6ième Ecole de Recherche Multimodale d'Information - TEchniques & Sciences

27, 28 et 29 septembre 2011

Sur l' Ile de Porquerolles,
« La Perle des Iles D'Or » du Var
http://glotin.univ-tln.fr/ERMITES11

ERMITES 2011

Décomposition Parcimonieuse, Abstraction et Structuration pour l'Analyse de Scènes Complexes

Intervenants :

- **J-P. HATON** (Université de Nancy - LORIA / IUF)
  Analyse de Scène et Reconnaissance Stochastique de la Parole
- **X. HALKIAS** (Université de Columbia – LabROSA)
  Detection and Tracking of Dolphin Vocalizations
- **M. KOWALSKI** (Université Paris 11 - L2S)
  *Sparsity and structure for audio signal: a *-lasso therapy*
- **J. RAZIK** (USTV - LSIS)
  Sparse coding: from speech to whales
- **Y. BENGIO** (Université de Montréal - LISA)
  Apprentissage Non-Supervisé de Représentations Profondes
- **S. MALLAT** (Ecole X - CMAP)
  Scattering & Matching Pursuit for Acoustic Sources Separation
- **O. ADAM** (Université Paris 6 - LAM / CNPS)
  Estimation de Densité de Population de Baleines par Analyse de leurs Chants
- **H. GLOTIN** (USTV - LSIS / IUF)
  Suivi & reconstruction du comportement de cétacés par acoustique passive

Comités :

- **H. Glotin**
  Président du comité de programme
- **S. PARIS**
  Président du comité d'organisation

Actes constitués sous la direction de :

- H. Glotin, F. Chamroukhi et S. Paris





ISBN 2-9524747-9-6
EAN 9782952474795

Soutenue par : LSIS, INRIA, ARIA, GDRI3, AFCP, CNRS, TPM et USTV

Photo des participants



(1) Joseph RAZIK, (2) Régis ABEILLE, (3) Sébastien PARIS, (4) Jean-Paul HATON, (5) Pierre-Henri MORIN, (6) Claude TOUZET, (7) Xanadu HALKIAS, (8) Stéphane MARCHAND-MAILLET, (9) Thibault MAILLOT, (10) Mohamed Amine SEHILI, (11) Joakim ANDEN, (12) Stéphane MALLAT, (13) Laurent PERRINET, (14) Arnaud BOUCHER, (15) Yoshua BENGIO, (16) Pierre MACHART, (17) Mathieu KOWALSKI, (18) Hervé GLOTIN, (19) Olivier ADAM, (20) Thomas PEEL, (21) Mathieu LAGRANGE, (22) Marie-Christine HATON, (23) Patrick MARMAROLI, (24) Mathias Rossignol, (25) Faicel CHAROUKHI, (26) Xabier Jaureguiberry.
(étaient hors photo Jean-Marc PREVOT et Rémy DELAYE).

Avant-propos

Pour sa 6e édition (2011), l'école ERMITES, constituée d'une trentaine de participants, met l'accent sur le couplage entre traitement du signal et classification, avec des perspectives cognitives. Le lien entre ces trois composantes fondamentales de la recherche d'information par le contenu est de plus en plus fort. C'est ce que démontrent les exposés de nos experts internationaux que nous remercions très chaleureusement d'avoir accepté notre invitation.

Nous aimerions exprimer toute notre gratitude à Y. Bengio qui traverse l'atlantique, à S. Mallat très pris par ailleurs, ainsi qu'à J.-P. Haton qui nous a toujours encouragé dans cette réflexion. Nous sommes aussi très heureux de pouvoir profiter de l'expertise de M. Kowalski. Nous pensons bien sûr aussi à X. Halkias, Phd de Columbia university, qui nous fait le plaisir d'ouvrir des perspectives bioacoustiques, qui seront suivies par un exposé de O. Adam, expert en la matière. Une synthèse parole – bioacoustique sera présentée par J. Razik, et d'autres challenges futurs seront discutés dans l'exposé de H. Glotin.

Enfin, tout le comité d'organisation vous souhaite un agréable séjour sur ce site paradisiaque de l'île de Porquerolles, dans le Parc National de Port-Cros. Si vous tendez l'oreille, peut-être entendrez-vous quelque sirène… Les cachalots sont très présents : nous avons encore enregistré trois titans de plus de quinze mètres au large de l'île il y a une semaine… Sans jamais les voir… Si J. Perrin – Prix Nobel de Physique – poussait à « Substituer du visible compliqué par de l'invisible simple », nous poussons dans cette édition à « Substituer de l'audible complexe par du visible simple »…

Nous remercions l'INRIA, l'ARIA, le GDRI3 et le CNRS, l'UMR LSIS, Toulon Provence Méditerranée (TPM), et l'USTV sans qui cette école n'existerait pas.

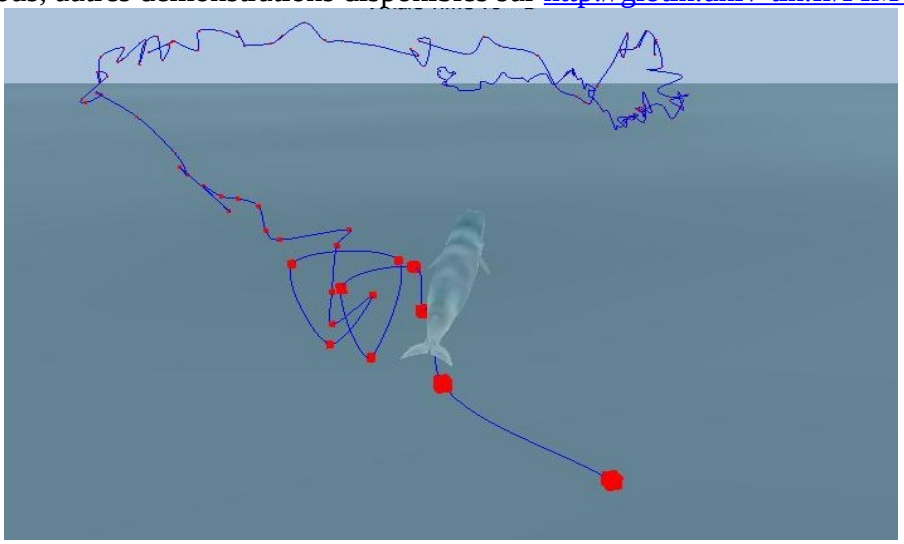
Comité de Programme :	Comité d'Organisation :
H. Glotin (prés.), Pr USTV & IUF, LSIS S. Paris, MC univ. Aix-Marseille, LSIS J. Razik, MC univ. Sud-Toulon Var, LSIS Y. Bengio, Pr univ. Montréal, LISA S. Mallat, Pr Ecole Polytechnique O. Adam, Pr univ. Paris VI, LAM J.-P. Haton, Pr univ. Nancy & IUF, LORIA M. Kowalski, MC univ Paris XI, L2S	S. Paris (prés.) J. Razik H. Glotin F. Chamroukhi, MC USTV, LSIS DYNI R. Abeille, doctorant USTV, LSIS DYNI M. Chouchane, doctorant, LSIS DYNI R. Delaye, ingénieur du son P. Machart, doctorant, LSIS DYNI & LIF F. Bénard, doctorant, LSIS DYNI J.-M. Prévot, IE USTV

A La Gardo, le 23 septembre 2011

Hervé Glotin et Sébastien Paris,
pour le CO et le CP.

INRIA : http://www.inria.fr/, ARIA : http://www.asso-aria.org/ , GDRI3 : http://www.irit.fr/GDR-I3/ , CNRS : http://www.cnrs.fr/, UMR LSIS : http://www.lsis.org/, Toulon Provence Méditerranéenne (TPM) : http://www.tpm-agglo.fr/jahia/jsp/index.jsp ,USTV : http://www.univ-tln.fr/

NB : Le thème de cette année s'inscrit dans le projet Dynamiques de l'Information du LSIS (http://lsis.org.dyni). Notamment dans son sous-projet PIMC (Plateforme d'Intégration Multimodale pour la Cétologie), dont un résultat majeur est la reconstruction de scène 3D de plongée du grand cachalot par acoustique passive (travaux réalisés dans le cadre de la thèse de F. Bénard et breveté PCT illustrés ci-dessous, autres démonstrations disponibles sur http://glotin.univ-tln.fr/PIMC/DEMO) :



Un autre projet DYNI d'analyse de scène manuscrite est connexe à cette édition d'ERMITES : l'ANR CONTINT COGNILEGO « Des pixels au sémantique » avec ses partenaires LNIA et A2IA. Nous yproposons de développer des modèles cognitifs d'intégration de l'information visuelle pour la transcription robuste de documents. Notre approche repose sur des traitements pyramidaux autoorganisés, à différentes échelles sémantiques comme illustré ci-dessous. Des approches DBN ou Scattering sont également envisagées. Le site de ce projet contenant références et objectifs est : http://cognilego.univ-tln.fr

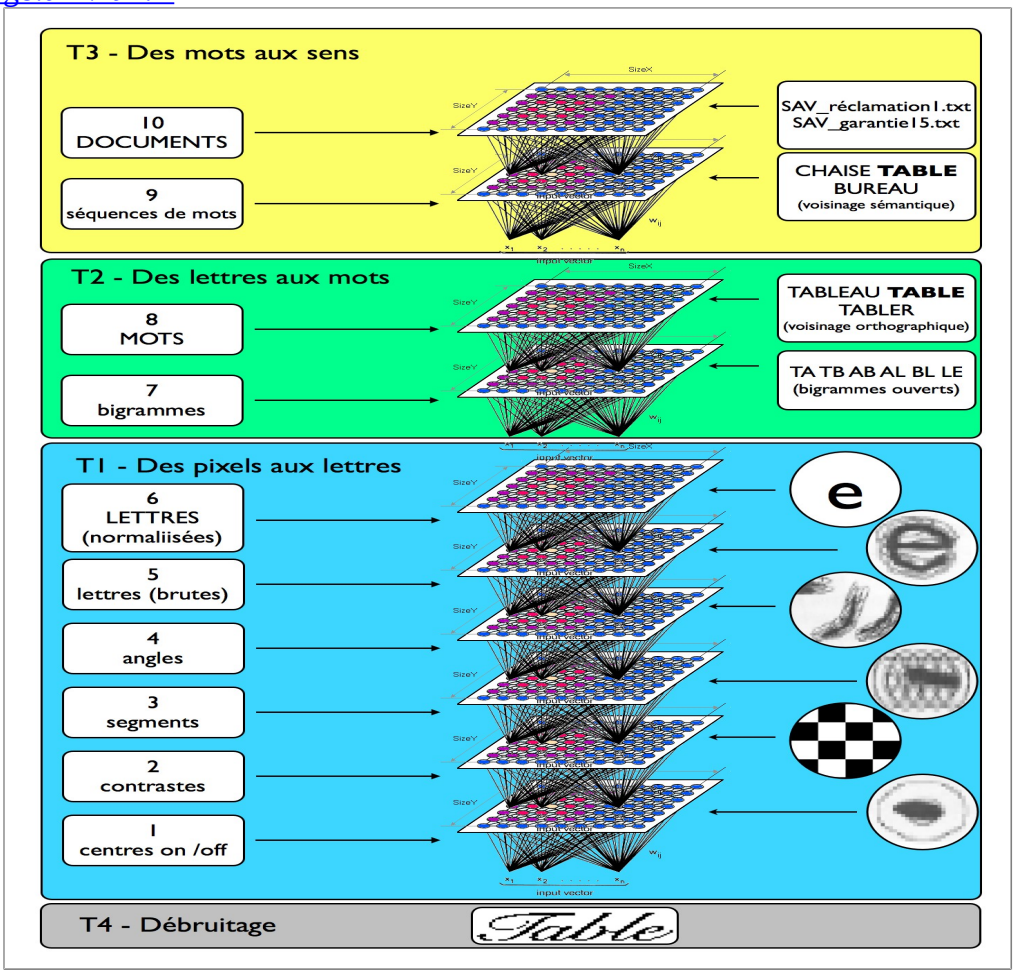


Illustration du schéma de l'intégration de l'informatio manuscrite proposée dans le projet ANR COGNILEGO.

Programme ERMITES 2011

DECOMPOSITION PARCIMONIEUSE, ABSTRACTION ET STRUCTURATION POUR L'ANALYSE DE SCENES COMPLEXES

Mardi 27 septembre

- **11h** *Café / réception*
- **11h30** H. Glotin - présentations des journées
- **12h** *Déjeuner*
- **13h15** J-P. Haton
- **15h15** *Pause café*
- **16h** X. Halkias
- **17h** M. Kowalski
- **19h** *Pause*
- **20h** *Dîner*

Mercredi 28 septembre

- **8h45** J. Razik
- **9h45** Y. Bengio (1/3)
- **10h30** P*ause café*
- **11h** Y. Bengio (2/3)
- **12h30** *Déjeuner*
- **14h** Y. Bengio (3/3)
- **15h15** *Pause café*
- **15h45** S. Mallat (1/2)
- **16h45** *Pause*
- **18h30** O. Adam
- **19h30** *Pause*
- **20h** *Dîner*

Jeudi 29 septembre

- **8h45** Ouverture avec présentations d'inscrits (Perrinet...)
- **9h30** S. Mallat (2/2)
- **10h30** *Pause café*
- **11H15** Discussion Scattering & DBN, Spiking & Event programming ...
- **12h30** *Déjeuner*
- **13h45** H. Glotin
- **15h** Discussion générale, projet édition...
- **16h** *Clôture, Pause café et Plage...*

Table des matières

J-P. Haton.....1

« Analyse de Scène et Reconnaissance Stochastique de la Parole »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Haton.mp4

X. Halkias.....79

« Detection and Tracking of Dolphin Vocalizations »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Halkias.mp4

M. Kowalski.....97

« Sparsity and structure for audio signal: a *-lasso therapy »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Kowalski.mp4

J. Razik.....145

« Sparse coding : from speech to whales »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Razik.mp4

Y. Bengio.....173

«Apprentissage Non-Supervisé de Représentations Profondes »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Bengio.mp4

O. Adam.....259

« Estimation de Densité de Population de Baleines par Analyse de leurs Chants »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Adam.mp4

S. Mallat.....301

« Scattering & Matching Pursuit for Acoustic Sources Separation »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Mallat.mp4

H. Glotin.....345

« Suivi & reconstruction du comportement de cétacés par acoustique passive »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Glotin.mp4

Exposés de participants :

L. Perrinet.....401

« Edge statistics in natural versus laboratory images »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Perrinet.mp4

M. Lagrange.....411

« From ASA to CASA, what does the C stands for anyway? »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_Lagrange.mp4

M-C. Haton.....417

« Parole audiovisuelle (Visible speech) »
http://lsis.univ-tln.fr/~glotin/ERMITES_11_MCHaton.mp4

Patrick Marmaroli.....423

« Suivi de véhicule routier par filtrage particulaire »

Liste des Participants.....425

Glossaire.....427

Jean-Paul Haton

Université de Nancy - LORIA / IUF

http://www.loria.fr/~jph/

"Analyse de scène et reconnaissance stochastique de la parole"

Les difficultés spécifiques de la reconnaissance de la parole sont nombreuses : nombre variable de locuteurs connus ou inconnus, microphones distants, bruit ambiant, effets liés aux locaux (écho, réverbération...). Cet exposé rappellera les principes de la reconnaissance automatique de la parole et présentera les niveaux fonctionnels d'un système : prise de son, paramétrisation, prise de décision. Nous insisterons sur les modèles stochastiques actuellement les plus performants pour la reconnaissance de la parole et présenterons les derniers développements du domaine. L'exposé sera illustré d'exemples pratiques.

Reconnaissance automatique de la parole

Jean-Paul Haton
Institut Universitaire de France,
LORIA - Université Henri Poincaré, Nancy 1

ERMITES'11 27-29 septembre 2011

Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

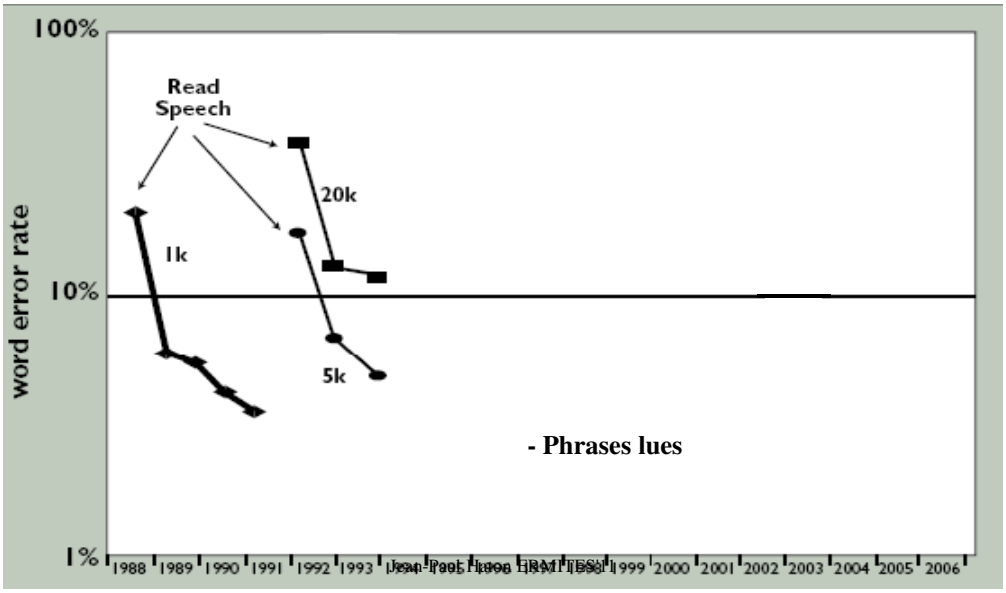
Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

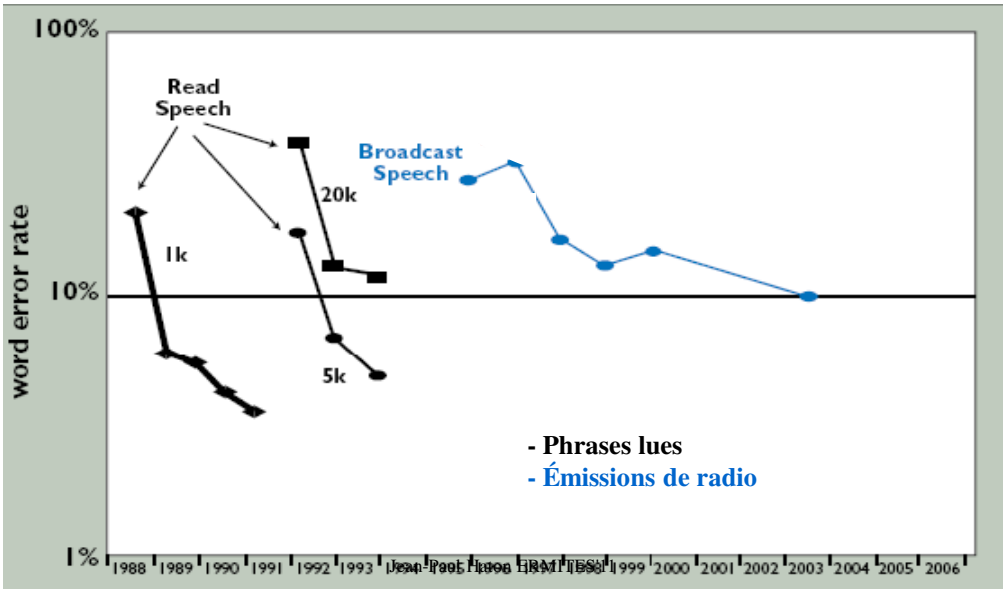
Le traitement automatique de la parole

- CODAGE ET TRANSMISSION
- SYNTHÈSE DE LA PAROLE
- RECONNAISSANCE DE LA PAROLE
- IDENTIFICATION DE LA LANGUE
- VÉRIFICATION DU LOCUTEUR

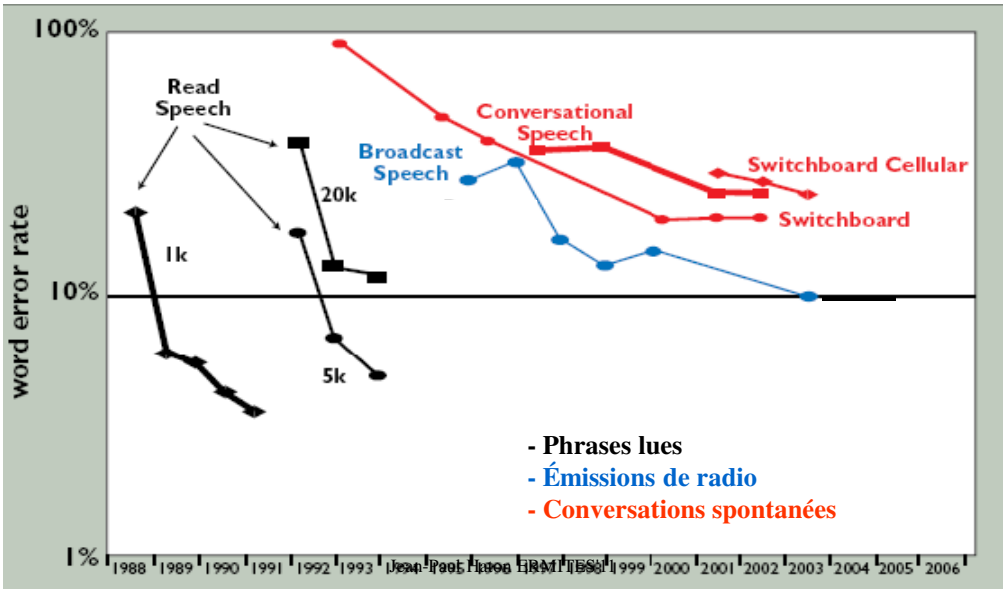
Résultats des évaluations de DARPA



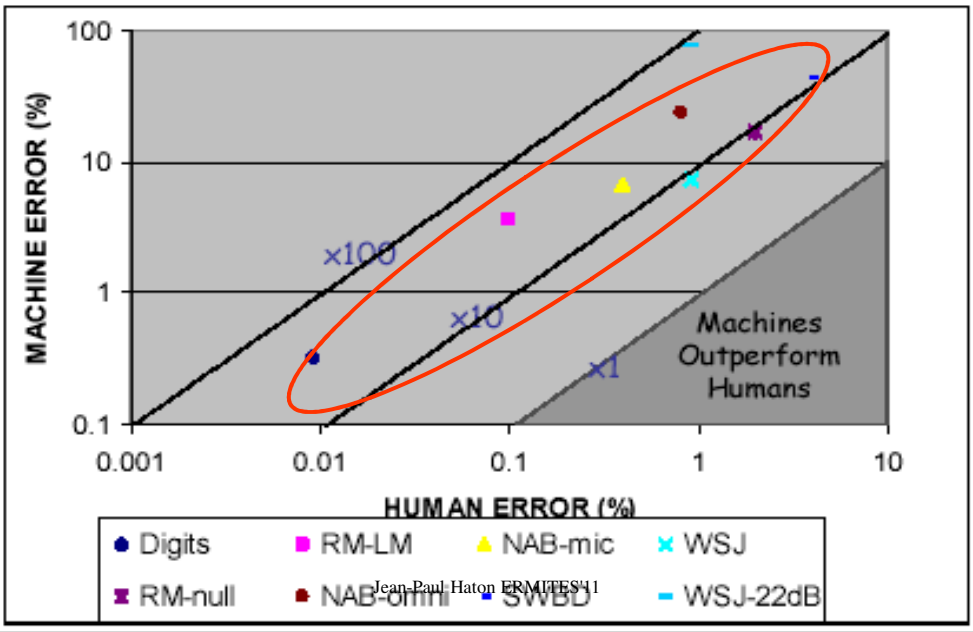
Résultats des évaluations de DARPA



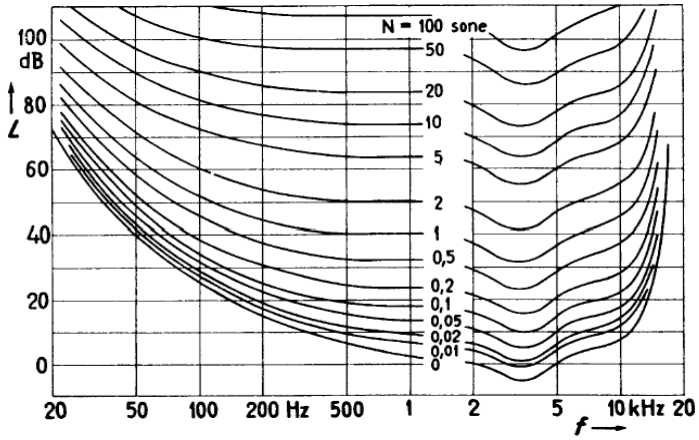
Résultats des évaluations de DARPA



La machine et l'être humain...



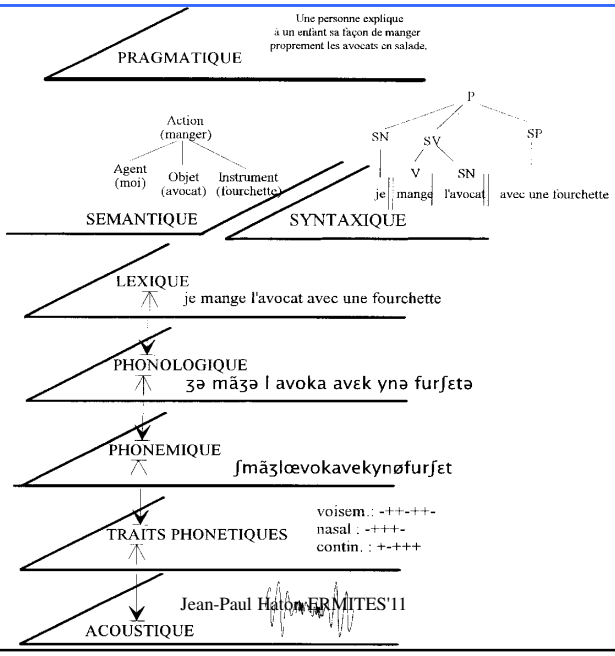
Perception auditive : sensibilité de l'oreille



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

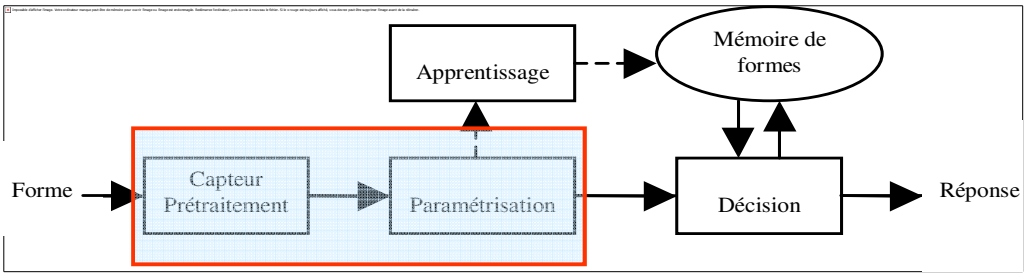
Niveaux de décodage de la parole



Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

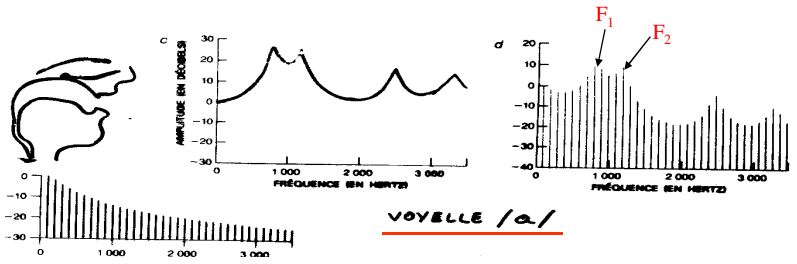
Principe de la reconnaissance des formes

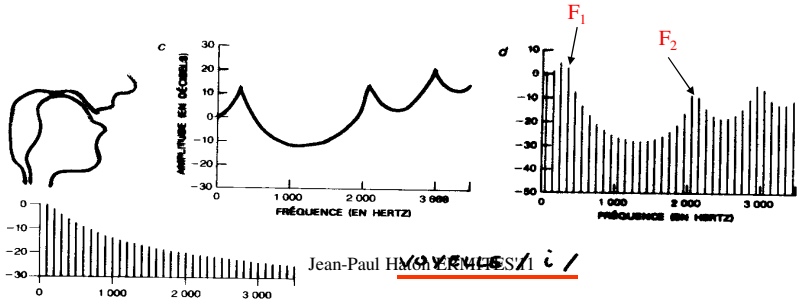


Paramétrisation du signal de parole

- Extraire du signal vocal des paramètres discriminants fondés sur certains critères, notamment perceptifs (aspects fréquentiels)
- Réduire le flux d'informations à traiter par le moteur de reconnaissance

Production des voyelles





Transformation de Fourier


Baron J. Fourier

Transformée de Fourier d'un signal x(t) :

$$X(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} x(t)\, e^{-j\omega t}\, dt$$

Projection de x(t) sur une base infinie de fonctions sinusoïdales

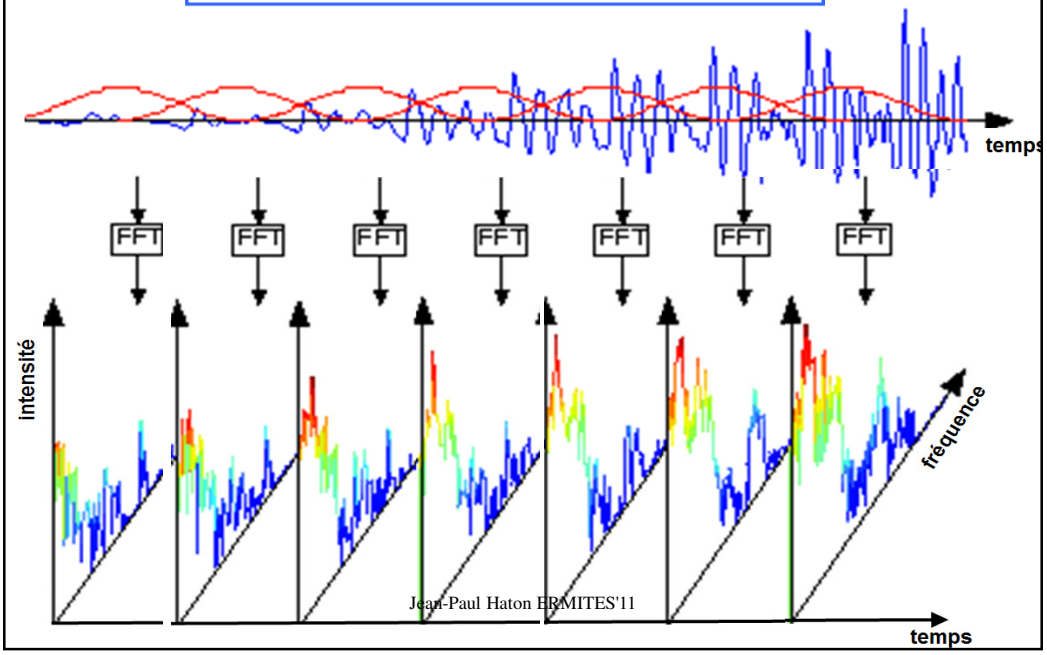
Transformée de Fourier discrète

Pour un signal numérisé discret : x[n] n = 0, N-1 :

$$X[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]\, e^{-j 2\pi k n / N}$$

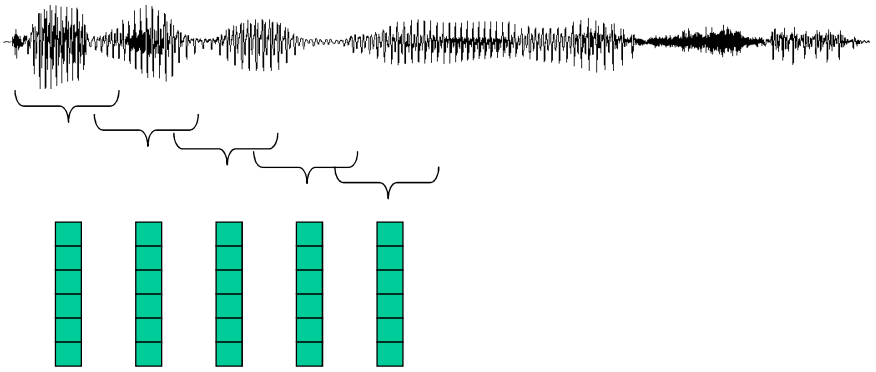
Aspects calculatoires : algorithmes FFT (*Fast Fourier Transform)*

Principe du spectrogramme



Paramétrisation et fenêtrage

- Fenêtrage... et recouvrement

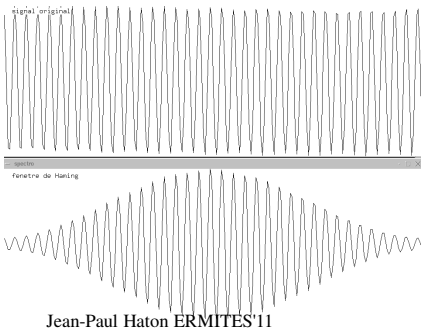


Paramétrisation et fenêtrage

- Fenêtre de Hamming :

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



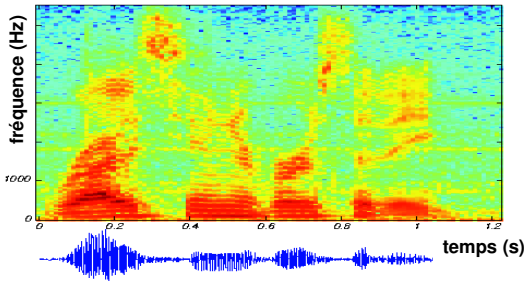


Le spectrogramme vocal ou « *sonagramme* »



W. Heisenberg

- *Visible Speech*



Bilinear Analyses

Limitation of spectrograms: lack of simultaneous resolution in time and frequency JPH1
(Heisenberg uncertainty principle: $\Delta f \, . \, \Delta t \geq 1/4\pi$ (equality for Gaussian window))

=>*Basis of time-frequency analysis revisited (Gabor, 1946)*

In particular, computation of spectrograms by bilinear distributions in which the analysis window is replaced by a "kernel" φ (« Cohen class »)

$$C_c(t,f,\varphi) = \iint_{\infty} \varphi(\xi;\tau,t,f).e^{j(\xi u - 2\pi f\tau - \xi t)} x(t-\frac{\tau}{2}) x^*(t+\frac{\tau}{2}) du d\xi d\tau$$

Diapositive 20

JPH1 Jean-Paul Haton; 04/01/1980

Examples of the Cohen class

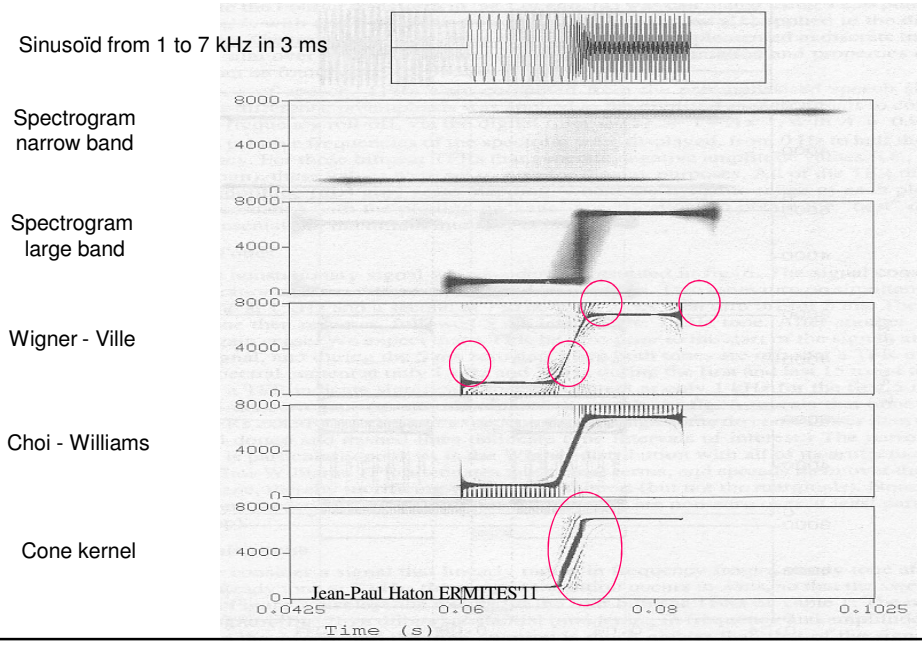
$$\varphi(.) = 1 \Rightarrow \text{Wigner Ville}$$

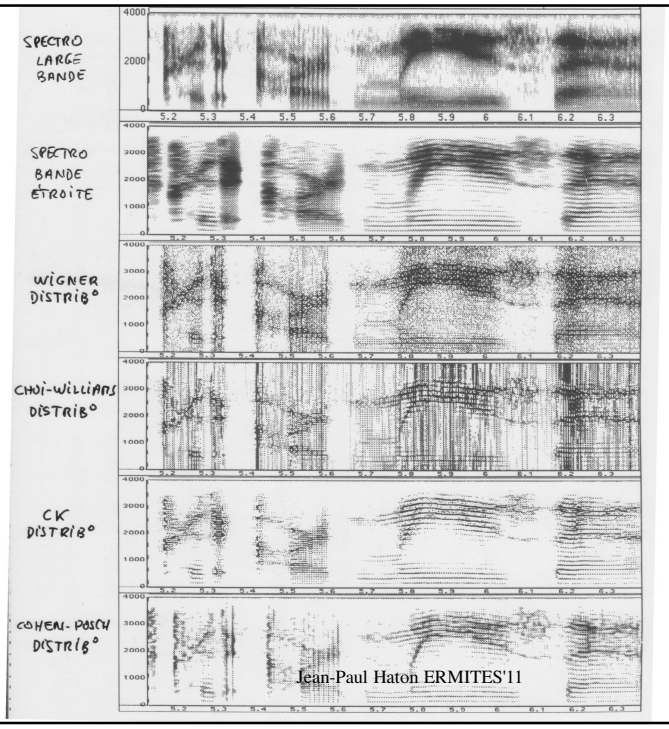
$$\varphi(.) = w(\frac{\tau}{2}).w(\frac{-\tau}{2}) \Rightarrow \text{Pseudo Wigner Ville}$$

$$\varphi(.) = \frac{1}{\sqrt{4\pi\tau^2/\sigma}}\exp(\frac{-(u-t)^2}{4\tau^2/\sigma}) \Rightarrow \text{Choï Williams}$$

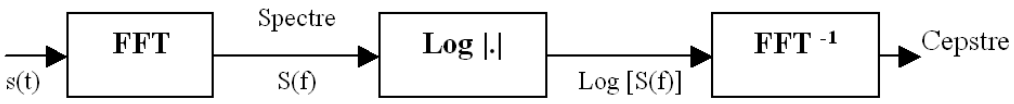
Better resolutions than conventional spectrograms,
but pertinence to human auditory perception questionable.

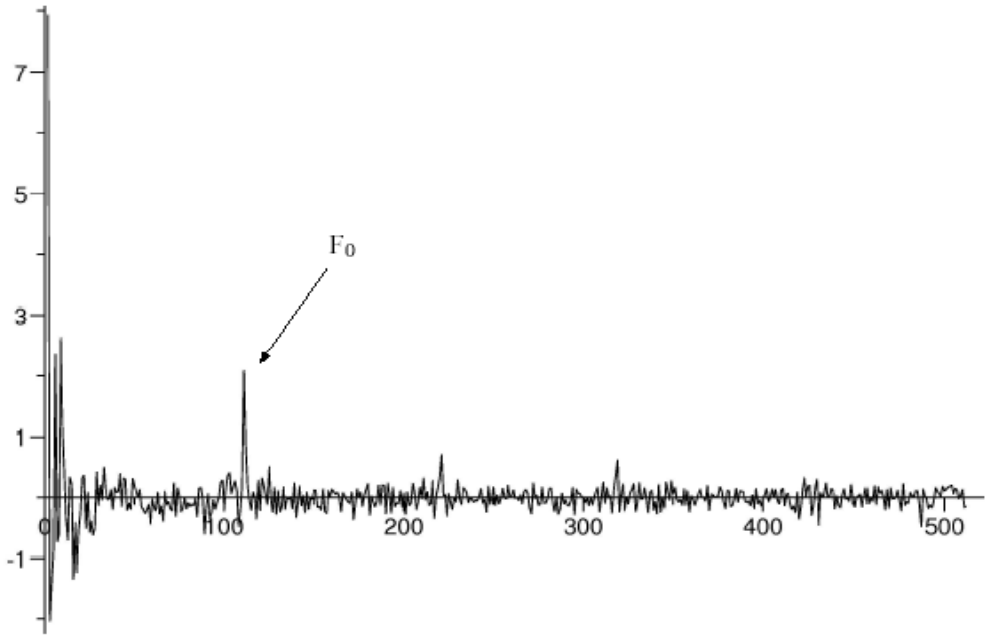
Artefacts...





Analyse homomorphique

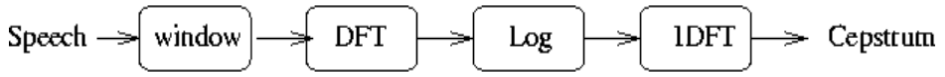




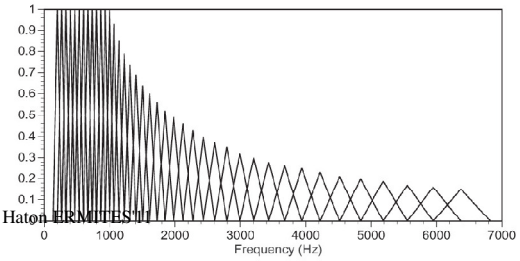
Analyse cepstrale

- Speech signal: convolution of an input signal and the impulse response of a production model.
- In order to caracterize the signal in terms of the parameters of the model, we must carry out a deconvolution.

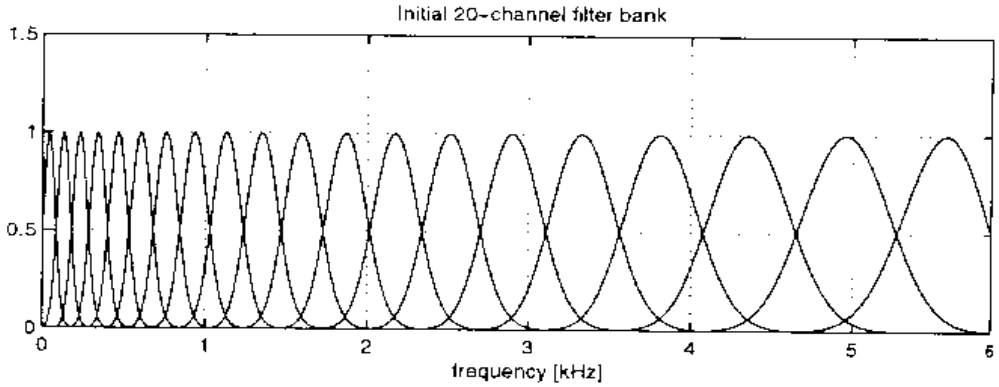
- Cepstral analysis is a common procedure used for de-convolution:



- Association with a Mel scale, MFCC (Mermelstein and Davis, 1980): most popular analysis method for ASR



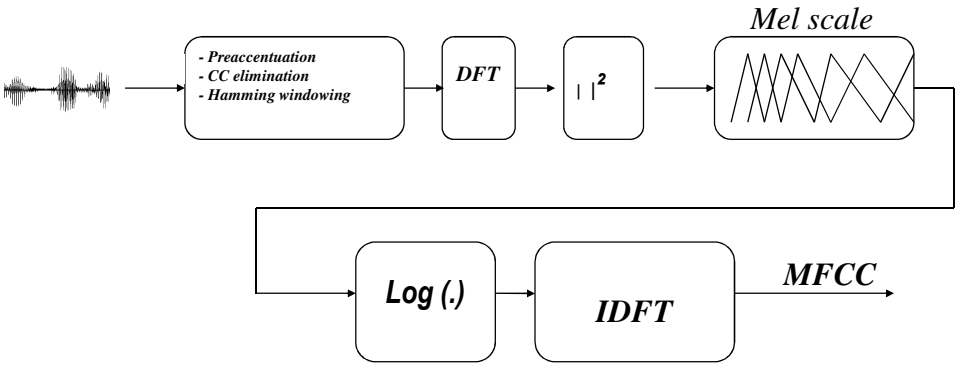
Banc de filtres (échelle Mel)



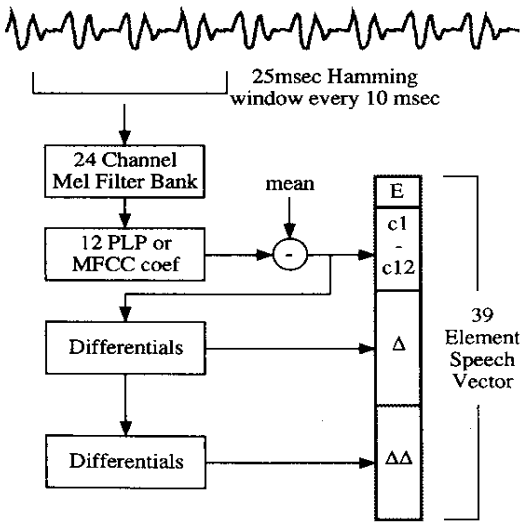
Paramétrisation MFCC

- Paramétrisation la plus utilisée : MFCC (*Mel Frequency Cepstral Coefficients*)
  - Transformation de Fourier pour décomposer le signal en ses fréquences constituantes
  - Analyse homomorphique pour séparer les contributions de la source et du conduit vocal
  - Filtres triangulaires inspirés du fonctionnement de l'oreille (échelle Mel)

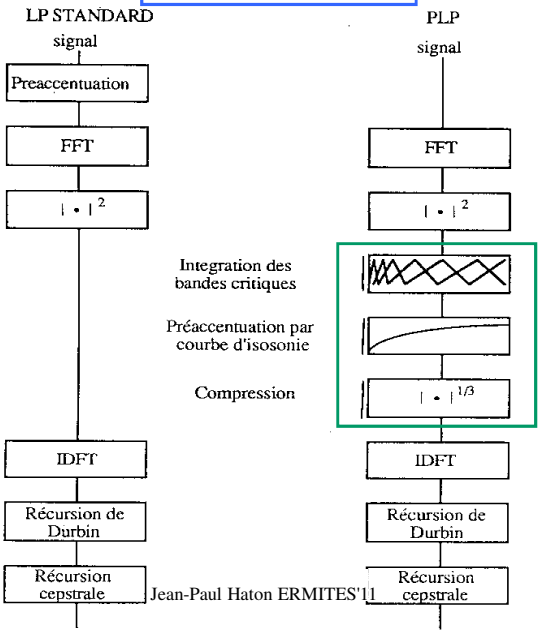
Mel Frequency Cepstrum Coefficients, MFCC



MFCC + dérivées



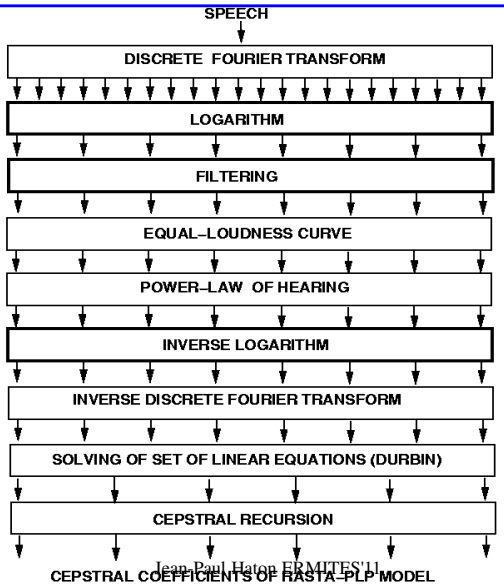
Analyse PLP



Analyse Rasta (RelAtive SpecTral Analysis)

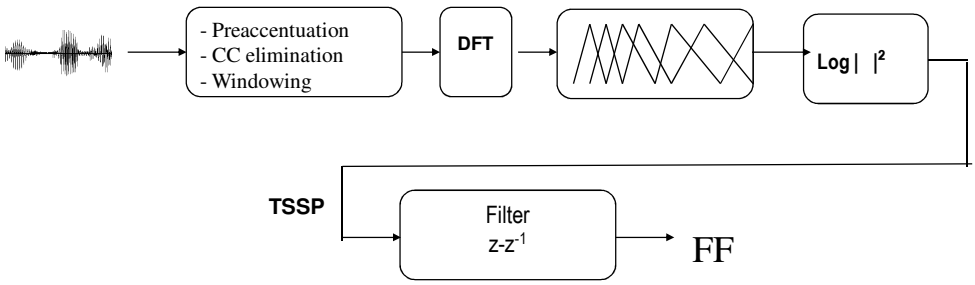
- Extension de l'analyse PLP, plus robuste au bruit
- Suppression par filtrage des variations temporelles trop lentes (zones stationnaires) ou trop rapides
- Se base sur le fait que l'oreille réagit aux valeurs relatives plus qu'aux valeurs absolues

Analyse RASTA-PLP



Frequency Filtered Parameters, FF
(Nadeu, 2001)

Idea: keep the advantages of
- spectral coefficients: interpretation (physical, physiological)
- cepstral coefficients: decorrelation



Temporel **S**equences of **S**pectral **P**arameters that model the time evolution of the spectral envelope

Autres méthodes

Modèles d'oreille

Paramètres fréquentiels filtrés

Ondelettes

Analyses à pas variable

etc.!

Modèles d'oreille

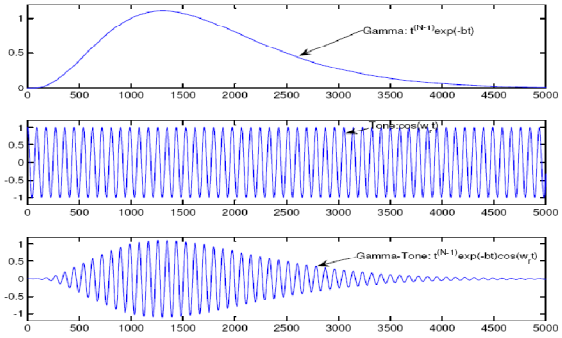
Idea: models inspired by the human auditory system

Example: Gammatone filter to mimick the cochlea

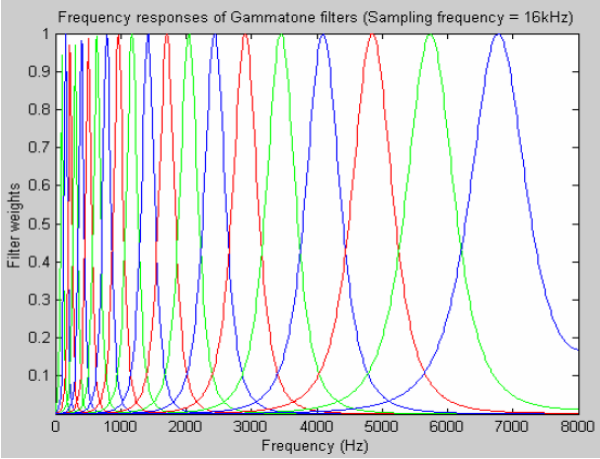
Gamma envelope

Cosine wave

Gammatone filter

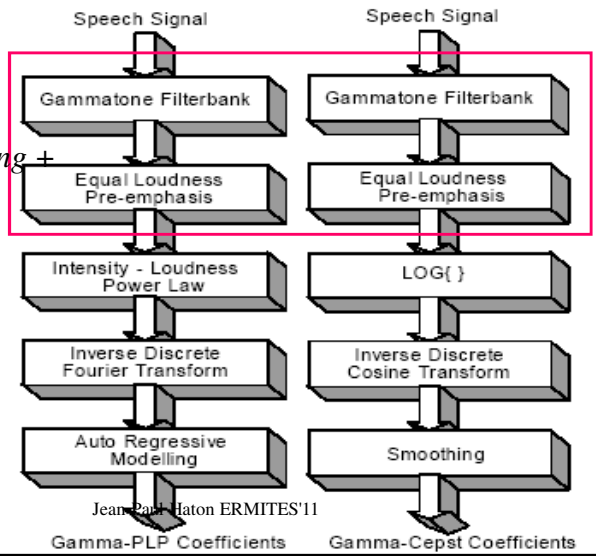


Example of Gammatone filterbank
(after Cheng et al., 2005)



Use of Gammatone filterbank
(after Abdulla, 2003)

FFT +
Gammatone filtering +
Preemphasis



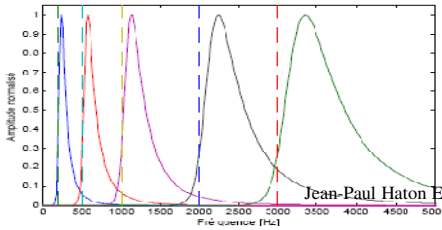
Gammachirp filters
(Irino and Patterson, 1997)

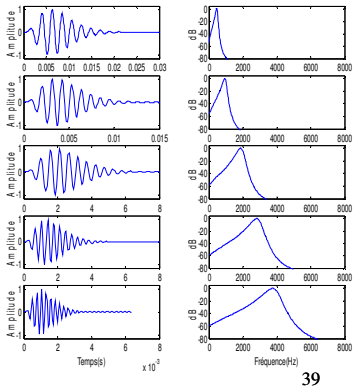
- The Gammachirp filter is constructed by adding a frequency modulation term to the Gammatone function.
This filter a good model of the asymmetric, level-dependent auditory filter.

$$g(t) = a_n t^{n-1} e^{-2\pi\beta t} e^{j(2\pi f_0 t + c\ln(t) + \varphi)}$$

Gamma Envelope *Tone carrier* *Chirp term*

with $C = 3.38 + 0.107\,P_s$





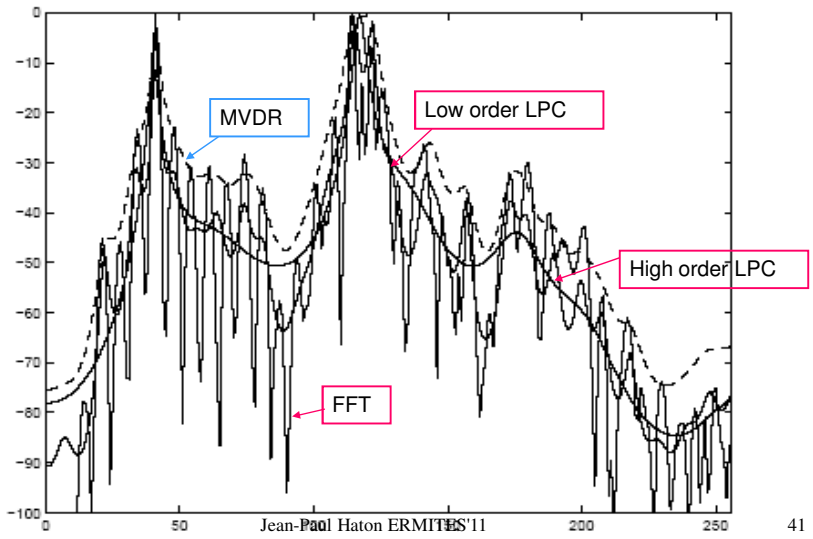
Spectral estimation: MVDR
(Minimum Variance Distortionless Response)

Principle: the signal power at a given frequency ω_l is determined by an FIR filter designed with the constraint that its response at this frequency has unity gain:

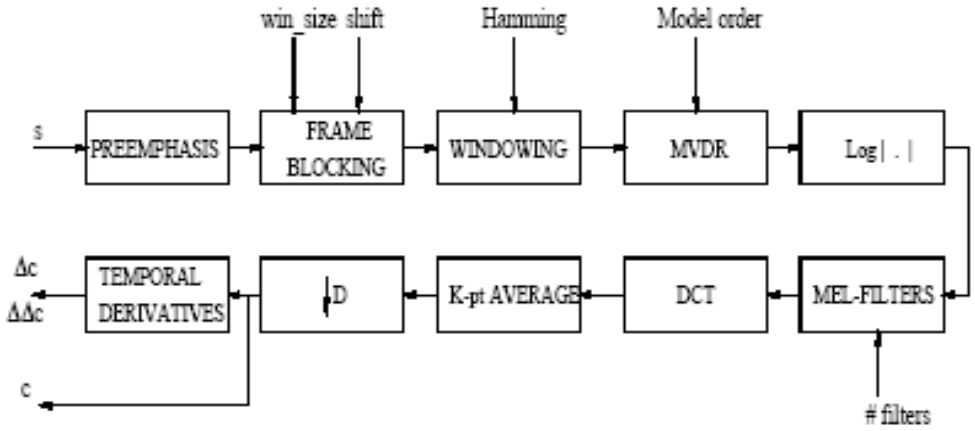
$$H(e^{j\omega_l}) = \sum_{k=0}^{M} h(k) e^{-j\omega_l k} = 1$$

Interest: MVDR eliminates the overemphasis of harmonic peaks typically seen in medium and high pitched voiced speech with classical spectral estimation.

Short-time spectrum and MVDR



Principle of MVDR/MFCC analysis for noisy ASR
(after Dharanipragada)



Non-linear analyses

- Deficiencies of the linear model of speech production.
- Speech analysis based on non-linear models
in order to (Kubin, 1995):
  - take into account phase information
  - cater non-Gaussian signals
  - deal with rapidly changing speech dynamics
  - reveal information about non linear speech production mechanisms

=>Interest for high order statistics: bispectrum, etc.

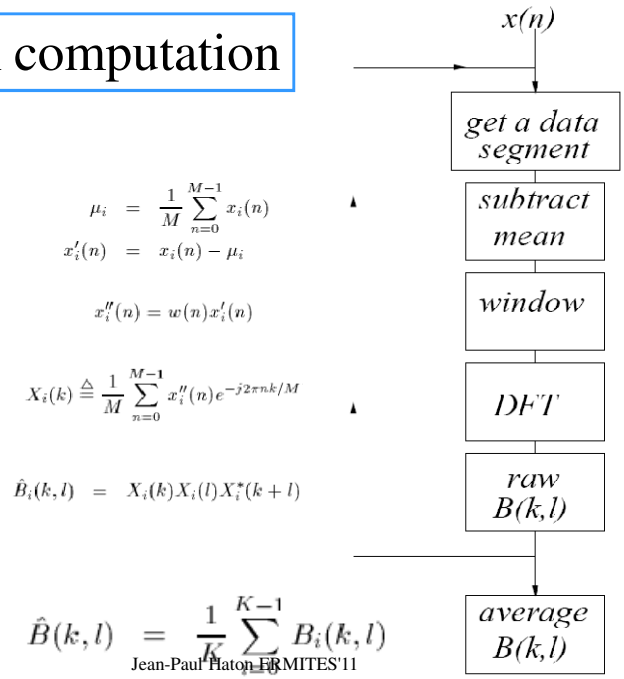
Bispectrum

Definition:

$$B(f_1,f_2)=\frac{1}{L}\sum_{j=1}^{L} X_j(f_1)X_j(f_2)X_j^*(f_1+f_2)$$

Usages:
- noise suppression
- voice activity detection
- noisy speech recognition (Onoue, 2008)

Bispectrum computation



$$\mu_i = \frac{1}{M}\sum_{n=0}^{M-1} x_i(n)$$

$$x_i'(n) = x_i(n) - \mu_i$$

$$x_i''(n) = w(n)x_i'(n)$$

$$X_i(k) \triangleq \frac{1}{M}\sum_{n=0}^{M-1} x_i''(n)e^{-j2\pi nk/M}$$

$$\hat{B}_i(k,l) = X_i(k)X_i(l)X_i^*(k+l)$$

$$\hat{B}(k,l) = \frac{1}{K}\sum_{i=0}^{K-1} B_i(k,l)$$

Wavelets

Time-Scale analysis vs Time-Frequency analysis

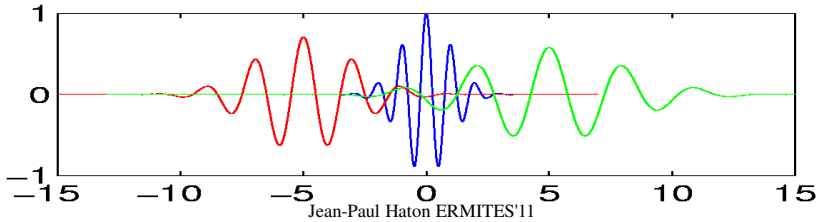
$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$$

Example: Real part of Morlet wavelet

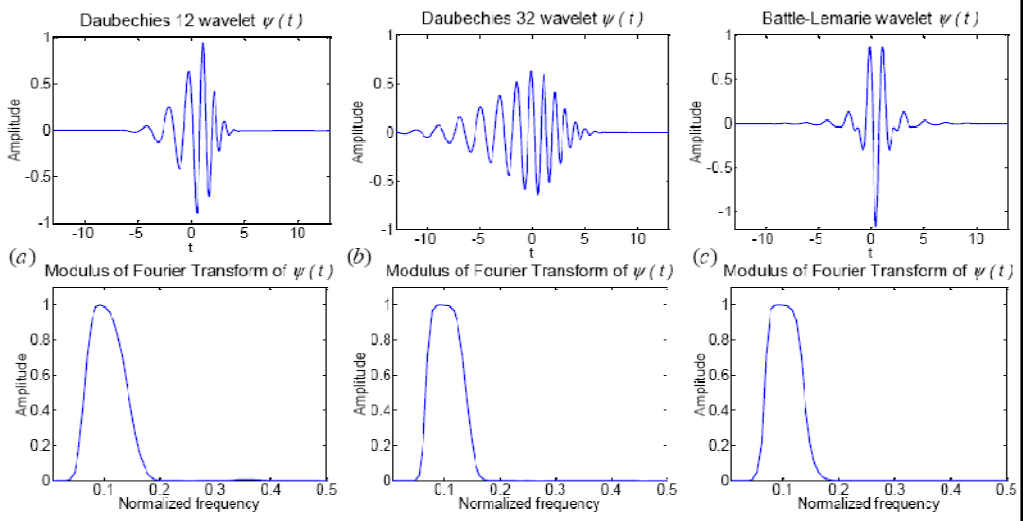
Blue : a = 1 and b = 0

Red : a = 2 and b = -5

Green : a = 3 and b = 5



Other wavelets



Applications of wavelets

Many uses in signal processing, particularly speech processing:

- coding
- denoising
- computing energy in sub-bands
- approximating the Mel scale
- replacing the cosine transform

But no decisive advantage for speech recognition...

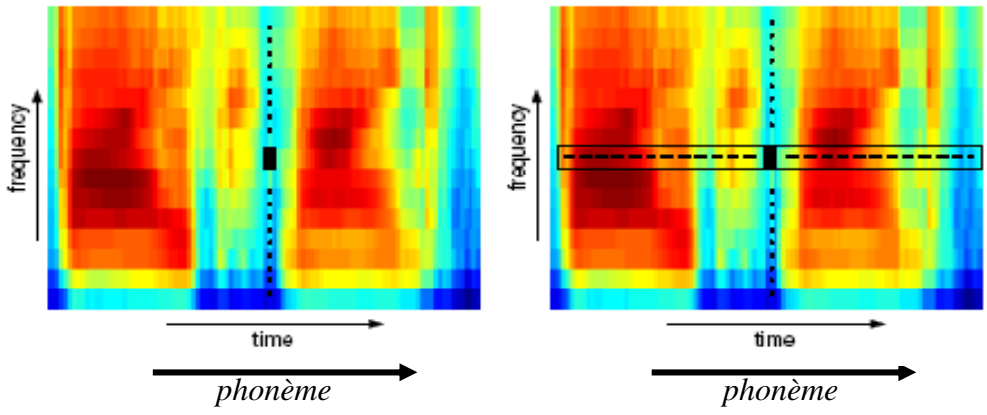
Rôle du temps

Constat : dégradation des performances humaines pour des segments courts (<100 ms)

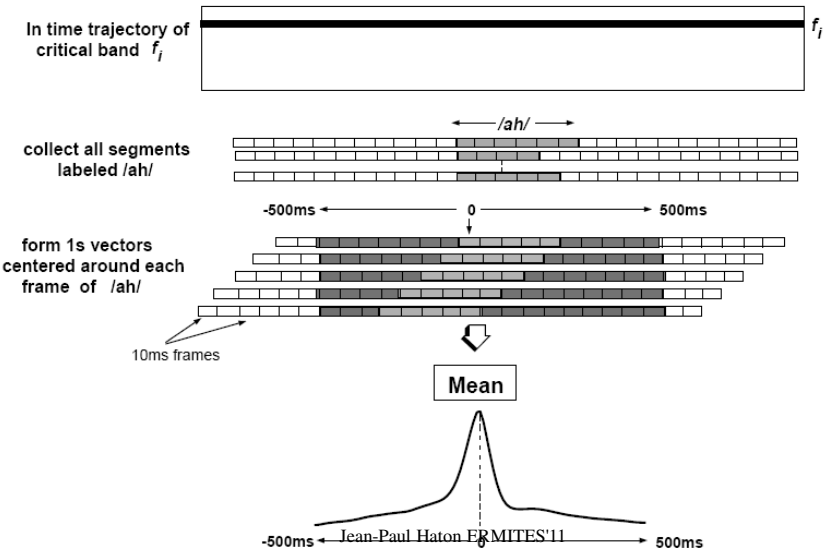
Solutions adoptées :

- incorporer des informations temporelles dans les vecteurs de paramètres à court-terme :
  - Coefficients "delta" (Furui, 1981) : ordre 1, 2, 3
  - *Cepstral mean normalisation*, CMN
  - RASTA
- *Temporal patterns*, TRAP (et variantes!)

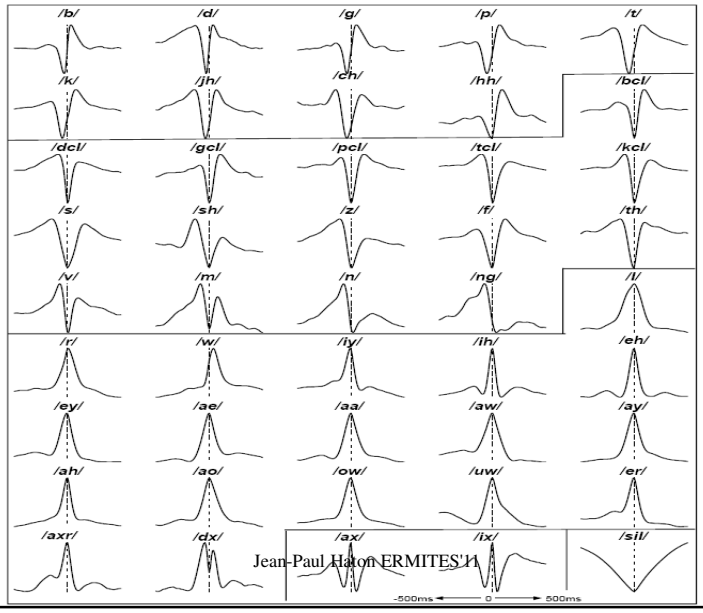
TempoRAl Pattern, TRAP
(Hermansky, Sharma 1998)

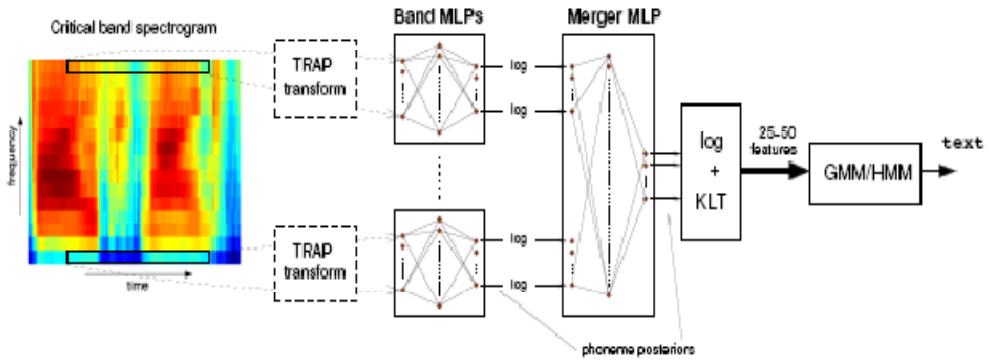


TRAP : évolution temporelle



Mean TRAP à 500Hz (5ème bande critique)

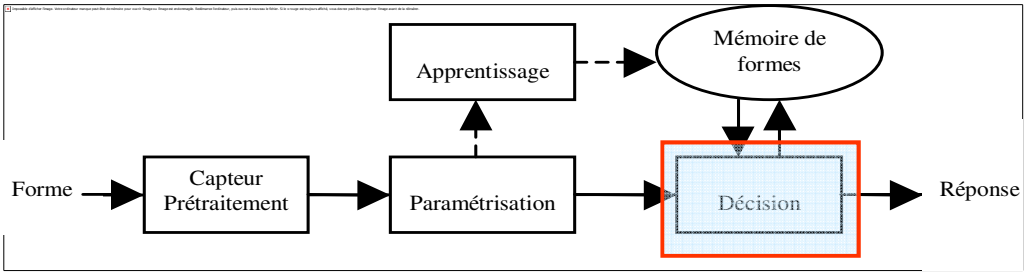




Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

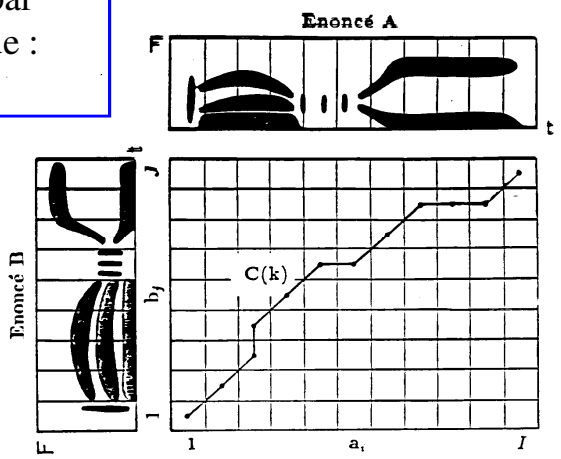
Principe de la reconnaissance des formes



Comparaison de formes par programmation dynamique : principe de la méthode

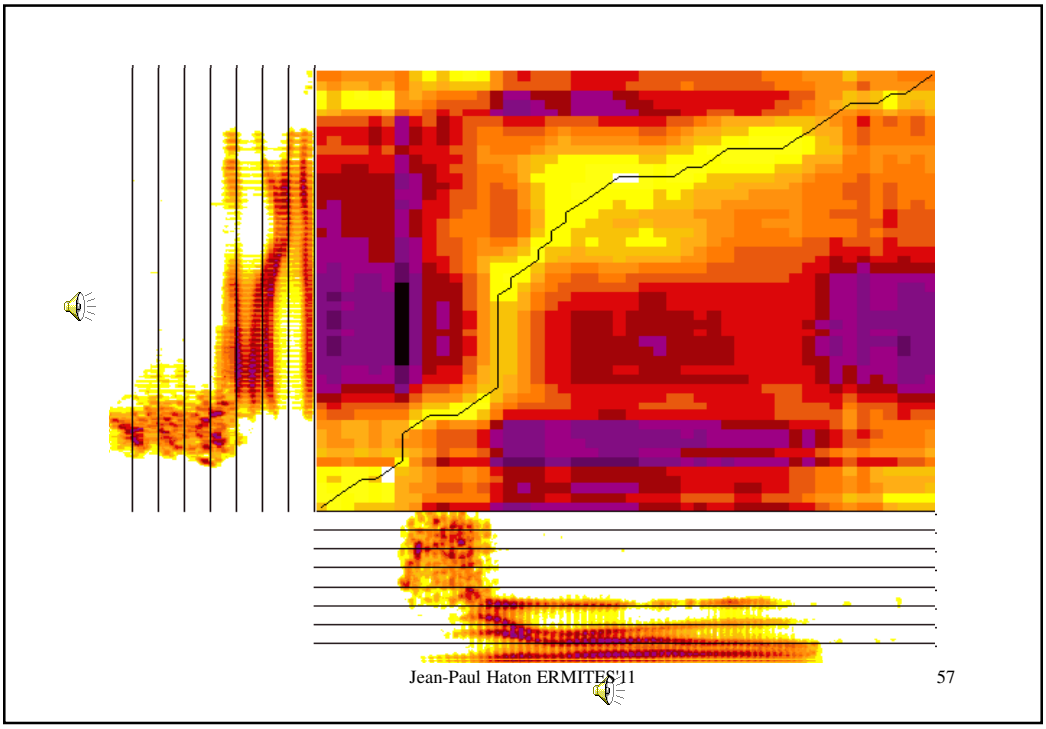


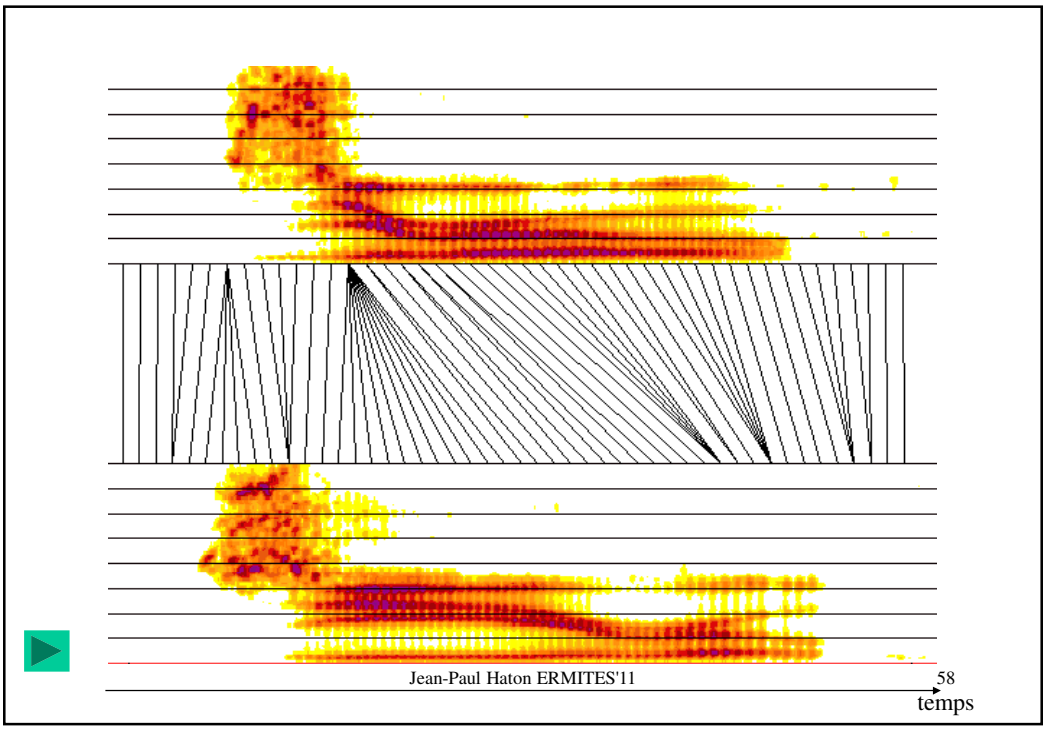
R. Bellman



Ressemblance entre les formes A et B :

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




temps

Décision probabiliste : règle de Bayes

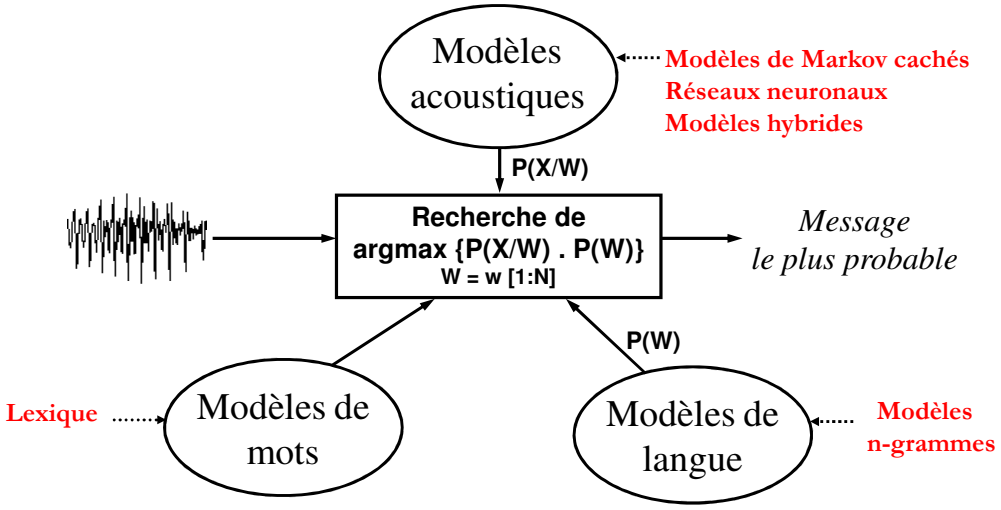
Soit un signal vocal paramétré X. On cherche la suite de mots prononcés la plus probable W connaissant X, soit *argmax { P(W/X) }*



T. Bayes

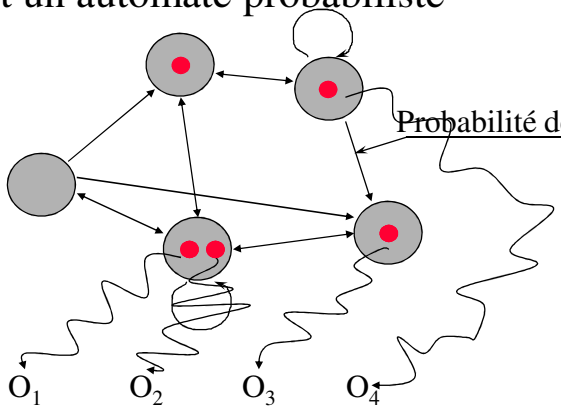
$$P(W / X) = P(X / W) \cdot P(W) / P(X)$$

Principe de la reconnaissance de la parole



Principe du modèle de Markov caché, HMM

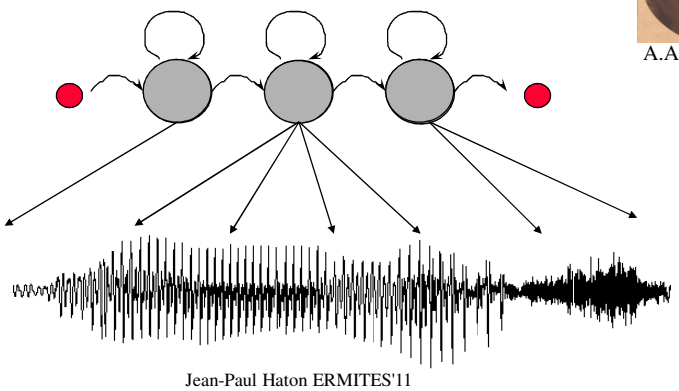
- c'est un automate probabiliste



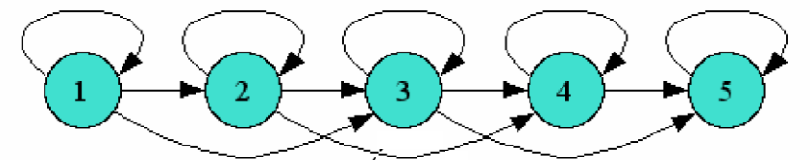
La modélisation stochastique de la parole

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

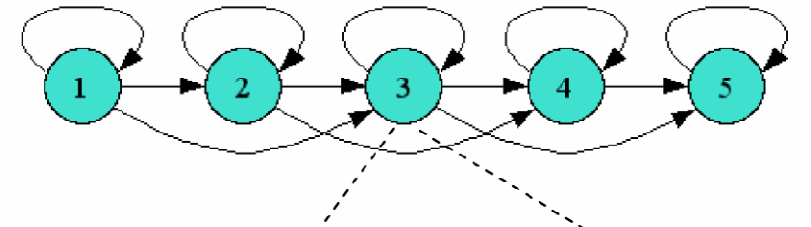

A.A. Markov



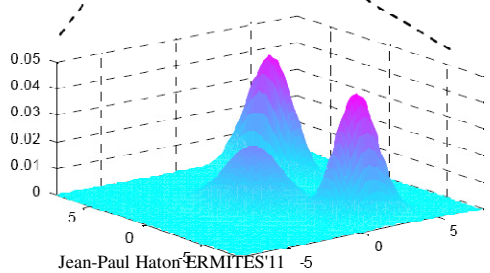
Un modèle de Markov caché (*HMM*)...



...et ses distributions de probabilités



$$b_j(\mathbf{x}) = \sum_{i=1}^{M} c_i \mathcal{N}(\mathbf{x}, \mathbf{m_{ji}}, \mathbf{C_{ji}})$$



Définition formelle

Pour définir un modèle de Markov il faut 3 distributions de probabilités :

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

Hypothèses simplificatrices

- P(M) peut être calculée séparément : modèle de langage calculé sur un corpus textuel
- P(O) est constant pour une observation donnée
- Modèle de Markov du premier ordre

$$P(q_t = i \mid q_{t-1}\, q_{t-2} \ldots q_1) \approx P(q_t = i \mid q_{t-1} = j)$$

- Indépendance des observations

$$P(o_t \mid q_1 .. q_t, o_1 o_2 \ldots o_{t-1}) = P(o_t \mid q_1 .. q_t)$$

- Probabilité d'émission ne dépendant que de l'état

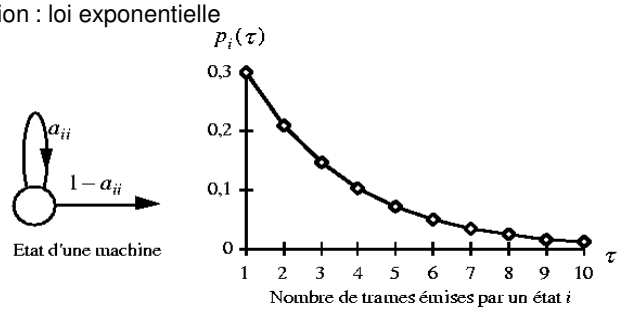
$$P(o_t \mid q_t q_{t-1} \ldots q_1) = P(o_t \mid q_t)$$

- Probabilité d'émission d'une observation modélisée par une somme de M fonctions gaussiennes

$$b_j(o) = \sum_{k=1}^{M} c_{jk} \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right)$$

Modélisation de la durée dans un HMM

Par construction : loi exponentielle

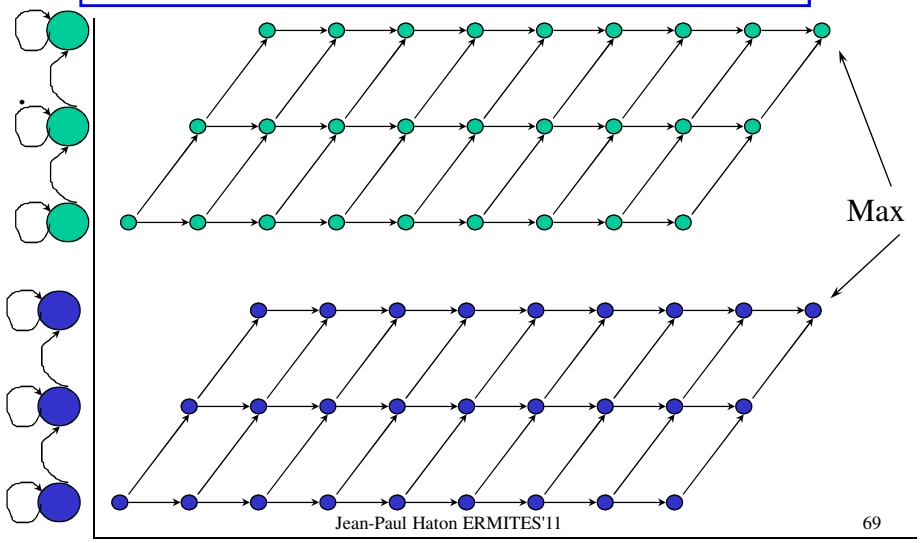


Diverses solutions possibles
-modèles à durée d'émission variable :
- modélisation explicite (Ferguson, 1980)
- HMM inhomogène (Ramesh, 1992)
- modèles probabiliste de la durée (Poisson, Gamma, etc.)
- modèles à durée d'émission bornée (Bush, 1987)
- autres solutions : analyse à débit variable, HMM maître-esclave.

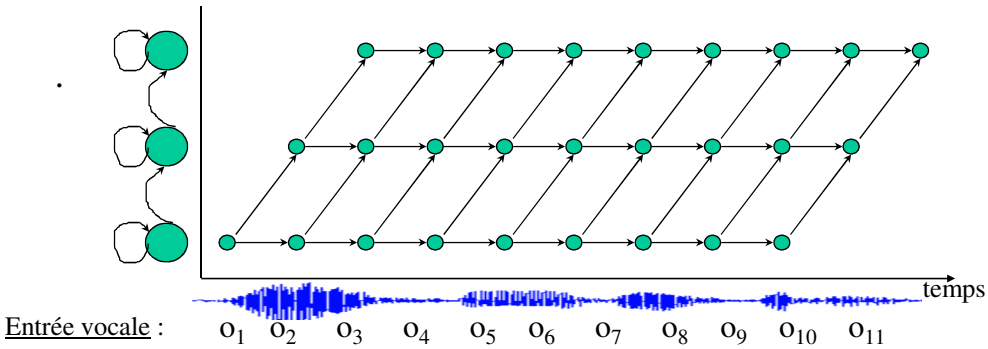
Deux problèmes à résoudre

- Reconnaissance : étant donné une observation $O=\{o_1..o_T\}$ et $\lambda=(\pi_i, a_{ij}, b_i)$, comment trouver le chemin optimal (la suite d'états q_i) selon un critère convenable.
- Apprentissage : étant donné des exemples d'apprentissages étiquetés, comment obtenir les paramètres optimaux $\lambda=(\pi_i, a_{ij}, b_i)$ des HMM.

Reconnaissance de mots isolés



Programmation dynamique et modèle de Markov



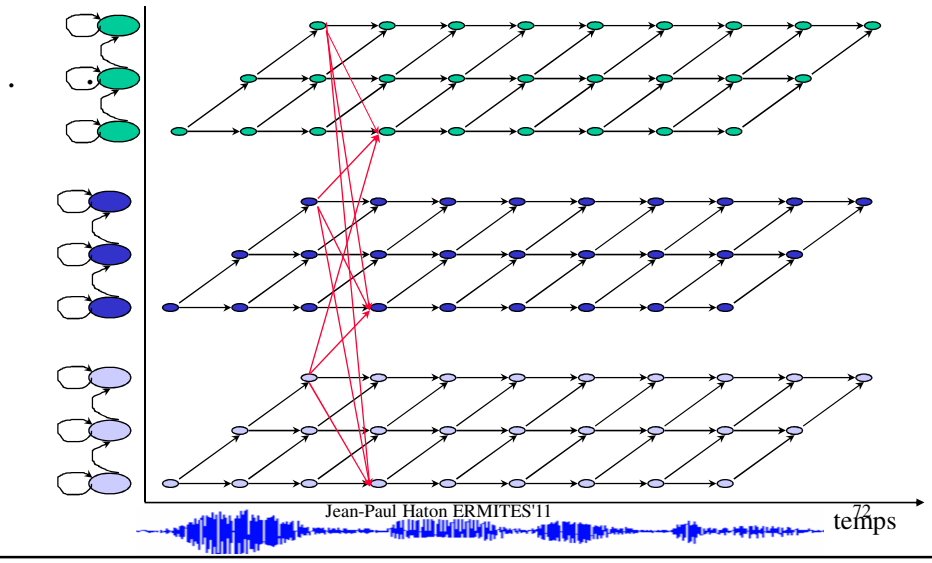
Entrée vocale : o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9 o_{10} o_{11}

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

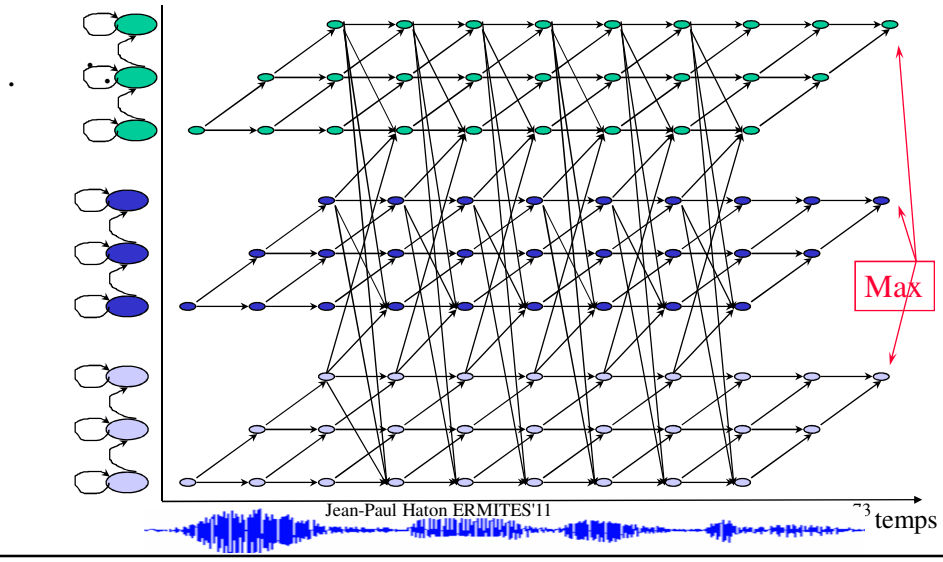
Algorithme de Viterbi en bref

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

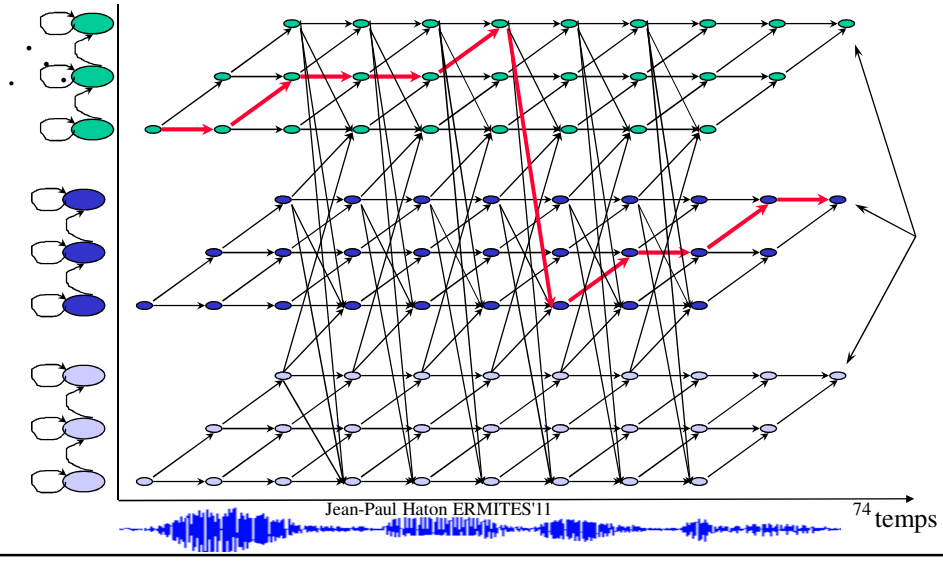
Algorithme pour une suite de mots enchaînés



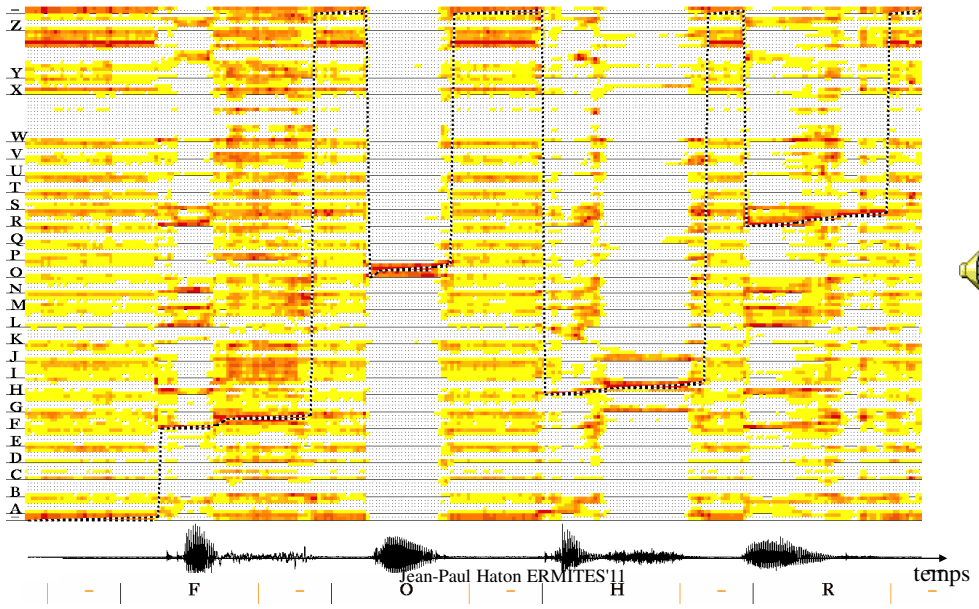
Graphe final et solution



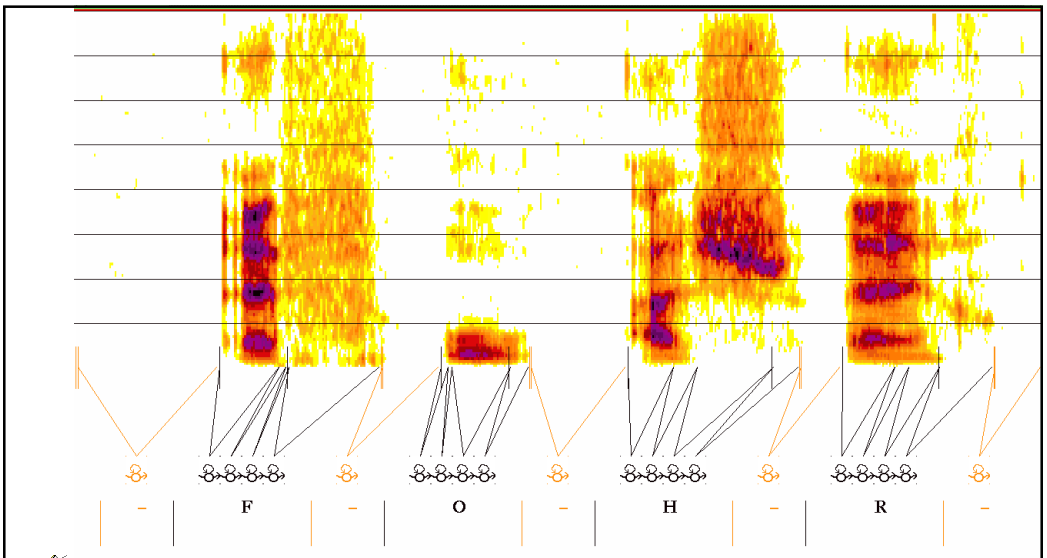
Solution obtenue par retour en arrière



Alignement temporel par *HMM* (D. Fohr)



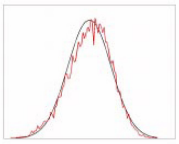
Jean-Paul Haton ERMITES'11

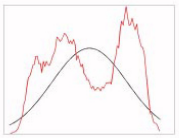


Alignement temporel par *HMM* (D. Fohr)

Jean-Paul Haton ERMITES'11

Représentation des distributions de probabilités

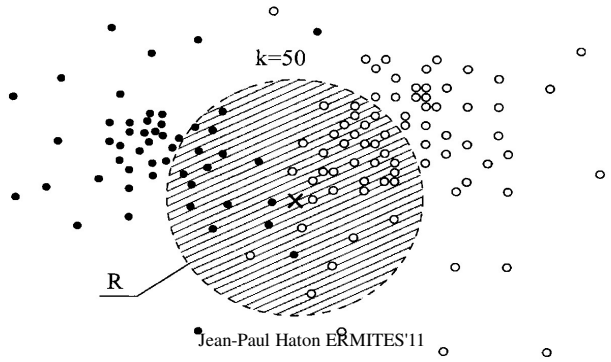




- Masses de probabilités discrètes
- Distributions continues :
  - Gaussiennes uniques
  - Mélanges de gaussiennes
- Réseaux neuromimétiques
- Modèles hybrides neuronaux-HMM
- Représentations non-paramétriques
- Représentation temps-fréquence : « HMM2 »

Représentation non paramétrique

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

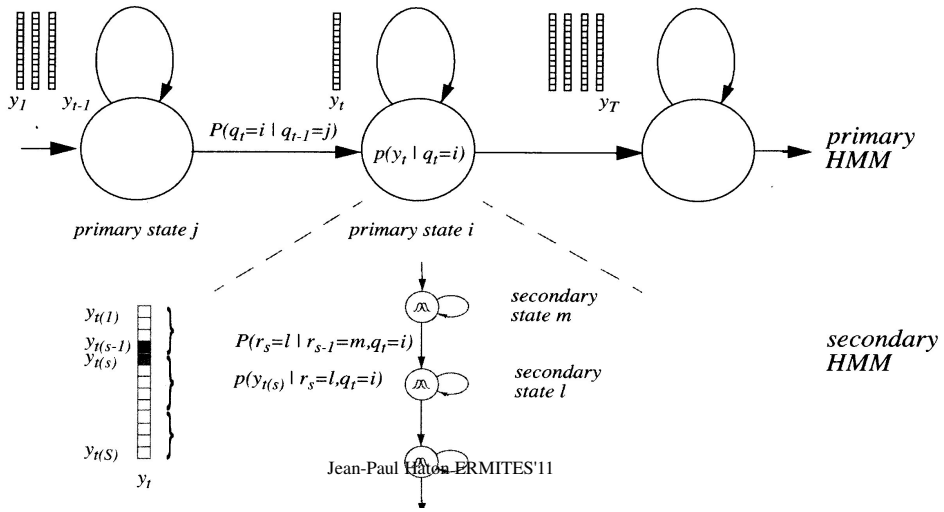


Représentation des distributions de probabilités

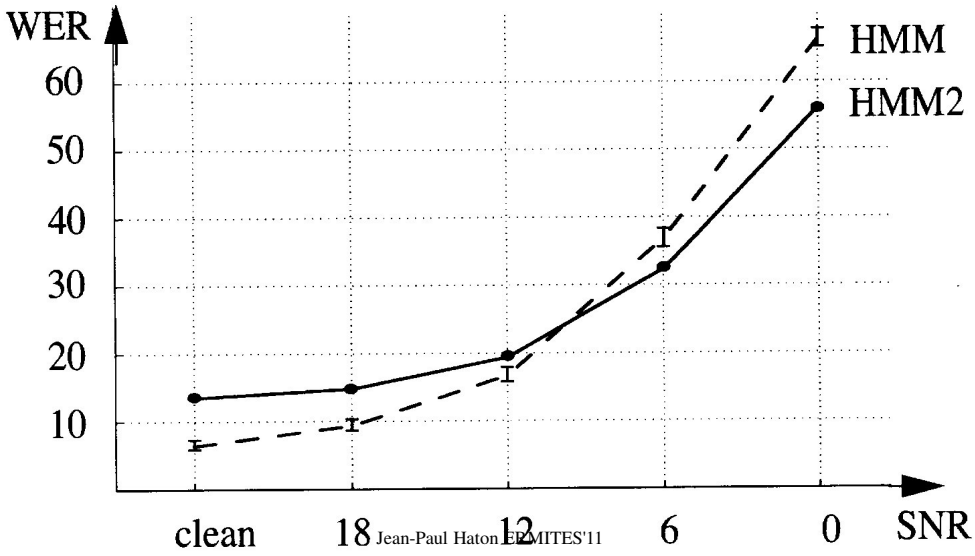
- Masses de probabilités discrètes
- Distributions continues :
  - Gaussiennes uniques
  - Mélanges de gaussiennes
- HMM du second ordre : Haton-Mari (1991)
- Réseaux neuromimétiques
- Modèles hybrides neuronaux-HMM
- Représentations non-paramétriques
- Représentation temps-fréquence : « HMM2 »

Principe des HMM2

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



HMM2 : performances



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

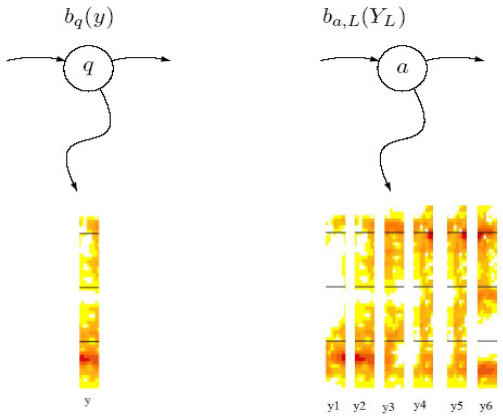
Principe

Exemples : Gong, Haton (1991), Ostendorf (1996), etc.

Résultats mitigés.

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

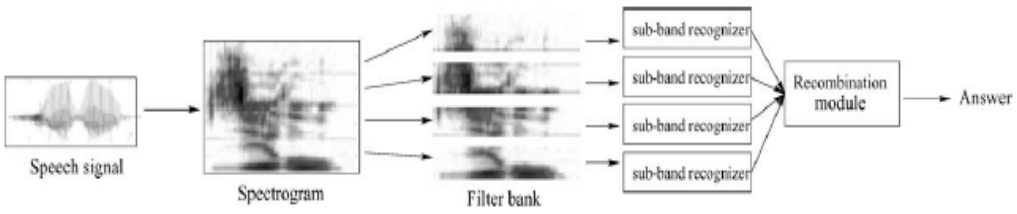
- Supprimer la notion d'échantillonnage fixe
- Un segment est un « atome » de parole
- Complexité importante mais bons résultats
- Application à d'autres champs de la RF



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 multi-bandes



Inspirés 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 (Glotin, 2001) (Zhu, 2003)
. HMM factoriels (Nock, 2003) (*cf.* Jordan (1997))
. Couplage d'états par réseaux bayésiens (Daoudi, 2003)

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

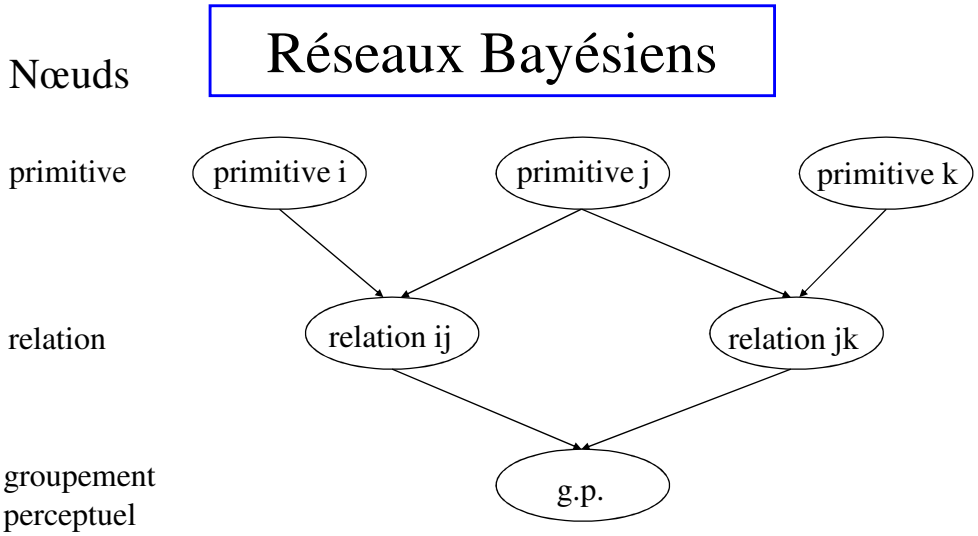
Modèles graphiques

- 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,…

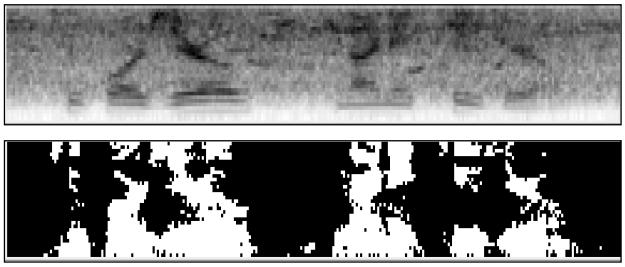
Réseaux Bayésiens



Solutions « nouvelles »

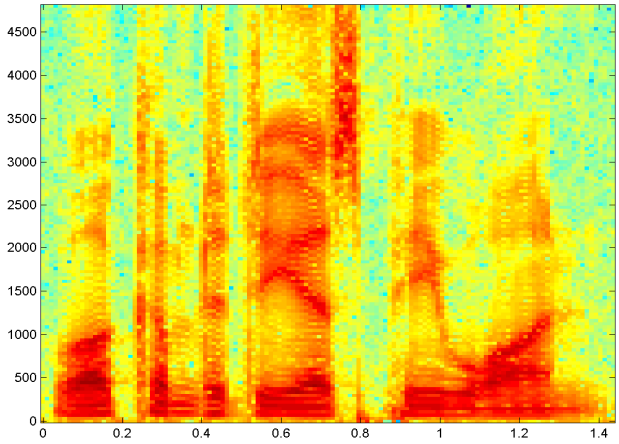
- Idée : rechercher de nouveaux formalismes :
  - des modèles mathématiques
  - des mécanismes d'apprentissage
  - la compréhension des mécanismes sous-jacents
- Tentatives « intéressantes » :
  - modèles segmentaux
  - modèles multibandes et multiflux
  - modèles graphiques
  - données manquantes

Données manquantes (Cooke et al., 2001)

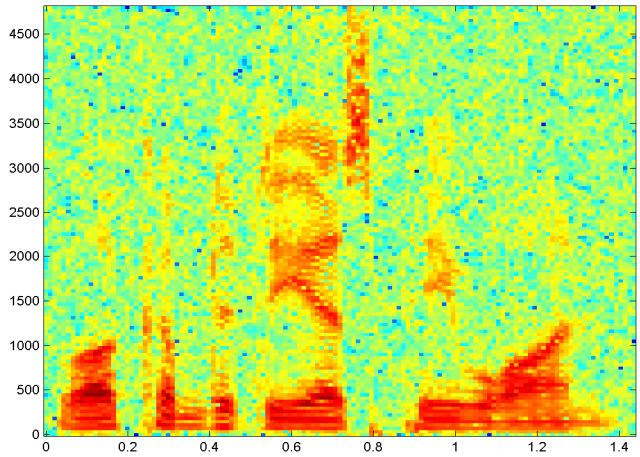


- On se place dans le domaine spectral
- Estimation et seuillage du RSB => masque
- « Marginalisation » des coefficients masqués lors de la reconnaissance

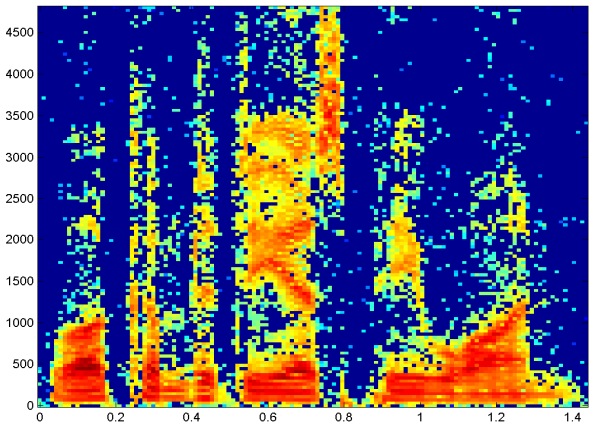
Spectrogramme initial



Superposition d'un bruit de 15 dB



Elimination des régions corrompues par le bruit



Detection des zones manquantes

(1) Critère d'énergie négative

$$| \hat{x}(w) |^2 = | x(w) + n(w) |^2 - | \hat{n}(w) |^2 < 0$$

(2) Critère de SNR

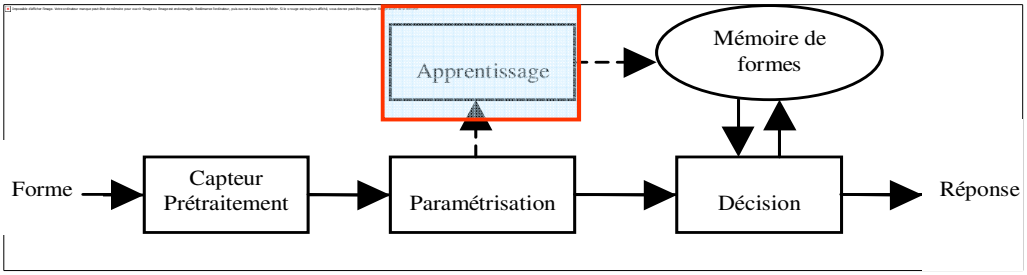
$$\log\left(\frac{| \hat{x}(w) |^2}{| \hat{n}(w) |^2} \right) < 0 \quad ou \quad | \hat{x}(w) |^2 < | \hat{n}(w) |^2$$

(3) Critère statistique (bruit supposé à distribution gaussienne)

$$P(SNR > 0) = \int_{-\infty}^{y(w)/2} \frac{1}{\sqrt{2\pi} \, | \hat{\sigma}_n(w) |} \exp\left[\frac{(x - \hat{u}_n(w))^2}{2\hat{\sigma}_n(w)^2} \right] dx$$

$$P(SNR > 0) < \theta$$

Principe de la reconnaissance des formes



Apprentissage

- Importance majeure en RAP statistique…
- Double aspect:
  - Apprentissage des modèles acoustiques
  - Apprentissage des modèles de langage
- Nécessité d'énormes quantités de données étiquetées

  => deux solutions possibles :
  - Apprentissage « légèrement » supervisé (si l'on dispose de données acoustiques avec transcription)
  - Apprentissage non supervisé

Influence du volume de données d'apprentissage

Training set	#States/Avg Components	eval03		dev04f	
		ML	MPE	ML	MPE
bntr-375h	7K/16	14.8	12.5	–	–
bntr-750h	7K/16	14.8	12.1	–	–
bntr-750h	7K/32	14.2	11.8	26.0	21.6
bntr-1050h	9K/32	13.8	11.4	25.0	20.3
bntr-1350h	9K/32	13.9	11.2	24.8	19.6
bntr-1790h	9K/32	13.7	11.0	24.4	19.3
bntr-2210h	9K/32	13.6	11.1	24.5	19.1

TABLE VI

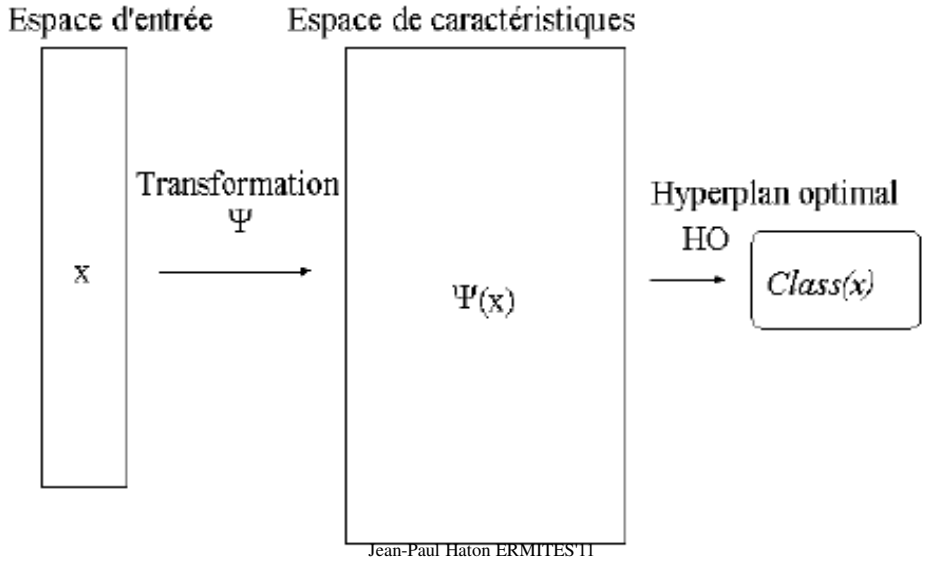
%WERs WITH THE GI ML/MPE MODELS WITH DIFFERENT TRAINING DATA SIZE. SINGLE PASS DECODING OF WB SEGMENTS WITH THE RT03 TRIGRAM LM. NB HYPOTHESIS USING THE RT03 NB MODELS.

(d'après Gales et al., Cambridge Univ.)

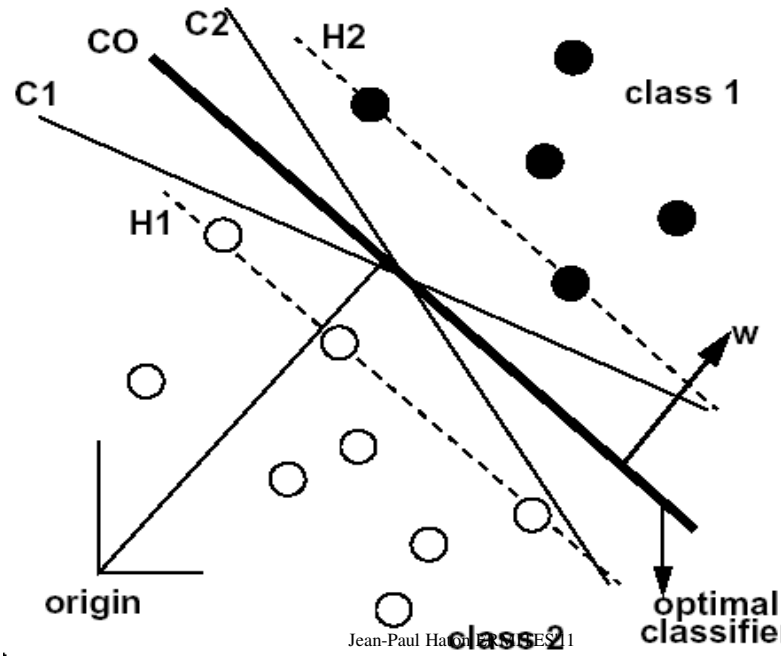
Apprentissage

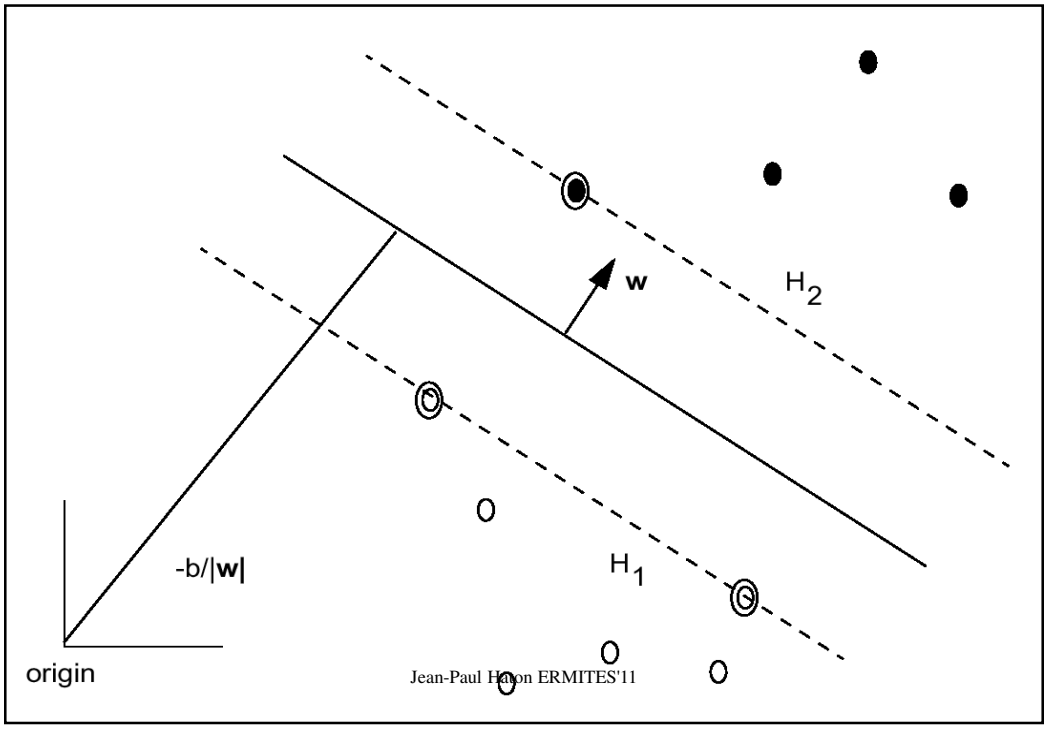
- Propriétés de l'apprentissage :
  - Non discriminant : ML (*max. de vraisemblance*)
  - Discriminant *:*
    - MCE (*min. d'erreur de classification*),
    - MMI (*max. d'information mutuelle*)
    - Réseaux neuromimétiques
    - SVM (Vapnik, 1995)

SVM : principe



Jean-Paul Haton ERMITES'11





Apprentissage

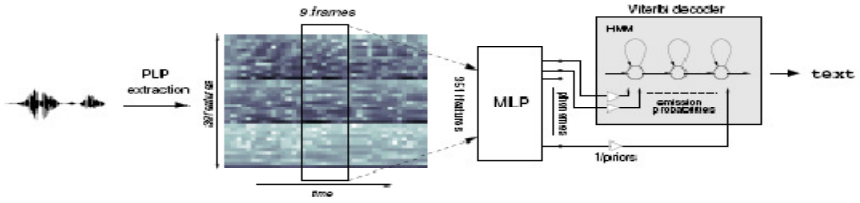
- Propriétés de l'apprentissage :
  - Non discriminant : ML (*max. de vraisemblance*)
  - Discriminant *:*
    - MCE (*min. d'erreur de classification*),
    - MMI (*max. d'information mutuelle*)
    - Réseaux neuromimétiques
    - SVM (Vapnik, 1995)
- Paramétrisation discriminante « probabiliste »

Paramètres probabilistes

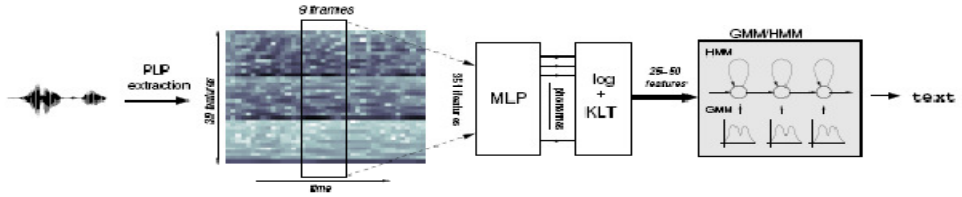
- Modélisation d'une distribution quelconque par un modèle neuronal type PMC. Les sorties du PMC peuvent approximer les probabilités *a posteriori* des classes phonétiques. Deux possibilités :
  - les utiliser dans les HMM à la place des GMM (Morgan, Bourlard)
  - les considérer comme des paramètres complémentaires des MFCC *:*
    - paramétrisation TANDEM (Hermansky)
    - « *bottleneck* »
    - méthode actuelle la plus efficace en parole continue

Paramètres probabilistes

Approche hybride classique (Morgan, Bourlard) :

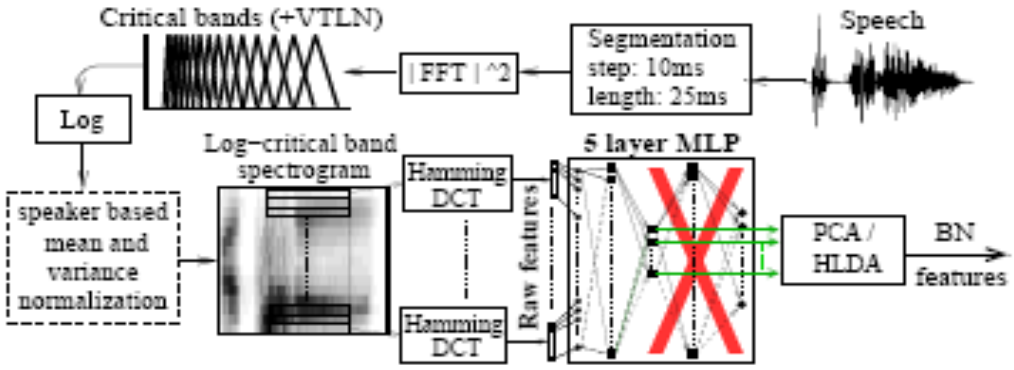


Approche tandem (Hermansky) :

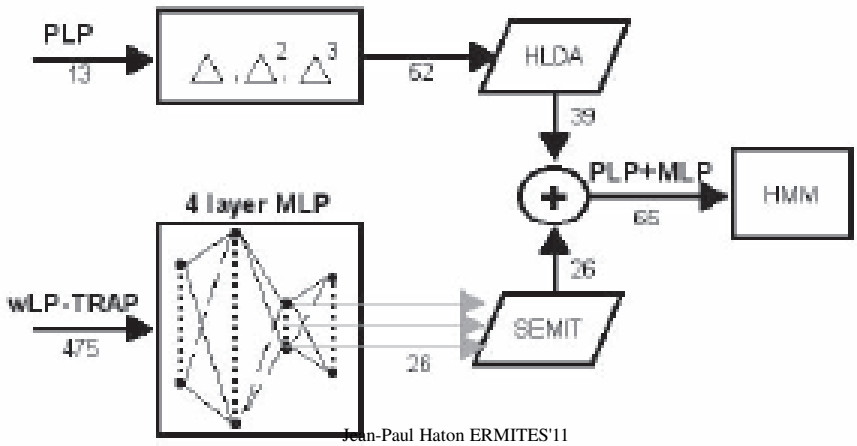


d'après (Fousek 2007)

« Bottle-neck » Features
(Gr´ezl, Fousek, 2007)



Exemple : Cambridge Univ. (Park *et al.*, 2009)



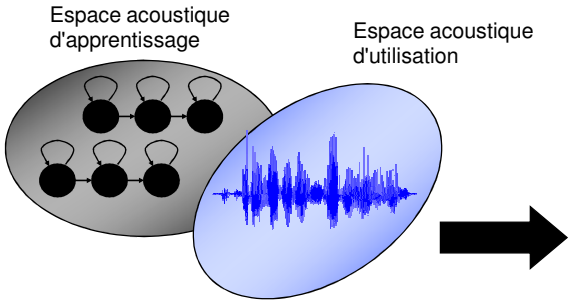
Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système

Robustesse et taux de reconnaissance



Resource Management	Erreur
Locuteurs natifs (USA)	3.6 %
Locuteurs non natifs	34.9%

Wall Street Journal 5000	Erreur
Locuteurs natifs (USA)	4.7 %
Locuteurs non natifs	29.1%

La disparité entre les deux espaces fait chuter les performances

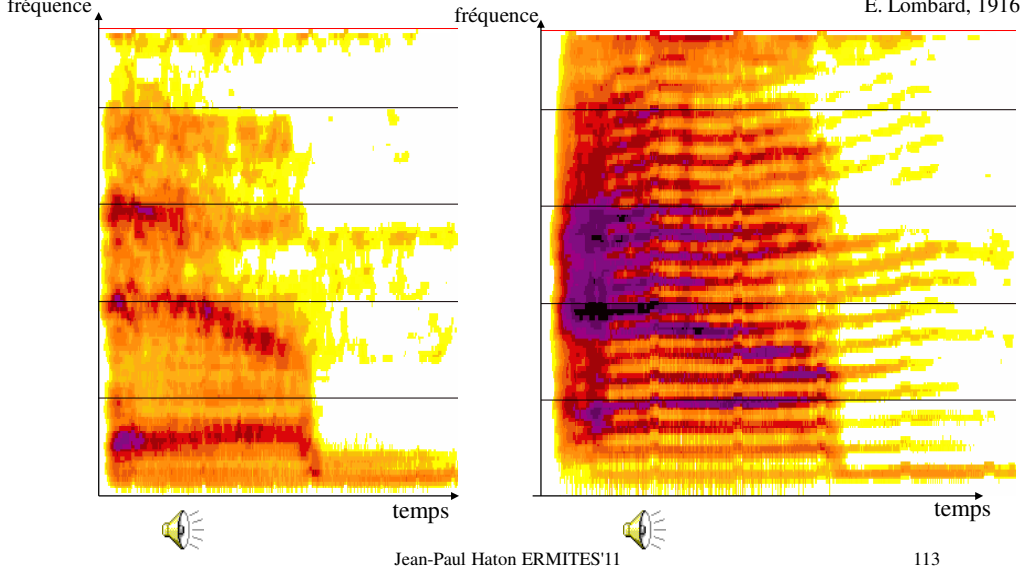
Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essoufflement, fatigue, effet Lombard, etc.)

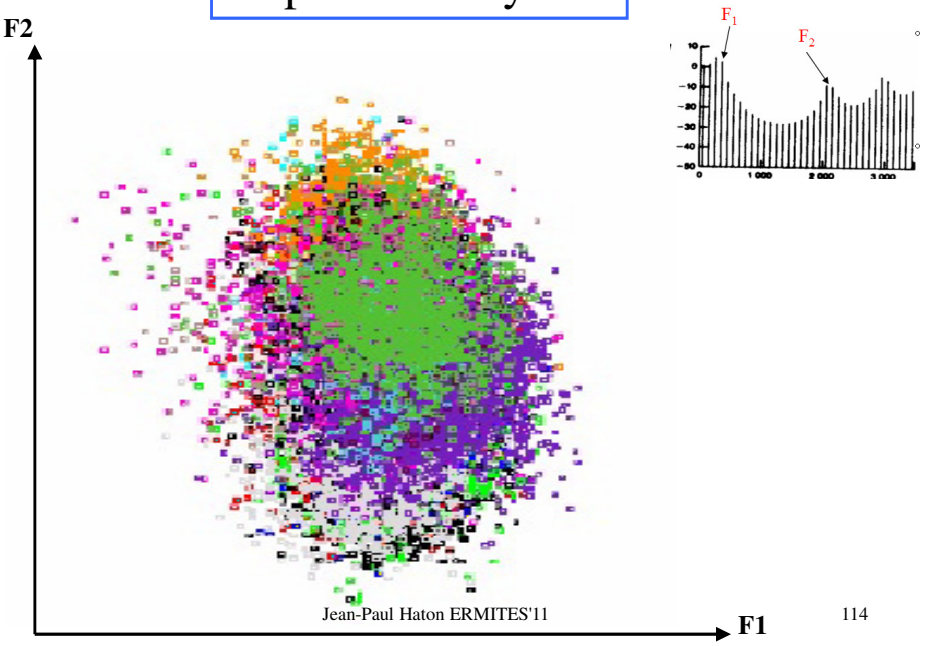
Effet Lombard

?

E. Lombard, 1916



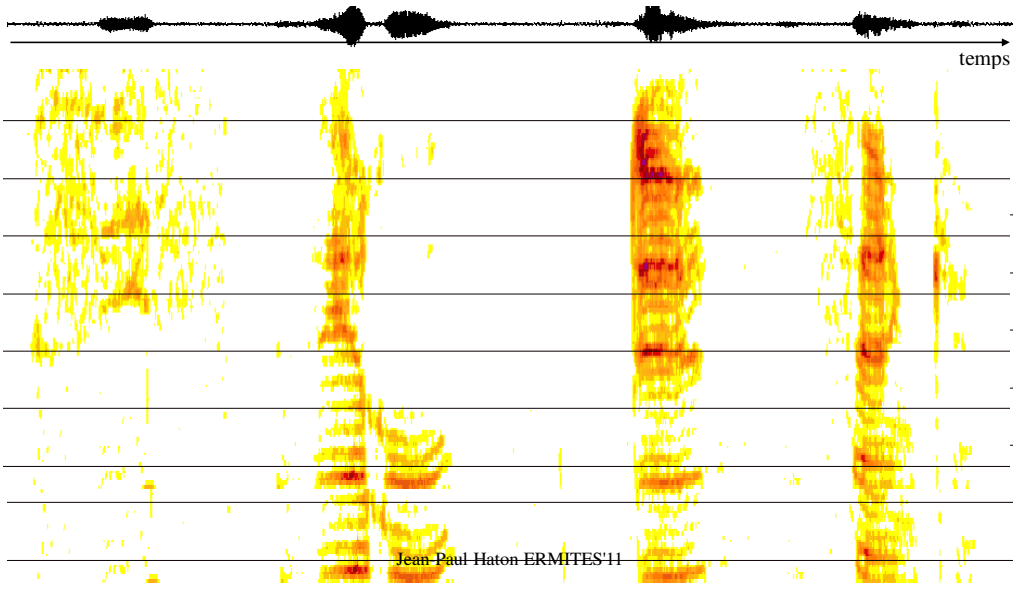
Le plan des voyelles



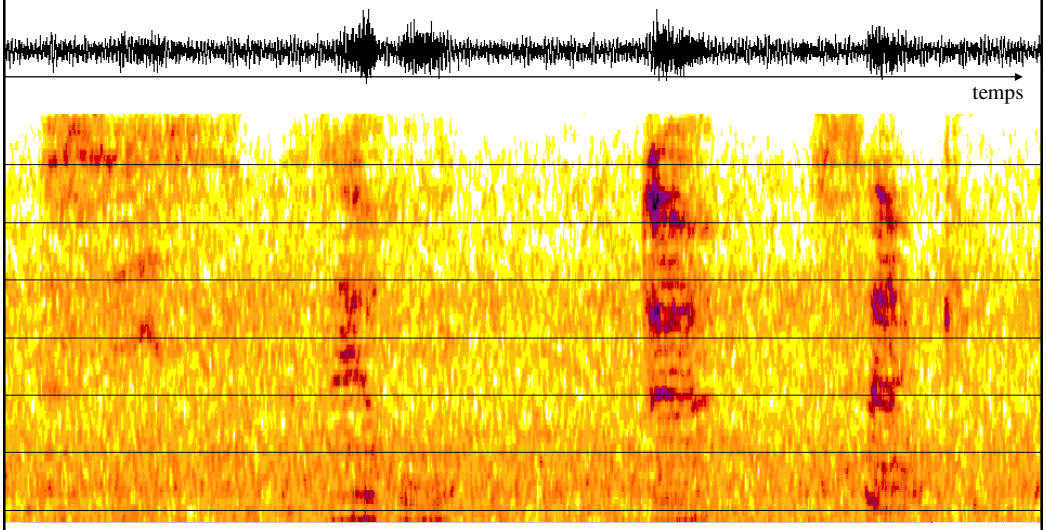
Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essoufflement, fatigue, effet Lombard, etc.)
  - à la prise de son (microphone, bruit ambiant, position, etc.)
  - au canal de transmission (distorsion, écho, bruit électronique)

Enregistrement à bord d'une voiture : à l'arrêt



Enregistrement à bord d'une voiture : à 90 km/h



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, essoufflement, fatigue, effet Lombard, etc.)
  - à la prise de son (microphone, bruit ambiant, position, etc.)
  - au canal de transmission (distorsion, écho, bruit électronique)
  - au contexte linguistique (coarticulation, assimilation, etc.)
- Résultat :
  - interactions complexes et effets cumulés!
  - nécessité de méthodes robustes à tous les niveaux

Augmenter la robustesse

Trois solutions, non exclusives :

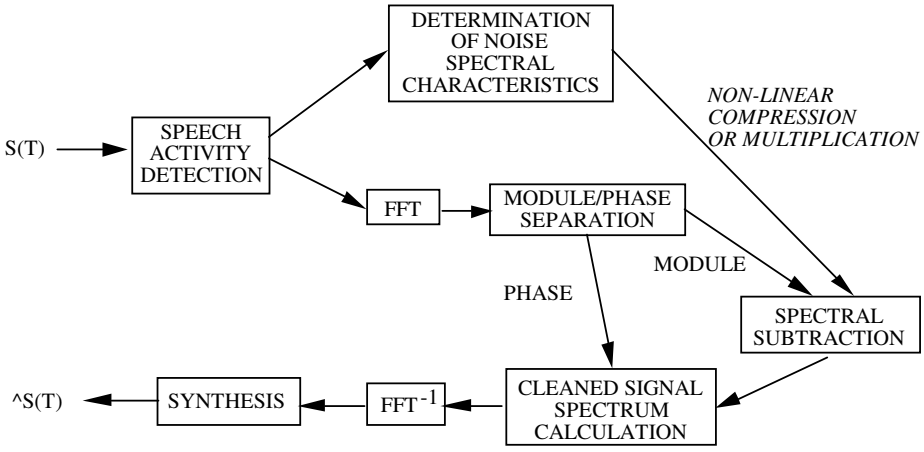
1. Prétraitement et débruitage du signal
2. Paramétrisation « robuste »
3. Adaptation des modèles acoustiques

Prétraitement

But : améliorer la qualité subjective du signal...

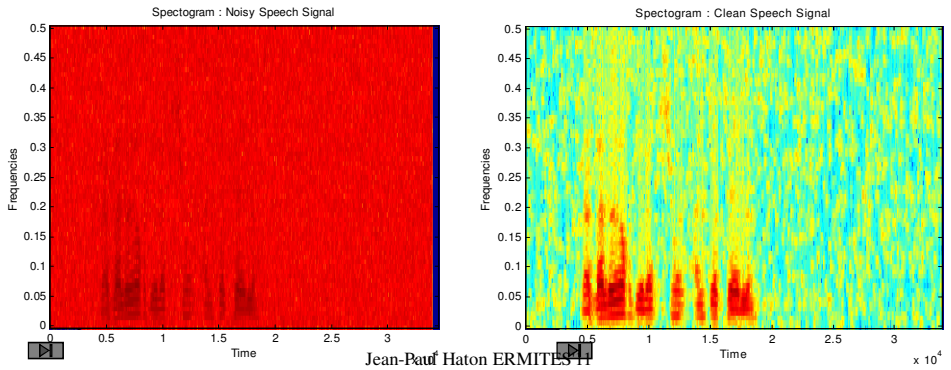
- Microphones : *noise cancellers*, antennes
- Soustraction spectrale (Berouti et al., 1979)
- Estimation d'erreur (*minimum mean square error estimation*, MMSE) (Ephraim and Malah, 1985)
- Wiener filtering (linear MMSE) (Wiener, 1949)
- Kalman filtering (Paliwal and Basu, 1987)
- Subspace methods (Ephraim and Trees, 1995)

Principe de la soustraction spectrale



Soustraction spectrale : Exemple

Minimum Square Error Algorithm (Ephraïm & Malah)



Robust parameterization

- Use of knowledge from perception, psychoacoustics, neurology, etc. :
  - ear models
  - noise masking
  - dynamic parameters
  - analyses PLP, RASTA, RASTA-PLP, etc.
- Channel Normalization

Channel Normalization

A simple, yet efficient technique, to reduce signal variability is channel normalization:

- *Cepstral Mean Normalization, CMN (Furui, 1981)* and variants:

Subtraction of cepstral mean to the cepstral coefficients
Reduce time-invariant distorsions (channel, recording device)

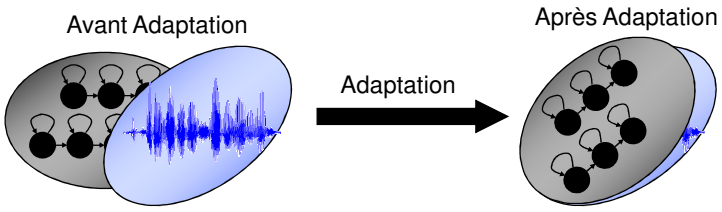
- *RASTA (Hermansky, 1994):*

Temporal filtering with a bandpass filter:

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})}$$

Associated with PLP: Rasta-PLP and JRASTA-PLP

Adaptation des modèles



Adaptation des modèles

Trois grandes catégories de méthodes :

– *Clustering* de locuteurs (voix propres, *eigenvoices*)

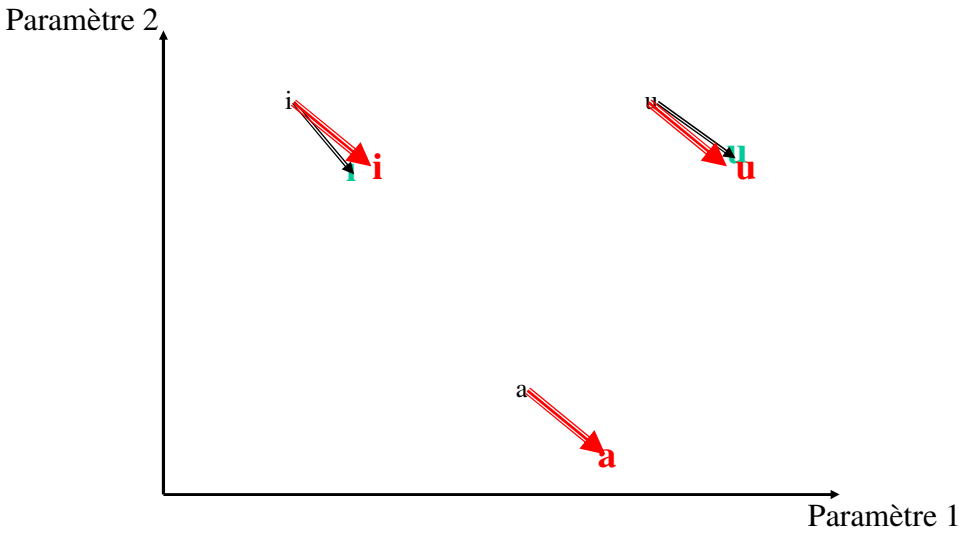
– Méthodes bayésiennes (Maximum a Posteriori, MAP)

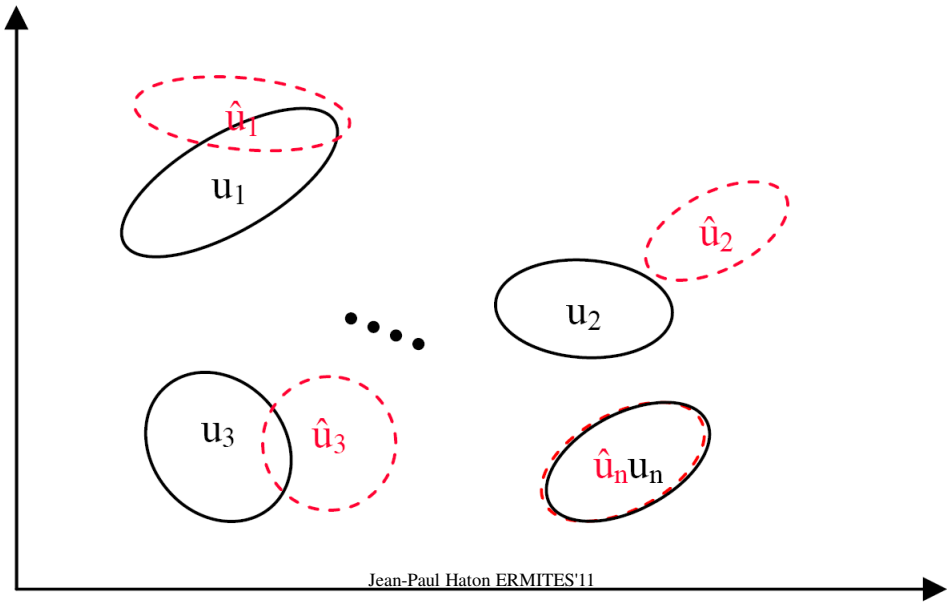
– Méthodes utilisant des transformations d'espaces (*Maximum-Likelihood Linear Regression,* MLLR)

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
- Un exemple simple : MLLR (Régression linéaire par maximum de vraisemblance)

MLLR : Exemple





MAP (maximum *a posteriori)*

- Apprentissage d'un HMM :

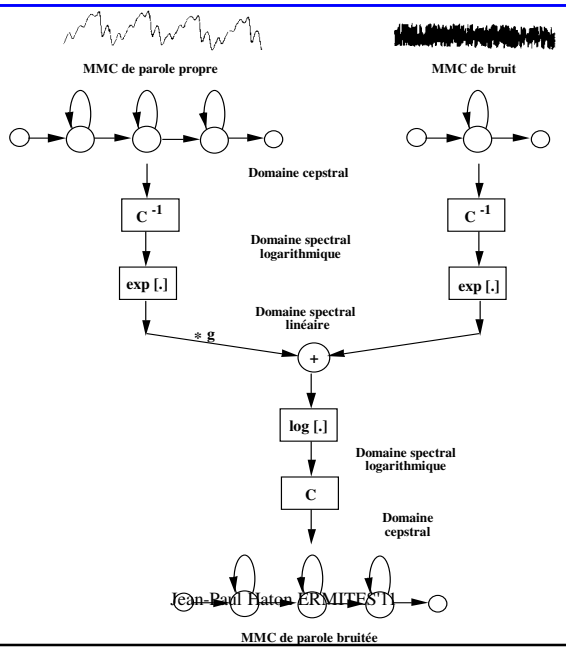
$$Arg\max_{\lambda} P(\lambda|O) = Arg\max_{\lambda} \frac{P(O|\lambda)P(\lambda)}{P(O)}$$

- 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

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

Domaines d'application

Contexte général :

- Évolution technologique
- Intégration des matériels et logiciels de TAP dans les postes de travail
- Convergence informatique, téléphone, télévision

Grands domaines :

- Machines à dicter
- Commandes d'appareils
- Handicapés
- Télécommunications : télématique vocale

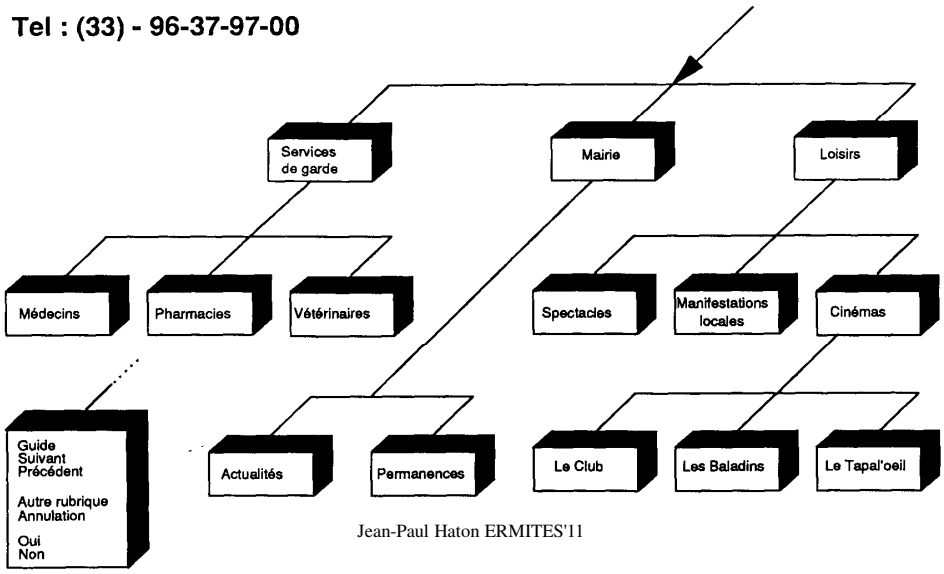
Télématique vocale

Les générations de systèmes de télématique vocale :

- <u>première génération (début vers 1990)</u>
  - . Amérique du Nord : ATT (*Call Collect*), Bell Northern (AABS)
  - . Japon : banques (ANSER)
  - . France : CNET (Mairievox), MACIF

Serveur vocal interactif MAIRIEVOX

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



Télématique vocale

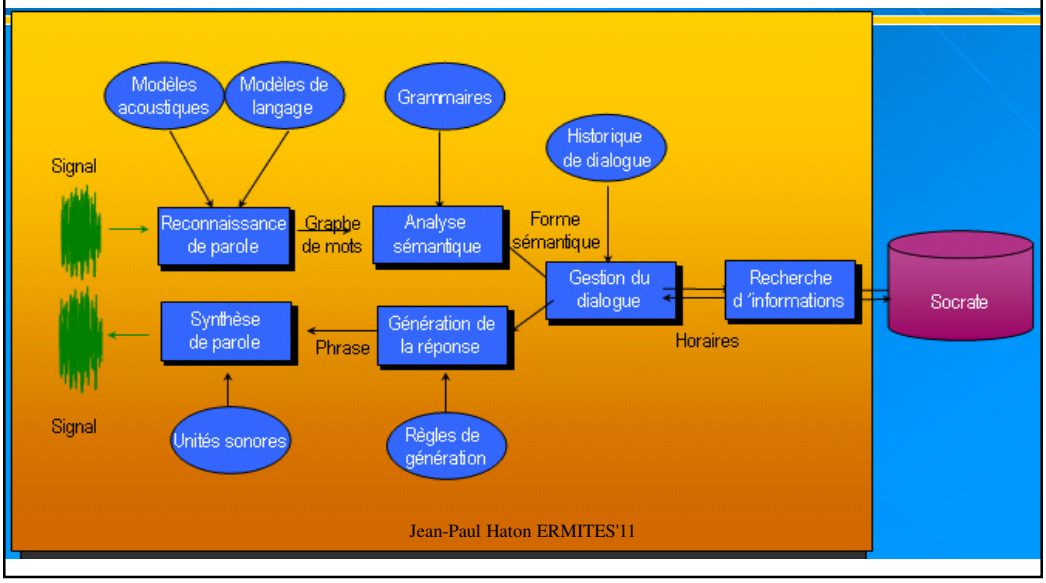
Les générations de systèmes de télématique vocale :

- <u>première génération (début vers 1990)</u>
  - . Amérique du Nord : ATT (*Call Collect*), Bell Northern (AABS)
  - . Japon : banques (ANSER)
  - . France : CNET (Mairievox), MACIF

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

- <u>générations futures</u> : intégration de services, traduction parole-parole, dialogue oral, identification de la langue

SNCF: système RECITAL



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 apprentissage...
- Exemples :
    - IBM *ViaVoice*
    - Nuance *Dragon Naturally Speaking*
    - Philips *Speech Magic*