La reconnaissance automatique de la parole

Jean-Paul Haton
LORIA-INRIA
Université Henri Poincaré, Nancy 1
Institut Universitaire de France
jph@loria.fr

Tutoriel TAIMA'2005 Hammamet, Tunisie

Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Traitement Automatique de la Parole

- CODAGE ET TRANSMISSION

- SYNTHÈSE DE LA PAROLE

- RECONNAISSANCE DE LA PAROLE

- IDENTIFICATION DE LA LANGUE

- VÉRIFICATION DU LOCUTEUR

Science Citation Index Publication (Speech and Language Research)

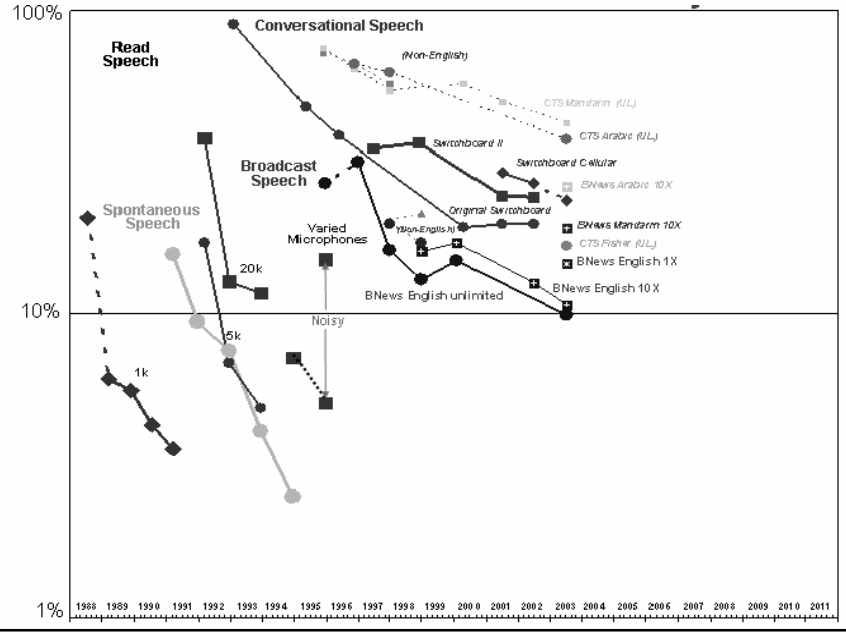
Automatic Speech Recognition and Natural Language Processing

Years	Speech Recognition	Speech Synthesis	Language Understanding	Language Processing
1981	16	14	2	11
1982	29	16	4	9
1983	46	22	6	21
1984	34	13	3	15
1985	41	7	1	25
1986	26	6	1	21
1987	36	6	0	18
1988	23	7	1	22
1989	30	7	2	21
1990	45	8	4	27
1991	171	26	38	241
1992	166	25	73	245
1993	162	37	88	323
1994	218	35	84	260
1995	272	40	142	393
1996	274	28	125	373
1997	285	34	116	388
1998	357	59	125	469
1999	393	36	152	553
2000	419	37	164	582
TOTAL	**3 043**	**463**	**1 131**	**4017**

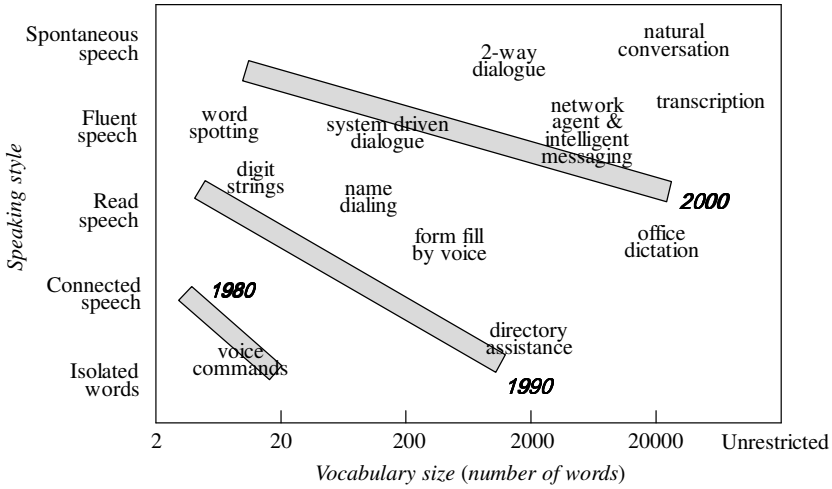
Tl

5

Campagnes d'évaluation NIST-DARPA



Types de tâches de reconnaissance



MACHINE VS. HUMAN WER
(*NON-MOBILE ENVIRONMENTS*)

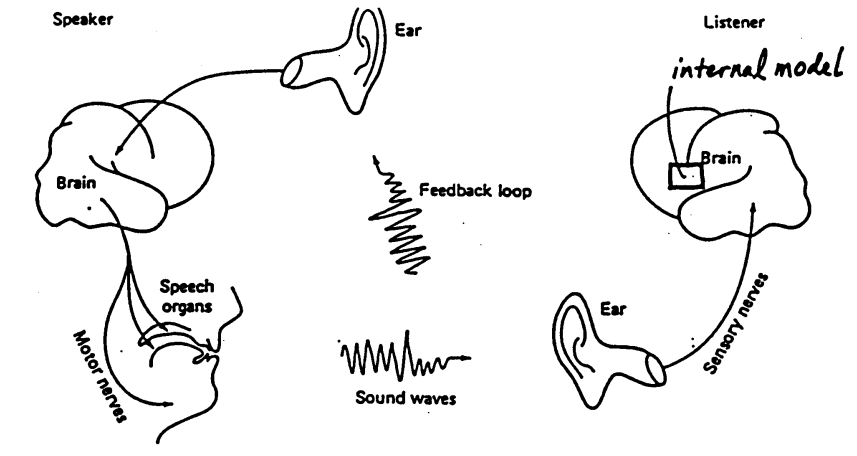
Task	Machine	Human
Connected digits	0.72%	0.009%
Letters	5.0%	1.60%
Transactional speech	3.6%	0.10%
Dictation	7.2%	0.9%
Conversational telephone	43.0%	5.0%

Source: R. Lippmann, Speech Communication, 1997

Plan de l'exposé

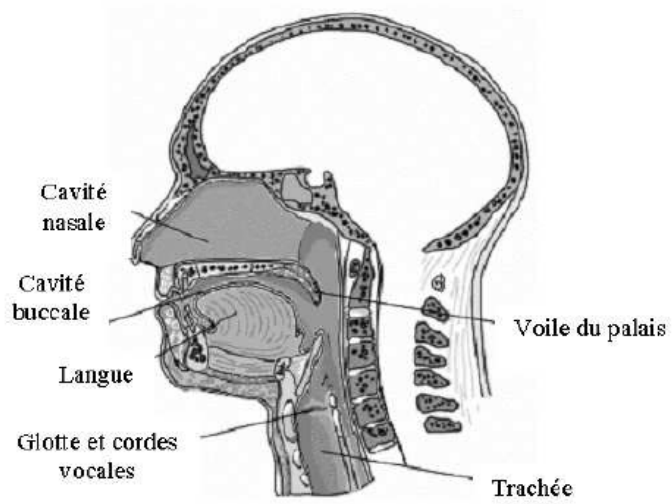
- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

La chaîne de communication parlée

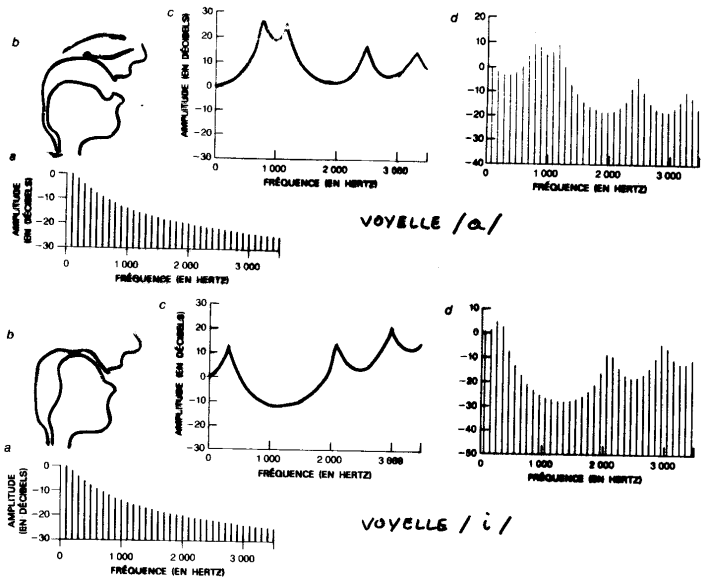


The speech chain (adapted from Peter Denes and Elliot Pinson, *The Speech Chain*)

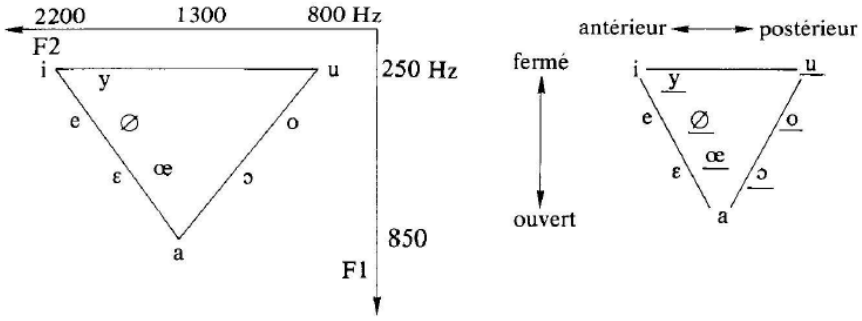
Le système phonatoire



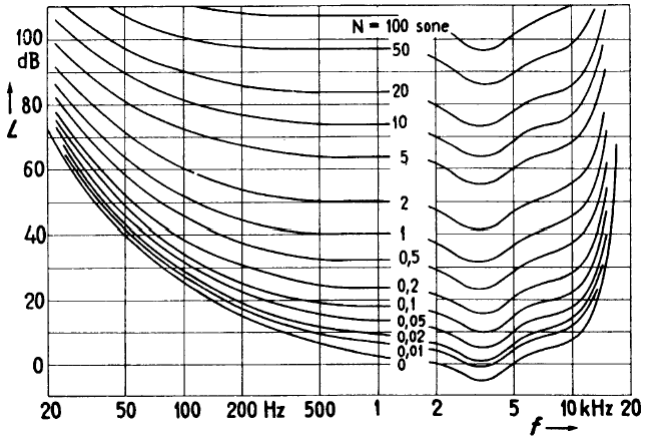
Production des voyelles



Le triangle des voyelles (d'après F. Lonchamp)

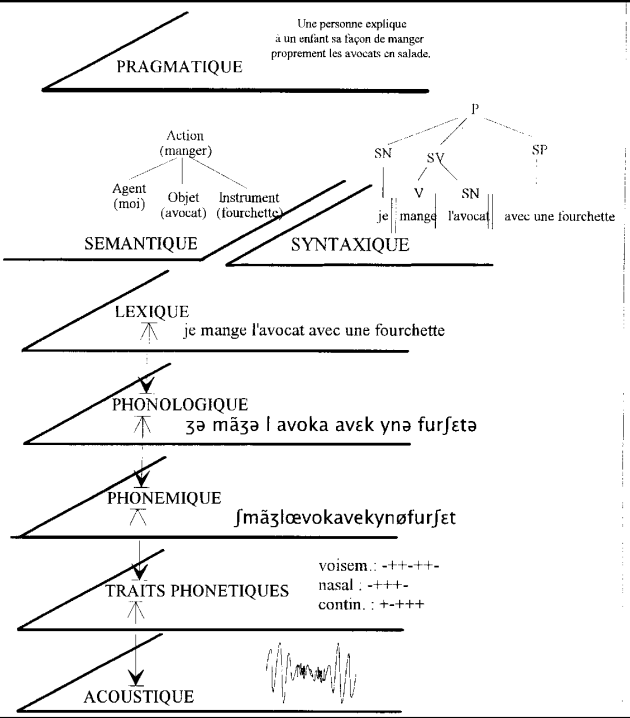


Perception auditive : sensibilité de l'oreille

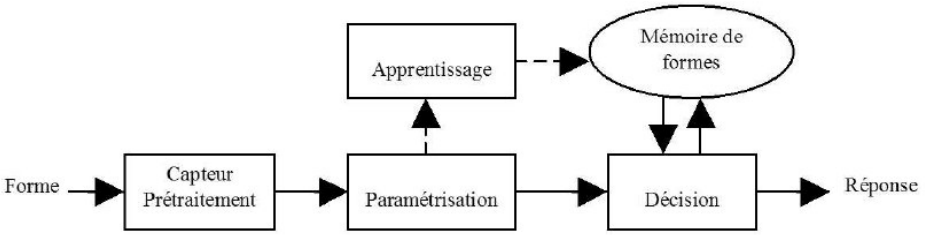


Echelle Bark : $B_{Bark} = 13\mathrm{A}rctg(\frac{0.76F_{Hz}}{1000}) + 3.5\mathrm{A}rctg(\frac{F_{Hz}}{7500})^2$

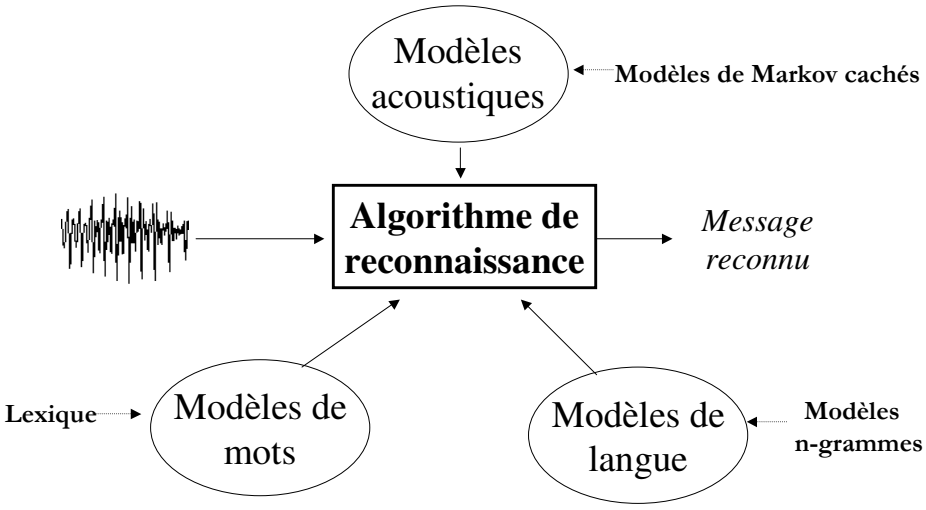
Echelle Mel : $M_{Mel} = 2595\log(1+\frac{F_{Hz}}{700})$



Principe de la reconnaissance des formes



Principe de la reconnaissance de la parole



Plan de l'exposé

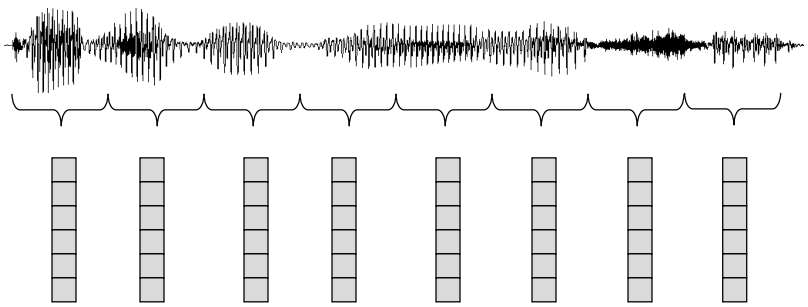
- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Paramétrisation

- transformer le signal brut en paramètres plus robustes et plus discriminants fondés sur certains critères, notamment perceptifs
- réduire le flux d'informations à traiter par le moteur de reconnaissance

Paramétrisation

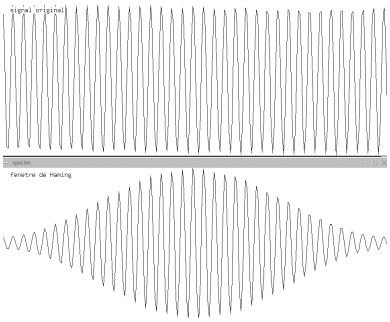
- Fenêtrage : spectre à court-terme



Paramétrisation

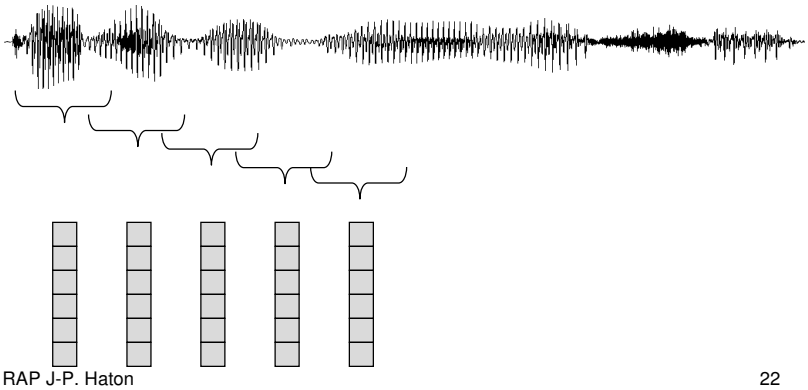
- Fenêtre de Hamming

$$h(n) = \begin{cases} 0,54 - 0,46\cos(2\pi\frac{n}{N-1}) & \text{si } 0 \leqslant n \leqslant N-1 \\ 0 \text{ sinon} \end{cases}$$



Paramétrisation

- Fenêtre de Hamming -> le centre est bien modélisé -> recouvrement



Transformation de Fourier

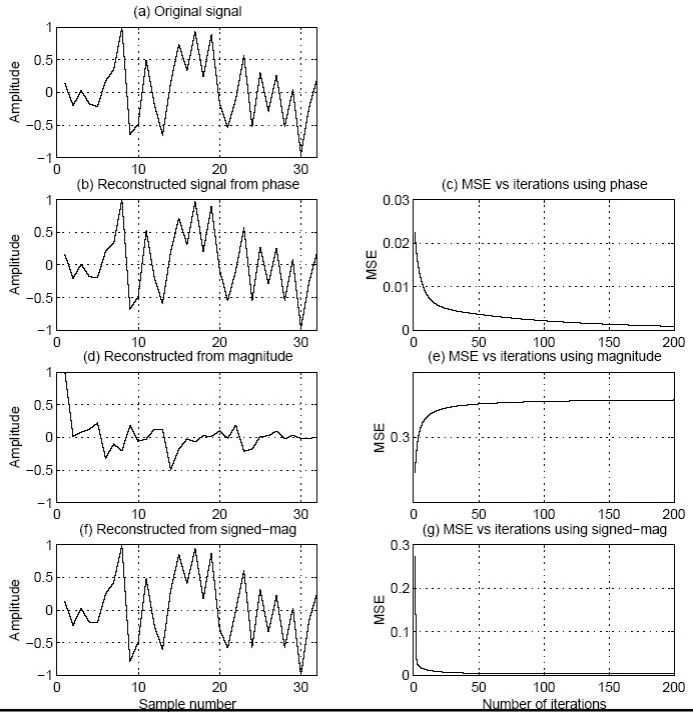
$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-i2\pi ft} dt$$

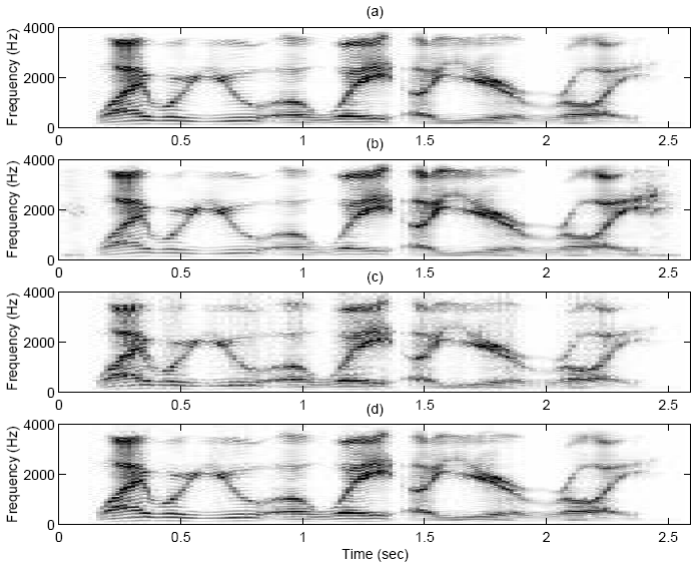
$$s(t) = \int_{-\infty}^{+\infty} S(f) e^{i2\pi ft} df$$

$$S(f) = R(f) + iI(f) = A(f) e^{i\Phi(f)}$$

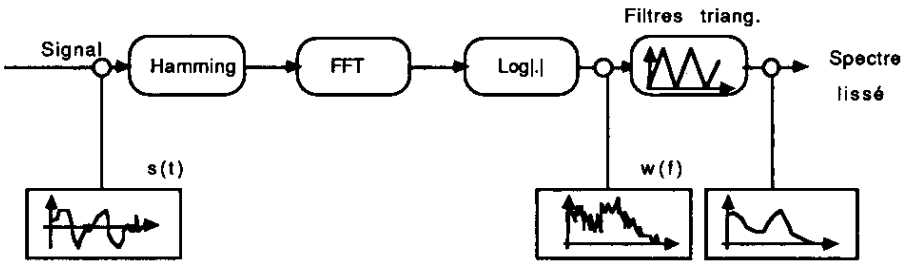
Spectre de puissance : $A^2(f) = R^2(f) + I^2(f)$

Spectre de phase : $\Phi(f) = \mathrm{A}rctg(I(f)/R(f))$

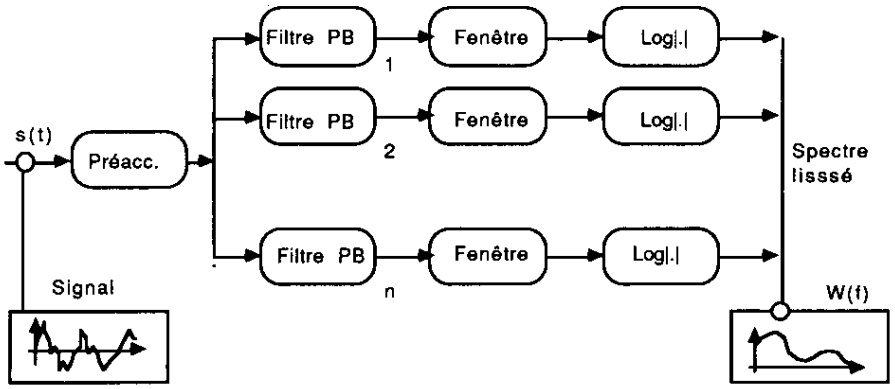




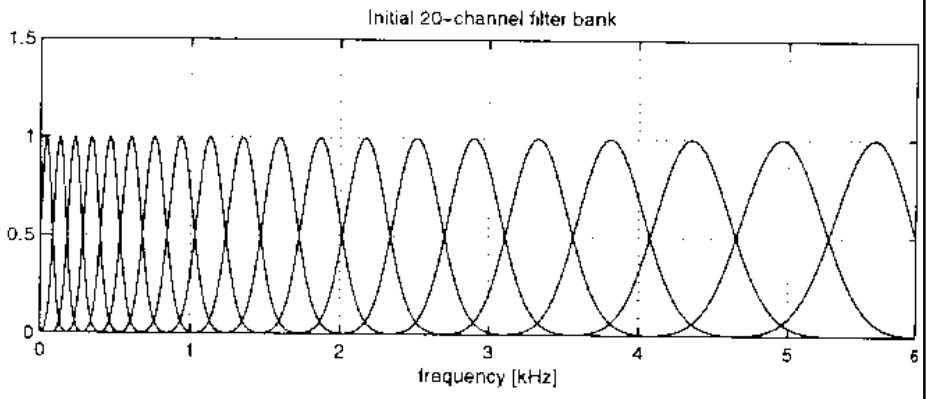
Analyse par transformée de Fourier



Analyse par banc de filtres



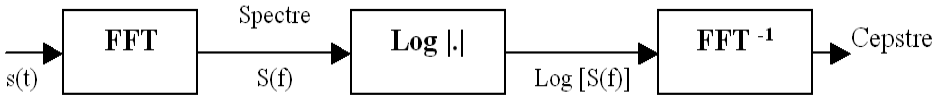
Banc de filtres (échelle Mel)



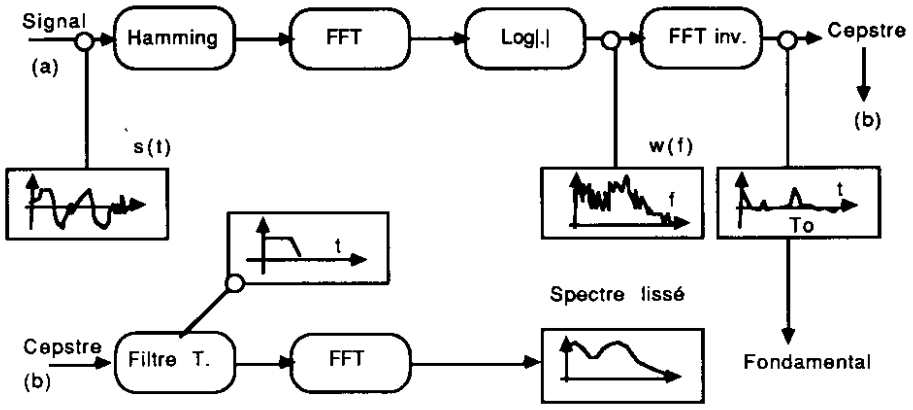
Paramétrisation

- paramétrisation la plus utilisée : MFCC (Mel Frequency Cespral Coefficients)
  - FFT pour décomposer le signal en ses fréquences constituantes
  - filtres triangulaires placés de façon à imiter le comportement de l'oreille (échelle Mel)

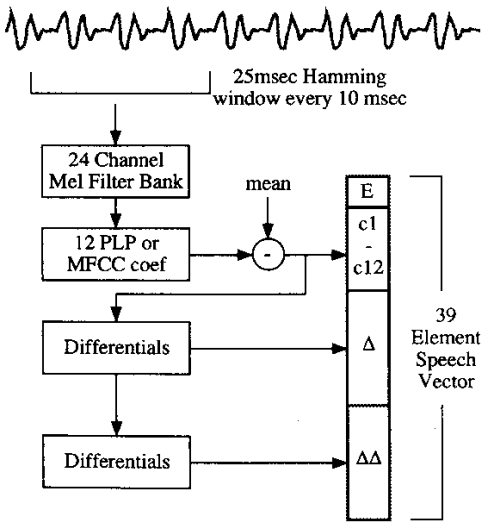
Analyse homomorphique



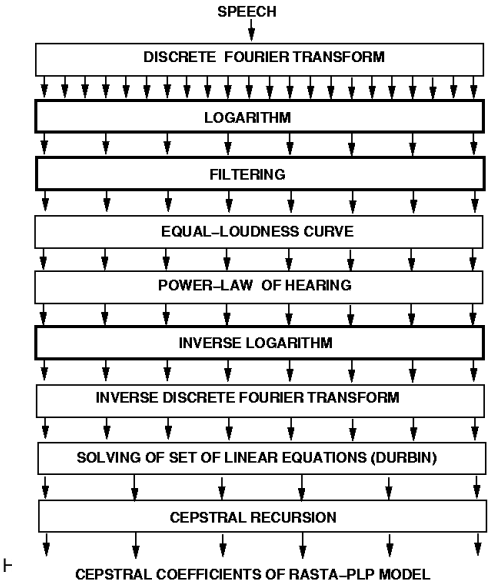
Analyse cepstrale



MFCC + dérivées



Analyse RASTA-PLP



Autres méthodes

Modèles d'oreille

Paramètres fréquentiels filtrés

Ondelettes

etc.!

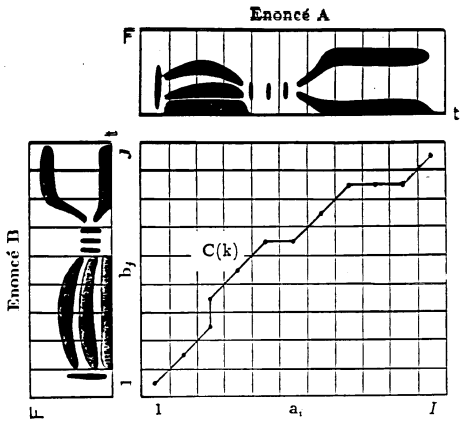
Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Approche statistique de la reconnaissance de la parole

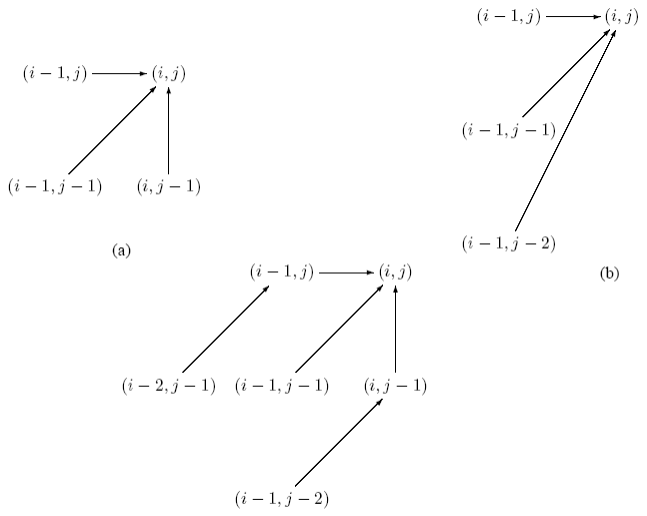
- Comparaison «élastique» de formes (« DTW »)
- Principe : règle de décision de Bayes
- Modélisation acoustique : trames vs segments
- Evolution des modèles :
  - modélisation de la durée
  - corrélation entre trames (HMM2, modèles AR, modèles contextuels)
  - modèles discrets, continus, mélanges de lois
  - partage de paramètres
  - apprentissage : MLE vs MMI

Principe de la programmation dynamique en RAP

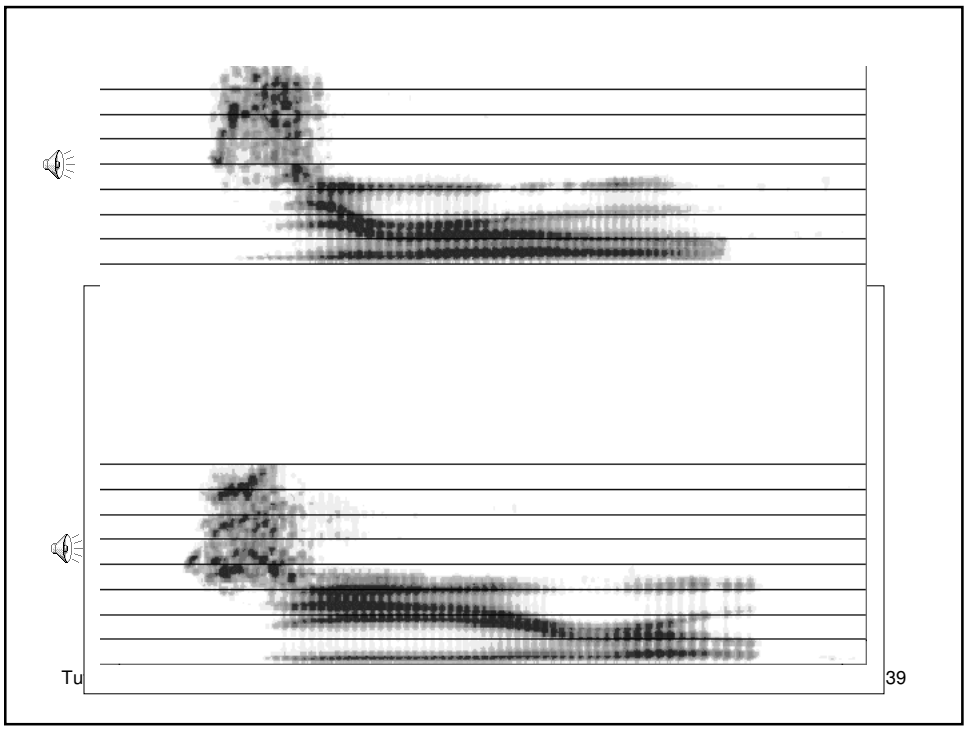


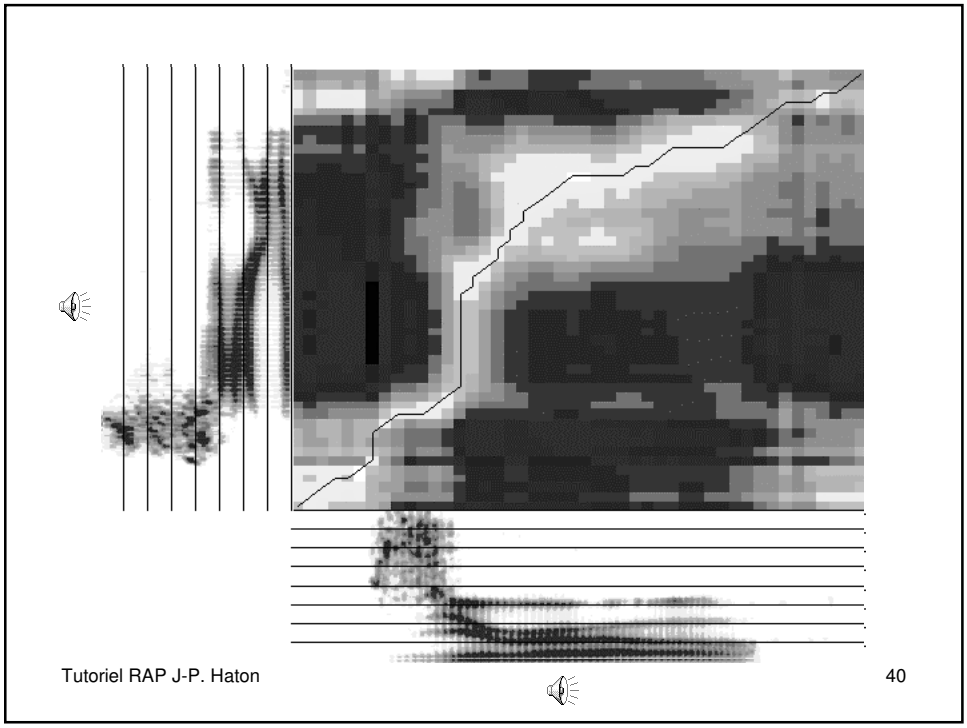
dissemblance entre A *et* B

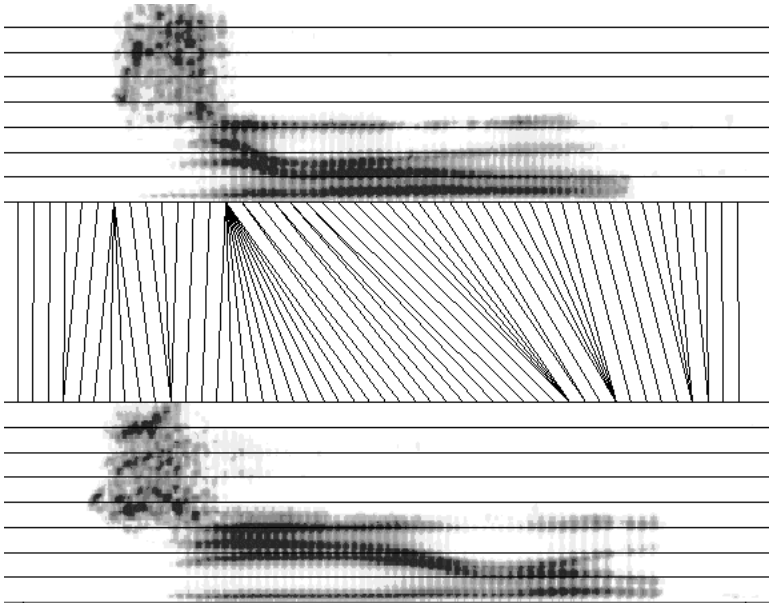
$$D(\mathbf{A},\mathbf{B}) = \min_{C}\left[\frac{\sum_{k=1}^{K} d(C(k))\omega(k)}{N(\omega)}\right]$$



$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases}$$



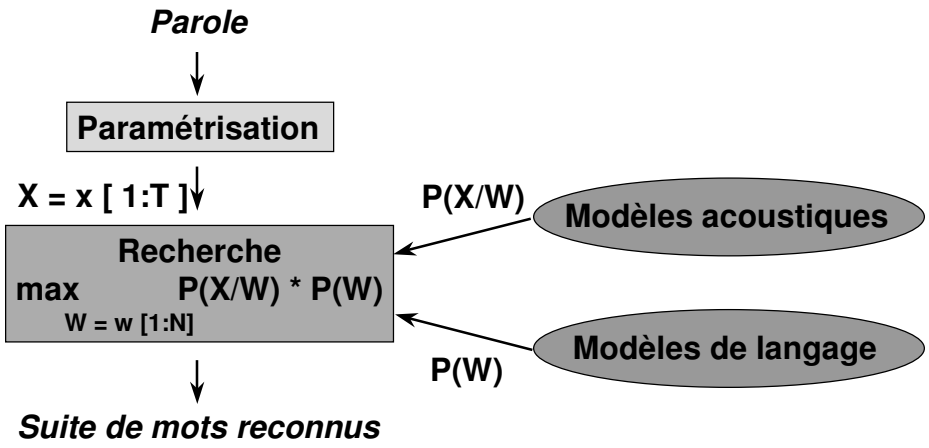




Approche statistique de la reconnaissance de la parole

- Comparaison «élastique» de formes (« DTW »)
- Principe : règle de décision de Bayes
- Modélisation acoustique : trames vs segments
- Evolution des modèles :
  - modélisation de la durée
  - corrélation entre trames (HMM2, modèles AR, modèles contextuels)
  - modèles discrets, continus, mélanges de lois
  - partage de paramètres
  - apprentissage : MLE vs MMI

Règle de décision de Bayes

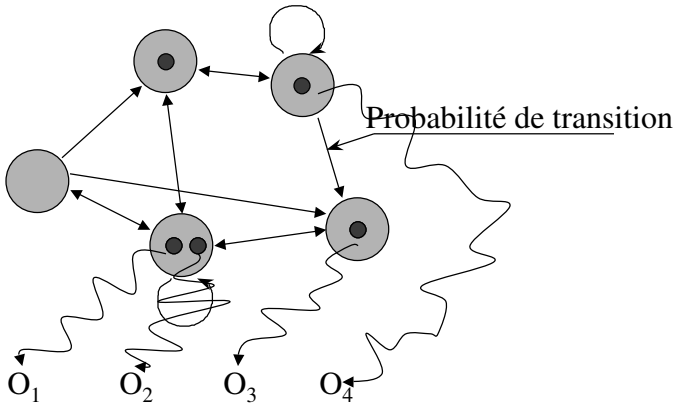


Approche statistique de la reconnaissance de la parole

- Comparaison «élastique» de formes (« DTW »)
- Principe : règle de décision de Bayes
- Modélisation acoustique : le HMM
- Evolution des modèles :
  - modélisation de la durée
  - corrélation entre trames (HMM2, modèles AR, modèles contextuels)
  - modèles discrets, continus, mélanges de lois
  - partage de paramètres
  - apprentissage : MLE vs MMI

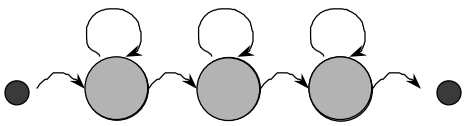
Qu'est-ce qu'un modèle de Markov caché, HMM ? (selon D. Fohr)

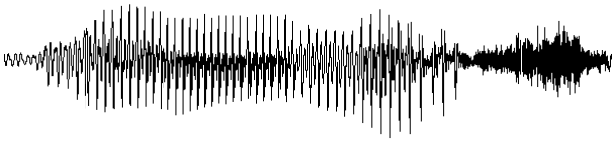
- c'est un automate probabiliste



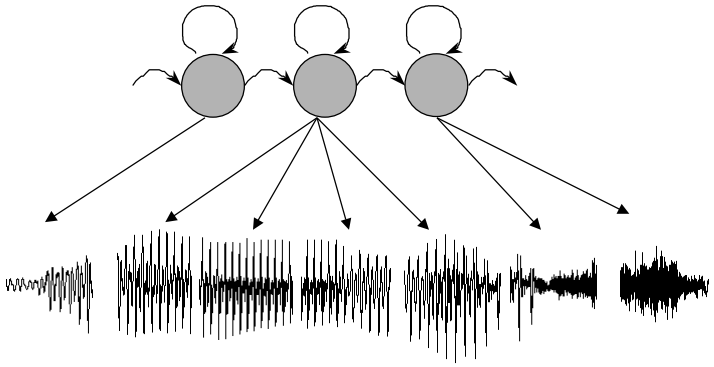
La parole modélisée par HMM

- On suppose que le système de production de la parole est un système markovien

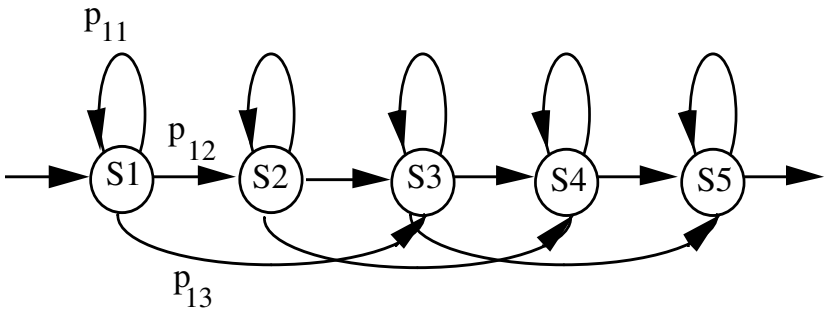




Modèle de Markov Caché



Modèle HMM de Bakis



Notations

- soit $O=(o_1,o_2,...,o_T)$ une suite d'observations de longueur T
- N : nombre d'états du modèle
- q : séquence d'états $q=(q_0,q_1,q_2...,q_T)$
- au temps t, le modèle
  - est dans l 'état q_t
  - engendre l 'observation o_t

Définition formelle

Pour définir un modèle de Markov il faut:

π_i : probabilité initiale : probabilité d'être à l'état i au temps 0 $\pi_i = P(q_0=i)$

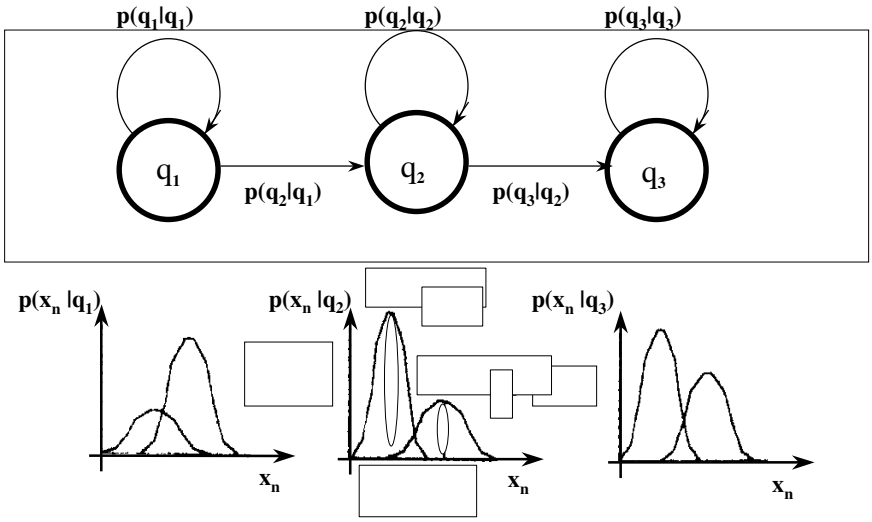
a_{ij} : probabilité de transition : probabilité d'aller de l'état i à l'état j $a_{ij}=P(q_t=j|q_{t-1}=i)$

b_i : densité de probabilité d'observation : probabilité d'observer o_t à l'état i

$b_i(o_t) = P(o_t|q_t=i)$

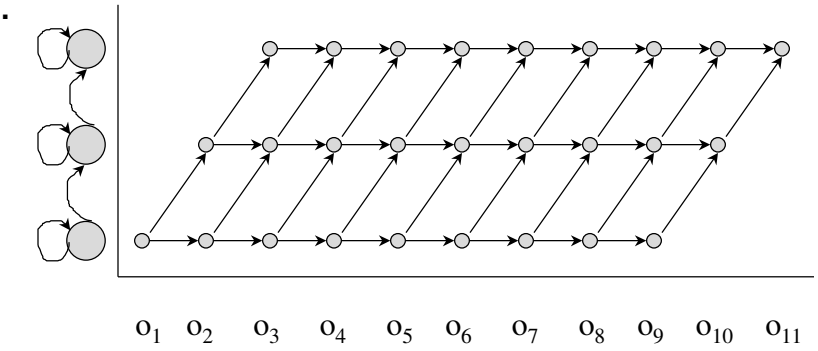
Apprentissage

- A l 'aide d'un corpus étiqueté d'exemples, il faut estimer:
  - les probabilités initiales π_i
  - les probabilités de transition a_{ij}
  - les probabilités d 'émissions $b_i(o)$ c 'est à dire les moyennes μ_i et les matrices de covariances Σ_i



HMM avec mélange de gaussiennes

Reconnaissance par HMM



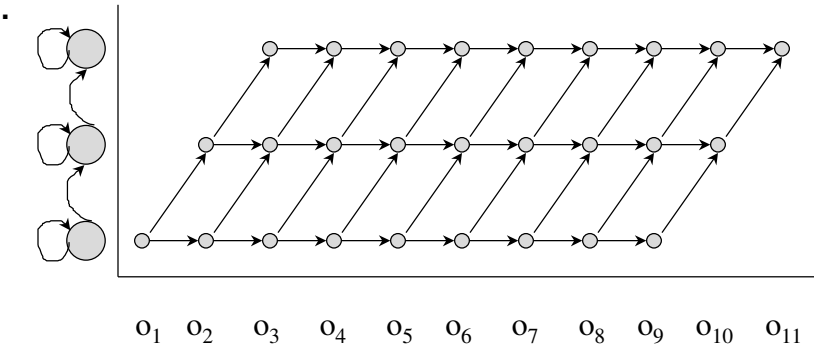
Algorithme de Viterbi

- But: trouver la meilleure séquence d'états q pour une observation O

soit : $$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = i | \lambda)$$

$\delta_t(i)$ est le meilleur score (plus grande probabilité) du chemin qui s'arrête à l'état i au temps t

calcul par récurrence



$$\delta_t(i) = \max \begin{cases} \delta_{t-1}(i) * a_{ii} * b_i(o_t) \\ \delta_{t-1}(i-1) * a_{(i-1)i} * b_i(o_t) \end{cases}$$

Algorithme de Viterbi en bref

- Initialisation

$$\delta_1(i) = \pi_i * b_i(o_1)$$

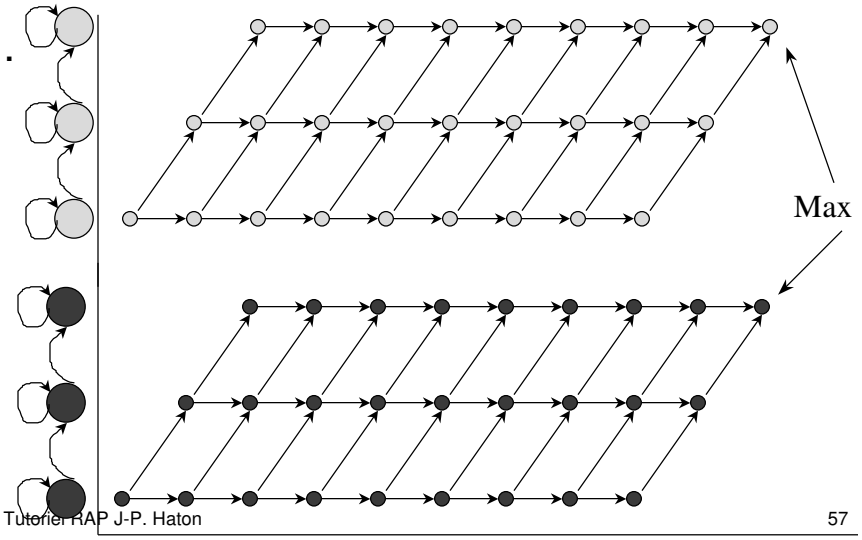
- Récursion

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) * a_{ij} \right] * b_j(o_t)$$

- Terminaison

$$P^* = \max_{1 \le i \le N} \left[\delta_T(i) \right]$$

Reconnaissance de mots isolés

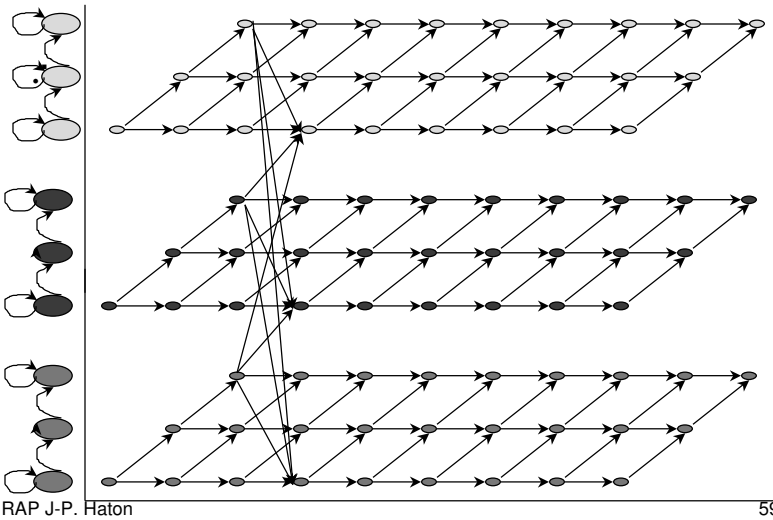


Cas des mots enchaînés

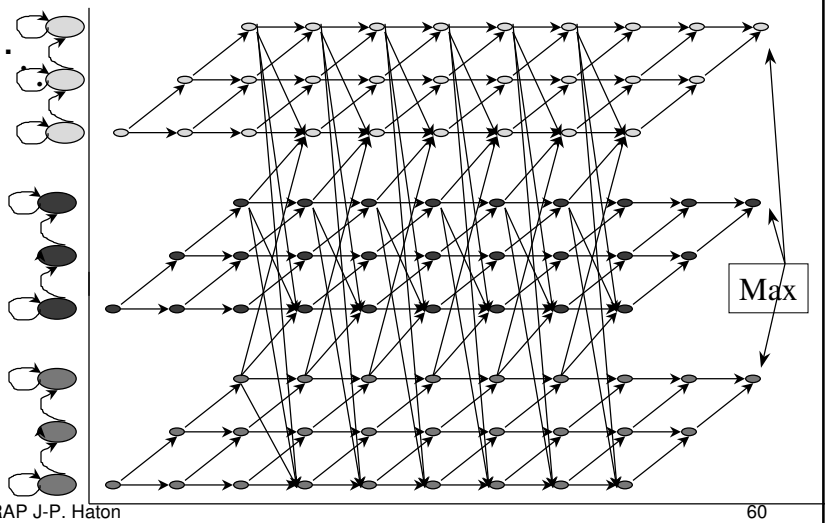
Soit une production vocale :

- on ne connaît pas le nombre de mots prononcés par le locuteur
- on ne sait pas où chaque mot commence et finit

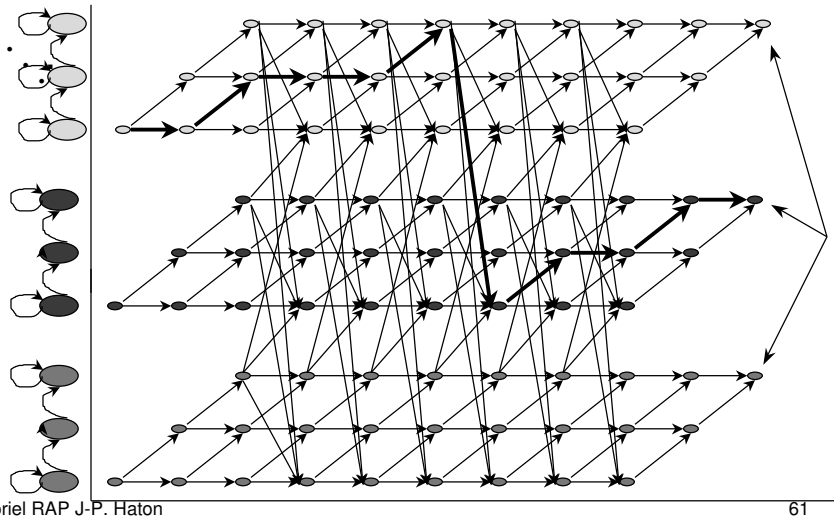
Algorithme efficace pour les mots enchaînés



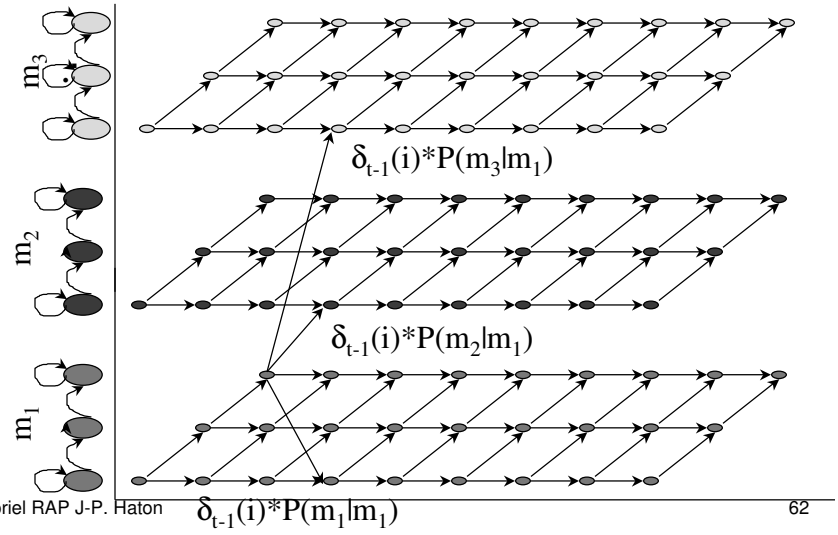
Graphe final et solution

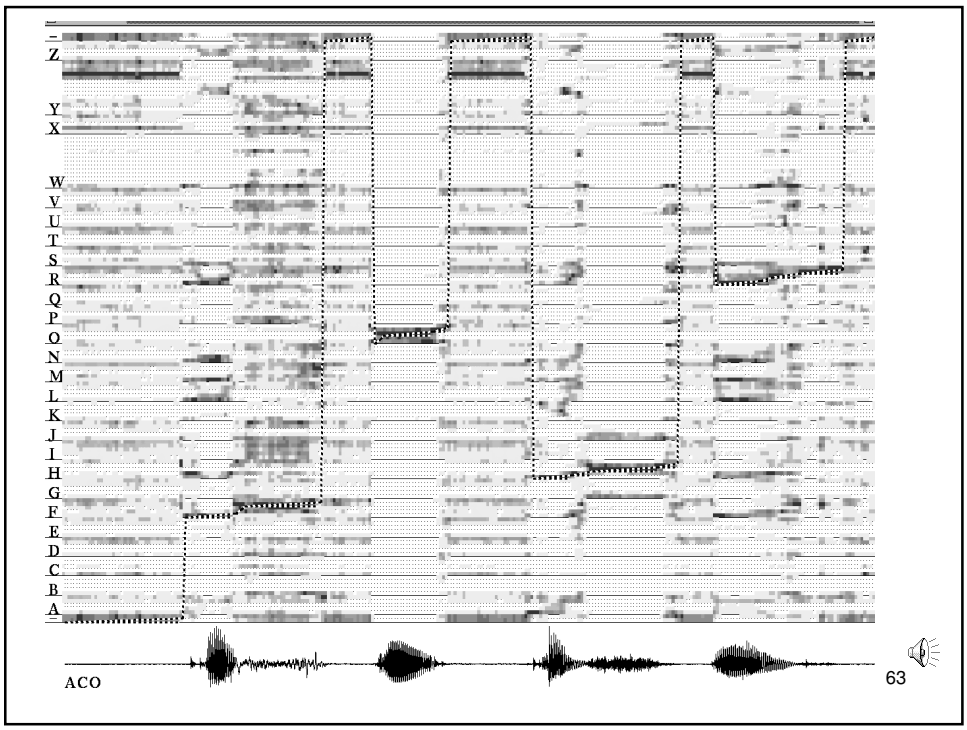


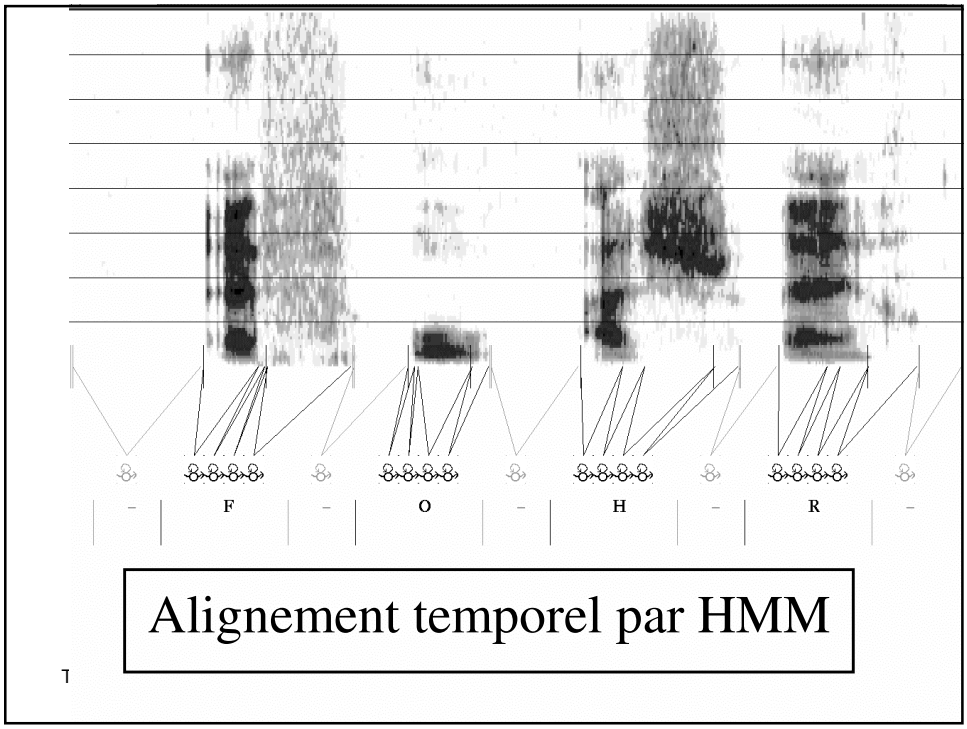
Solution obtenue par retour-arrière



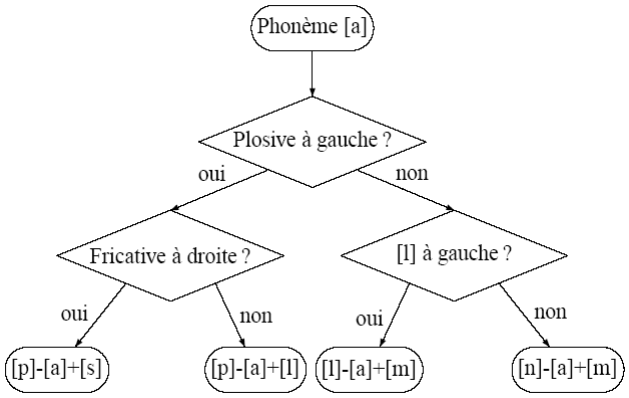
Introduction d ’une grammaire




Z
Y
X
W
V
U
T
S
R
Q
P
O
N
M
L
K
J
I
H
G
F
E
D
C
B
A
ACO
63

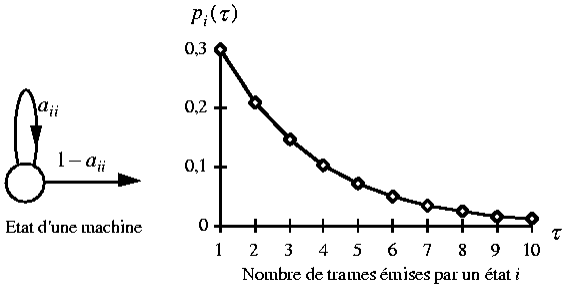

F
O
H
R
Alignement temporel par HMM

Partage de paramètres



Modélisation de la durée dans les HMM

Par construction : loi exponentielle



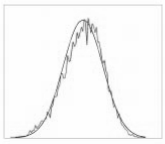
Diverses solutions possibles

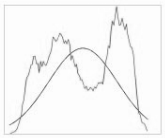
Amélioration des modèles : solutions actuelles

- Idée : dépasser le cadre statistique standard
- Accroître le volume de données d'apprentissage
- Améliorer la représentation des distributions de probabilités

Représentation des distributions de probabilités

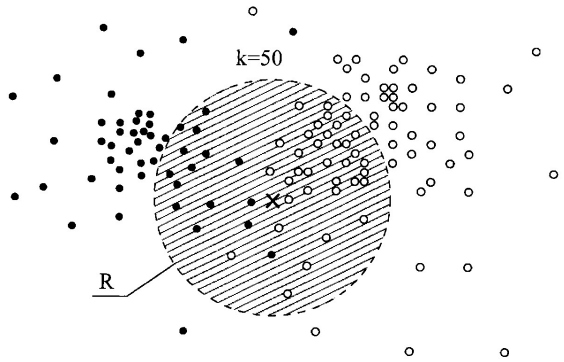
- Masses de probabilités discrètes
- Distributions continues :
  - Gaussiennes uniques
  - Mélanges de gaussiennes
- Réseaux neuromimétiques
- Modèles hybrides neuronaux-HMM
- Représentation non-paramétriques
- Représentation temps-fréquence : « HMM2 »





Représentation non paramétrique

cf. Lefèvre (2000) :
- Estimation des distributions de probabilités par la méthode des plus proches voisins
- Résultats mitigés : nécessité de nouvelles topologies…

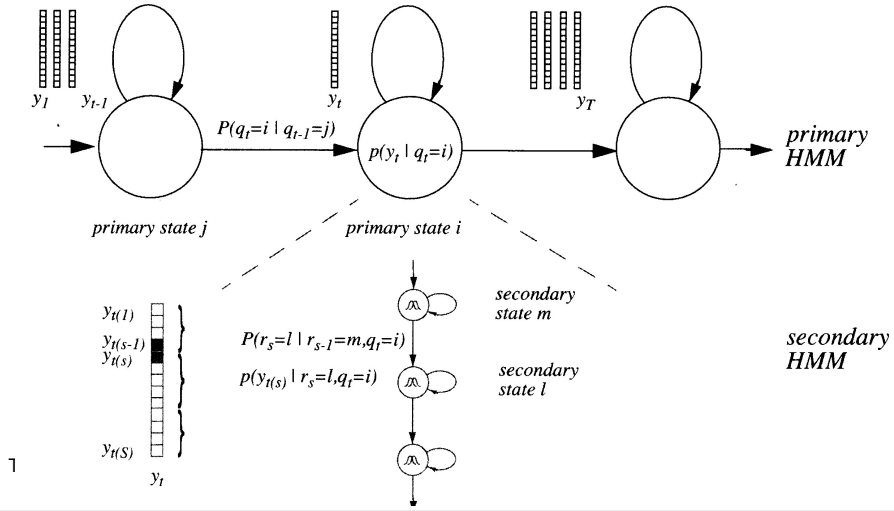


Représentation des distributions de probabilités

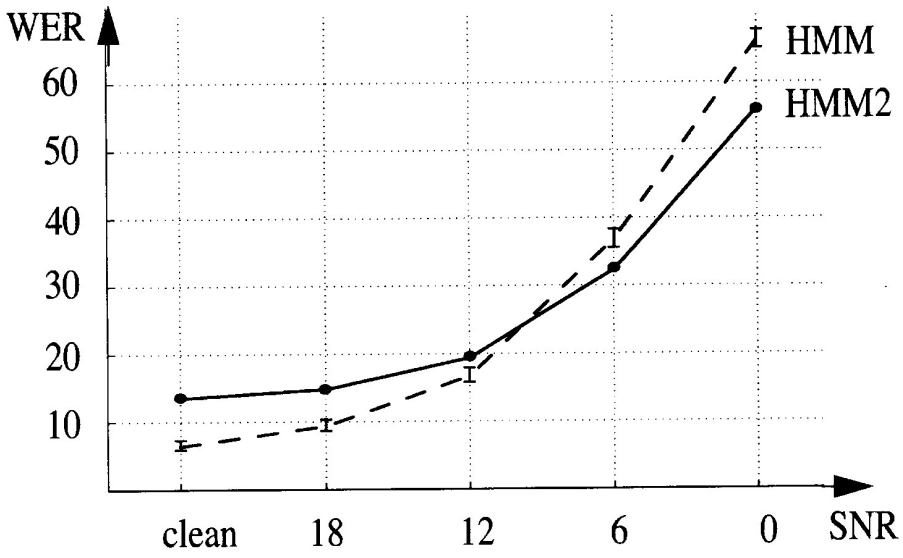
- Masses de probabilités discrètes
- Distributions continues :
  - Gaussiennes uniques
  - Mélanges de gaussiennes
- Réseaux neuromimétiques
- Modèles hybrides neuronaux-HMM
- Représentation non-paramétriques
- Représentation temps-fréquence : « HMM2 »

Principe des HMM2

cf. Weber, Bourlard (2000): parole et Levin (1993):vision



HMM2 : performances

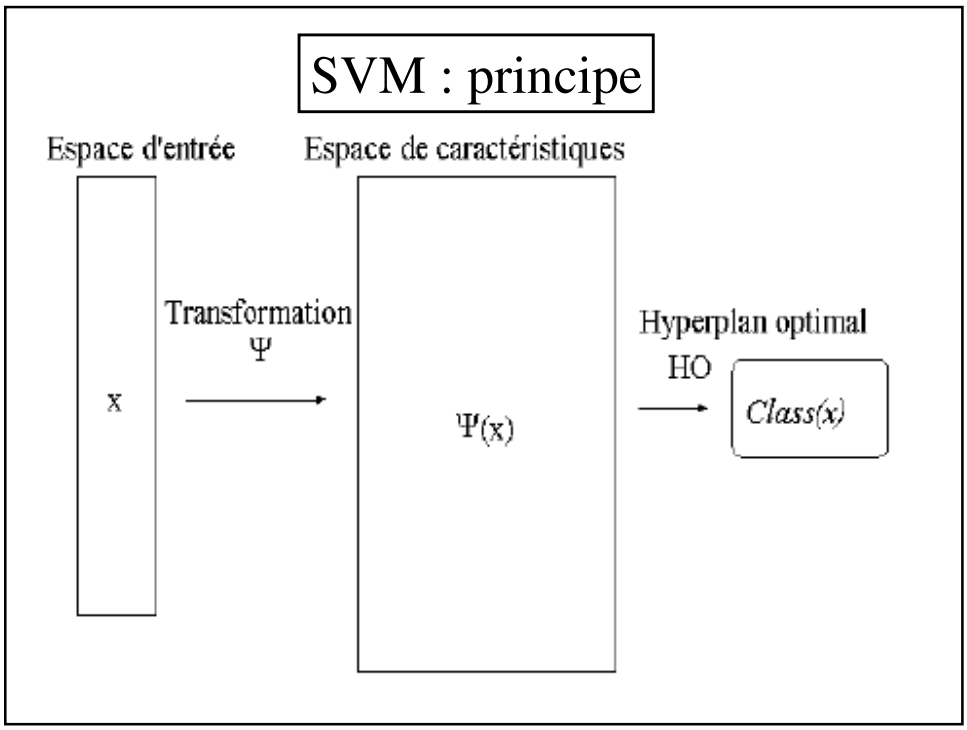


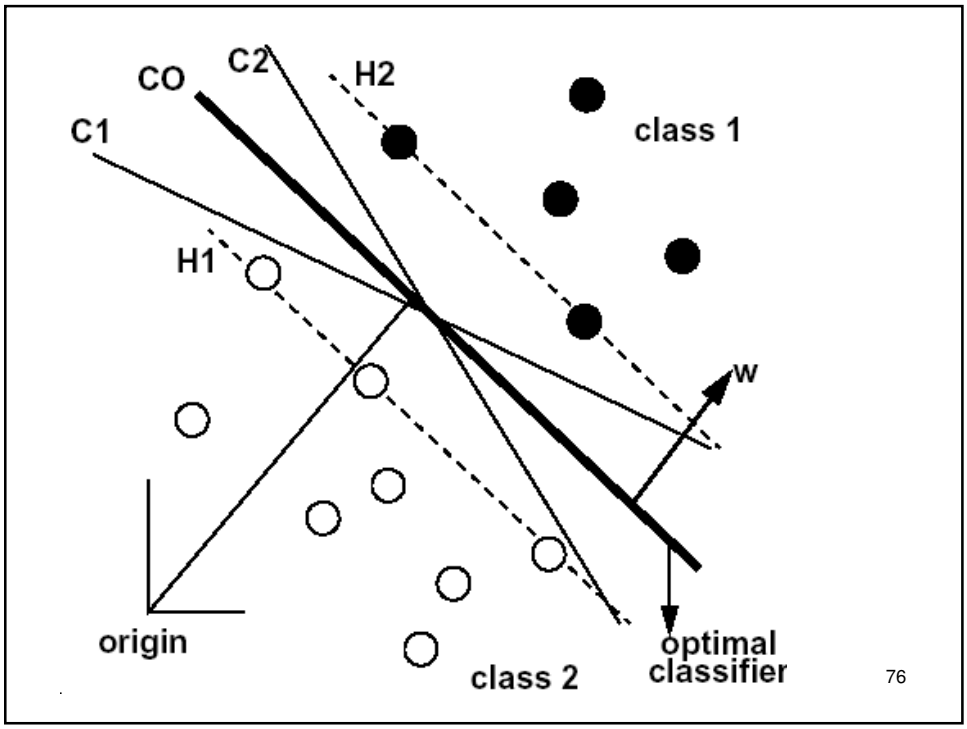
Solutions actuelles

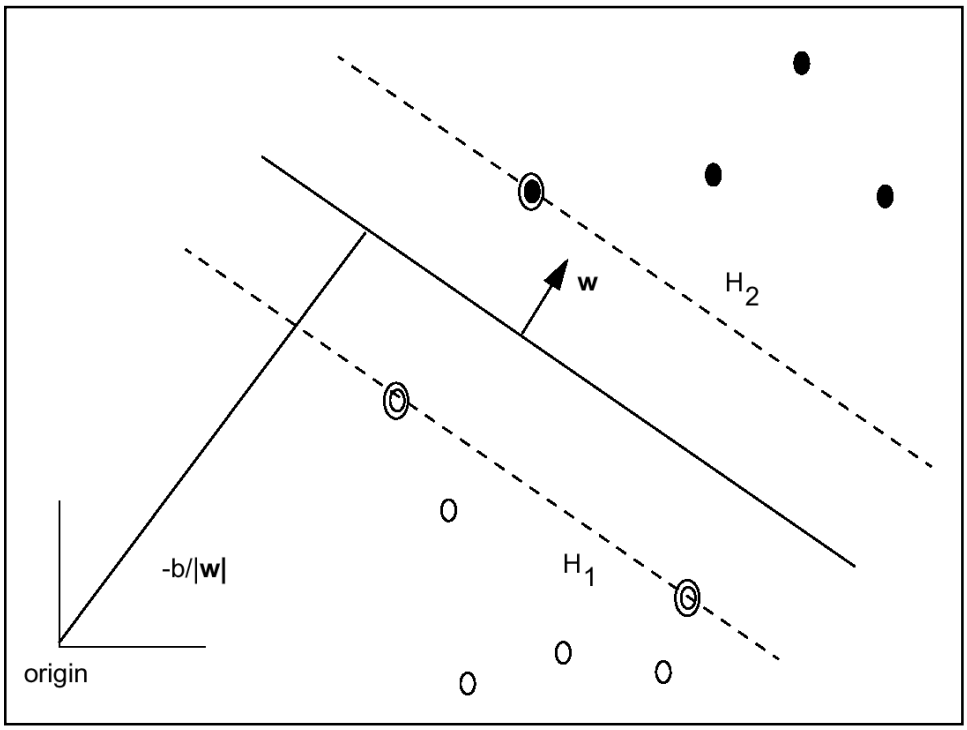
- Idée : dépasser le cadre statistique standard
- Accroître le volume de données d'apprentissage
- Améliorer la représentation probabiliste
- Complexifier les modèles :
  - HMM du second ordre (HMM2): Haton-Mari (1991)
  - Adaptation : MLLR, MAP, *eigenvoices, etc.*

Solutions actuelles

- Idée : dépasser le cadre statistique standard
- Accroître le volume de données d'apprentissage
- Améliorer la représentation probabiliste
- Complexifier les modèles :
  - HMM du second ordre (HMM2): Haton-Mari (1991)
  - Adaptation : MLLR, MAP, *eigenvoices, etc.*
- Complexifier l'apprentissage :
  - MCE (*min. d'erreur de classification*), MMI (*max. d'information mutuelle*) vs ML (*max. de vraisemblance*)
  - Réseaux neuromimétiques et modèles hybrides
  - SVM (Vapnik, 1995)


SVM : principe
Espace d'entrée
Espace de caractéristiques
x
Transformation
Ψ
Ψ(x)
Hyperplan optimal
HO
Class(x)


CO
C2
H2
C1
class 1
H1
w
origin
class 2
optimal
classifier
76



Solutions « nouvelles »

- Idée : rechercher de nouveaux formalismes :
  - des modèles mathématiques
  - des mécanismes d'apprentissage
  - la compréhension des mécanismes sous-jacents
- Tentatives intéressantes :
  - modèles segmentaux
  - modèles multibandes
  - modèles graphiques
  - données manquantes

Solutions « nouvelles »

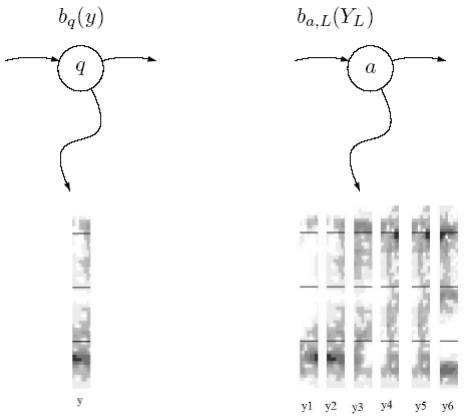
- Idée : rechercher de nouveaux formalismes :
  - des modèles mathématiques
  - des mécanismes d'apprentissage
  - la compréhension des mécanismes sous-jacents
- Tentatives intéressantes :
  - modèles segmentaux
  - modèles multibandes
  - modèles graphiques
  - données manquantes

Modèles segmentaux

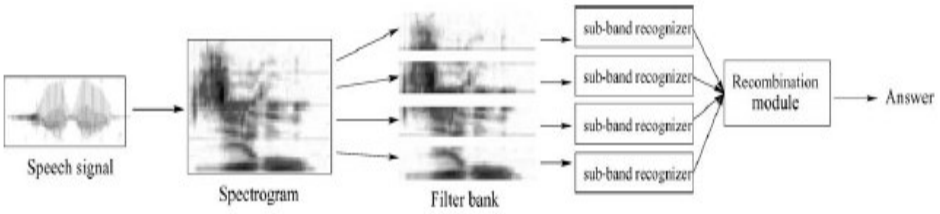
Principe
Exemples : Gong, Haton (1991), Ostendorf (1996), etc.
Résultats : mitigés …

Extension : SUMMIT (Zue, 2000), (Glas, 2003)

- Supprimer la notion d'échantillonnage fixe
- Un segment est un « atome » de parole
- Complexité important mais bons résultats
- Application à d'autres champs de la RF



Modèles multi-bandes



Inspiré des travaux de Fletcher sur l'audition (*cf.* Allen, 1994)
Exemples : Bourlard (1996), Cerisara, Haton (1996), Hermansky (1996), Okawa (1998), Mirghafori (1999), etc.
Améliorations du couplage inter-bandes:

- Pondération des bandes (Zhu, 2003)
- HMM factoriels (Nock, 2003) (*cf.* Jordan (1997))
- Couplage d'états par réseaux bayésiens (Daoudi, 2003)

Modèles graphiques

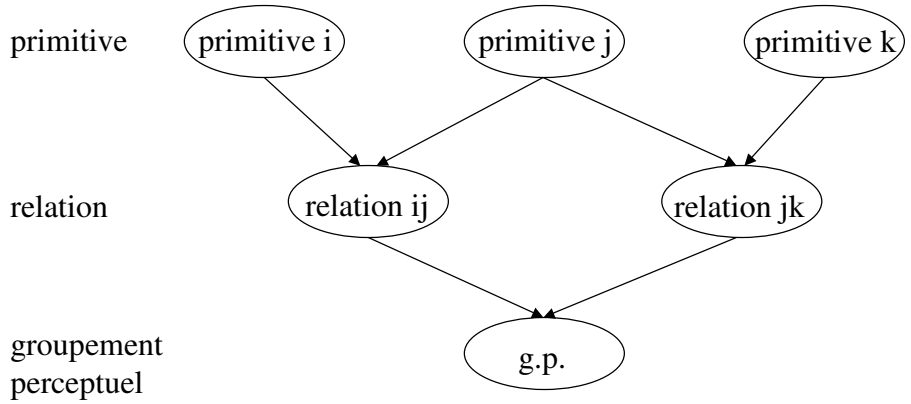
- Principe : modèles graphiques probabilistes
  - graphes non orientés : champs de Markov
    application à la parole : Gravier (1998)
  - graphes orientés : réseaux bayésiens (Pearl, 1988)
    (Jordan, 1999) et RB dynamiques

Réseaux bayésiens

- Principes :
  - Graphes acycliques :
    nœuds : variables aléatoires
    arcs : indépendances conditionnelles entre nœuds
    -> représenter et exploiter la causalité entre variables
- Inférence : calcul des probabilités conditionnelles de certaines variables (algorithmes JLO, Dawid)
- Applications : images, parole, diagnostic, robotique,..

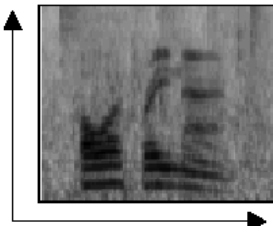
Réseaux Bayésiens

Nœuds

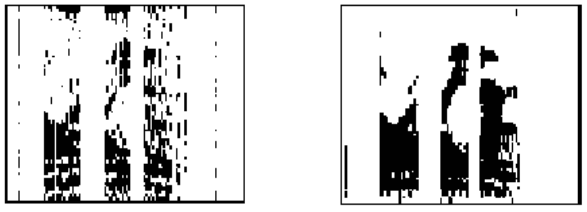


Principe des données manquantes

Initial spectrogram (digit sequence in factory noise)



Reliable regions



Modèles de langage

- Calcul de la probabilité d'une suite de mots P(w1...wn)
- Méthode courante : modèles *n-grammes* (bi- ou tri-)
  - calcul de P(w/w1...wk) avec k=1 (bi) ou 2 (tri) pour un mot w
  - calcul de la probabilité d'une séquence de mots
  - prédiction du mot suivant

Modèles de langage

- Nombreux autres modèles
  - Modèles n-classes (syntaxiques ou sémantiques)
  - modèles n-grammes avec caches
  - modèles multigrammes (suites de mots)
  - modèles hybrides (combinant plusieurs modèles)
- Modèles stochastiques et linguistiques
- Apprentissage : nécessité de très gros corpus

Modèles de langage

- Amélioration des modèles :
  - problème de la limitation des corpus d'apprentissage
  - méthodes d'interpolation
- Evaluation des modèles
  - mesure de la qualité de représentation d'un langage
  - valeur moyenne : perplexité ou entropie

Combinaison des modèles acoustiques et de langage (K. Smaili)

Fudge factor

- On ne peut pas multiplier simplement les probabilités provenant des deux modèles
- Il faut opérer une pondération : P(X/W) << P(W)
- Ajout d'un coefficient appelé : linguistic weight ou fudge factor

$$\hat{W} = \arg\max P(X/W) \times P(W)^{lw}$$

- Lw est déterminé empiriquement (7)

Combinaison des modèles acoustiques et de langage

Pénalité linguistique

- Pour ne pas favoriser certaines phrases composées de peu de mots longs ou au contraire de séquences composées de nombreux mots courts.
- Ajuster le système par une pénalité linguistique

$$\hat{W}=\arg\max P(X/W)\times lp^{N(W)}P(W)^{lw}$$

- N(W) le nombre de mots de la phrase.
- lp est déterminé empiriquement

Loi de Zipf

La loi de Zipf

Lorsqu'on s'intéresse à la distribution des fréquences de mots, on s'aperçoit qu'elle obéit à une loi dite loi de Zipf qui reste valable quelque soit le corpus.

$f * r = C$

Rang

Fréquence

Le mot de rang 100 est 10 fois moins fréquent qu'un mot de rang 10

Autres modèles de langage

- **Modèle Cache**
- L'idée de base du cache est si un mot apparaît dans un texte, il a de forte chance de réapparaître.
- Modèle introduit par Kuhn et De Mori (90). Il permet de renforce un mot lorsque celui-ci a été rencontré dans l'historique.

$$P_{cache}(w_i / w_{i-M}^{i-1}) = \frac{1}{M}\sum_{m=1}^{M}\delta(w_i / w_{i-m}) \quad avec \delta(x,y) = \begin{cases} 1 & si \ \ x = y \\ 0 & \text{sinon} \end{cases}$$

- Le cache est combiné avec un modèle de base :

$$P(w_1 \cdots w_n) = \lambda P(w_i / h) + (1-\lambda) P_{cache}(w_i / w_{i-M}^{i-1})$$

Modèle Cache

- **Inconvénients du cache**
- La position du mot n'est pas prise en compte pour le renforcement de la probabilité du mot courant.
- Aller vers un cache dépendant de la position : le poids du mot décroît avec sa distance par rapport au mot courant.
- Modèle appelé Decaying history

$$P_{cache}(w_i / w_{i-M}^{i-1}) = \beta \sum_{m=1}^{M} \delta(w_i / w_{i-m}) e^{-\alpha(i-m)}$$

- α : Paramètre de décroissance
- β : Constante de normalisation

Autres types de modèles de langage

Les modèles distants

- La portée des modèles de langage de type n-gramme est très limitée.
- Augmenter n pour couvrir des modèles plus importants.
- Difficulté de réalisation
- Prise en compte de modèles n-grammes distants ou à trous
- On s'intéresse dans ce cas à : $P(w_i/h)=P(w_i/h_d)$
- $h_d = w_1 \ldots w_{i-d}$

Les modèles distants

Cas du modèle bigramme distant :

Il ne peut pas venir aujourd'hui.

Le metteur en scène a raté sa prise.

La tentation du Christ a été interdit

Formulation du modèle distant :

$$P(w_i/h)=\sum_{d=1}^{K}\lambda_d P_d(w_i/w_{i-d})$$

λ : coefficients d'interpolation et *d* distance du modèle.

$$P_d(w_1\cdots w_n)=\frac{N(w_1\cdots w_{n-1},w)}{N(w_1\cdots w_{n-1})}$$

d mots séparent les 2 termes

Les modèles n-classes

- Pour atténuer le problème de manque de données, on peut concevoir les modèles de langage autrement.
- Regroupement de mots dans les classes.
- Introduction de connaissances linguistique.
- Regroupement peut se faire sur des critères syntaxiques ou sémantiques :
- **Exemples :**
  - ARD = {Le, La, les, l'}
  - VEC ={mange, parle, tire, regarde, …}
  - Jour = {Dimanche, Lundi, ….}
  - Fruits = {Poire, Pomme, …}

Les modèles n-classes

- Formulation : Plusieurs possibilités.

$$P(w_3/w_1w_2)\cong\begin{cases}P(w_3/C_3)P(C_3/w_1w_2)\\P(w_3/C_3)P(C_3/w_1C_2)\\P(w_3/C_3)P(C_3/C_1C_2)\\P(w_3/C_1C_2)\end{cases}$$

- Les classes sont déterminées manuellement ou automatiquement.

Les modèles à horizon variable

- N-gramme : Contexte de longueur fixe

$$P(w_1 \cdots w_n) = \prod_{i=1}^{n} P(w_i / w_{i-k+1} \cdots w_{i-1})$$

- P(abcd)=P(a/##).P(b/#a).P(c/ab).P(d/bc)

- Multigramme : contexte de longueur variable.
- Indépendance entre successives.
- W = $w_1 w_2 \ldots w_n$
- Soit S une décomposition de W en Qs sous-séquences
- $S = S_1 S_2 \ldots S_{Qs}$ avec long(S_k) <= m

Les modèles à horizon variable

- Le modèle m-multigramme est approché par :

$$P(w) = \max_{\{S\}} \prod_{j=1}^{Q_s} P(S_j)$$

Approximation sur toutes les segmentations

- Soit w = abcd

$$P(w) = \max \begin{cases} P(abc)P(d) \\ P(a)(bcd) \\ P(ab)P(cd) \\ P(ab)P(c)P(d) \\ P(a)P(bc)P(d) \\ P(a)P(b)P(cd) \\ P(a)P(b)P(c)P(d) \end{cases}$$

Segmentation : Trouver la segmentation de vraisemblance maximale
Algorithme de VITERBI

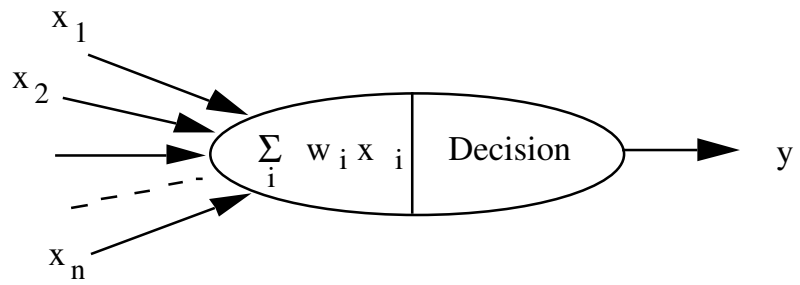
Exemple de systèmes de RAP à grand vocabulaire

– CUED-HTK (Woodland et al., 1998) : construit à l'université de Cambridge, GB, avec la boîte à outils *HMM Toolkit, HTK* (Woodland et Young, 1993)
– BBN Byblos (Kubala et al., 1998).
– CMU Sphinx (Lee et al., 1990) (Seymore et al., 1998)
– Dragon system (Wegmann et al., 1998).
– IBM system (Chen et al., 1998).
– OGI CSLU (Yan et al., 1998).
– Philips Research system (Beyerlein et al., 1998).
– SRI (Sankar et al., 1998).
– Microsoft Whisper (Huang et al., 1995).
– ISIP (Zheng et al., 2001) : boîte à outil gratuite et téléchargeable sur Internet.
– Julius (Lee et al., 2001) : Julius est un moteur de reconnaissance développé au Japon. Il est téléchargeable gratuitement sur Internet.
En France :
– LIMSI (Gauvain et al., 1994)
– Sirocco (Gravier et al., 2002) : plusieurs laboratoires français ont développé en commun ce décodeur à grand vocabulaire.

Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

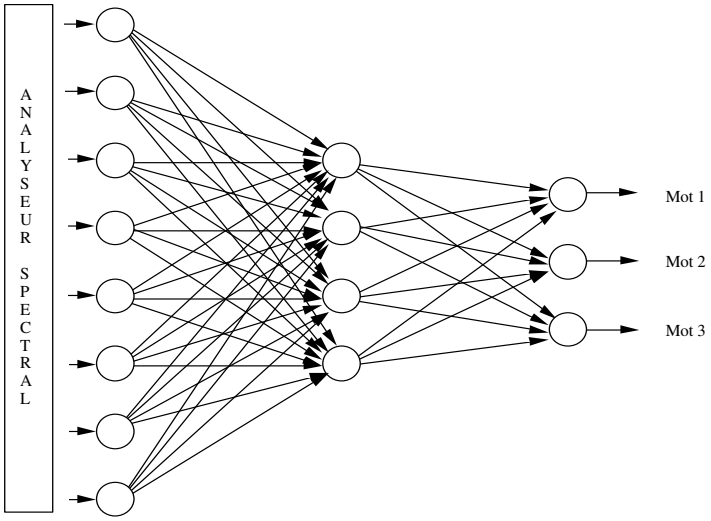
Le neurone «machine»



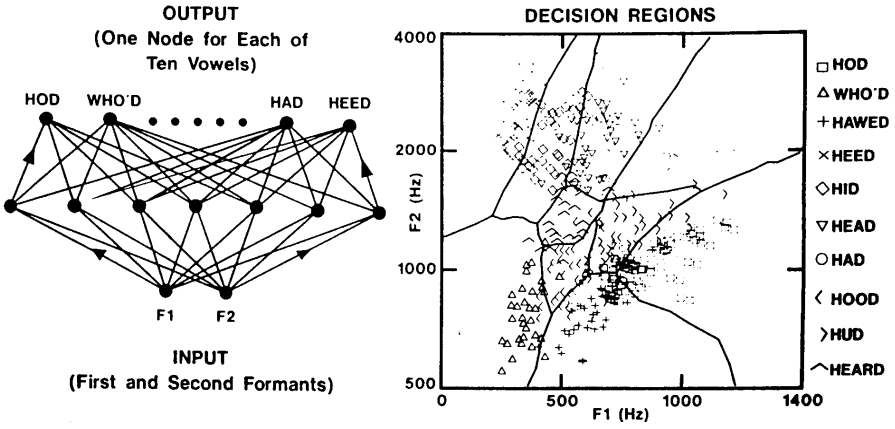
Modèles les plus courants en RAP

- RESEAUX MULTICOUCHES : PERCEPTRONS
- RESEAUX A CONNEXION COMPLETE
- RESEAUX RECURRENTS
- CARTES AUTO-ORGANISATRICES

Le perceptron multicouches

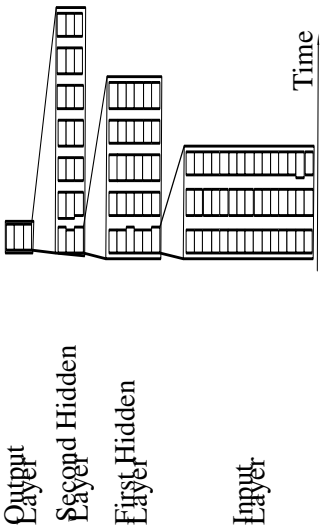


Reconnaissance de voyelles



- TRAIN WITH BACK PROPAGATION (50,000 Trials)
- ERROR RATE SAME AS BEST CONVENTIONAL CLASSIFIER
- DECISION REGIONS SAME AS DRAWN BY EXPERT

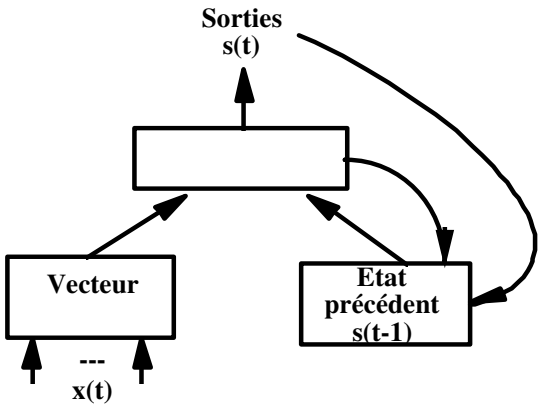
Exemple de TDNN (Waibel *et al.*)



RESEAUX RECURRENTS

- Principe pour un réseau multicouche
- Réseaux partiellement récurrents :
  - modèle de Jordan
  - modèle de Elman
  - modèle de Robinson
- Apprentissage : rétropropagation étendue avec "dépliage temporel" (Watrous, Shastri)
- Autre modèle : machine de Boltzmann (Prager, Fallside)
- Réseau récurrent hiérarchique, HRNN (Chen *et al.*)

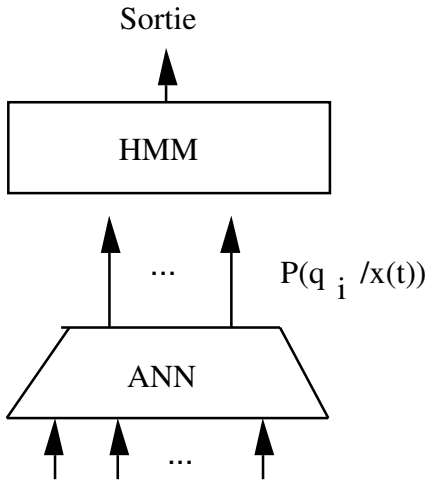
Réseaux récurrents



Modèles «biologiques»

- MODÈLE À PROPAGATION GUIDÉE (Béroule, 1985)
- APPRENTISSAGE PAR SELECTION (Changeux *et al.*, 1987)
- COLONNE CORTICALE (Burnod, 1988) (Alexandre *et al.*, 1990)
- COLONNE DE MÉMORISATION (Ans, 1990)
- NEURONES INTÉGRATEURS À FUITES (Wang-Arbib, 1990)
- etc.

Modèle hybride MCN-HMM



Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Approches fondées sur des connaissances explicites

- Décodage acoustico-phonétique
  - décodage et raisonnement
  - exemple : le système APHODEX
- Inversion de modèles articulatoires
- Reconnaissance multi-bandes
- Modèles multi-agents

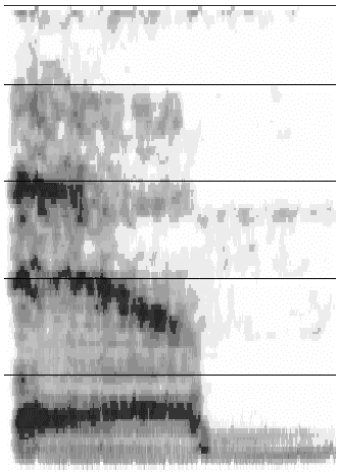
Plan de l'exposé

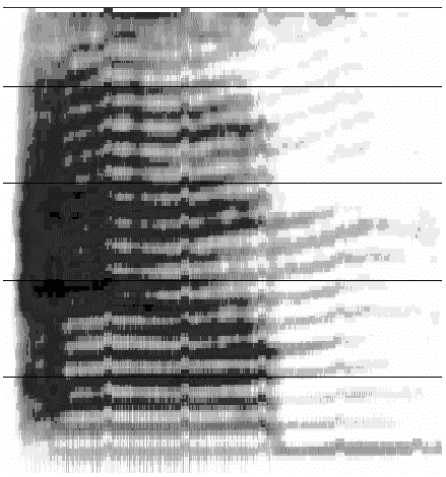
- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

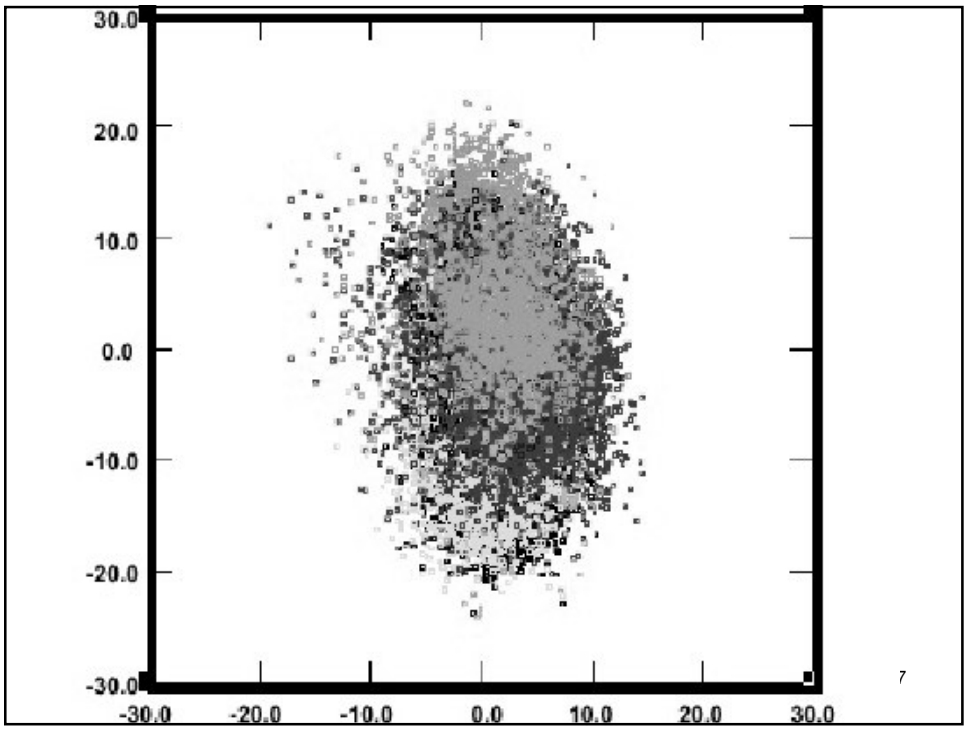
Robustesse des systèmes

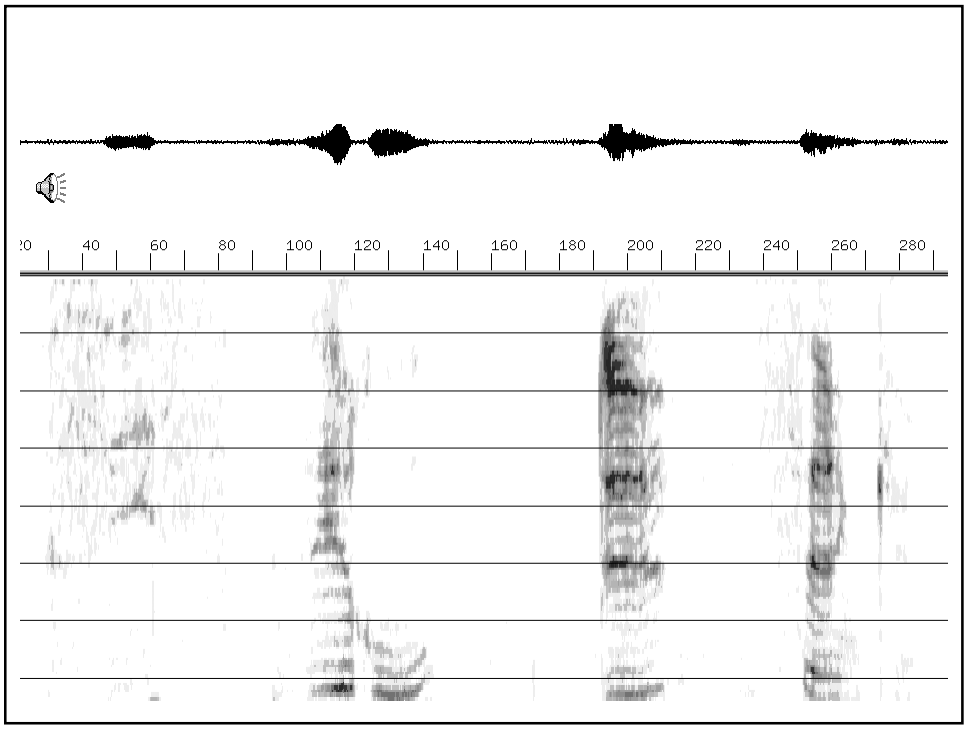
- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essouflement, fatigue, effet Lombard, etc.)
  - à la prise de son (microphone, bruit ambiant, position, etc.)
  - au canal de transmission (distorsion, écho, bruit électronique)
  - au contexte lingusitique (co-articulation, assimilatioin, etc.)
- Résultat :
  - interactions complexes et effets cumulés!
  - nécessité de méthodes robustes à tous les niveaux

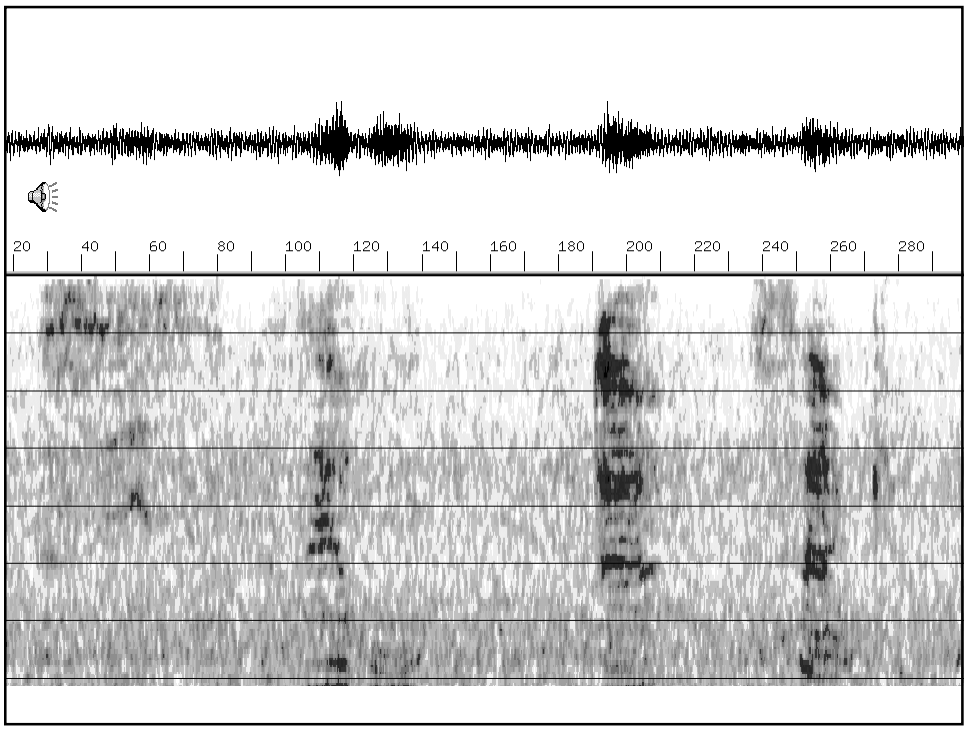
Effet Lombard



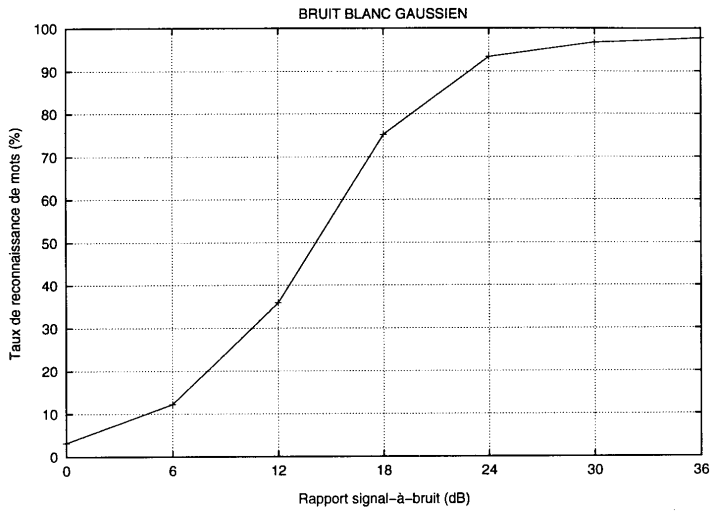



30.0
20.0
10.0
0.0
-10.0
-20.0
-30.0
-30.0
-20.0
-10.0
0.0
10.0
20.0
30.0
7

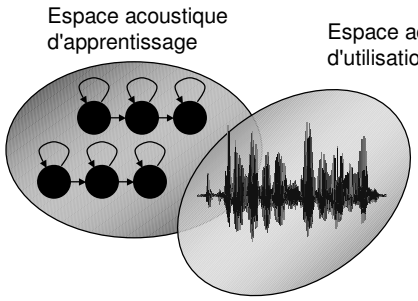

40
60
80
100
120
140
160
180
200
220
240
260
280



Influence du bruit sur le taux de reconnaissance



Robustesse et reconnaissance

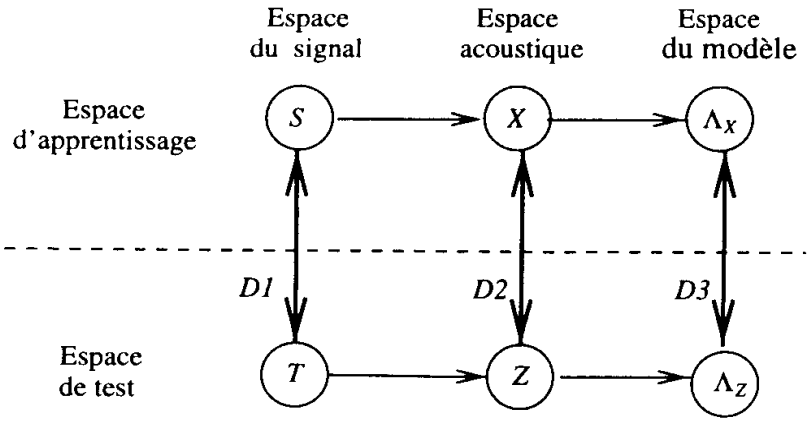


RM Task	WER
Native Speakers	3.6 %
Non-Native Speakers	34.9 %
Telephone Channel	

WSJ 5K Task – SI 84	WER
Native Spk. (Nov92)	4.7 %
Non-Native Speakers	29.1 %

La différence entre les deux espaces fait chuter les performances

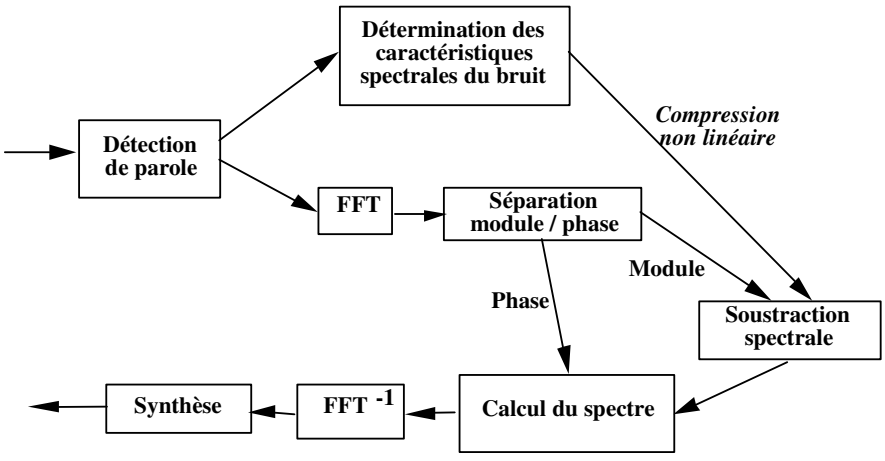
Nécessité de méthodes robustes



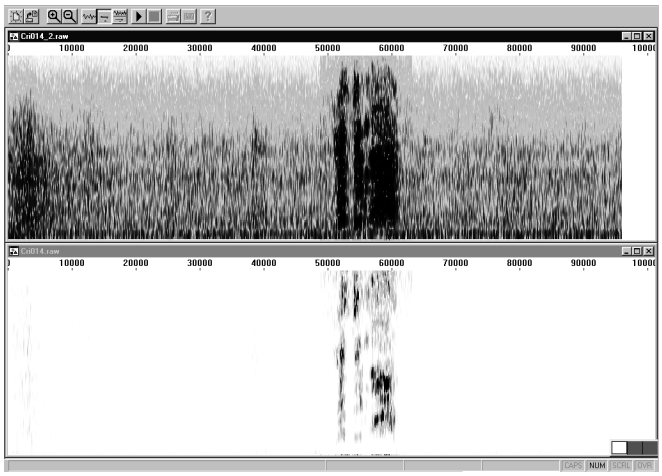
Traitement du signal

- Microphones : réducteurs, antennes
- Filtrage adaptatif
- Soustraction spectrale :
  - Principe : linéaire / non linéaire
  - Exemple : détection de cris de détresse (RATP)

Soustraction spectrale



Détection de cris (RATP)

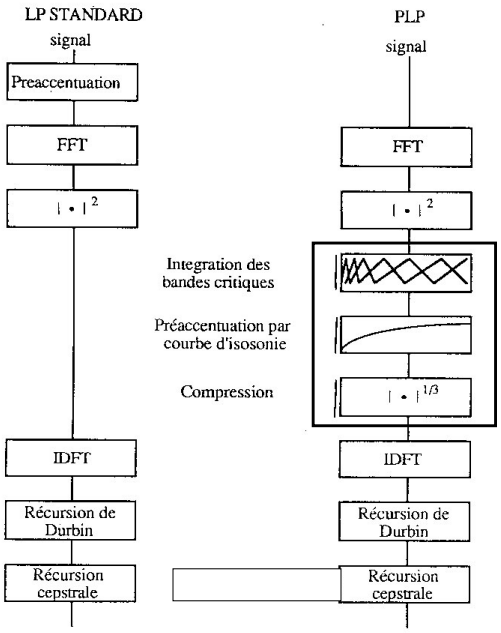


Tutoriel RAP J-P. Haton 125

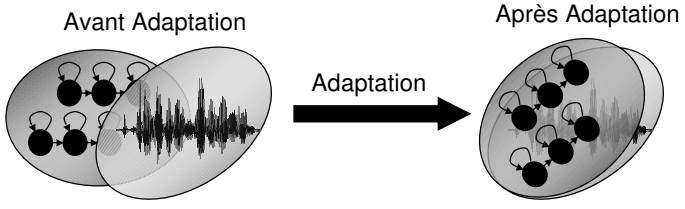
Paramétrisation robuste

- Utilisation de connaissances en perception, psychoacoustique, neurologie, etc. :
  - modèles de l'oreille
  - masquage de bruits
  - paramètres dynamiques
  - analyses PLP, RASTA, RASTA-PLP, etc.
- Normalisation cepstrale

Tutoriel RAP J-P. Haton 126



Adaptation des modèles

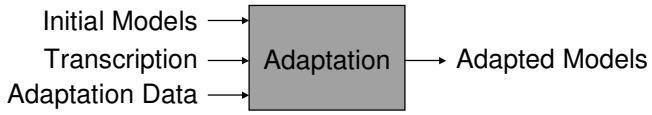


Adaptation: motivations et modes

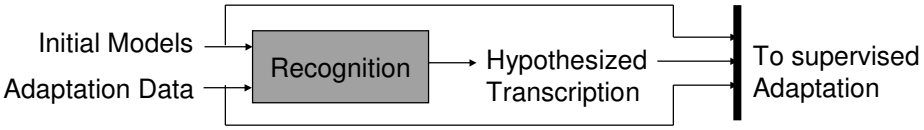
- But : améliorer les performances d'un système indépendant du locuteur
- L'adaptation doit
  - fonctionner avec peu de données
  - être rapide
- Modes d'adaptation
  - Supervisée
  - Non supervisée
  - Incrémentale

Scenario d'adaptation

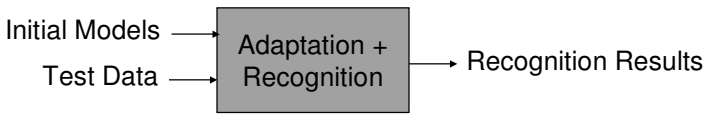
- ***Batch supervised adaptation***



- ***Batch unsupervised adaptation***



- ***Online (unsupervised) adaptation***



Transformations linéaires

- Principe : on obtient les paramètres du modèle adapté en appliquant une transformation linéaire aux paramètres du modèle initial
- Les transformations sont obtenues par maximisation de la vraisemblance des données d'adaptation
- Exemple le plus courant : MLLR *(Maximum Likelihood Linear Regression)*

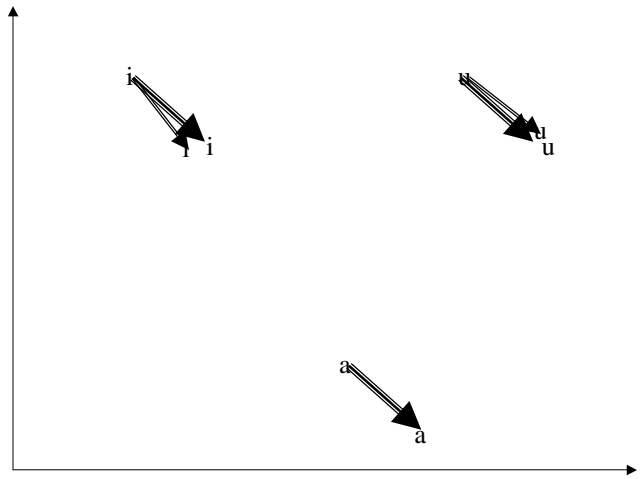
MLLR

- La moyenne μ_s d'une distribution *s* est adaptée en lui appliquant la transformation linéaire :

$$\hat{\mu}_s = A_s \mu_s + b_s$$

  - μ_s est la moyenne issue de l'apprentissage
  - $\hat{\mu}_s$ est la moyenne adaptée

MLLR : Example



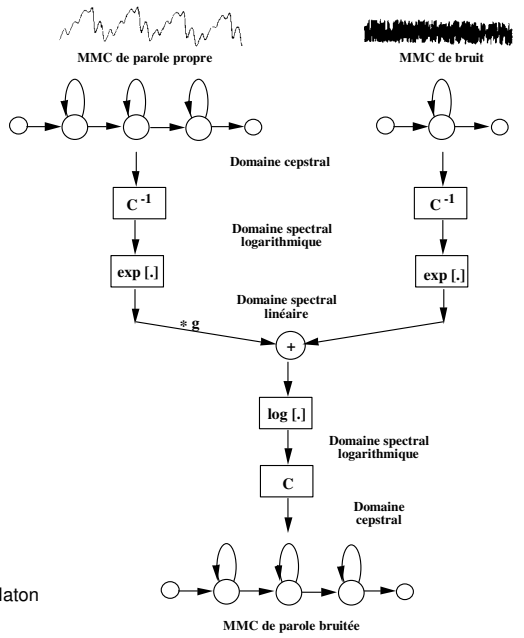
MAP (maximum *a posteriori)*

- Apprentissage d'un HMM :

$$Arg\max_{\lambda} P(\lambda|O) = Arg\max_{\lambda} \frac{P(O|\lambda)P(\lambda)}{P(O)}$$

- Si on a aucune connaissance sur P(λ) :
  argmax P(O|λ) : maximum de vraisemblance
- Dans le cas de MAP, on considère les modèles indépendants du locuteur comme probabilité *a priori*
- Problème: adapte uniquement les paramètres rencontrés pendant l'adaptation

Combinaison de modèles, PMC (Gales, 1995)



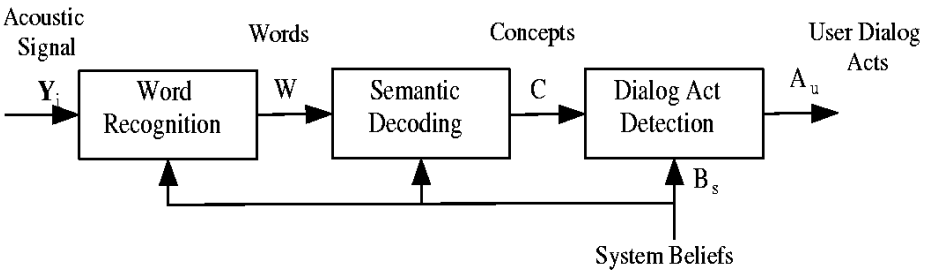
Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

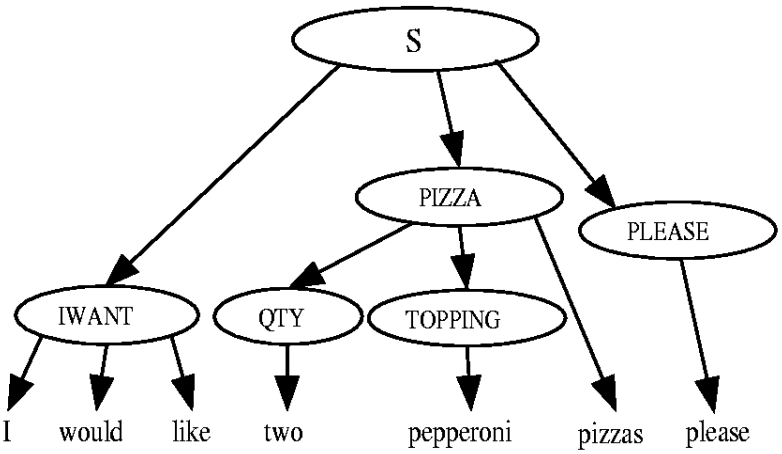
Dialogue oral homme-machine

- Rôle fondamental pour la compréhension
- Forme très diverse selon la tâche (traitement de textes, centre de renseignements grand public, commande de systèmes complexes)
- Pour simplifier : 3 niveaux de dialogue

Compréhension de la parole : approche statistique



Exemple d'arbre d'analyse sémantique (d 'après S. Young)



Tu

Niveaux de dialogue

- Dialogue par mots-clés
  - simple, très dirigé, le plus courant
- Dialogue dirigé en langue «naturelle»
  - liberté d'expression dans un cadre prédéfini
- Dialogue naturel
  - initiative au locuteur, rôle de l'IA

Plan de l'exposé

- **Introduction**
- **La communication parlée**
- **Analyse du signal acoustique**
- **Approche statistique de la reconnaissance**
- **Utilisation de modèles neuromimétiques**
- **Approches fondées sur des connaissances**
- **Robustesse des systèmes**
- **Compréhension et dialogue homme-machine**
- **Application de la RAP**
- **Conclusion et perspectives d'avenir**

Domaines d'application (1)

Contexte général :
- évolution technologique de la microinformatique
- intégration des logiciels de TAP dans les postes de travail
- extension des réseaux interconnectés et de la télématique

- TÉLÉCOMMUNICATIONS
. SERVICES DE TÉLÉMATIQUE VOCALE (SVI)
. COMPOSITION DE NUMÉROS
. SERVICES SPÉCIAUX (call collect, Mairievox, banques, etc.)
. BORNES INTERACTIVES
. ACCÈS À DES BANQUES D'INFORMATIONS

Domaines d'application (2)

Les générations de systèmes de télématique vocale :

- première génération (début vers 1990)
  - . Amérique du Nord : ATT (VRCP), Bell Northern (AABS)
  - . Japon : banques (ANSER)
  - . France : CNET (Mairievox), MACIF

- deuxième génération (début vers 1994)
  - . automatisation partielle des services de renseignements (Bell Canada, ATT)
  - . annuaires vocaux d'entreprises (CSELT, CNET)
  - . répertoires et composeurs vocaux personnalisés (NYNEX, Sprint)
  - . cas particulier des téléphones mobiles

- génération future : intégration de services, traduction, dialogue oral, identification de la langue

Domaines d'application (3)

- ENTRÉE DE DONNÉES
- **. SAISIE D'INFORMATIONS**
- **. MACHINE À DICTER ET BUREAUTIQUE**

- COMMANDE DE MACHINES ("mains libres")
- **. AVIONS, HELICOPTERES**
- **. SYSTEMES DE GUIDAGE D'AUTOMOBILES (GPS)**

- TRADUCTION AUTOMATIQUE (cf. ATR)

- JOUETS

- AIDE AUX HANDICAPÉS

RAP: performances actuelles

Task	Vocabulary	Style	Channel	Acoustics	% Word error
Air travel information system	2 000	Spontaneous, human to machine	High bandwidth	Clean	2.1
North American business news	60 000	Read	High bandwidth	Clean	6.6
Broadcasting news	60 000	Various	Various	Various	27.1
Switchboard	23 000	Spontaneous, conversational	Telephone	Clean	35.1

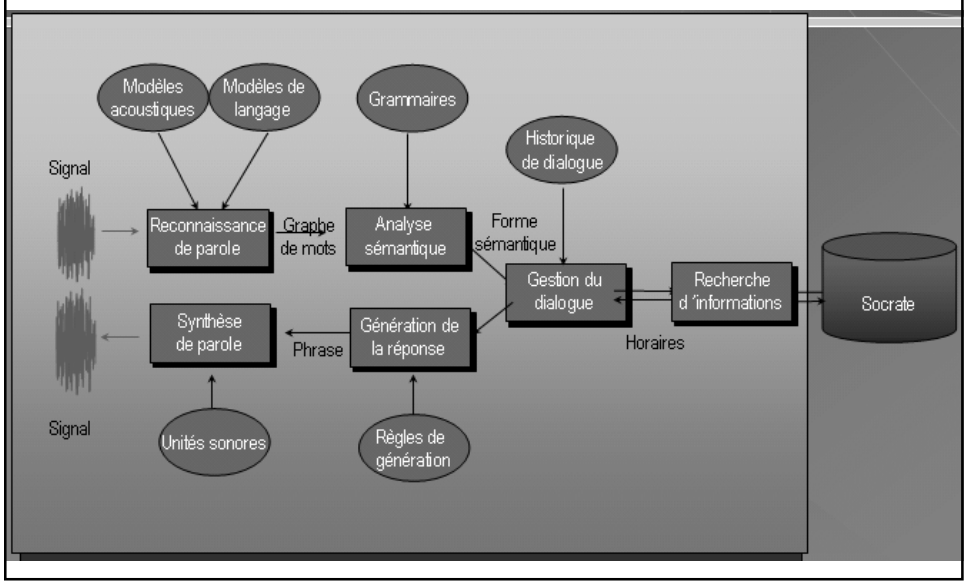
Domaines d'application et produits en reconnaissance de la parole

TYPE de FONCTION	DESCRIPTION	EXEMPLES
Contrôle/Commande	**Commande vocale d'appareils ou de logiciels :** **- chaise roulante,** **- appareillage,** **- commandes à un système d'exploitation d'ordinateurs,** **- numérotation téléphonique.** ***Mots isolés, petits vocabulaires***	**- Différents logiciels ou composants reconnaissant des mots isolés,** **- Téléphones avec numérotation vocale (Matra, Northern Telecom, Uniden),** **- Dragon Dictate et Via Voice comportent des commandes vocales au système d'exploitation.**
Saisie de données	**Entrée à la voix de données dans un ordinateur (remplissage de formulaires, contrôle de qualité, passage d'une commande, etc..).** ***Mots isolés, petits ou moyens vocabulaires***	**- Plusieurs prototypes dans différents domaines, pas de produits commercialisés.**
Télématique vocale	**Messagerie vocale.** **Accès à une base de données ou un centre de renseignements.** **Opérations bancaires.** ***Mots isolés, multilocuteurs, vocabulaires moyens.***	**- Système d'assistance au téléphone d'ATT,** **- Voice FONCARD de Sprint.**
Dictée vocale	**Production de lettres ou de documents écrits par dictée.** ***Parole continue, grands vocabulaires (plusieurs dizaines de mots), adaptation au locuteur.***	**- Natutally Speaking de Dragon,** **- Via Voice d'IBM,** **- Speech Magic de Philips,** **- Voice Xpress de Lernout et Hauspie**

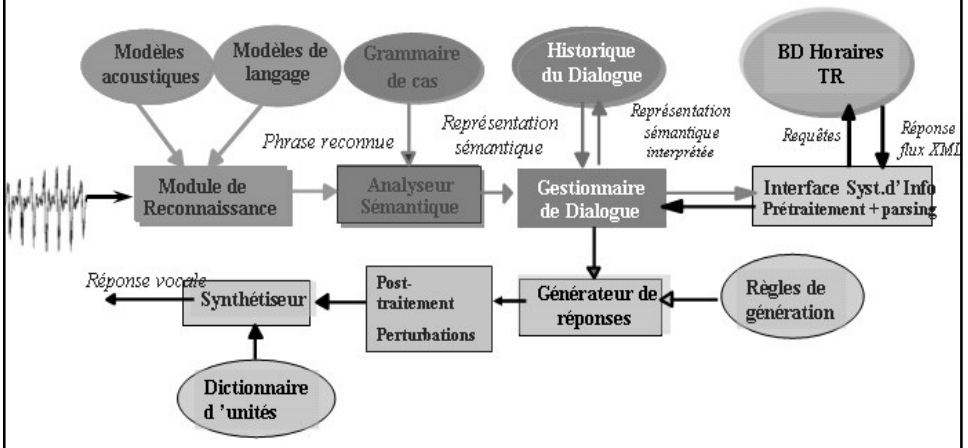
ASR Application Fields

-Telecommunications and telematics
- Car
- Avionics
- Oral control of machines
- Data entry, dictation machine
- Handicaps
- Toys
- Indexation and transcription
- Speech-to-speech translation

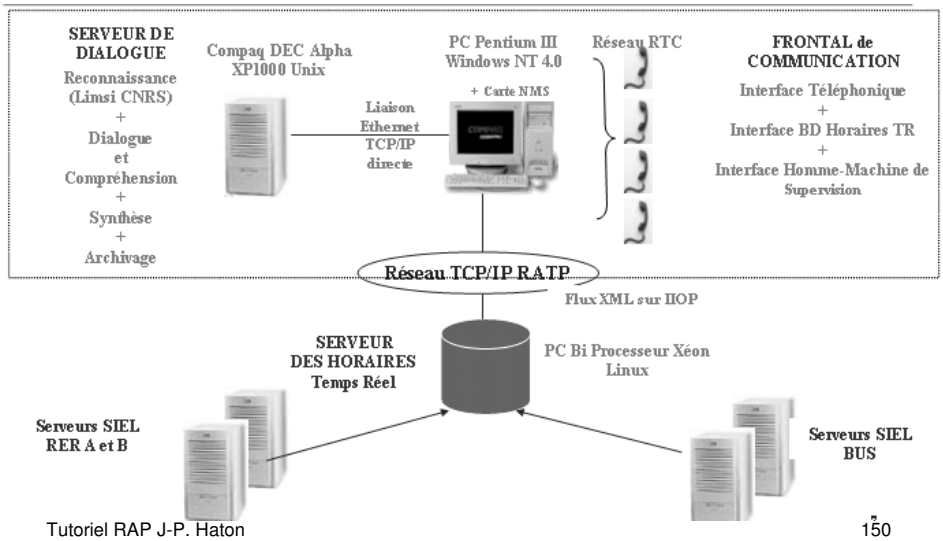
SNCF: RECITAL



RATP: SIEL



SIEL: architecture matérielle



Telematic Applications

Three generations of systems:

- *First generation (circa 1990)*

North America : ATT (VRCP), Bell Northern (AABS)
Japon : banks (ANSER)
France : CNET (Mairievox), MACIF

- *Second generation (circa 1994)*

Information systems (Bell Canada, ATT)
Company directories (CSELT, CNET)
Customized dialing systems (NYNEX, Sprint)
Cellular phones

- *Next generations*

Service integration
Language identification
Translation
Spontaneous dialog

Principe de l'annulation d'échos

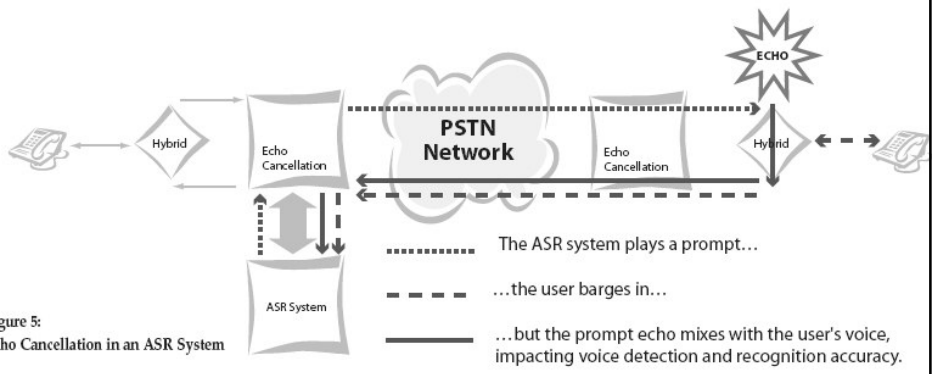
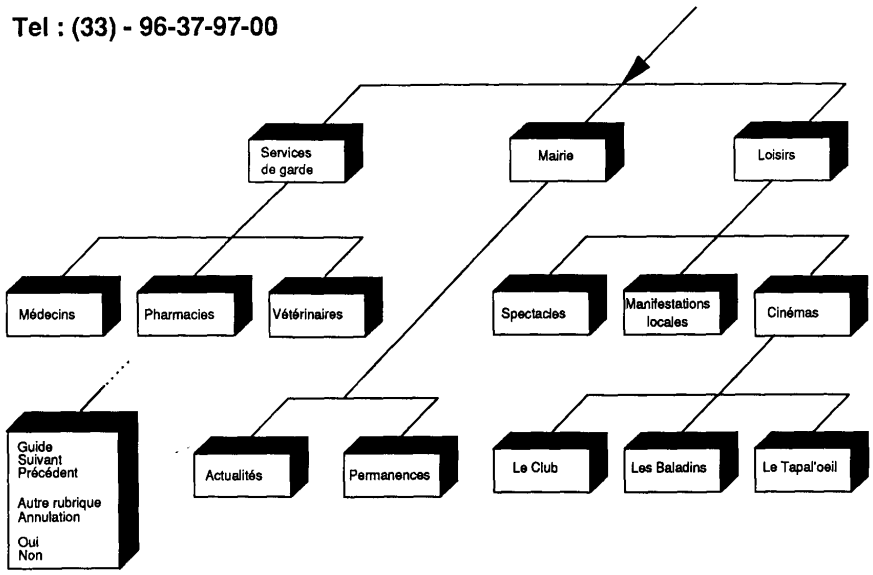


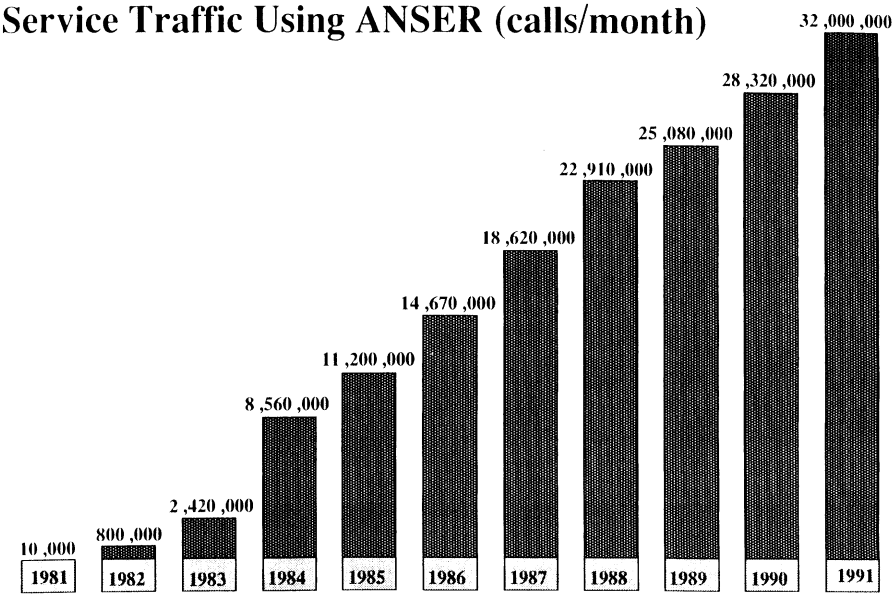
Figure 5:
Echo Cancellation in an ASR System

Serveur vocal interactif MAIRIEVOX

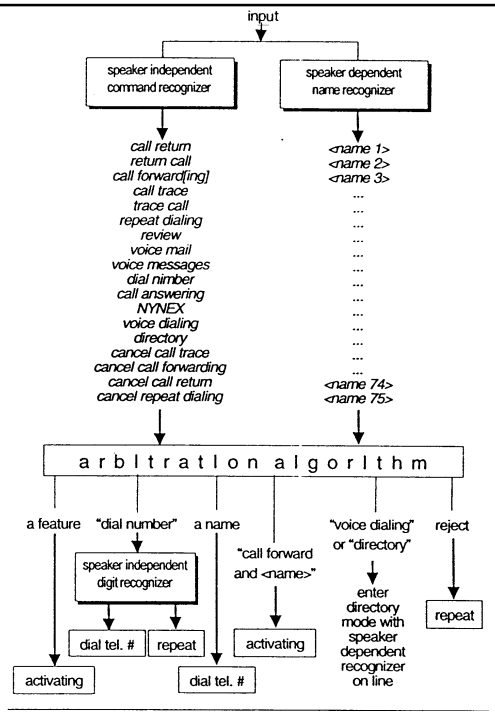
Tel : (33) - 96-37-97-00



Service Traffic Using ANSER (calls/month)



Le système de NYNEX



Machines à dicter

- Produits commercialisés dans différentes langues :
  allemand, anglais, espagnol, français, italien, mandarin, ...
- Applications : courrier commercial, comptes rendus médicaux, textes juridiques
- Bonnes performances, surtout après adaptation
- Exemples :
  - IBM *ViaVoice*
  - Dragon *Naturally Speaking*
  - Philips *Speech Magic*
  - Lernout et Hauspie *Voice Xpress*
  - …puis *ScanSoft*

Transcription/Indexation de documents audio

- Pourquoi ?
  - Explosion de la quantité de documents sonores
  - Exploitation manuelle impossible
- Pour quelles applications ?
  - consultation d'audiothèques
  - filtrage des émissions de radio
  - commerce de la musique sur le Web
  - accès aux contenus audiovisuels
  - post production de film, …
- Comment ?
  - extraction d'information représentative du contenu
  - organisation et structuration de l'information

Exemple de transcription automatique : RFI

neuf heures trente à paris sept heures trente en temps universel vous écoutez radio France internationale le journal jacques Alix bonjour bonjour à tous

les titres de l' actualité avec un espoir de dénouement de la crise des transports en France certains syndicats

DE ROUTIERS PARLENT

DONT IL PARLE

de progrès significatifs après de nouvelles négociations

à LA CLÉ L' éventuelle levée du blocus des dépôts de carburants

à LAQUELLE éventuelle levée du blocus des dépôts de carburants

le conseil de sécurité de l' ONU veut renforcer la capacité d' intervention des missions de PAIX de TÊTE

les chefs d' état se sont engagés à l' occasion du sommet du millénaire à New York

les ministres des finances de l' union européenne se retrouvent à Versailles

aujourd'hui l' Euro était au plus bas hier

EN CORSE UN nouvel assassinat hier soir sur les lieux mêmes du meurtre

ENCORE CE nouvel assassinat hier soir sur les lieux mêmes du meurtre

il y a tout juste un mois du responsable nationaliste Jean Michel Rossi

Conclusion et perspectives

- Progrès importants des recherches et produits actuels performants (machines à dicter, télématique, etc.)
- ... mais grande variabilité des taux d'erreur (valeurs en laboratoire, indépendants du locuteur) :
  - 0,3 % (suite de chiffres)
  - 5 % (dictée en continu, vocabulaire de 20 000 mots)
  - 8 % (lettres épelées)
  - 55 % (conversations téléphoniques)
- Nécessité d'augmenter la robustesse des systèmes à tous les niveaux de traitement pour des applications réalistes :
  - utilisateurs occasionnels
  - terminaux mains libres, ambiances bruitées
  - systèmes conversationnels
- Au-delà des HMM !

POUR EN SAVOIR PLUS...

OUVRAGES

R. BOITE et coll., « Traitement de la parole », Presses polytechniques et universitaires romandes, 2000.

J.P. HATON et coll., « La reconnaissance de la parole : du signal à son interprétation », Dunod, 2006.

J.C. JUNQUA and J.P. HATON, « Automatic Speech Recognition in Adverse Conditions: Fundamentals and Applications » , Kluwer Academic, 1995.

X. HUANG et al., « Spoken Language Processing », Prentice-Hall, 2001.

K.F. LEE, « Automatic Speech Recognition, the Development of the SPHINX System », Kluwer Academic, 1989.

L. RABINER, B. HUANG, "Fundamentals of Speech Recognition", Prentice-Hall, 1993.

REVUES

Computer Speech and Language

IEEE Transactions on Speech and Audio

Int. Journal of Pattern Recognition and Artificial Intelligence

Journal of the Acoustical Society of America

Speech Communication

Traitement du signal