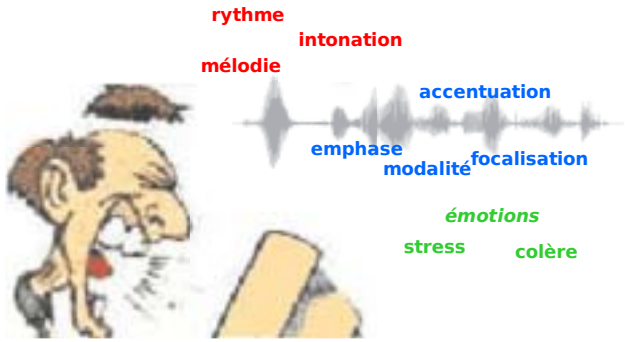
Difficulté : la frontière entre les mots n'est pas facile à trouver quand on ne connait pas la langue.

Utiliser l'inventaire des mots d'une langue impose de disposer d'importantes ressources lexicales, qui ne sont pas forcément faciles à obtenir (langues rares ou bien langues ne disposant pas de transcriptions textuelles).

Si l'on veut pouvoir rajouter une langue facilement à un systme, cette source d'information n'est pas privilégiée car elle demande des ressources coûteuses ou bien demandant l'utilisation d'expertises.

Quelques travaux ont été réalisés en utilisant partiellement cette ressource ([Hieronymous 96], [Adda 98]

Sources d'information : prosodie (1/3)



Sources d'information : prosodie (3/3)

Langues accentuelles
- Anglais
- Néerlandais
- Polonais
- …

Langues syllabiques
- Espagnol
- Italien
- Français
- Catalan
- …

Langues moraïques
- japonais

Partie II

Modélisation

Modélisations

Modélisations sans prise en compte de l'enchaînement temporel
- Loi simple de probabilité
- Mélange de gaussiennes
- Machines à vecteur support

Modélisations stochastiques
- Modèles de Markov Cachés

Modélisations de suites temporelles
- N-gram
- N-multigram

Loi de probabilité

Soit un ensemble de langues à identifier :

L = {L1, L2… Ln}

Et O une observation.

P(L/O)

En utilisant la règle de Bayes :

P(O/L)=P(L/O) / P(L)

En supposant les langues équiprobables il reste à définir :

P(O/L)

L*=argmax(P(O/L))

L

Mélanges de lois Gausiennes

$Pr(O_o|L_i)$

En utilisant l'indépendance temporelle des observations :

$$Pr(O_o|L_i) = \prod_{k=1}^{n_o} Pr(o_k|L_i)$$

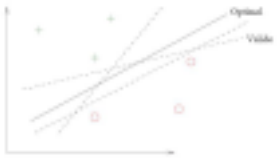
$$Pr(o_k|L_i) = N(\mu^i, \Sigma^i, Q_i^i)$$

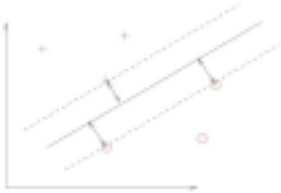
$$Pr(o_k|L_i) = \sum_{j=1}^{Q_i^i} \frac{\omega_j^i}{(2\pi)^{n/2}\sqrt{|\Sigma_j^i|}} \exp\left(-\frac{1}{2}(o_k - \mu_j^i)'(\Sigma_j^i)^{-1}(o_k - \mu_j^i)\right)$$

SVM (1/9)

Il s'agit ici de définit la frontière entre deux classes, il s'agit d'une modélisation discriminative et non générative (comme les GMM).

But : trouver un hyperplan de séparation optimal





Marge : distance du point le plus proche à l'hyperplan

SVM (2/9) : maximisation de la marge

Cas linéaire $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$

Déf. De l'hyperplan $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$

Distance d'un point au plan $d(\boldsymbol{x}) = \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$

Maximiser la marge revient à minimiser $\|\boldsymbol{w}\|$ sous contraintes

SVM (3/9) : implémentation

Un point $(\boldsymbol{x}_i, y)$ est bien classé si et seulement si : $y f(\boldsymbol{x}) > 0$

Comme le couple $(\boldsymbol{w}, b)$ est défini à un coefficient multiplicateur près, on impose : $y f(\boldsymbol{x}) \geq 1$

Rappelons que $d(\boldsymbol{x}) = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$

On obtient le problème de minimisation sous contraintes

$$\begin{cases} \min \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \forall i,\ y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq 1 \end{cases}$$

SVM (4/9) : problème dual

On passe au problème dual en introduisant des multiplicateurs de Lagrange pour chaque contrainte.

Ici on a une contrainte par exemple d'apprentissage

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \\ \forall i, \alpha_i \geq 0 \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Problème de programmation quadratique de dimension n (nombre d'exemples)

Matrice Hessienne : $(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)_{i,j}$

SVM (5/9) : Propriétés

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i$$

Seuls les α_i correspondants aux points les plus proches sont non nuls. On parle de vecteur de support.

Fonction de décision : $f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b$

Cas non séparables : On introduit des variables pour assouplir les contraintes

$$\begin{cases} \min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \forall i,\ y_i(\mathbf{w}.\boldsymbol{x}_i + b) \geq 1 - \xi_i \end{cases}$$

On pénalise le dépassement de la contrainte. Le problème dual a la même forme, on fixe juste une borne supérieure sur les α

SVM (6/9) : Cas non séparable

On introduit des variables pour assouplir les contraintes

$$\begin{cases} \min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \forall i,\ y_i(\mathbf{w}.\boldsymbol{x}_i + b) \geq 1 - \xi_i \end{cases}$$

On pénalise le dépassement de la contrainte.

Le problème dual a la même forme, on fixe juste une borne supérieure sur les α

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \\ \forall i, 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

SVM (7/9) : Espace intermédiaire

Au lieu de chercher un hyperplan dans l'espace des entrées, on passe d'abord dans un espace de représentation intermédiaire (*feature space*) de grande dimension.

$$\begin{aligned} \Phi : \mathbb{R}^d &\to \mathcal{F} \\ \boldsymbol{x} &\mapsto \Phi(\boldsymbol{x}) \end{aligned}$$

On doit donc résoudre

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j) \\ \forall i, 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Et la solution a la forme

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}) + b$$

SVM (8/9) : Fonction noyau

Le problème et sa solution ne dépendent que des produits scalaires $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$

Plutôt que de choisir la fonction non linéaire $\Phi : \mathcal{X} \to \mathcal{F}$ on choisit une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ appelée fonction noyau.

Elle représente un produit scalaire dans l'espace de représentation intermédiaire. Elle traduit donc la répartition des exemples dans cet espace.

$$k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$$

Lorsque k est bien choisie, on n'a pas besoin de calculer la représentation des exemples dans cet espace pour calculer la fonction

Permet d'utiliser des représentations non vectorielles

Le noyau matérialise une notion de proximité adaptée au problème.

SVM (9/9) : exemple de noyaux

Linéaire

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}'$$

Polynomial

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}')^d \text{ ou } (c + \boldsymbol{x} \cdot \boldsymbol{x}')^d$$

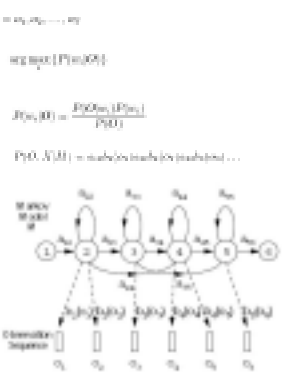
Gaussien

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\sigma}$$

Laplacien

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x} - \boldsymbol{x}'\|_1/\sigma}$$

HMM



N-gram

N-gram = sous séquence de n éléments

$P(w_i \mid w_1,\ldots,w_{i-1}) = P(w_i \mid w_{i-(n-1)},w_{i-(n-2)},\ldots,w_{i-1})$

Ex : trigrammes (n=3)

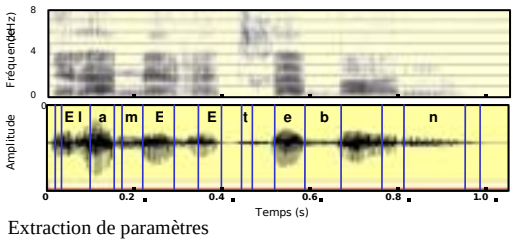
$$P(w_i \mid w_1,\ldots,w_{i-1}) = P(w_i \mid w_{i-2},w_{i-1})$$

$$P(w_i \mid w_1,\ldots,w_{i-1}) = P(w_i \mid w_{i-2},w_{i-1})$$

Apprentissage :comptabiliser les séquences.

Modélisation alternative : n-multigrammes (faire varier la taille des séquences).

Un cas d'étude : prosodie (1/5)



Extraction de paramètres

Segmentation du signal (algo. Divergence Forward-Backward [André-Obrecht 1988])

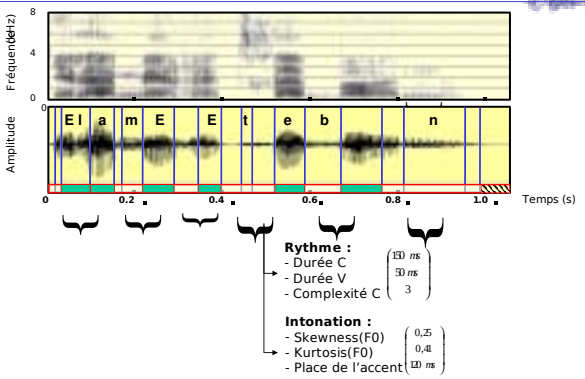
Détection d'activité vocale

Classification consonne-voyelle

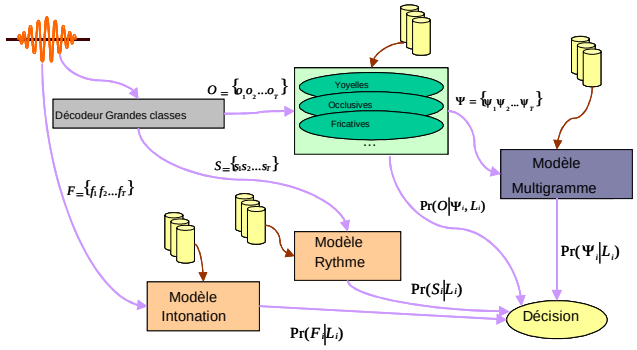
Calcul de caractéristiques

- Rythme
- Intonation

Un cas d'étude : prosodie (2/5)



Fusion d'information



Notes

Ecole RMITES
Septembre 2006

Reconnaissance du Locuteur & Indexation de Documents Audio

Jean-François Bonastre

jean-francois.bonastre@univ-avignon.fr
www.lia.univ-avignon.fr

Plan

I Pourquoi des documents audio ?
II Reconnaissance du locuteur
III Logiciels, précautions et exemples

I Pourquoi des documents audio ?

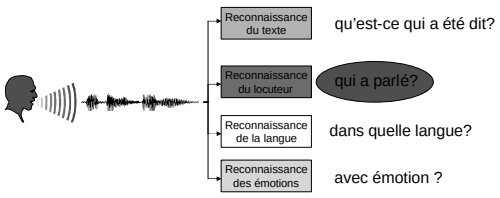
- Diversité
  - Radio, télévision (broadcast)
  - Réunion, conférence, téléphone («conversationnel »)
  - Surveillance locaux, médicale, fabrication (« événement »)
- Volume !
- Spécificité = richesse
  - Aspects conversationnels et personnels
  - Spontanéité
- Problème
  - Diversité de forme et de fond
  - Référentiel non défini
  - Complexité en terme de calcul

II Reconnaissance du locuteur

Contexte
Les différentes tâches
Cadre applicatif
Mesure de performances et contraintes
Technique de reconnaissance
Les performances
Applications
Autres approches
Problèmes majeurs

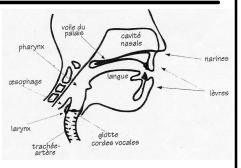
Contexte (1)

A partir d'un signal de parole, des informations de natures très différentes peuvent être extraites:

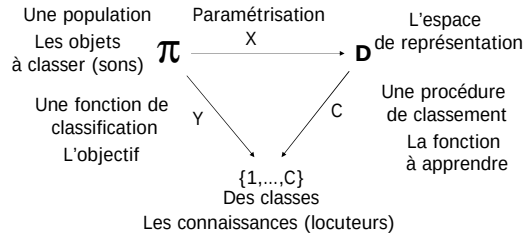


Contexte (2) Production de la parole

- Appareil vocal
  - Poumons et trachée-artère
    - production d'un souffle d'air
  - Larynx
    - vibration des cordes vocales Conduit vocal
    - pharynx, cavité buccale, cavité nasale
    - organes articulateurs
      - mâchoire, lèvres, langue
- Sources sonores résonant dans le conduit vocal
  - Vibrations quasi-périodiques des cordes vocales
  - Bruits d'écoulement d'air
  - Occlusions rapides



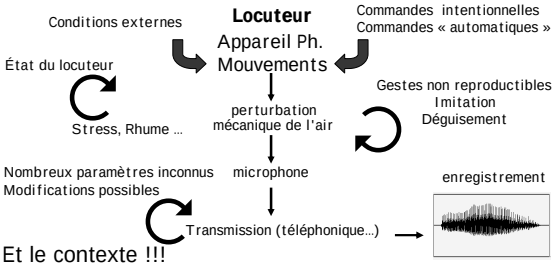
Contexte (3)
Un problème de classification automatique



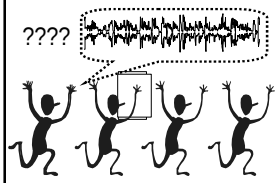
Contexte (4)
Caractéristiques du locuteur

- Les humains utilisent différentes sources d'information
- Pas de caractères exclusifs pour l'identité d'un locuteur
- Types d'informations (avec recouvrement)
  - Anatomie de l'appareil phonatoire
  - Prosodie : rythme, vitesse, intonation, volume, modulation
  - Phonétique : cibles phonémiques
  - Accents régionaux
  - Linguistique : syntaxe, grammaire, sémantique
  - Diction, prononciation
  - Emotionnelle, pathologique

Contexte (5)
Information captée



Les différentes tâches (1)



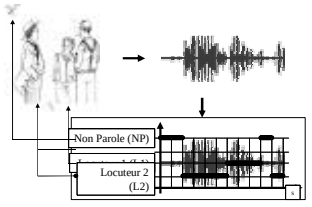
Un message vocal

Identification
« Qui parmi ces personnes
a prononcé le message ? »

- Les locuteurs (client) doivent être connus du système
- Les locuteurs sont coopératifs
- La base peut être fermée (id. = 1 parmi n) ou ouverte
- Le contenu du message peut être contrôlé ou non

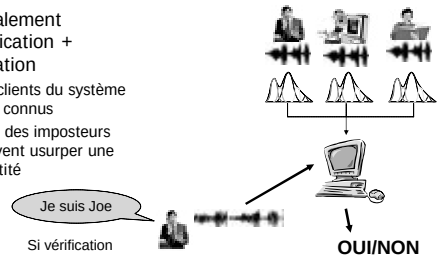
Les différentes tâches (2)
Segmentation et suivi

- « Qui parle et quand ?»
- Segmentation
  - Pas d'information a priori sur les locuteurs
  - Trouver le nombre de locuteurs
  - Trouver les tours de parole associés à chacun
- Suivi
  - Locuteurs connus a priori
  - Trouver leur tours de parole



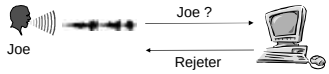
Cadre applicatif

- Généralement Identification + vérification
  - Les clients du système sont connus
  - Mais des imposteurs peuvent usurper une identité

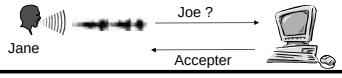


Mesure de performances et contraintes (1)
2 types d'erreurs en vérification

- Le client est rejeté alors que l'identité proposée est la sienne
  - Joe prétend être Joe mais le système d'authentification le rejette = **Faux Rejet (FR)** **ou Miss probability**

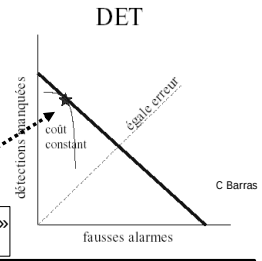


- Le client est accepté alors que l'identité proposée n'est pas la sienne
  - Jane prétend être Joe mais le système d'authentification l'accepte = **Fausse Acceptation (FA)**



Mesure de performances et contraintes (2)

- Identification = % de tests réussis
- Vérification
  - Courbe DET (Martin) = ROC + échelle en déviation de loi normale
  - EER -> FA=FR (ou miss)
  - CDF -> un coût de fonctionnement donné, défini par une fonction de FA et FR



Un seuil = Décision « hard »
= 1 Point (1 coût)

Mesure de performances et contraintes (3)
Dépendance aux messages

- Systèmes dépendants du texte
  - Messages fixés (mots de passe, uniques ou personnalisés)
  - Messages promptés
  - → Meilleures performances (dues maj. à la diminution de la variabilité)
- Systèmes indépendants du texte
- Dépendance à la durée des messages
  - *Lors de :*
    - *l'enrôlement des clients (facteur principal)*
    - *des tests*

Mesure de performances et contraintes (4)
Dépendance à l'environnement

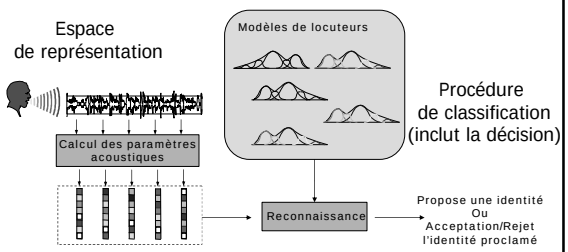
☺ Environnement contrôlé :
- Un local
- Un micro/un canal
- Pas de stress

→ Bonne qualité et CONSTANCE

☹ Environnement « libre »
- Multiples lieux
- Multiples « micros »/canaux
- Bruits, stress, multiples locuteurs

→ Qualité pouvant être mauvaise et VARIABLE

Technique de reconnaissance (1)
Schéma général



Technique de reconnaissance (2)
Le rapport d'hypothèse Bayésien (a)

- Etant donné un signal de test *X*, et une identité *I*
  - H_0 : *X* a été généré par le locuteur d 'identité *I (OK)*
  - H_1 : *X* a été généré par un locuteur imposteur *(Imposture)*

- Test statistique d 'hypothèses (classique) qui consiste à comparer

$$P_0(X)=P(H_0|X) \text{ et } P(X)=P(H_1|X)$$

- En fait, on va comparer les log.

$$\log P_0(X) - \log P_1(X) \underset{reject}{\overset{accept}{\gtrless}} Seuil$$

Technique de reconnaissance (3)
Le rapport d'hypothèse Bayésien (b)

- En log, après Bayes, et en intégrant les apriori dans le seuil

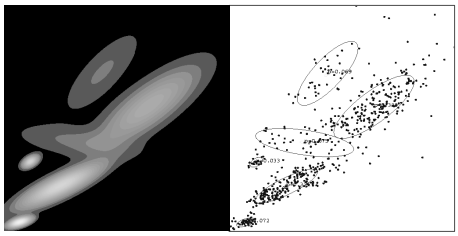
$$\log(p(X|H_0)) - \log(p(X|H_1)) \geq Seuil$$

- Modélisation acoustique d'un locuteur client
  - On a des données « représentatives » du client
  - Approche statistique : modèle génératif
- Modélisation de l 'imposture
  - Cohorte : on modélise l'imposture par un ensemble de locuteurs
  - modèle « du monde » : on modélise l'imposture par un modèle génératif appris à partir de données provenant de n locuteurs, en général n'incluant pas les utilisateurs potentiels

Technique de reconnaissance (4)
Modélisation d'une classe acoustique

- Créer des modèles de H_0 et H_1 à partir de données les représentant
  - Données manquantes -> modèle paramétrique
- Calcul de l'appartenance des données à une classe par le principe de la vraisemblance
- Méthode majoritaire
  - formalisme Markovien
  - Mixture de Gaussiennes (GMM) = 1 état
    - Estimateur de densité de probabilité par une moyenne pondérée de lois gaussiennes

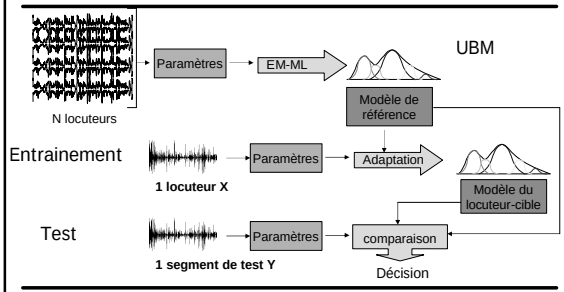
Technique de reconnaissance (5)
Les mixtures de gaussiennes - exemple



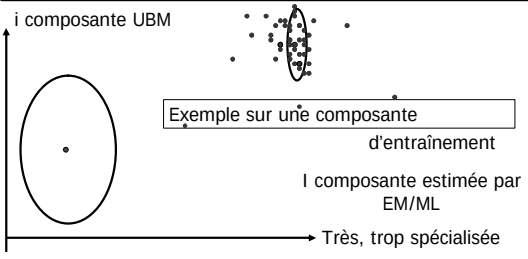
Technique de reconnaissance (6)
Les GMM en RAL (a)

- Introduit par Reynolds
- Etat de l'art (indépendant du texte)
- Modèles de clients dérivés par MAP
  - D'un modèle générique (world)
  - Uniquement les moyennes
- Pas mal de recettes peu expliquées ou explicables
  - Matrices de covariances diagonales
  - Un grand nombre de composantes (~2000)
  - Initialisation
  - Contrôle de la variance
- Modèle du monde joue un rôle très important !!

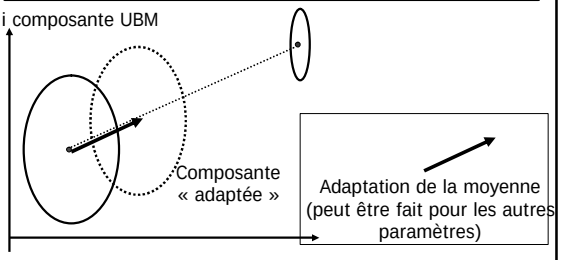
Technique de reconnaissance (7)
Les GMM en RAL (b)



Technique de reconnaissance (8)
Les GMM en RAL (c)



Technique de reconnaissance (9)
Les GMM en RAL (d)



Les performances

- Une « référence » actuellement : évaluations NIST
- Bases de données disponibles :
  - Locuteurs américains (hommes & femmes)
  - Conversations téléphoniques réelles
    - coopération implicite
  - Diverses sources de variabilités
  - Tests d'impostures simulés : tests croisés
- Protocoles standards pour la comparaison des systèmes

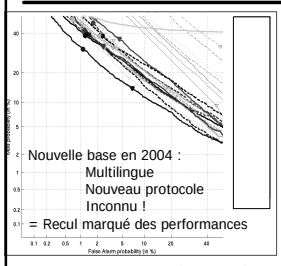
Les performances (1)

- Identification sans rejet :
  - moins de 1% d'erreurs en parole préparée « studio », parmi 630 locuteurs (6s enrol., 3s test)
  - 40% d'erreurs en parole téléphonique spontanée
- Vérification/détection, pourcentage d'égale erreur
  - 0,1% parole propre, prompt fixé
  - 1% parole téléphonique, prompt fixé
  - 10% parole téléphonique spontanée
    - -> 5% NIST 2005
    - Moins de 1% avec env. 30 minutes
  - 25% parole radio bruitée spontanée
- Importance de la durée d'apprentissage et de test

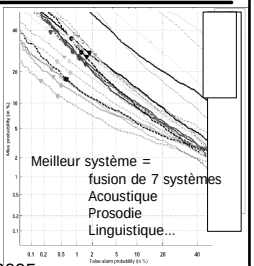
Les performances (2)
Les évaluations NIST

- Depuis 1996
- Evaluation sur de la parole téléphonique
- Tâches orientées par le sponsor....
- Inscription gratuite
  - Donne accès aux données
- Un protocole solide
  - Détection de locuteur
  - Règles précises
  - Evaluation "en aveugle"
- Grand nombre de test
  - Env 5000 clients
  - Env 45000 "imposteurs"
- Tests
  - Sans connaissance des autres clients
  - Sans normaliation inter-tests

Les performances (3)
NIST04 et NIST05 (one side - one side)



Nist 2004 results.v6



2005 NIST_SRE05-a/pdf/nist_nist05.pdf

Les performances (4)
Interprétation des performances

- Performances d'un système de RAL fortement dépendantes :
  - des locuteurs de la base et de leur nombre
  - des conditions d'enregistrement & canaux de transmission

Résultats peu transposables d'une application à l'autre

- Validité statistique
  « Règle des 30 » = 30 exemples d'une erreur !
  1% EER -> 3 000 tests « client » (par client ?)
  0.001% EER -> 3 000 000 tests « client » !

Applications (1)

- Sécurité
  - contrôle d'accès (en complément d'un code, d'un badge)
    - banques, voitures, entreprises...
    - consultation de compte bancaire par téléphone...
- Police criminelle (identification de suspects) ?
  - filtrage de voix suspectes (avec validation humaine)
  - ...pas assez fiable pour utiliser comme preuve !
    - Position de l'AFCP
- Transcription automatique
  - adaptation des modèles acoustiques à la voix du locuteur
- Indexation multimédia
  - indexation par locuteur

Applications (2)
Utilisation en conditions réelles

- Au niveau théorique :
  - On ne sait pas modéliser correctement le modèle de rejet en vérification (connaissance *a priori* des imposteurs)
  - Approche statistique : que modélise-t-on réellement ?
- Evaluations réalisées :
  - Locuteurs coopératifs ou neutre
  - Résistance aux vraies impostures inconnue
  - Pas encore d'évaluation avec des imitateurs
  - Conditions environnementales connues
- Non transposition des résultats

Autres approches...

- HMM = extension des GMM
  - Peu d'avantages, à part reconnaître le texte
- Classifieurs classiques (Réseaux de neurones, polynomiaux)
- DTW pour des systèmes (très) dépendant du texte
- Support Vector Machines
  - Approche discriminante
  - Complexité dans le noyau (noyau polynomial de Campbell)
  - Assez surprenant (1 vs 1000)
  - Proche des GMM à partir de 2 minutes de test, meilleur ensuite
- En fait, les GMM avec adaptation MAP des moyennes sont discriminants (voir Mariethoz)

Problèmes majeurs

- estimation de la 2ème hypothèse
- variabilité due au locuteur
  - émotion, fatigue, stress
- conditions d'enregistrement variables
  - microphone, bruit ambiant
- conditions de transmission variables
  - canal téléphonique
- nouveaux problèmes
  - GSM : codage, bruit évoluant au cours du temps

Logiciel, précautions et exemples

Un système complet (1)
ALIZE/LIA_SpkDet

- Tout le logiciel présenté est en « libre »
  - Documenté
  - Maintenu
  - Disponible en ligne
- Basé sur ALIZE (LGPL), un toolkit intégrant les fonctionnalités « bas niveau »
  - LLK
  - EM
  - Viterbi
- http://www.lia.univ-avignon.fr/heberges/ALIZE/

Un système complet (2)
Mais aussi

- Demos, mode distant
- LIA_Seg : Segmentation en locuteur
  - Broadcast news, conversations, meetings
  - Testé sur NIST RT et ESTER
- LIA_AcouSeg : Segmentation en classes acoustiques
- Speeral(s) : Transcription de la parole (*)
- Recherche d'information (*)

(*) disponibles sous conditions

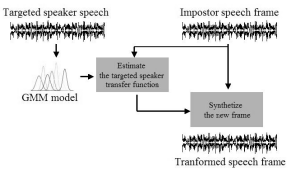
Précautions

- Locuteurs coopératifs !
- De nombreux facteurs incontrôlés
- Déterminer avec certitude si la ressemblance entre deux enregistrements provient du locuteur ou d'autres facteurs n'est pas possible à ce jour

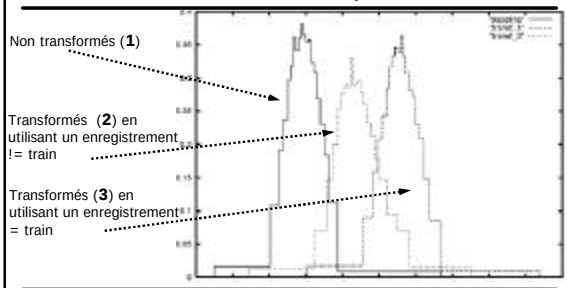
JEP 04 : **J.F. Bonastre, F. Bimbot, L.J. Boe, J. P. Campbell, D. A. Reynolds, I. Magrin-Chagnolleau**, Authentification des personnes par leur voix : Un necessaire devoir de précaution, 2004 Journées d'Etude de la Parole, Fèz (Maroc)

Exemple : transformation de la voix d'un imposteur pour tromper un système

- Si on connaît
  - La méthode utilisée
  - Un exemple de la voix d'une personne cible
- Est-il possible de tromper le système en question ? i.e. le système reconnaît l'imposteur comme étant la cible
- Méthode simple, basée sur les GMM



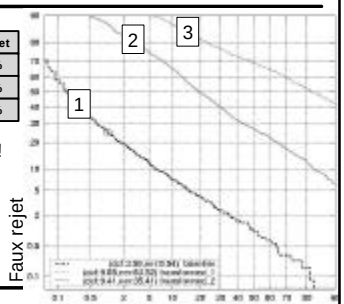
Résultats (1)
Distribution des scores imposteurs



Résultats (2)
Det et taux d'erreurs

	F. Accept.	Faux Rejet
1 - baseline	0.88 %	27.45 %
2 - !=	49.72 %	27.45 %
3 - =	96.55 %	27.45 %

De 0.88 % à 96.55 % !!
(seuil identique)



Exemples

Original d'Alain Passerel

Driss Redragui Fabrice Drouelle Franck Mathevon Joel Collado

	7396			8049		
NCFB_A	-1.94	4.84		0.46	5.47	

Démo de recherche information / navigation

Transcription automatique de la parole

pour l'analyse multimédia

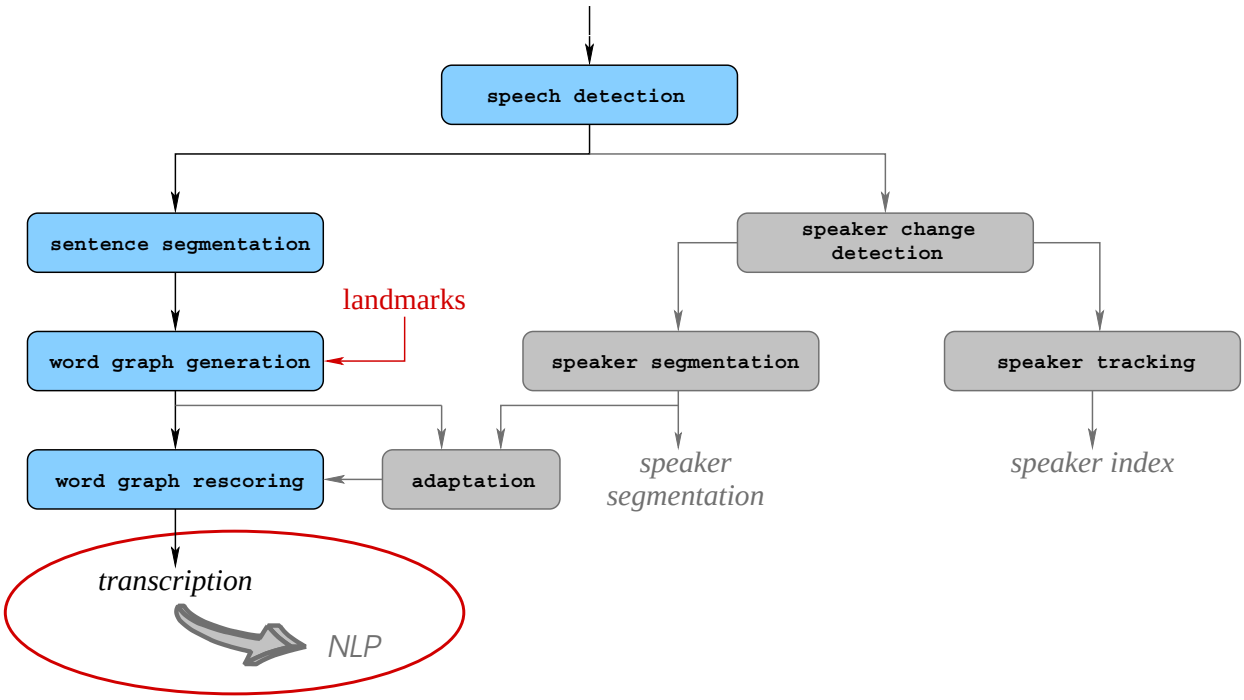
Guillaume Gravier

guillaume.gravier@irisa.fr

IRISA, Équipe Modélisation et Expérimentation pour le traitement de l'Information et des Signaux Sonores

http://www.irisa.fr/metiss

Contexte



Plan de l'exposé

1. Quelques généralités
2. Transcription automatique
   - Modélisation acoustique
   - Grammaires et modèles de langage
   - Intégration modèles acoustique et de langage
   - Décodage
3. Évaluation des performances
4. Segmentation du flux sonore
5. Couplage transcription / TALN
6. Outils et ressources libres
7. Conclusion

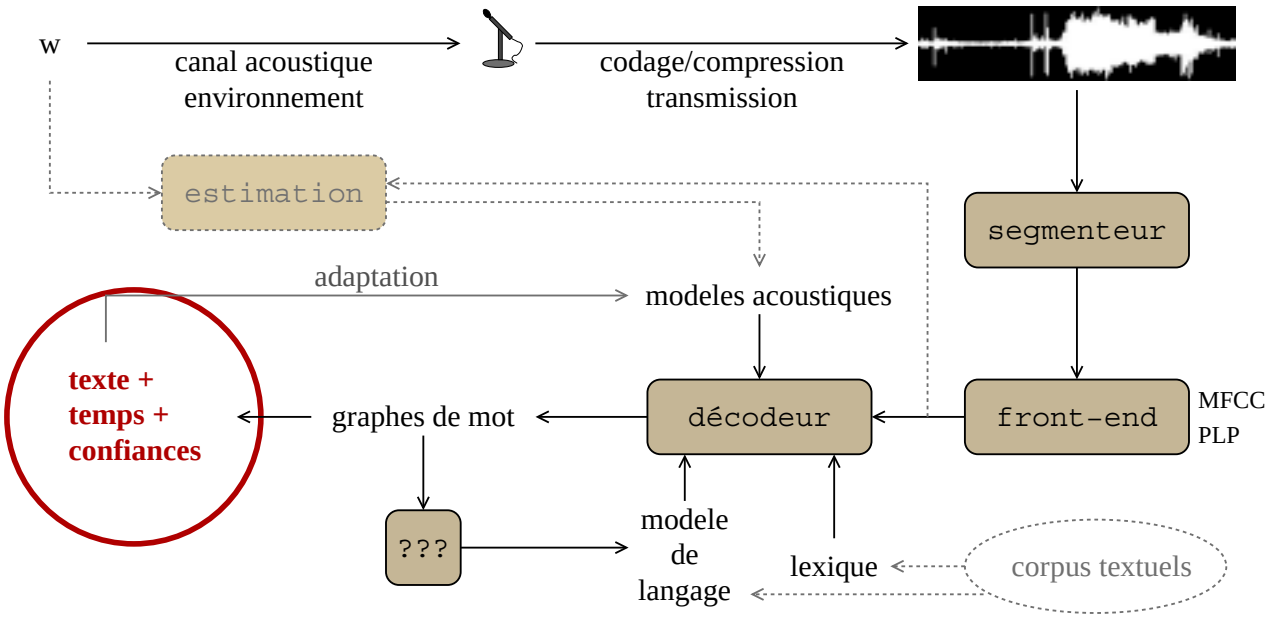
Pourquoi transcrire?

La transcription n'est pas une fin en soi, mais une description du document nécessaire à son analyse, à son traitement par des methodes d'analyse de texte (RI, structuration, indexation, etc...), voire à sa compréhension.

La parole possède un fort contenu sémantique!!!!!!

TRANSCRIRE $\neq$ COMPRENDRE

Architecture d'un système



Modes de fonctionnement

locuteur

- un seul locuteur
- plusieurs locuteurs
- n'importe quel locuteur

niveau de difficulté

type de parole

- mots isolés
- mots connectés
- détection de mots clés
- parole continue

Variabilité

La parole est soumise à plusieurs sources de variabilités qui ont un impact fort sur la transcription :

- variabilité inter- et intra- locuteurs
  - → un locuteur < indépendant du locuteur — VTLN, VNorm, SAT
- variabilité des conditions acoustiques
  - → qualité des capteurs — CMN, VNorm
  - → studio ou téléphone — modèles spécifi ques
  - → bruits ambients — fi ltrage, CDCN
- variabilité dans la grammaire et le vocabulaire
  - → taille et choix du vocabulaire — ML
  - → contraintes de la grammaire

$\Rightarrow$ approche statistique du problème

Formulation statistique

$$\widehat{w} = \arg\max_{w} \; p(y|w) \; P[w]$$

modèle acoustique — modèle de langage

- modèles de Markov cachés
- modèles multiflux (audiovisuels)
- modèles de segments/trajectoires
- réseaux bayésiens dynamiques

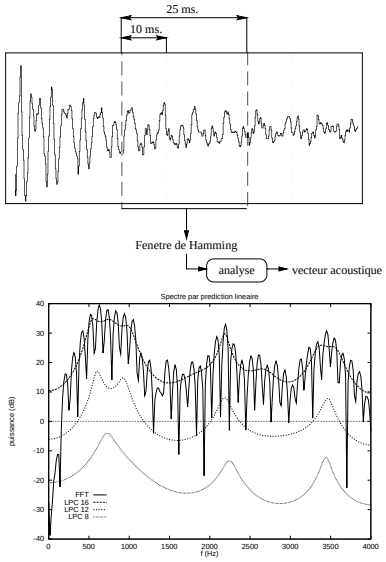
- grammaires stochastiques
- modèles N-grammes (et ses variantes)
- modèles N-classes
- modèles caches, trigger, . . .

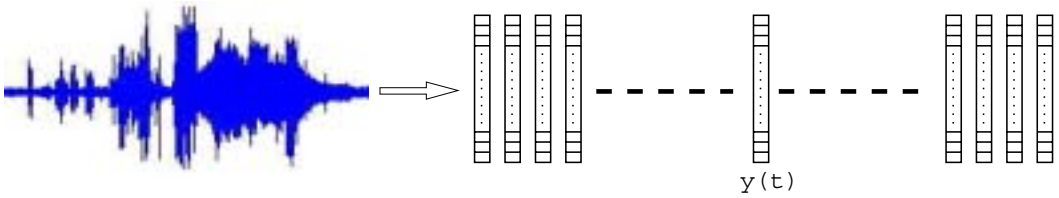
En pratique,

$$\widehat{w} = \arg\max_{w} \ln\left(p(y|w)\right) + \beta \ln\left(P[w]\right) + p|w| \; .$$

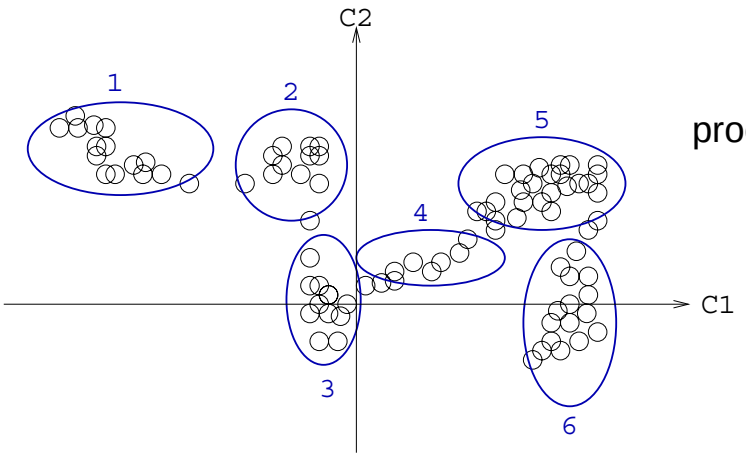
Représentation du signal

- Analyse à court terme
- Représentation de l'enveloppe spectrale
  - Mel Frequency Cepstral Coeffs
  - (Perceptual) Linear Prediction Cepstral Coeffs
- Dynamique de l'enveloppe (Δ et $\Delta\Delta$)
- Normalisations :
  - soustraction de la moyenne à long terme (CMN)
  - normalisation de la variance





Modèle de Markov cachés (MMC)



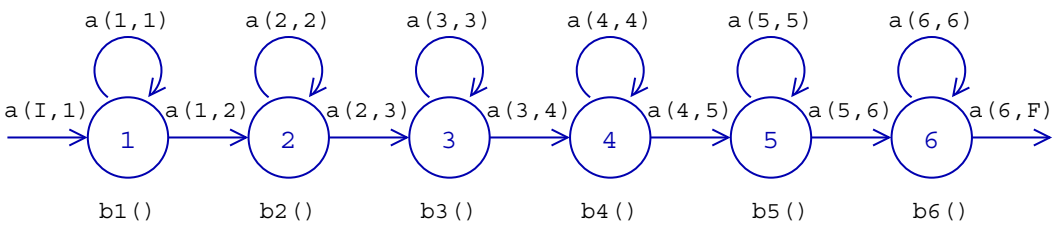
processus localement stationnaire

⇓

modèle à état

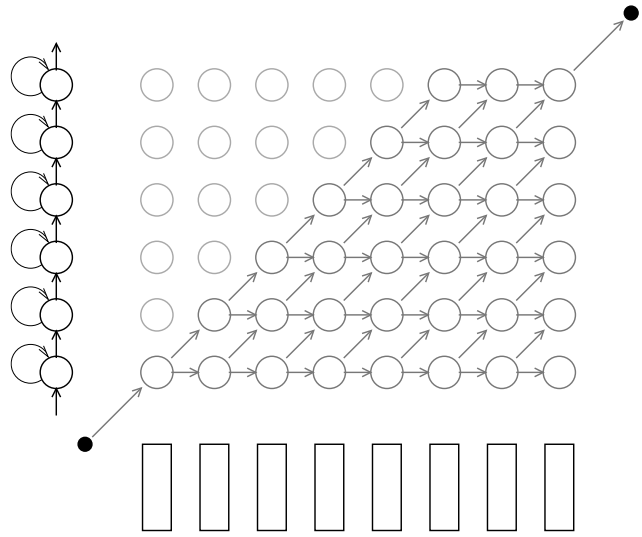
⇓

modèle de Markov cachés



En pratique, b() = modèle de mélange de gaussiennes.

Algorithme de Viterbi



Approximation de Viterbi:

$$p(y|w) = \sum_s p(s|w)\, p(y|s,w)$$
$$\simeq \max_s p(s|w)\, p(y|s,w)$$

⇓

recherche du meilleur chemin dans un graphe orienté valué

$$S(j,t) = \max_i S(i,t-1) + \ln(a_{ij}) + \ln b_j(y_t)$$

Mais il existe aussi d'autres algorithmes que Viterbi, par exemple A*.

Estimation des paramètres

- Paramètres
  - probabiliés de transitions
  - poids, moyennes et variances des lois conditionnelles

- Estimation des paramètres d'un modèle à partir d'exemples (nombreux) selon un critère à optimiser
  - Maximum de vraisemblance
  - Maximum a posteriori (adaptation)
  - Maximum d'information mutuelle
  - Erreur de classification minimum

Nécessite une grande collection de données d'apprentissage!!!!!

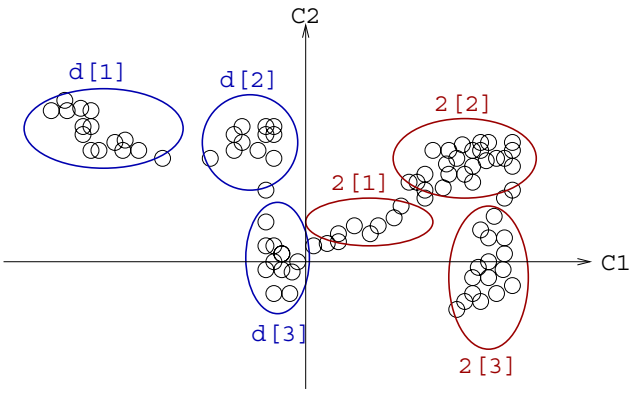
Choix des unités

- Petit vocabulaire (< 100 mots)
  - un modèle par mot (nombre d'état dépendant de la longueur du mot)
- Moyen et grand vocabulaire (> 100 mots)
  - impossible d'apprendre un modèle pour chaque mot
  - décomposition des mots en une suite d'unités élémentaires
    - phone, syllabe
    - modèles contextuels pour la coarticulation (⇒ partage de paramètres)
  - lexique de prononciation

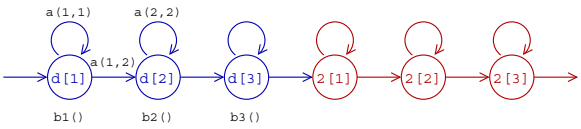
un	U
deux	d 2
...	...
six	s i
six	s i s
six	s i z
...	...

un	*-U +*
deux	*-d+2 d-2+*
...	...
six	*-s+i s-i+*
six	-*s+i s-i+s i-s+*
six	*-s+i s-i+z i-z+*
...	...

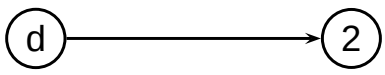
Choix des unités (suite)



de = d + 2



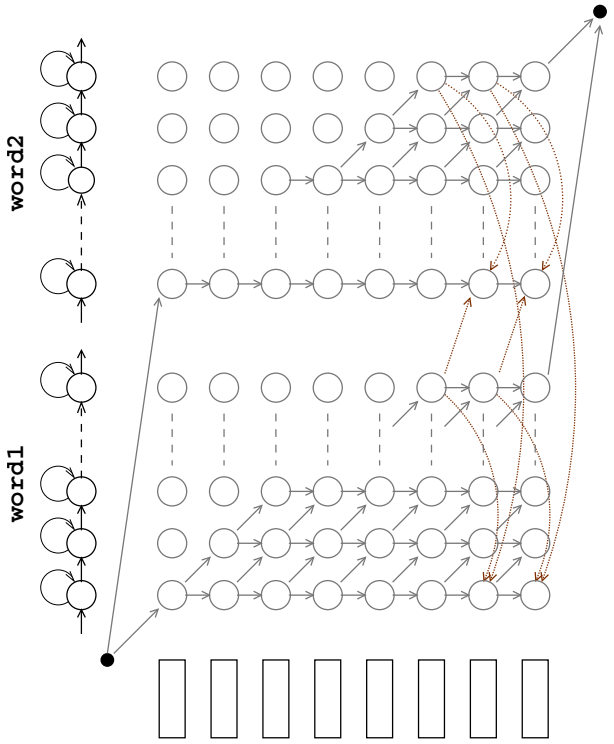
Le modèle est le résultat de la composition du graphe lexical pour le mot 'de' , soit



avec les graphes des MMC des deux phones [d] et [2].

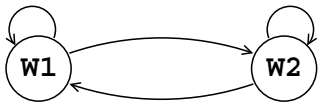
En pratique, les phones sont représentés par un modèle à 3 états, éventuellement avec saut.

Extension à la parole continue



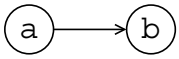
Recherche du meilleur chemin dans un graphe résultant de la composition

1. du graphe représentant la grammaire

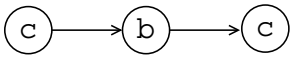


2. des graphes lexicaux

W1



W2

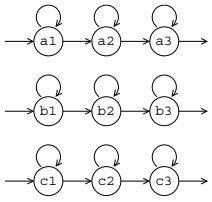


3. et des graphe acoustique des unités de base

a

b

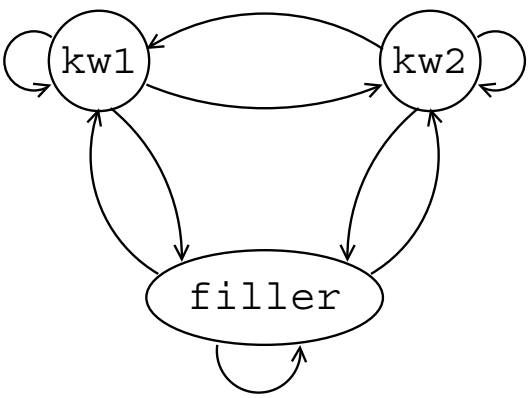
c



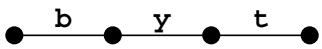
Détection de mots clés

La détection de mots-clés suit le même principe de recherche dans un graphe représentant la grammaire

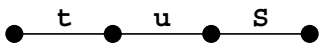
< $keyword | $filler >



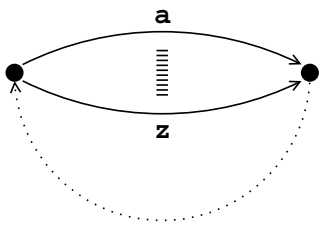
mot-clé 1



mot-clé 2



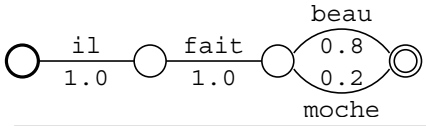
filler



Grammaires et modèles de langage

Objectif : contraindre les séquences de mots w et (éventuellement) assigner une probabilité $P[w]$.

- Grammaire (stochastique)
  - séquence de mots acceptables
  - détection de mots clés
  - grammaire “context-free”




```
$mc1 = but; $mc2 = touche; ...
$fi ller = a | b | s | ... | z;
< $mc1 | $mc2 | ... | $fi ller >
```

- Modèles de langage statistiques
  - distribution de probabilités sur les séquences de mots
  - pas de contraintes sur les séquences de mots
  - désambiguïsation des homophones : “Il était une fois l’été.”

Modèles de langage statistiques

Il est impossible d’estimer toutes les probabilités du produit

$$P[w_1 \dots w_N] = \prod_{n=1}^{N} P[w_n | \underbrace{w_1 \dots w_{n-1}}_{\text{historique, h}}]$$

$\Rightarrow$ modèles portant sur un historique restreint, $f(h)$.

- Modèle N-gramme :
  - limiter l’historique au $N-1$ mots précédents (modèle de Markov)

$$P[w_n|w_1 \dots w_{n-1}] = P[w_n|w_{n-N+1} \dots w_{n-1}]$$

  - prise en compte indirecte de la syntaxe (dans une certaine mesure)
  - en pratique, $N \in [2, 4]$ la plupart du temps (dépendance courte)
- Modèles cache/trigger :
  - favoriser des mots en fonction du passé, *e.g.* favoriser ’coup franc’ si les mots ’but’, ’corner’, . . . apparaissent dans l’historique (peu utilisés en pratique)

Estimation des modèles N-grammes

Estimation des probabilités par comptage sur un (grand) corpus d'apprentissage *représentatif de l'application visée*, soit pour un trigramme

$$P[w_k|w_i\ w_j] = \frac{C(w_i, w_j, w_k)}{C(w_i, w_j)}$$ ⇒ beaucoup de probabilités nulles !!!!

→ interpolation : interpoler les N-grammes de tous ordres

$$P[w_k|w_i\ w_j] = \lambda_1 P[w_k|w_i\ w_j] + \lambda_2 P[w_k|w_j] + \lambda_3 P[w_k]$$

→ décomptage et repli : enlever de la masse de probabilités aux événements fréquents, et la redistribuer sur les événements peu fréquents en se basant sur les N-grammes d'ordre inférieur.

$$P[w_k|w_i\ w_j] = \begin{cases} \frac{C^*(w_i,w_j,w_k)}{C(w_i,w_j)} & \text{si} \quad C(w_i, w_j, w_k) > K \\ \lambda(w_i, w_j) P[w_k|w_j] & \text{sinon} \end{cases}$$

Le modèle N-classes

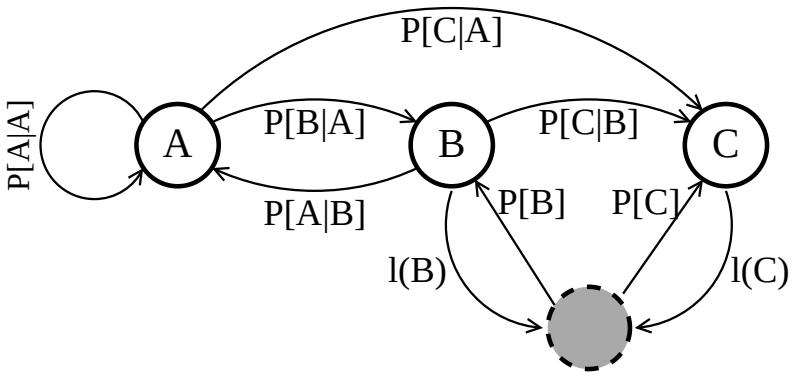
Associer les mots du vocabulaire à des classes, le modèle N-gramme portant sur les classes

$$P[w_k|w_i\ w_j] = P[c_k|c_i\ c_j]\ P[w_k|c_k]$$

- Type de classes
  - classes génériques : nom propre, ville, jour, mois, année, durée, ...
  - classes morpho-syntaxiques : nom masculin pluriel, adverbe, etc...
  - classes statistiques
- Avantages/Inconvénients
  - ∧ meilleure estimation des probabilités N-grammes (car moins de classes que de mots)
  - ∧ plus grande généricitée par rapport aux *entités nommées*
  - ∨ pas aussi bien qu'un N-gramme sur les mots si l'on dispose d'un corpus d'apprentissage suffi samment grand

Décodage avec un modèle N-gramme

Bonne nouvelle : le modèle N-gramme avec repli peut se représenter sous la forme d'un graphe orienté valué.

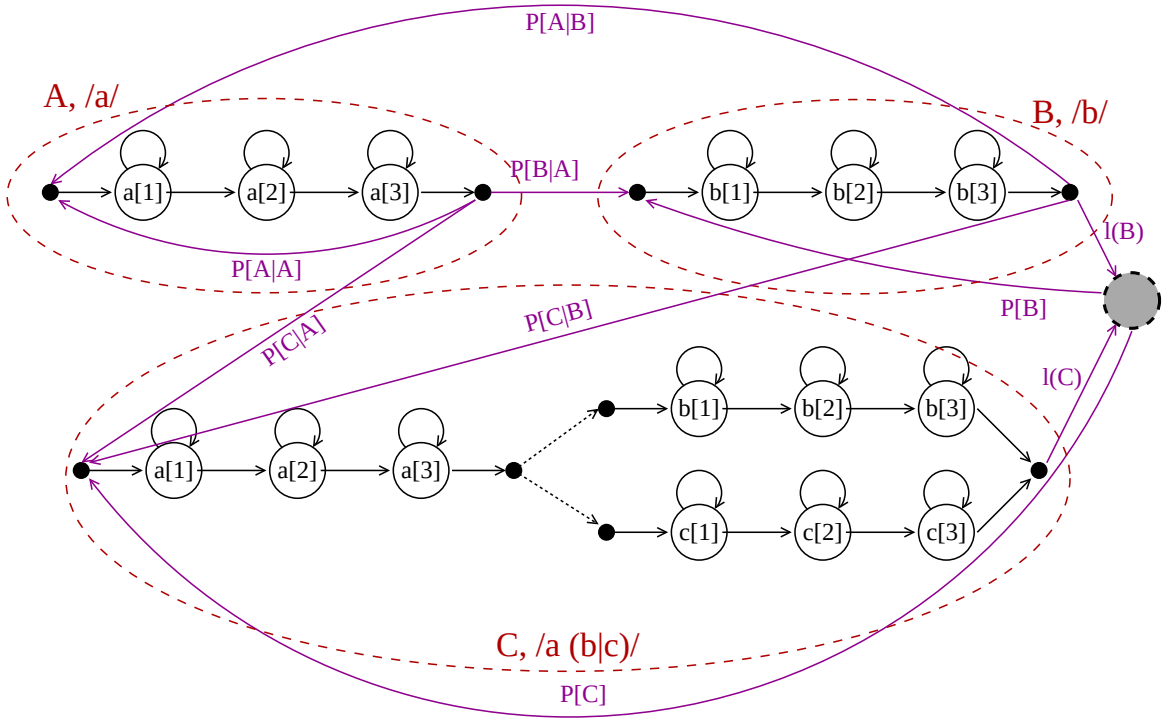


Quelques ordres de grandeurs (pour un vocabulaire de 65k mots) :

- #bigrams=8.5M, #trigrams=7.3M
- #nodes=8M, #arcs=24M

⇒ décodage dans le graphe résultant de la composition du modèle de langage, du lexique et des MMC

Décodage avec un modèle N-gramme (suite)



La taille du graphe de décodage est énorme!

Stratégies de décodage

La recherche exhaustive du meilleure chemin dans le graphe de décodage est impossible $\Rightarrow$ élagage des mauvais chemins

$$S(i,t) < -\alpha + \max_j S(j,t)$$

Stratégies de décodage :

1. construction statique du graphe de décodage
   - ○ composition, determinisation et minimisation
   - $\wedge$ décodage extrêmement rapide
   - $\vee$ construction de l'automate déterministe minimal complexe
2. construction dynamique du graphe de décodage
   - ○ lexique arborescent
   - $\wedge$ mise en œuvre (relativement) facile
   - $\wedge$ ressources mémoires limitées
   - $\vee$ décodage +/- lent suivant l'effi cacité des structures de données

Temps de calcul typique pour un système 65k, MMC contextuels : 5-20 x RT.

Représentation des hypothèses de phrases

Un système de transcription permet de trouver non seulement la meilleure hypothèse de phrase mais aussi

- ○ un ensemble d'hypothèses concurrentes représentées de manière plus ou moins compacte
- ○ des mesures de confiances associées aux mots reconnus

$\Rightarrow$ une meilleure interface pour le TALN

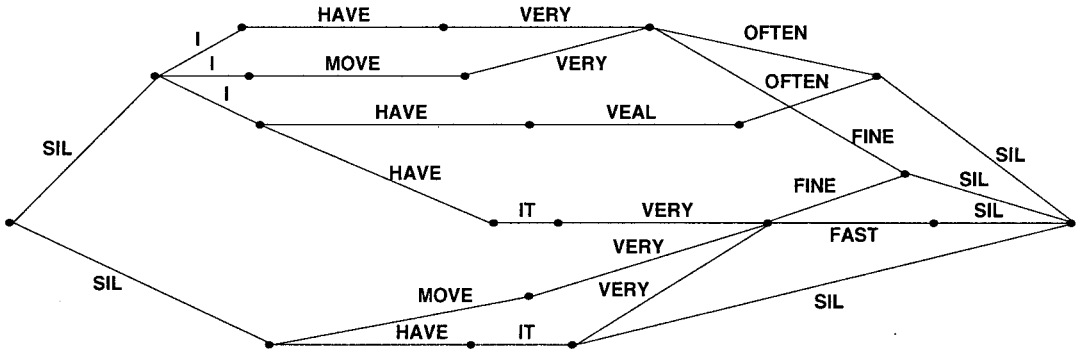
- Liste des N-meilleures hypothèses `[play]`

```
<s>     c'   est     le journal de franck mathevon </s>
<s> -   si   -       le journal de franck mathevon </s>
<s> -   ici  -       le journal de franck mathevon </s>
<s> -   -    ces     le journal de franck mathevon </s>
<s> -   et   si      le journal de franck mathevon </s>
<s> -   -    laissez le journal de franck mathevon </s>
<s> -   ses  -       le journal de franck mathevon </s>
<s> et  c'   est     le journal de franck mathevon </s>
```

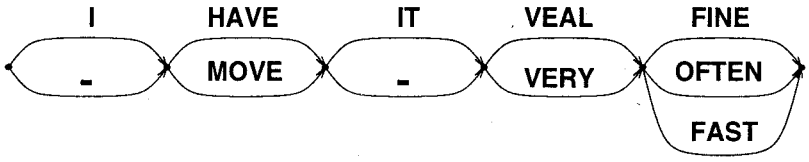
Représentation des hypothèses de phrases (suite

- Graphe de mots et réseaux de confusions (a.k.a. saucisses)

(a) Input lattice ("SIL" marks pauses)



(b) Multiple alignment ("-" marks deletions)



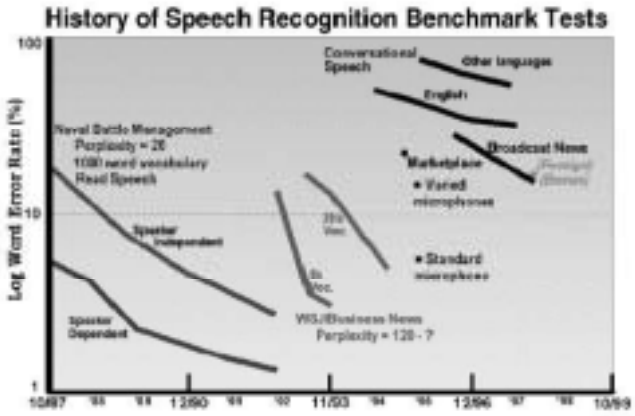
Évaluation des performances

- Évaluation par comparaison à une référence

```
REF: quatre millions d' électeurs envolés en ****** sept ans
HYP: quatre millions d' électeurs ******* en voyant ces  temps
```

⇒ 2 confusions, 1 insertion, 1 omission

- Facteurs à prendre en compte
  - diffi culté de la tâche
  - taille et diffi culté du vocabulaire
  - conditions acoustiques
  - type de parole
  - données apprentissage (1 ans ≃ +4%)

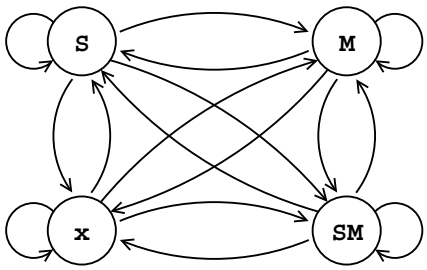


- Typologie des erreurs
  - "dérapages" liés à des conditions acoustiques particulières
  - fautes d'orthographe (accord, participe passé / infi nitif)
  - insertions/omissions de mots outils courts (le, la, etc.)

Segmentation du flux sonore

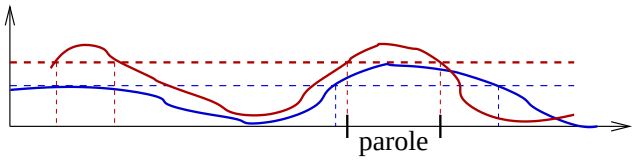
- Détection parole(/musique)

MMC



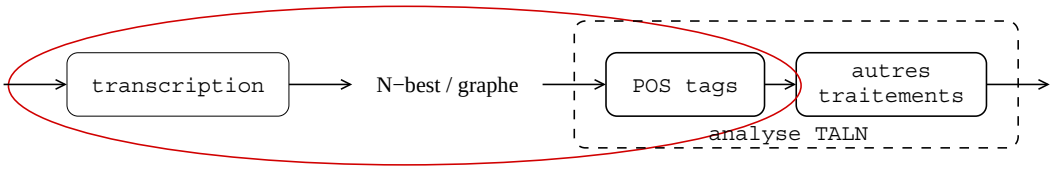
Classification

- Segmentation a priori ou BIC
- Descripteurs : énergie, modulation à 4 Hz, ZCR, centroïde spectral, “rolloff” spectral, . . .
- Classifi eur : heuristique, GMM, MLP, SVM, . . .



- Segmentation locuteur et/ou en conditions acoustiques $\Rightarrow$ BIC
- Segmentation en “phrases”
  - détection des pauses “longues” basée sur l’énergie
  - décodage acoustico-phonétique : `<$phone | $filler>`
  - segmentation intégrée au décodeur (taux de fi llers dans les meilleurs hypothèses)

Transcription et TALN



1. étiquetage morpho-syntaxique de chaque hypothèse selon le modèle N-classes, $\widehat{c} = \arg\max_c P[c]P[w|c]$, *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. révaluation selon le score linguistico-acoustico-syntaxique

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

```
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
```

	%WER		%SER	
	No POS	POS	No POS	POS
Sys. 1	20.0	19.2	61.9	59.7
Sys. 2	13.1	12.4	51.3	41.8

Outils et ressources

- Traitement du signal
  - SPro, http://www.irisa.fr/metiss/guig/spro
  - SPTK, http://kt-lab.ics.nitech.ac.jp/ tokuda/SPTK
  - HTK 3.x, http://htk.eng.cam.ac.uk
- Segmentation, partitionnement
  - audioseg, http://gforge.inria.fr/audioseg (prochainement)
  - Alize, http://www.lia.univ-avignon.fr/heberges/ALIZE
- Logiciels pour la transcription
  - CMU Sphinx, http://cmusphinx.sourceforge.net
  - HTK 3.x, http://htk.eng.cam.ac.uk
  - Sirocco, http://gforge.inria.fr/sirocco
- Logiciels pour les modèles de langages
  - SRI LM, http://www.speech.sri.com/projects/srilm
  - CMU LM, http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html
  - HTK 3.x, http://htk.eng.cam.ac.uk
- Ressources : bases de données, lexiques, etc.
  - Linguistic Data Consortium, http://www.ldc.upenn.edu
  - ELRA/ELDA, http://www.elda.org
  - ESTER Repository, http://www.afcp-parole.org/ester
  - ISIP Resources, http://www.cavs.msstate.edu/hse/ies/projects/speech/index.html

Conclusions

Ce qu'il faut retenir

1. La transcription permet un accès sémantique à un document
2. Bonnes performances dans des conditions contrôlées, proches des données de développement ($\Rightarrow$ il n'existe pas de système prêt à l'emploi)
3. Beaucoup d'interactions possibles avec d'autres domaines du multimédia :
   - reconnaissance audiovisuelle
   - couplage entre transcription et TALN
   - intégration du flux textuel (transcrit) dans l'analyse multimodale
   - mais aussi avec les aspects locuteurs

Quel avenir?

- vers des systèmes plus robustes : aux conditions d'utilisation, au type de parole
- intégration de connaissance provenant d'autres analyse, voire d'autres médias
- couplage avec les autres composantes d'un système multimédia

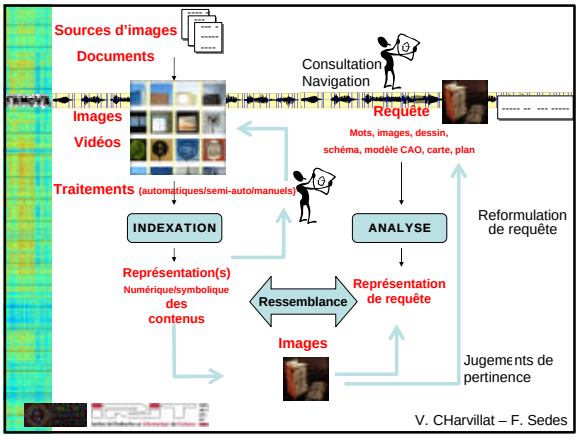
Notes

Segmentation et thématisation de séquences vidéo
Similarité de contenus visuels

Ph. Joly

Sommaire

- Similarité entre documents audiovisuels par le contenu
  - Similarité visuelle
  - Distance, mesure de similarité
  - Analyse de collections ou d'un flux
- Similarité entre documents audiovisuels par la structure
  - Segmentation
    - Plans
    - Macrosegments
  - Organisation structurelle
  - Similarité structurelle



V. CHarvillat – F. Sedes

Similarité entre documents audiovisuels par le contenu

- Comparaison de caractéristiques vidéo
  - Couleur
  - Texture
  - Formes, régions, blobs
  - Mouvement (d'objet, de caméra)
  - Structure
- Comparaison de caractéristiques audio
  - MFCC
  - Energie à 4 Hz, …

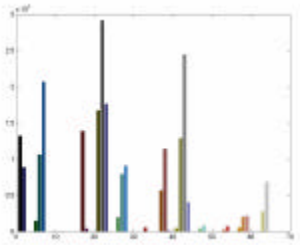
Similarité entre documents audiovisuels par le contenu

- Extraction de descripteurs servant de support aux requêtes
  - Robustesse
  - Invariance
  - Utilisabilité
  - Discriminance
  - Echelonnabilité
  - Interprétation

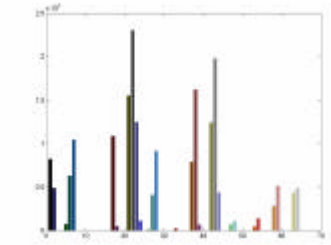
Similarité entre documents audiovisuels par le contenu

- Comparaison de caractéristiques vidéo
  - MPEG-7
    - Descripteurs normalisés
    - Distances non-normalisées
    - Validations indépendantes des applications
    - Vers une technologie de l'indexation multimédia
  - Couleur

Descripteur : histogramme de structure de couleur



Descripteur : histogramme de structure de couleur



Couleur échelonnable - Utilisabilité

- Utilisabilité
  - Histogramme de couleur en HSV après Transformée de Haar
  - Distance de type L1

$$D=\sum_i |R_i - Q_i|$$

Couleurs dominantes - Utilisabilité

- Distance

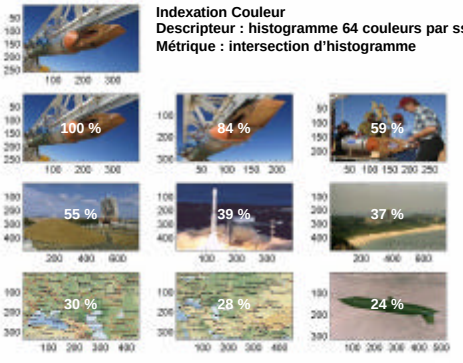
$$RR = \sum_i \sum_j \frac{p_{1i}.p_{1j}.e^{\sum_{c=r,g,b} (ColQ_{i,c}-ColQ_{j,c})^2/(VarQ_{i,c}-VarQ_{j,c})}}{2\Pi\sqrt{2\Pi \sum_{c=r,g,b}(VarR_{i,c}-VarR_{j,c})}}$$

$$QQ = \sum_i \sum_j \frac{p_{2i}.p_{2j}.e^{\sum_{c=r,g,b} (ColQ_{i,c}-ColQ_{j,c})^2/(VarQ_{i,c}-VarQ_{j,c})}}{2\Pi\sqrt{2\Pi \sum_{c=r,g,b}(VarQ_{i,c}-VarQ_{j,c})}}$$

$$RQ = \sum_i \sum_j \frac{p_{2i}.p_{2j}.e^{\sum_{c=r,g,b} (ColR_{i,c}-ColQ_{j,c})^2/(VarR_{i,c}-VarQ_{j,c})}}{2\Pi\sqrt{2\Pi \sum_{c=r,g,b}(VarR_{i,c}-VarQ_{j,c})}}$$

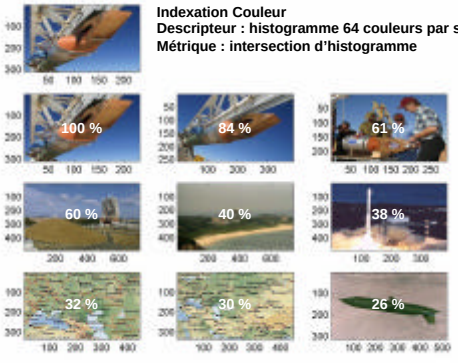
$$D=RR+QQ-2*RQ$$

Série sur une pièce



Indexation Couleur
Descripteur : histogramme 64 couleurs par ss-ech
Métrique : intersection d'histogramme

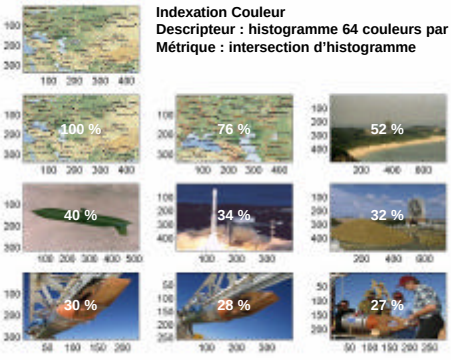
Série sur une pièce



Indexation Couleur
Descripteur : histogramme 64 couleurs par ss-ech
Métrique : intersection d'histogramme

Distinguer les cartes du reste de la base

Indexation Couleur
Descripteur : histogramme 64 couleurs par ss-ech
Métrique : intersection d'histogramme



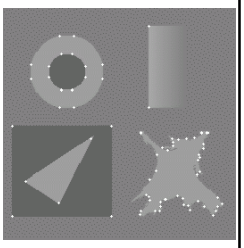
Similarité entre documents audiovisuels par le contenu

- Texture (énergie du spectre quantifié + somme des écarts)

Forme - Invariance

Les points d'intérêts

- Retrouver une image à partir d'un exemple partiel
- Repérage des "coins" (points d'intérêt)
  - Moravec (1977), Harris (1988)
- Invariance aux transformations linéaires
  - Changements d'échelle
  - Nombre de points



Similarité entre documents audiovisuels par le contenu

- Régions, blobs

(J.F. Omhover – Syst. Strict)

- Formes
  - Représentation de la courbure

Distances

intersection

$$d(h1,h2) = \sum_{i=0}^{N-1} \min(h1[i], h2[i])$$

χ^2

$$d(h1,h2) = \sum_{i=0}^{N-1} \frac{\left(h1[i] - \frac{h1[i]+h2[i]}{2}\right)^2}{\frac{h1[i]+h2[i]}{2}}$$

Kullback-Leibler

$$d(h1,h2) = \sum_{i=0}^{N-1} h1[i] \log\left(\frac{h1[i]}{h2[i]}\right)$$

EMD

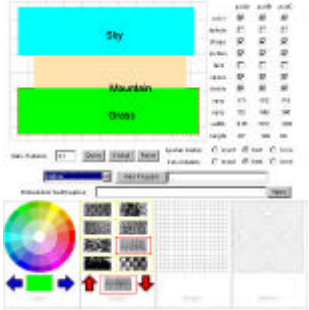
$$d(h1,h2) = \frac{\sum_i \sum_j g_{ij} d_{ij}}{\sum_i \sum_j g_{ij}}$$

Jensen

$$d(h1,h2) = \sum_{i=0}^{N-1} h1[i] \log\left(\frac{h1[i]}{h2[i]}\right) + h2[i] \log\left(\frac{h2[i]}{h1[i]}\right)$$

Outils intégrés

- VisualSeek



Logiciels libres

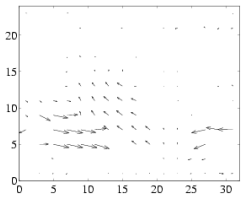
- ImgSeek

Réseau européen

- Schema

Mouvement - Extraction

- Dérivé de l'encodage



Similarité entre documents audiovisuels par le contenu

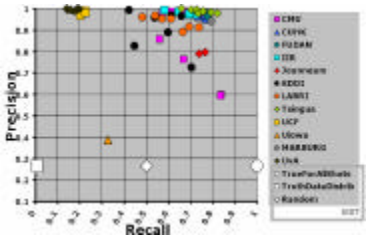
- Mouvement de caméra
  - hypothèse du mouvement dominant
  - vérité terrain
- Mouvement d'objet
  - Distance entre trajectoires

$$d(T1,T2) = \omega_P \sum_i \frac{(P1_i - P2_i)^2}{\Delta t_i} + \omega_V \sum_i \frac{(V1_i - V2_i)^2}{\Delta t_i} + \omega_A \sum_i \frac{(A1_i - A2_i)^2}{\Delta t_i}$$

- Utilité ?

Similarité entre documents audiovisuels par le contenu

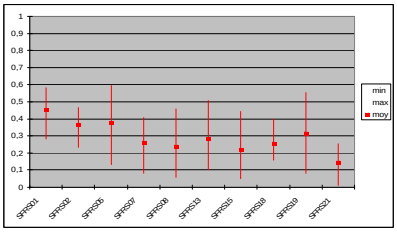
- Mouvement de caméra : plongée/pano/zoom



D'après W. Kraaij, T. Ianeva – Trec Vid 2005

Similarité entre documents audiovisuels par le contenu

- Résultats par vidéos : (moyenne 0,29)



D'après E. Kijak – Argos 2006

Similarité entre documents audiovisuels par le contenu

- Calcul de la matrice
  - S1={6 7 6 7 1 4 1 5} et S2 ={4 1 4 1 5 1 5 1}. Fixons tmax=4 et tmin=2

$[\min(S1)\max(S1)] = [1\,7]$

$[\min(S2)\max(S2)] = [1\,5]$

$[1\,7] \cap [1\,5] \neq \varnothing$

$S1 = \{6\,7\,6\,7\} \cup \{1\,4\,1\,5\} = S11 \cup S12$

$S2 = \{4\,1\,4\,1\} \cup \{5\,1\,5\,1\} = S21 \cup S22$

Similarité possible entre :
- (S12, S21)
- (S12, S22)

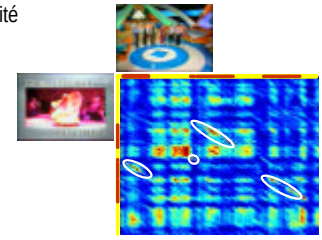
Similarité entre documents audiovisuels par le contenu

- Calcul de la matrice
  - Taux de couverture > 50 %

	6 7	6 7	1 4	1 5
4 1	0	0	0.75	0
4 1	0	0	0	0.75
5 1	0	0	0	0
5 1	0	0	0	0

Similarité entre documents audiovisuels par le contenu

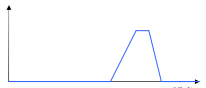
- Matrice de similarité

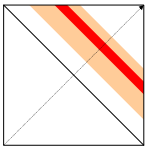


Matrice de similarité

Similarité entre documents audiovisuels par le contenu

- Distance, mesure de similarité globale

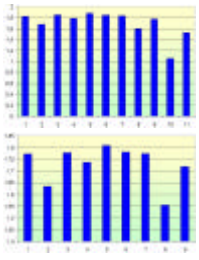




$$M = \max_{k=1}^{Nbdiag}(W_k \cdot m_k)$$

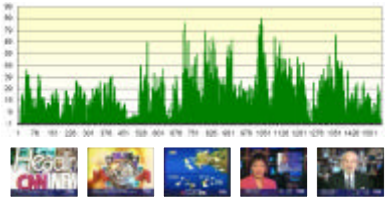
Similarité entre documents audiovisuels par le contenu

- Application 1 : analyse de collections
  - Mesure inter-classe
  - Mesure intra-classe



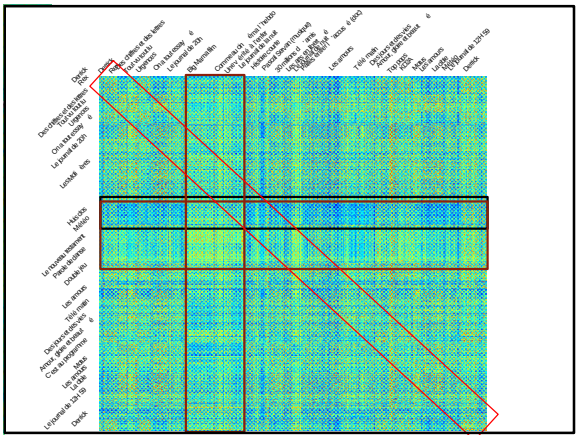
Similarité entre documents audiovisuels par le contenu

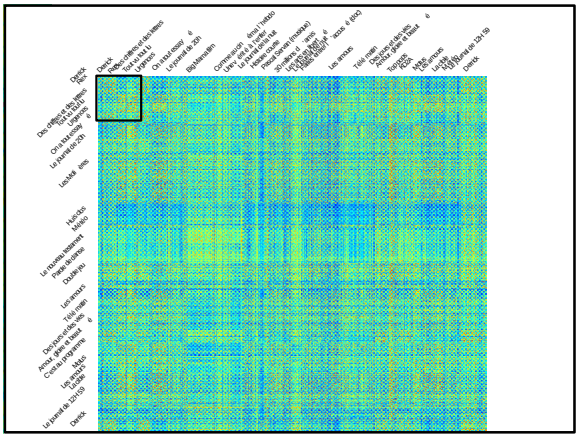
- Application 1 : analyse de collections

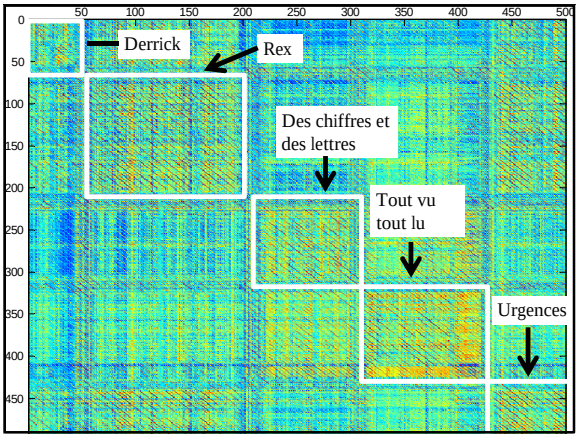


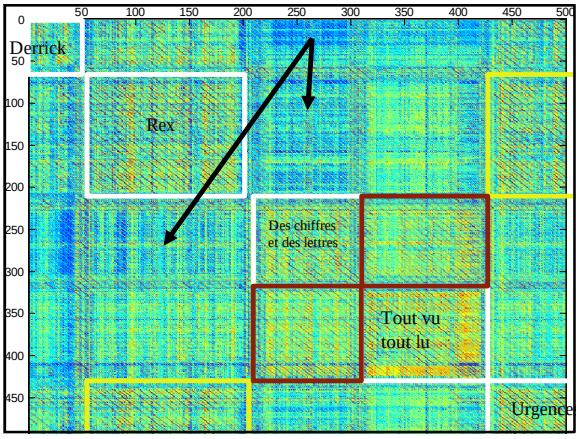
Similarité entre documents audiovisuels par le contenu

- Application 2 : analyse d'un flux



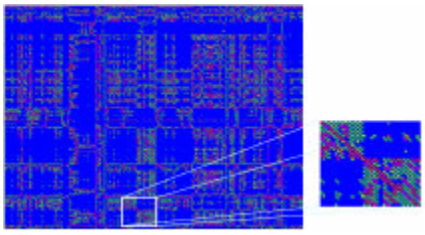






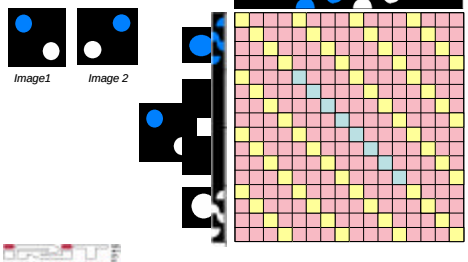
Similarité entre documents audiovisuels par le contenu

- Extension à l'audio

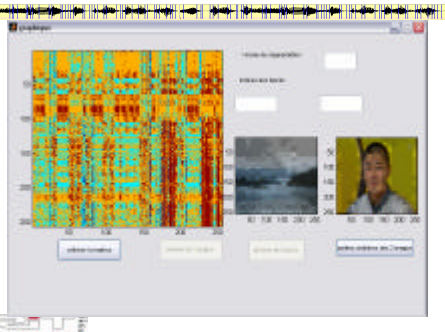


Similarité entre documents audiovisuels par le contenu

- Extension à l'image fixe



Similarité entre documents audiovisuels par le contenu

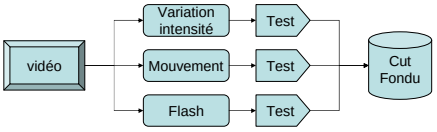


Similarité entre documents audiovisuels par la structure

- Segmentation
- Motivation
  - Document audiovisuel = document structuré
    - XML/HTML (cf. Jacques Le Maître)

Similarité entre documents audiovisuels par la structure

- Segmentation en plans
  - L'approche classique



Similarité entre documents audiovisuels par la structure

- Cuts

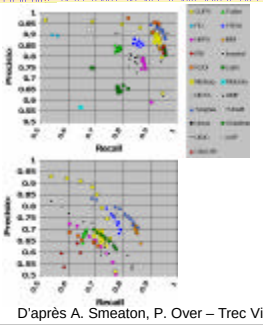
Bien traité

Vers le « rough indexing »

- Effets progressifs

Améliorable

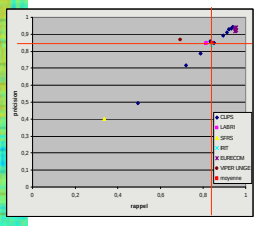
(Amélioration 2004-05 = 1 à 2 %)



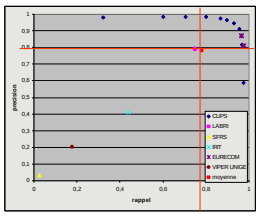
D'après A. Smeaton, P. Over – Trec Vid 2005

Similarité entre documents audiovisuels par la structure

- Courbe rappel/précision sur l'ensemble du corpus



Mesure ARGOS

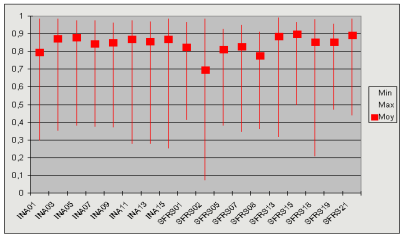


Mesure TREC

D'après E. Kijak – Argos 2006

Similarité entre documents audiovisuels par la structure

- Moyenne sur l'ensemble du corpus : 0.846
  - Mesure ARGOS



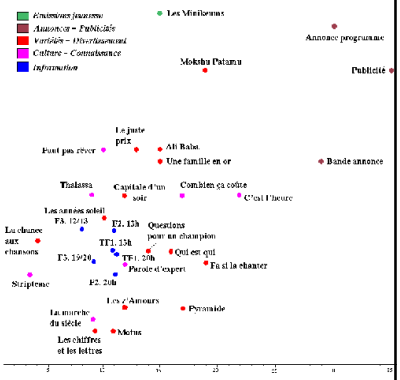
D'après E. Kijak – Argos 2006

Similarité entre documents audiovisuels par la structure

- Segmentation en plans
  - Les variantes
    - nombre de contours
    - variance de l'intensité
    - …
  - Notion de segment ?

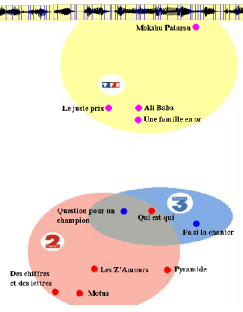
Similarité entre documents audiovisuels par la structure

- Fouille de données



Similarité entre documents audiovisuels par la structure

- Fouille de données



Similarité entre documents audiovisuels par la structure

- Macrosegmentation
  - Définition
  - Formats très contraints : JT
  - Formats contraints : retransmissions sportives
  - Formats libres (longs métrages, fictions, documentaires, …)
  - Evaluation ?

Similarité entre documents audiovisuels par la structure

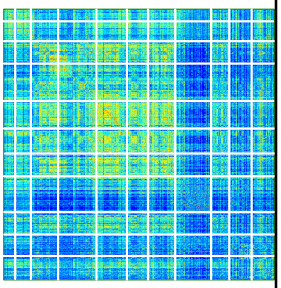
- Macrosegmentation
  - Journaux télévisés – retransmissions sportives
    - Modélisation : HMM
    - Cf. Guillaume Gravier / Patrick Gros
    - Exemples :

Similarité entre documents audiovisuels par la structure

- Macrosegmentation
  - Clustering (distance visuelle)
  - Grammaires généralistes
  - Auto-similarité

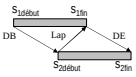
Similarité entre documents audiovisuels par la structure

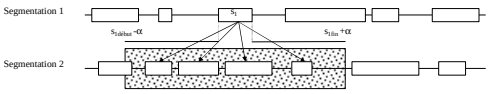
- Matrice d'autosimilarité
  - Foote & Cooper
  - Delezoïde



Similarité entre documents audiovisuels par la structure

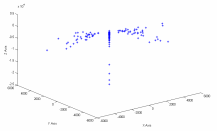
- Organisation structurelle

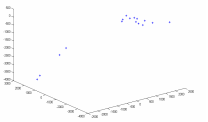


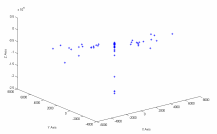


Similarité entre documents audiovisuels par la structure

- Organisation structurelle
  - entre 2 joueurs
  - joueur-présentateur

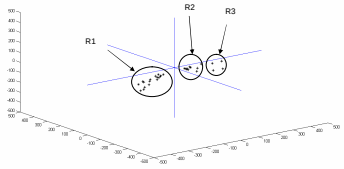






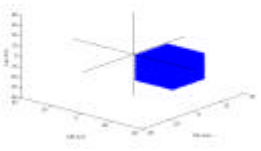
Similarité entre documents audiovisuels par la structure

- Organisation structurelle
  - Identification de classes de relation



Similarité entre documents audiovisuels par la structure

- Organisation structurelle
  - Identification de classes de relation
    - « overlap »



Similarité entre documents audiovisuels par la structure

- Organisation structurelle
  - Opérateurs algébriques sur
    - les relations
    - les classes de relation
    - les TRM
  - Exemple : conjonction entre relations $\&_R$

R1(db1, de1, lap1) $\&_R$ R2(db2, de2, lap2) = R3(db3, de3, lap3)
avec db3 = db1 + db2,
de3 = b1 + b2 et
lap3 = lap1 – de2 = db1 + lap2

Conclusion

- Recherche de contenus audiovisuels
  - Panoramique
  - Montgolfière
  - Statue de la liberté
  - R2D2 et C3P0 sur la même image
  - Montagnes (mais pas en arrière-plan)
  - Personnes faisant du ski nautique
  - Personnes parlant devant un drapeau américain
  - Cerf avec ses bois
  - Ronald Reagan parlant
  - Atterrissage d'un hélicoptère
  - Plans avec des monologues
- Décollage de la navette spatiale… Discovery
- Véhicule lunaire circulant sur la lune
- La planète Jupiter
- Le rover lunaire des missions Apollo

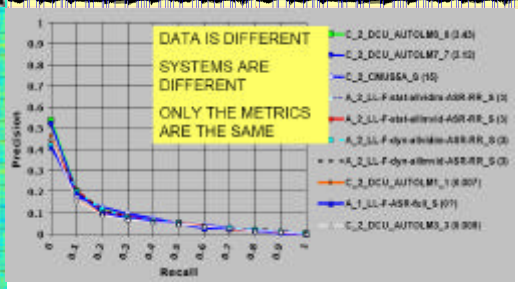



Conclusion

- Recherche de contenus audiovisuels
  - Condoleeza Rice, présidents, …
  - Carte de l'Irak, avec la position de Bagdad indiquée, sans que ce soit une carte météo
  - Personnes entrant ou sortant d'un bâtiment
  - Buts marqués pendant un match de foot
  - …

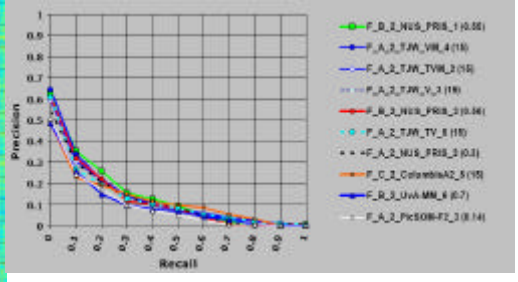



Recherche (automatique – 2004)



D'après A. Smeaton, T. Ianeva – Trec Vid 2005

Recherche (automatique – 2005)



D'après A. Smeaton, T. Ianeva – Trec Vid 2005

Conclusion

« votre définition du type d'information que vos systèmes analysent/recherchent »

- Un segment AV qu'on peut modéliser
  - à l'aide de mesures qui le caractérisent
  - de manière aussi générique que possible
  - de manière échelonnable
    - Exploitabilité de la méthode et des résultats

Références

- [1] Foote, Cooper. Summarizing popular music via structural similarity analysis. http://www.fxpal.com/publications/FXPAL-PR-03-204.pdf#search=%22foote%20cooper%20similarity%22
- [2] B. Delezoide. Modèles d'indexation multimédia pour la description automatique de films de cinéma. Thèse. http://mediatheque.ircam.fr/articles/textes/Delezoide06c/
- [3] S. Haidar, P. Joly, B. Chebaro. Mining for Video Production Invariants to Measure Style Similarity. Dans : International Journal of Intelligent Systems (IJIS), Wiley, V. 21 N. 7, p. 747-763, juillet 2006.
- [4] B. Moulin, Conceptual Graph Approach for the representation of temporal information in discourse. In Knowledge based systems, Vol 5, Num 3, 1992, pp 183-192.
- [5] Zein Al Abidin Ibrahim, Isabelle Ferrané, Philippe Joly. Conversation Detection in Audiovisual Documents: temporal Relation Analysis and Error Handling. Dans : 11th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'2006), Paris, France, 02/07/2006-07/07/2006, EDK Editions médicales et scientifiques, (CD/ROM), 2006.

Notes

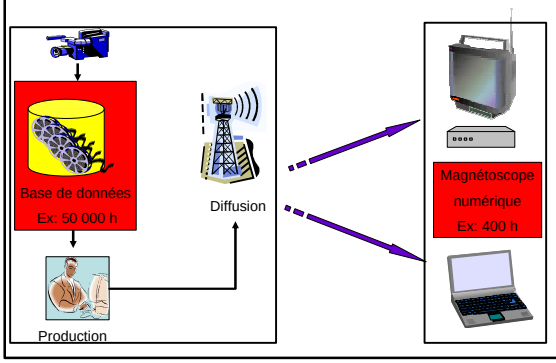
Structuration multimodale de vidéos de sports

Scénario et mise en scène: *Guillaume Gravier*
Production: *Patrick Gros*

Camera 1: Ewa Kijak
Camera 2: Manolis Delakis

Une co-production Thomson Multimedia

Contexte



Pourquoi le sport?

- Pourquoi se restreindre au sport?
  - $$$$
  - Besoins professionnels: archives, PVR, repurposing
  - Structure forte, aisément compressible dans le temps
    - règles de production spécifiques
    - syntaxe de montage forte
    - ensemble fini d'éléments sémantiques
- Sports étudiés
  - Football : détection des buts [Assfalg 03]
  - Football américain [Babaguchi 03]
  - Baseball [Hua 02]
  - Basketball : détection des paniers, des tirs au panier [Nepal 01]
  - Tennis : classification des coups joués [Sudhir 98]

Spécificités du sport

Contraintes temporelles	Contraintes structurelles
(foot, basket, etc.)	(tennis, baseball, etc.)
⇩	⇩
Détection d'événements	Structuration
=	=
Table des index	Table des matières
⇩	⇩
Détection d'événements isolés	Structuration / Segmentation

⇨ Distinction pas toujours pertinente

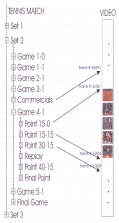
Nécessité d'une analyse multimodale pour la structuration!

Stratégies d'intégration multimodale

- Fusion de décisions
  - décisions indépendantes selon chaque modalité
  - règles, décisions en cascade, réseaux bayésiens
  - souvent employée en détection d'événements
- Fusion des descripteurs
  - projection dans un espace multimodal (concaténation, LDA, etc.)
  - SVM, MaxEnt Model, ANN, HMM
  - segmentation, classification, détection
- Modélisation conjointe
  - modèle conjoint des flux d'informations
  - MMC multiflux, MMC couplés, réseaux bayésiens dynamiques
  - segmentation, classification

Syntaxe du tennis

- Informations a priori
  - liées au domaine : règles et structure du tennis
  - liées aux règles de production
    - nombre fixe de caméras
    - sélection des points de vue (montage)
- Syntaxe
  - règles de composition des plans, produisant des motifs répétitifs et formant des scènes caractéristiques



Objectif : Structuration la vidéo en scènes caractéristiques

Scènes caractéristiques

- Premier service + Échange
  - vue du terrain de courte durée + gros plan de courte durée + vue du terrain + gros plans éventuels
  - balles + silence + balles + applaudissements
- Échange
  - vue du terrain + gros plans éventuels
  - balles + applaudissements
- Temps morts
  - succession de plans non relatifs au terrain d'une durée totale longue (publicités, interviews, changement de cote, etc.)
  - présence éventuelle de musique
- Rediffusions
  - plans encadrés de transitions spéciales

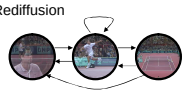
Modélisation par MMC

Premier service, échange

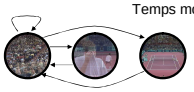
Échange

1 état = 1 plan vidéo

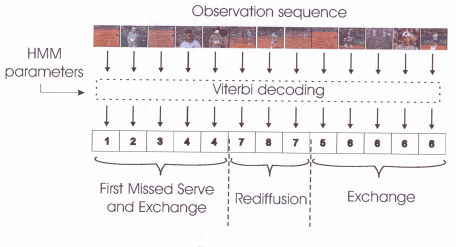
Rediffusion



Temps morts



Modélisation par MMC (décodage)



Caractérisation des plans

- Similarité à une vue globale de référence

✓ Extraction automatique d'une image clé de référence K_{ref}

✓ Mesure de similarité entre K_t et K_{ref}, $v_t = v(K_t, K_{ref})$, fonction de

? la cohérence spatiale

? la couleur dominante

? l'activité du plan

✓ Discrétisation sur 10 bins


Open de Paris


US Open


Roland Garros

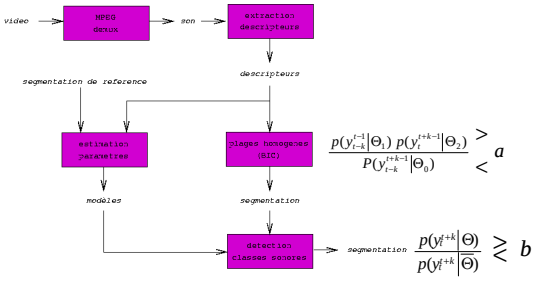
- Durée du plan

✓ Discrétisation sur 10 bins

- Plan fondu enchaîné

Caractérisation du son

Détection de sons clés : balles, applaudissements, musique



Intégration multimodale (principe)

Ajouter des descripteurs décrivant le contenu sonore des plans

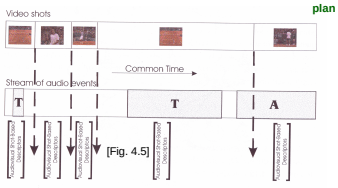
=

intégration au niveau des descripteurs (equiv. MMC multi flux synchrone)

A	B	M
1	1	0

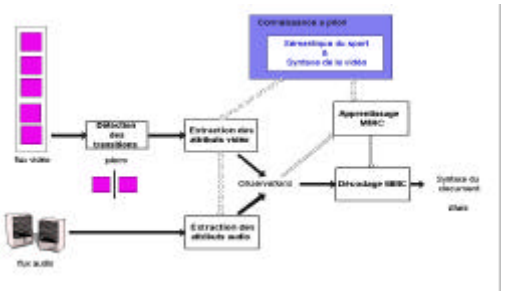
Présence d'applaudissements et de bruits de balle dans le plan

Absence de musique et silence dans le plan



[Fig. 4.5]

Intégration multimodale (système)



Intégration multimodale (résultats)

Descripteurs $O_t^{av} = \begin{bmatrix} O_t^v, & \text{visual similarity} \in [1,10] \\ O_t^l, & \text{shot length} \in [1,10] \\ O_t^d, & \text{dissolve} \in [0,1] \\ O_t^a, & \text{applause} \in [0,1] \\ O_t^b, & \text{ball hits} \in [0,1] \\ O_t^m, & \text{music} \in [0,1] \end{bmatrix}$

supposés indépendants $\ln b(O_t^{av}) = \sum_x \ln(b(O_t^x))$

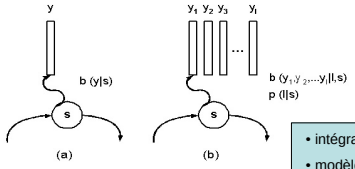
	C	F	P	R	F+
MMC V	76.3	77.4	82.0	73.4	**59.5**
MMC VA	80.2	82.1	84.7	79.7	**66.4**
MMC VA*	84.7	90.5	92.2	88.8	**78.3**

Des MMC aux modèles de segments…

Limites des MMC :

- Intégration du son au niveau des plans peu satisfaisante
- Difficulté d'intégrer d'autres informations (scores, parole, etc.)

Modèles de segments



Modèle de Markov caché

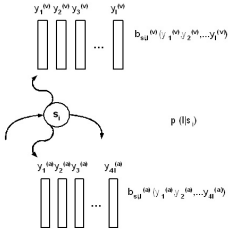
Modèle de segment

- intégration au niveau de la scène
- modèle de durée de la scène
- modélisation indépendante des flux d'information dans une scène

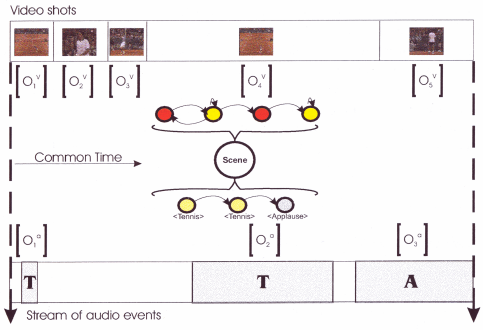
Des MMC aux modèles de segments…

Un segment =

- un modèle de durée p(l|s)
- un modèle de séquence pour chaque flux d'information



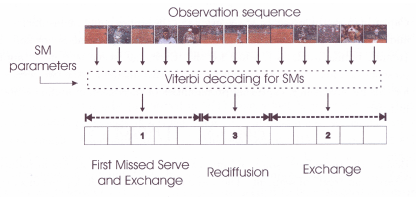
Des MMC aux modèles de segments…



Des MMC aux modèles de segments…

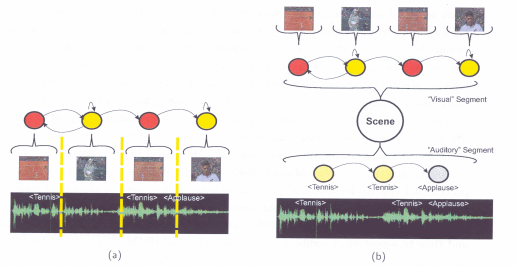
$$(L_{1:N}^*, Q_{1:N}^*) = \arg\max_{L_{1:N}, Q_{1:N}} p(Q_{1:N}) + p(L_{1:N} \mid Q_{1:N}) + \sum_{x \in \{a,v\}} b(O^{(x)} \mid L_{1:N}, Q_{1:N})$$

avec $b(O^{(x)} \mid L_{1:N}, Q_{1:N}) = \sum_i b(O^{(x)}_{s(Li)} \cdots O^{(x)}_{e(Li)} \mid Q_i)$



Résolution par un algorithme de Viterbi modifié

Des MMC aux modèles de segments…



Quels modèles pour les séquences audio et vidéos?

Modélisation des segments vidéos

- MMC sur les attributs visuels O_t^v, O_t^l et O_t^d.
  - MMC utilise pour le calcul de la probabilité conditionnelle $b(O^{(v)}_{s(Li)} \cdots O^{(v)}_{e(Li)} \mid Q_i)$ ou $O_t^{(v)} = \{O_t^v, O_t^l, O_t^d\}$
  - pas pour la recherche de la meilleure séquence d'états!
- Réseau de neurones (LSTM)

	C	F	P	R	F+
MMC VA	80.2	82.1	84.7	79.7	**66.4**
SM MMC VA	84.4	82.6	86.2	79.3	**69.3**
SM LSTM	81.3	80.4	84.1	77.1	**65.2**

Intégration audiovisuelle

- MMC audiovisuels

  $b(O^{(av)}_{s(Li)} \cdots O^{(av)}_{e(Li)} \mid Q_i)$ avec $O_t^{(av)} = \{O_t^v, O_t^l, O_t^d, O_t^a, O_t^b, O_t^m\}$

- Modèles de séquences d'événements audio

  - unigramme $b(O^{(a)}_{s(Li)} \cdots O^{(a)}_{e(Li)} \mid Q_i) = \prod_t \prod_{x \in \{a,b,m\}} b(O_t^x \mid Q_i)$

  - bigramme $b(O^{(a)}_{s(Li)} \cdots O^{(a)}_{e(Li)} \mid Q_i) = \prod_t b(O_t^x \mid O_{t-1}^x, Q_i)$

  P(BBA) = P(B|<s>)P(B|B)P(B|A)P(A|</s>)

- MMC sur des attributs audio bas niveaux
  - coefficients cepstraux
  - éviter les erreurs de détection des événements sonores

Intégration audiovisuelle (résultats)

	C	F	P	R	F+
MMC VA	80.2	82.1	84.7	79.7	**66.4**
SM VA	84.4	82.6	86.2	79.3	**69.3**
SM A1g	80.1	79.4	83.9	75.3	**63.4**
SM A2g	81.8	81.7	84.1	79.4	**66.8**
SM cepstre	79.9	79.6	84.6	75.2	**63.6**
SM VA + A2g	84.7	82.9	84.1	81.7	**69.7**

- Fusion asynchrone
  - bons résultats avec le modèle bigramme
  - attributs audio bas niveaux pas très performants
- Intégration des descripteurs meilleure que intégration conjointe
  - **hypothèse d'indépendance fausse ?**
  - le mieux est encore de faire les deux…

Intégration audiovisuelle (résultats)

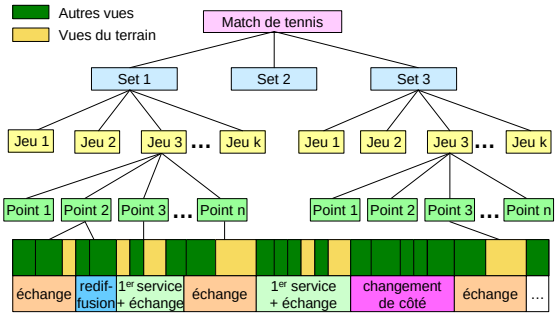
- Intégration des descripteurs ou modélisation asynchrone?

	C	F	P	R	F+
SM VA	84.4	82.6	86.2	79.3	**69.3**
SM 2g	81.8	81.7	84.1	79.4	**66.8**
SM Vhmm Ahmm	81.5	82.3	87.8	77.5	**67.5**
SM (LD)hmm Chmm Ahmm	78.9	80.2	86.3	75.0	**63.8**

- Et dans un monde parfait?

	C	F	P	R	F+
MMC VA*	84.7	90.5	92.2	88.8	**78.3**
SM VA*	89.6	91.1	93.1	89.2	**82.1**
SM A*2g	82.6	84.6	89.1	80.6	**70.5**

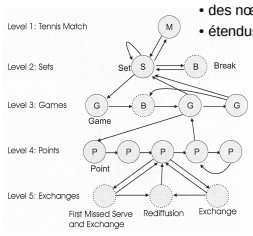
Approche hiérarchique



Approche hiérarchique

Les nœuds en pointilles sont

- des nœuds « émetteurs » dans le cas des SM
- étendus au MMC correspondant dans le cas des MMC



Pas d'amélioration des performances de classification et de segmentation, probablement du fait du manque de données pour estimer les probabilités de transitions dans la partie hiérarchique.

Intégration du score

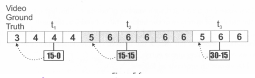
Comment exploiter cette information?

1. **ajouter un descripteur a chaque plan/scène**
   - descripteur binaire, présence/absence de score
2. **modifier l'algorithme de décodage**
   - trouver une solution cohérente avec les indications de score
   - nombre de points marqués entre deux affichages

Intégration du score (algorithme)

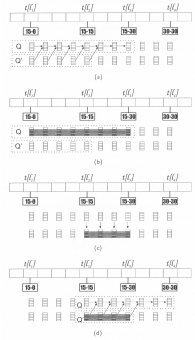
Trouver le meilleur chemin cohérent avec le nombre de points marqués entre deux affichages du score.

- Scène contenu entre deux labels adjacents



La scène correspondant au score 15-15 a forcement lieu entre t1 et t3.

- Recherche locale des N-meilleurs chemins
  - sur le segment [t1,t3]
  - un chemin pour chaque nombre de points
  - pénalisation des chemins non cohérents
- Mise a jour de la recherche globale
  - à partir des optimisations locales pénalisées



Intégration du score (résultats)

On suppose une détection parfaite des scores affichés.

	C	F	P	R	F+
MMC	80.2	82.1	84.7	79.7	**66.4**
+ descripteurs score	80.8	83.0	85.7	80.4	**67.7**
+ décodage contraint	82.2	82.9	83.4	82.4	**68.3**
SM	81.7	81.7	84.1	79.4	**66.8**
+ descripteurs score	81.9	81.9	84.2	79.7	**67.1**
+ décodage contraint	86.0	84.1	84.9	83.4	**71.8**

- intégration des scores comme descripteur peu efficace
- décodage contraint particulièrement efficace avec le modèle de segments
- robustesse aux erreurs de détection (par apprentissage)

Conclusions

- Conclusions
  - modèles de segments adaptés à la modélisation multimodale
  - bon modèle pour une intégration asynchrone (à l'intérieur des scènes)
  - avantage de la fusion de descripteurs
- Perspectives
  - intégrer de nouveaux flux faiblement synchronisés (e.g. mots-clés dans le football)
  - modéliser les dépendances entre flux d'information
  - un vrai modèle de synchro entre flux d'information

Exploitation du cadre des réseaux bayésiens

Notes

Recherche d'information textuelle dans des documents XML

Jacques Le Maitre
LSIS
Université du Sud Toulon-Var

Qu'est-ce qu'un document structuré

- Nous appelons ***document structuré*** tout document contenant des informations multimédia (textes, images, vidéo, son…) dont la structure logique est décrite explicitement de façon à pouvoir être manipulée par des programmes informatiques.
- Par exemple, les documents :
  - LaTEX
  - HTML
  - XML
  - PDF
  - …

Typologie des requêtes

- INEX (INitiative for the Evaluation of XML Retrieval) distingue deux types de requêtes textuelles sur les documents structurés :
  - Sur le contenu uniquement - *Content Only* (**CO**) *Queries*
    - L'utilisateur ne connaît pas la structure des documents interrogés et donc ne peut pas s'en servir pour spécifier les fragments de ces documents sur lesquels porte sa requête.
  - Sur le contenu et la structure - *Content And Structure* (***CAS***) *Queries*
    - L'utilisateur connaît la structure des documents et s'en sert pour spécifier les chemins permettant d'atteindre les fragments de ces documents sur lesquels porte sa requête.

Requêtes *Content and Structure*

- Deux interprétations sont distinguées :
  - Stricte (**SCAS** *Queries*)
    - La localisation structurelle de l'information recherchée peut être strictement déduite des chemins spécifiés dans la requête. Par exemple, si un utilisateur demande qu'un élément `<titre>` soit retourné, il doit être retourné par le moteur de recherche.
  - Vague (**VCAS** *Queries*)
    - L'utilisateur peut avoir une bonne connaissance de la structure des documents interrogés mais en ignorer certains aspects. Par exemple, quel est le nom exact des balises repérant chaque élément de cette structure. La recherche d'information ne peut donc pas se baser strictement sur les chemins exprimés dans la requête.

Objectifs du cours

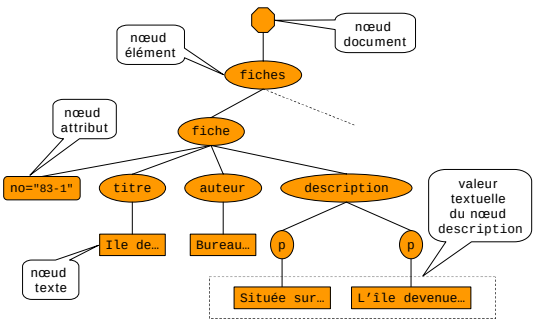
- Ce cours concerne la RI textuelle dans les documents XML par des requêtes de type SCAS.
- L'objectif est de présenter le langage **XQuery Full-Text** en cours de développement au sein du W3C pour étendre XQuery par des fonctionnalités de RI textuelle.
- Pourquoi XQuery Full-Text ? Parce que c'est un langage :
  - très complet,
  - en voie de standardisation,
  - compatible avec tous les langages de manipulation XML car construit sur le modèle XDM,
  - formellement bien fondé : sémantique fonctionnelle,
  - offrant un cadre pour prendre en compte les requêtes VCAS :
    - le degré de pertinence d'une requête Full-Text est accessible pour être manipulé par la requête XQuery dans laquelle elle est insérée.

Exemple de document XML

(Enregistré dans le fichier *fiches.xml*)

```
<?xml version="1.0"?>
<fiches>
<fiche id="83-1">
<titre>Ile de Porquerolles</titre>
<auteur>Bureau d'Informations de Porquerolles</auteur>
<description>
<p>Située sur le même parallèle que le Cap Corse ce qui en fait le point le
   plus méridional de la Côte Provençale. L'île de Porquerolles est la plus
   grande des trois îles d'Or avec ses 1254ha de superficie. Elle forme un
   arc orienté Est-Ouest, aux bords découpés, de 7.5km de long sur 3km de
   large. Son pourtour est d'une trentaine de kilomètres. L'île culmine au
   sémaphore à 142m.</p>
<p>L'île devenue site classé en 1988, réserve aux visiteurs de magnifiques
   promenades pédestres et cyclables, d'attrayantes plages de sable et de
   superbes points de vue le long des falaises du sud dominant une mer aux
   couleurs chatoyantes. Son climat tempéré ajoute au charme de chaque
   saison.</p>
</description>
</fiche>
...
</fiches>
```

Le modèle XDM : arbre de document



Langages de manipulation XML

- XPath
  - langage d'adressage des nœuds d'un arbre de document
- XSLT
  - langage de transformation d'un arbre de document
- XQuery
  - langage de requêtes à la SQL
    - centré données
    - centré documents

Exemple de requête XPath

- Numéros des fiches dont le 1er paragraphe contient la sous-chaîne de caractères « plages » ?
  - ```
    /fiches
    /fiche[contains(.//p[1], "plages")]
    /@no
    ```

Exemple de requête XQuery

- Numéro et titre des fiches contenant un paragraphe contenant la chaîne de caractères « plages », classées par ordre alphabétique de numéro.
  - ```
    for $f in doc("fiches.xml")/fiche
    let $nof = $f/@no
    where $f//p[fn:contains(., "plages")]
    order by $nof
    return
       <fiche no="{$nof}">
       $f/titre/text()
       </fiche>
    ```

De XQuery à XQuery Full-text

- En XQuery les seuls opérateurs disponibles pour interroger le texte d'un document sont des opérateurs classiques sur les chaînes de caractères :
  - extraction d'une sous-chaîne de caractères,
  - recherche d'une sous-chaîne de caractères (`fn:contains`),

  insuffisantes pour une véritable RI textuelle.
- Le W3C a donc entrepris le développement d'une extension de XQuery :
  - **XQuery Full-Text** (XQuery 1.0 FullText)
- Cette extension s'applique aussi à XPath (XPath 2.0 Full-Text).

XQuery Full-Text

- XQuery Full-Text étend XQuery par :
  - l'ajout d'un nouvel opérateur : `ftcontains` permettant d'exprimer une requête textuelle sous la forme d'une expression appelée *FTContains*,
  - l'extension de l'opérateur FLWOR pour prendre en compte le degré de pertinence d'une requête textuelle,
  - l'extension du modèle XDM par le concept de *AllMatches* afin de pouvoir manipuler des suites de mots d'un texte éventuellement regroupés en phrases ou pargraphes.

Un exemple de requête XQuery Full-Text

- Numéros des fiches dont la description parle d'îles et de plages de sable ?
  - `for $f in fn:doc("fiches.xml")//fiche where $f/description ftcontains ("ile" with stemming) && "plages de sable" return $f/@no`

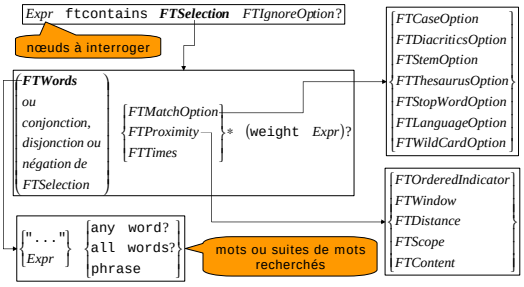
Expression *FTContains*

- Une expression *FTContains* a la forme suivante :

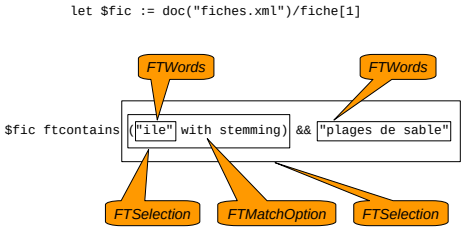
  Expr `ftcontains` *FTSelection FTIgnore*?

  où :
  - *Expr* est une expression XQuery dont la valeur est la séquence de nœuds à interroger,
  - *FTSelection* spécifie la condition que doit vérifier la valeur textuelle des nœuds à interroger,
  - *FTIgnore* spécifie les nœuds descendants des nœuds à interroger dont la valeur textuelle doit être ignorée.
- Sa valeur est :
  - `true`, si la valeur textuelle d'au moins un des nœuds interrogés vérifie la *FTSelection*,
  - `false`, sinon.

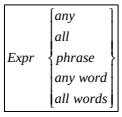
Syntaxe d'une expression *FTContains*



Imbrication des *FTSelection*



FTSelection élémentaire (*FTWords*)



- *Expr* est une expression XQuery dont la valeur est une séquence de chaînes de caractères appelées ***requêtes textuelles***.
- Soit *T* la suite de mots contenue dans la valeur textuelle d'un des nœuds interrogés :
  - `any` : la suite de mots d'au moins une des requêtes doit être présente dans *T*,
  - `all` : la suite de mots de chaque requête textuelle doit être présente dans *T*,
  - `phrase` : les suites de mots de chaque requête sont concaténées en une seule suite de mots qui doit être présente dans *T*,
  - `any word` : au moins un des mots d'une des requêtes doit être présent dans *T*,
  - `all words` : tous les mots des requêtes doivent être présents dans *T*.

Exemples de requêtes (1)

- `let $fic := doc("fiches.xml")/fiche[1]`
- La requête :
  - `$fic//p[. ftcontains "Ile" || "Porquerolles"]`

  retourne les 2 nœuds p.
- Les requêtes :
  - `$fic//p[. ftcontains {"plages", "falaises"} all`
  - `$fic//p[. ftcontains "plages falaises"} all words`

  retournent le 2e nœud p.
- La requête :
  - `$fic//p[. ftcontains $fic/titre]`

  retourne le 1er nœud p.

Exemples de requêtes (2)

- La requête :
  - `$fic/titre ftcontains "ile" not in "Ile de Porquerolles"`

  retourne `false`.
- La requête :
  - `$fic//p[. ftcontains "cycle" with stemming]`

  retourne le 2e nœud p.
- La requête :
  - `$fic//p[. ftcontains "plage*" && "falaise*" with wildcards with stop words ("de", "des", "et", "le") distance at most 5]`

  retourne le 2e nœud p.

Exemples de requêtes (3)

- La requête :
  - `$fic ftcontains "Corse" && "Porquerolles" same sentence`

  retourne `false`.
- La requête :
  - `$fic//p[. ftcontains ("ile*" with wildcards) occurs at least 3 times]`

  retourne le 1er nœud p.
- La requête :
  - `$fic ftcontains "îles d'Or" without content description`

  retourne `false`.

Sémantique de XQuery Full-Text

- XQuery Full-Text interroge des documents XML dont le contenu textuel est décomposé en une suite de mots (*tokens*).
- Les requêtes textuelles (*FTWords*) sont elles-mêmes décomposées en une suite de mots.
- Afin de représenter et de composer les résultats des *FTSelection* le concept de *AllMatches* a été rajouté au modèle XDM.
- Un *AllMatches* représente l'ensemble des suites de mots du document interrogé qui vérifient une *FTSelection*.

Décomposition en mots, phrases et paragraphes des documents interrogés

- La valeur textuelle des fragments de documents interrogés doit être décomposée en une suite de mots (*tokens*).
- Ces mots peuvent être regroupés en phrases et les phrases en paragraphes.
- Cette décomposition peut être réalisée :
  - soit au moment de la compilation en XDM des documents qui contient les fragments interrogés,
  - soit au début de l'évaluation de l'expression *FTContains* et ne concernait que les fragments interrogés.
- Chaque mot, phrase ou paragraphe est identifié par sa position relative dans cette décomposition.
- La position relative de la phrase et du paragraphe qui le contient est affectée à chaque mot.
- La façon de réaliser cette décomposition est **dépendante de l'implantation**.

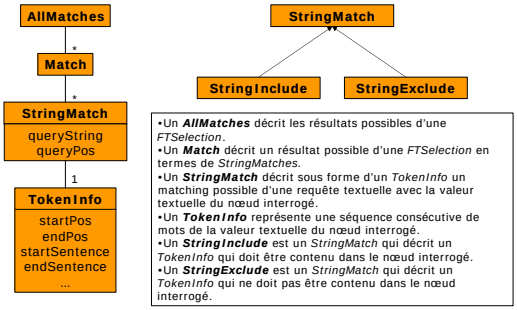
Décomposition en mots et en phrases de la 1ère fiche du document *fiches.xml*

```
<fiche no="83-1">
<titre>[Ile(1) de(2) Porquerolles(3)](1)</titre>
<auteur>[Bureau(4) d(5) Informations(6) de(7) Porquerolles(8)](2)</auteur>
<description>
<p>[Située(9) sur(10) le(11) même(12) parallèle(13) que(14) le(15) Cap(16)
   Corse(17) ce(18) qui(19) en(20) fait(21) le(22) point(23) le(24)
   plus(25) méridional(26) de(27) la(28) Côte(29) Provençale(30)](3)[L(31)
   île(32) de(33) Porquerolles(34) est(35) la(36) plus(37) grande(38)
   des(39) trois(40) îles(41) d(42) Or(43) avec(44) ses(45) 1254ha(46)
   de(47) superficie(48)](4)[Elle(49) forme(50) un(51) arc(52) orienté(53)
   Est(54) Ouest(55) aux(56) bords(57) découpés(58) de(59) 7.5km(60) de(61)
   long(62) sur(63) 3km(64) de(65) large(66)](5)[Son(67) pourtour(68)
   est(69) d(70) une(71) trentaine(72) de(73) kilomètres(74)](6)[L(75)
   île(76) culmine(77) au(78) sémaphore(79) à(80) 142m(81)](7)</p>
<p>[L(82) île(83) devenue(84) site(85) classé(86) en(87) 1988(88),
   réserve(89) aux(90) visiteurs(91) de(92) magnifiques(93) promenades(94)
   pédestres(95) et(96) cyclables(97) d(98) attrayantes(99) plages(100)
   de(101) sable(102) et(103) de(104) superbes(105) points(106) de(107)
   vue(108) le(109) long(110) des(111) falaises(112) du(113) sud(114)
   dominant(115) une(116) mer(117) aux(118) couleurs(119)
   chatoyantes(120)](8)[Son(121) climat(122) tempéré(123) ajoute(124)
   au(125) charme(126) de(127) chaque(128) saison(129)](9)</p>
</description>
</fiche>
```

Décomposition en mots des requêtes textuelles

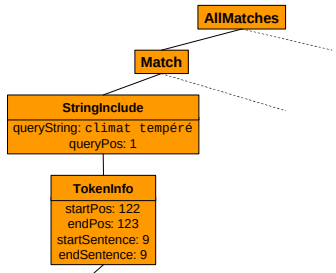
- Chaque requête textuelle d'une expression *FContains* est identifiée par sa position relative dans la clause *FTSelection* de cette expression.
- Chaque requête textuelle d'une *FTSelection* élémentaire est décomposée en mots et la position relative de cette requête est affectée à chacun de ces mots.
  - Ceci afin de pouvoir imposer l'ordre dans lequel doivent apparaître les mots ou les suites de mots recherchés dans les fragments interrogés.

Structure d'un *AllMatches*



- Un ***AllMatches*** décrit les résultats possibles d'une *FTSelection*.
- Un ***Match*** décrit un résultat possible d'une *FTSelection* en termes de *StringMatches*.
- Un ***StringMatch*** décrit sous forme d'un *TokenInfo* un matching possible d'une requête textuelle avec la valeur textuelle du nœud interrogé.
- Un ***TokenInfo*** représente une séquence consécutive de mots de la valeur textuelle du nœud interrogé.
- Un ***StringInclude*** est un *StringMatch* qui décrit un *TokenInfo* qui doit être contenu dans le nœud interrogé.
- Un ***StringExclude*** est un *StringMatch* qui décrit un *TokenInfo* qui ne doit pas être contenu dans le nœud interrogé.

Exemple de *AllMatches*



[Son(121) **climat**(122) **tempéré**(123) ajoute(124) au(125) charme(126) de(127) chaque(128) saison(129)]**(9)**

Evaluation d'une expression *FTContains*

- L'évaluation d'une expression *FTSelection* est réalisée en commençant par les *FTSelection* élémentaires (*FTWords*).
- L'évaluation d'un *FTWords* produit un ensemble de *AllMatches* : un pour chaque requête textuelle.
- Une *FTSelection* non élémentaire s'applique à un *AllMatches* et produit un *AllMatches*.
- Une expression *FTContains* a la valeur :
  - `true` si le *AllMatches* résultat de sa *FTSelection* :
    - n'est pas vide,
    - ou comporte au moins un *Match* sans *StringExclude*,
  - `false`, sinon.

Conjonction, disjonction et négation de *AllMatches*

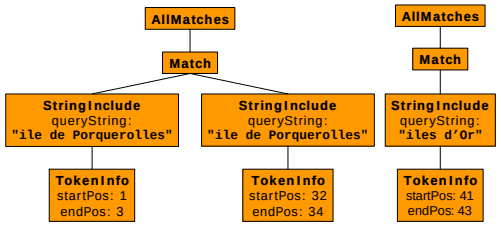
- Soit a, a', a_1, a_2 des *AllMatches* :
  - a_1 || a_2 = a tel que
    Matches(a) = *Matches*(a_1) $\cup$ *Matches*(a_2)
  - a_1 && a_2 = a tel que
    Matches(a) = *Matches*(a_1) $\otimes$ *Matches*(a_2)
  - !a = a' tel que
    Matches(a') = *Matches*(a)
    après inversion des *StringInclude* en *StringExclude* et inversement.

Options de *matching*

- Le *matching* d'une requête textuelle est effectué conformément aux options spécifiées dans la *FTSelection* à laquelle elle appartient :
  - prise en compte de la casse (*FTCaseOption*),
  - prise en compte des diacritiques (*FTDiacriticsOption*),
  - spécification de la langue (*FTStemOption*),
  - comparaison sur les racines des mots (*FTThesaurusOption*),
  - élimination des mots vides (*FTStopWord*),
  - utilisation d'un thésaurus (*FTLanguage*),
  - utilisation de wildcards (*FTWildcards*).

Exemple d'évaluation d'une expression *FTContains* (1)

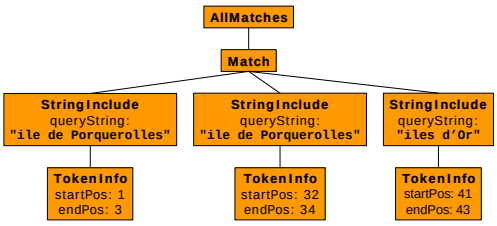
`$fic ftcontains "ile de Porquerolles" || "iles d'Or"`



Après évaluation de **"ile de Porquerolles"** et de **"iles d'Or"**

Exemple d'évaluation d'une expression *FTContains* (1)

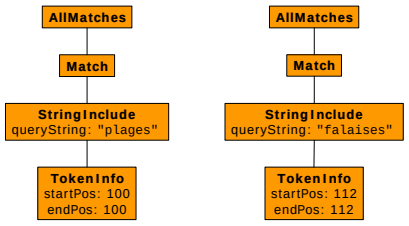
`$fic ftcontains "ile de Porquerolles" || "iles d'Or"`



Après évaluation de "ile de Porquerolles" || "iles d'Or"
fcontains retourne true car il y a au moins un *Match* sans *StringExclude*

Exemple d'évaluation d'une expression *FTContains* (2)

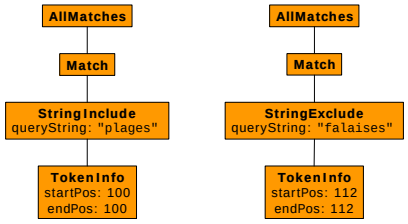
`$fic/description/p[2] ftcontains "plages" && !"falaises"`



Après évaluation de "plages" et de "falaises"

Exemple d'évaluation d'une expression *FTContains* (2)

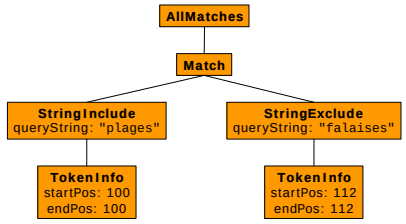
```
$fic/description/p[2] ftcontains "plages" && !"falaises"
```



Après évaluation de `!"falaises"`

Exemple d'évaluation d'une expression *FTContains* (2)

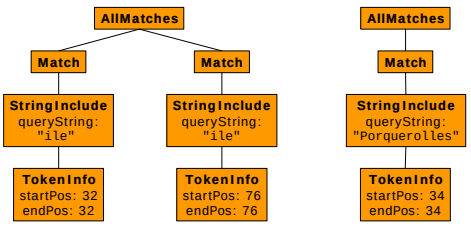
```
$fic/description/p[2] ftcontains "plages" && !"falaises"
```



Après évaluation de `"plages" && !"falaises"`
`ftcontains` retourne `false` car il y a pas de *Match* sans *StringExclude*

Exemple d'évaluation d'une expression *FTContains* (3)

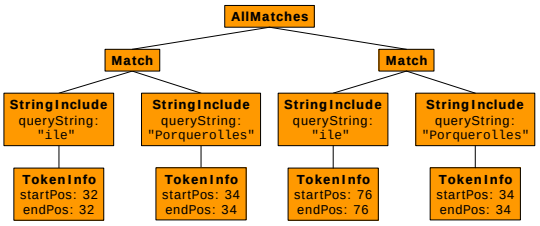
```
$fic//p[1] ftcontains "ile" && "Porquerolles" distance at most 1
```



Après évaluation de `"ile"` et de `"Porquerolles"`

Exemple d'évaluation d'une expression *FTContains* (3)

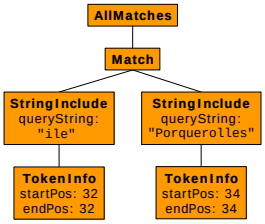
```
$fic//p[1] ftcontains "ile" && "Porquerolles" distance at most 1
```



Après évaluation de `"ile" && "Porquerolles"`

Exemple d'évaluation d'une expression *FTContains* (3)

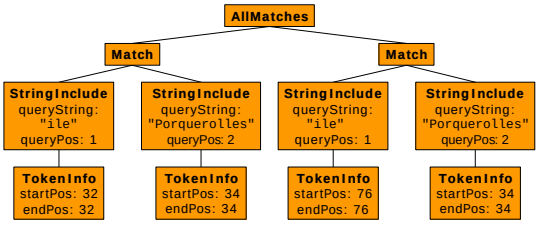
```
$fic//p[1] ftcontains "ile" && "Porquerolles" distance at most 1
```



Après évaluation de `"ile" && "Porquerolles" distance at most 1`
`fcontains` retourne `true` car il y a au moins un *Match* sans *StringExclude*

Exemple d'évaluation d'une expression *FTContains* (4)

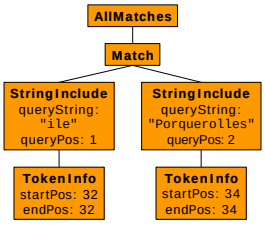
```
$fic//p[1] ftcontains "ile" && "Porquerolles" ordered
```



Après évaluation de `"ile" && "Porquerolles"`

Exemple d'évaluation d'une expression *FTContains* (4)

```
$fic//p[1] ftcontains "ile" && "Porquerolles" ordered
```



Après évaluation de `"ile" && "Porquerolles" ordered`
`fcontains` retourne `true` car il y a au moins un *Match* sans *StringExclude*

Prise en compte du degré de pertinence

- L'évaluation d'une clause *FTSelect* peut retourner le degré de pertinence (ou poids) attaché au résultat.
- Ce degré de pertinence peut être utilisé de deux façons :
  1. en posant une contrainte de minimalité sur ce degré dans la clause *FTSelect* elle-même :
     `… weight` *Expr*
  2. en liant le degré de similarité d'un nœud à une variable spécifiée dans la clause `for` ou la clause `let` qui sélectionne ce nœud :
     `for $x score $s in …/x[`*FTContains Expr*`]`
- Le mode de calcul du degré de pertinence est dépendant de l'implantation.

Exemple de prise en compte du degré de pertinence (1)

- `let $fiches = fn:doc("fiches.xml")//fiche`
- Quelles sont les fiches dont la description est conforme à la la requête « plages de sable fin » avec un degré de pertinence d'au moins 0.7 ?
  - `$fiches[description ftcontains "plages de sable fin" weight 0.7]`

Exemple de prise en compte du degré de pertinence (2)

- Numéros des fiches, classées par degré de pertinence décroissant, dont la description est conforme à la la requête « plages de sable fin » avec un degré de pertinence d'au moins 0.5 ?
  - ```
    for $f score $s in $fiches[description
    ftcontains "plages de sable fin"]
    where $s = 0.5
    order by $s descending
    return $f/@no
    ```

Problèmes ouverts

- Passer des requêtes strictes (SCAS) aux requêtes vagues (VCAS) :
  - Comment calculer les degrés de pertinence :
    - booléen étendu ?
    - vectoriel ?
    - probabiliste ?
  - Munir l'opérateur *FTSelect* d'une sémantique floue ?
  - Le découpage sémantique d'un document est-il équivalent à son découpage structurel ?
  - S'inspirer des travaux menés sur XPath flou et dans le cadre de l'initiative INEX (`http://inex.is.informatik.uni-duisburg.de/`).

  ...

Bibliographie

- S. Amer-Yahia et al, XQuery 1.0 and XPath 2.0 Full-Text, W3C Working Draft, http://www.w3.org/TR/2006/WD-xquery-full-text-20060501/, 2006.
- R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
- S. Boag et al., XQuery 1.0: An XML Query Language. W3C Candidate Recommendation, http://www.w3.org/TR/2006/CR-xquery-20060608/, 2006.
- P. Bosc et al, *Gradualité et imprécision dans les bases de données. Ensembles flous, requêtes flexibles et interrogation de données mal connues*, Ellipses, Paris, France, 2004.
- J. Clark and S. DeRose, XML Path Language (XPath) Version 1.0, W3C Recommendation 16 November 1999. http://www.w3.org/TR/1999/REC-xpath-19991116
- N. Fuhr, K. Großjohann, XIRQL: A Query Language for Information Retrieval from XML Documents, *Proceedings of SIGIR 2001*, New Orleans, Louisiana, USA, September 2001, pp. 151-158.
- J. Le Maitre, Indexing and Querying Content and Structure of XML Documents According to the Vector Space Model, *Proceedings of the IADIS International Conference WWW/Internet 2005*, Lisbon, Portugal, October 2005, vol. II, pp. 353-358.
- B. Trotman and B. Sigurbjörnsson, Narrowed Extended XPath I (NEXI), *Advances in XML Information Retrieval, 3rd International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004)*, Dagstuhl Castle, Germany, December 6-8, 2004, LNCS 3493, Springer 2005, pp. 16-40.
    ```

Notes