

Reconnaissance Automatique de la Parole

LINARES Georges

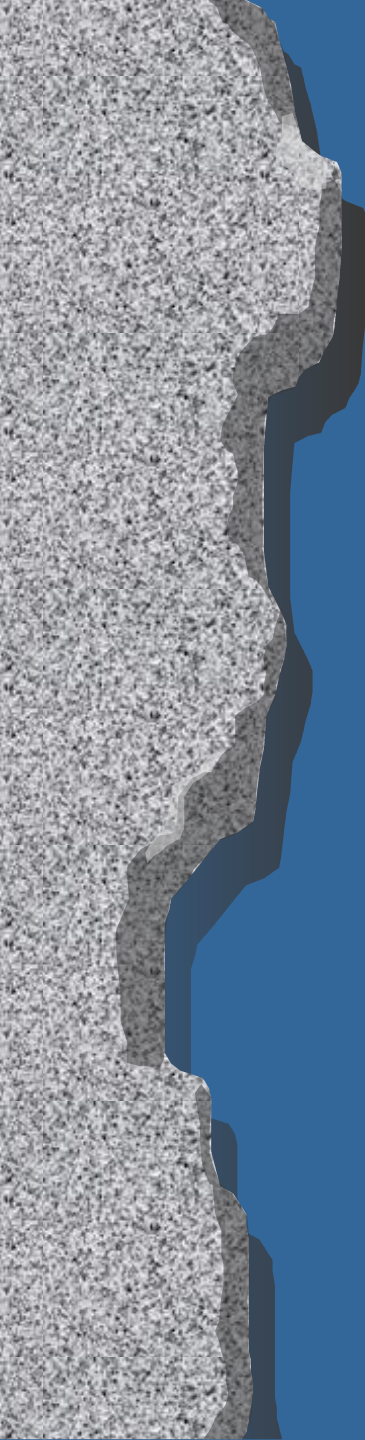
Laboratoire d'Informatique d'Avignon
Université d'Avignon et des Pays de
Vaucluse

Objectifs du cours

- ✓ Comprendre les problèmes liés au traitement automatique de la parole
- ✓ Connaître les types de modèles et d'algorithmes utilisés
- ✓ Connaître les capacités et les limites des systèmes actuels

Plan

- ✓ Introduction : la RAP dans l'IA
- ✓ Objectifs et Applications
- ✓ Problématique
- ✓ La paramétrisation
- ✓ Modélisation acoustique : DTW et HMM
- ✓ Modélisation linguistique
- ✓ Algorithmes de recherche
- ✓ Mise en oeuvre d'un SRAP



Une machine à notre image ?

- ✓ *Frankestein ou le prométhé moderne* (M. Shelley, 1818)
- ✓ HAL (*2001 Odyssée de l'espace*, A. C. Clarke)
- ✓ *La trilogie d'Ender* (O.S. Card, 1986)



Les objectifs de l'I.A.

- ✓ Simulation des comportements humains
 - ✓ Perception
 - ✓ Raisonnement
 - ✓ Décision
- ✓ Applications industrielles



Approche anthropomorphique

- ✓ Implantation en machine des mécanismes de la pensée
 - ✓ Ces mécanismes sont mal connus
- ✓ Cerveaux artificiels
 - ✓ Différences structurelles

Approche pragmatique

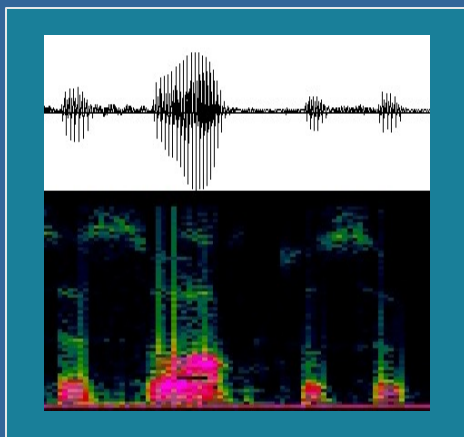
- ✓ Elaborer des systèmes adaptés aux machines
 - ✓ Acquisition des connaissances (apprentissage)
 - ✓ Représentation des connaissances (modèles)
 - ✓ Exploitation des connaissances (algorithmes)

Reconnaissance de la parole et I. A.

- ✓ Tâche de perception : reconnaissance des formes
- ✓ Modélisation des connaissances
 - ✓ Conception des modèles
 - ✓ Paramétrisation (apprentissage)
- ✓ Algorithmes de décodage

RAP : Objectifs

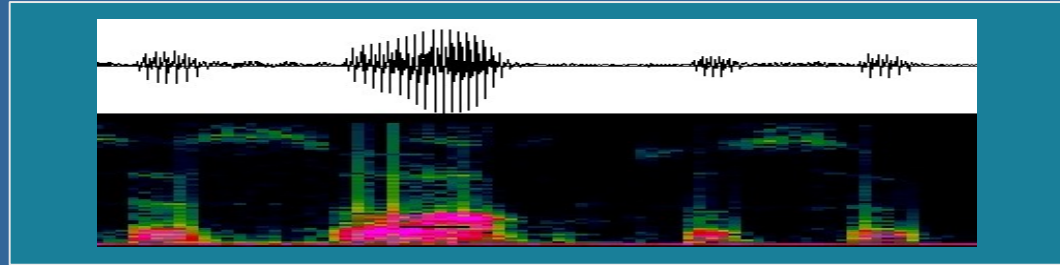
- ✓ Communication homme-machine
- ✓ Traitement de documents audios



Applications

- Commande vocale
 - Machines, véhicules, appareils ménagers, ordinateurs
- Aides aux handicapés
- Interaction avec des systèmes complexes
 - Dialogue, serveurs interactifs, bornes multimodales
- Application aux Télécommunications
 - Téléphone mains libres, composeur vocal,
- Indexation de documents audios

Dictée vocale : du son au texte



Signal continu et concret
Information : 100 000 bits par seconde



CE SOIR TOUS AU

Mots
Information : 60 bits par seconde

Difficultés du TAP

- Grande quantité d'informations
- Grande variabilité des informations
 - Type de parole,
 - Environnement,
 - Locuteur,
 - Coarticulation.
- Diversité des sources de connaissance
 - Acoustiques
 - Linguistiques
 - Extra-linguistiques

Acoustique

- Identification des unités acoustiques (phonèmes)
 - *Médecin* -> *métcin*
- Insuffisant :
 - *oo li on dd oo rr*
 - *(oo li) (on dd) (oo rr)*
 - *(oo li on dd) (oo rr)*

Prosodie

- Intonation, durée des phonèmes
 - Modalité
 - *tu y vas?*
 - *tu y vas!*
 - Désambiguisation
 - *La belle ferme le voile*
 - *Le boucher sale la tranche*

Linguistique

- ✓ Information lexicale
 - ✓ *Oli ondor*
 - ✓ *Ila man get*
- ✓ Information syntaxique
 - ✓ *Il l'a manger*
 - ✓ *Le président a par les.*

Sémantique

- ✓ Sémantique (sens)

- ✓ *La pierre m'a demandé l'heure*

- ✓ *La couturière tisse des fils de soie/soi*

- ✓ Pragmatique (contexte)

- ✓ *Son petit tamis est tombé dans la piscine*

- ✓ *Les deux partis font appel*

Contraintes Générales en RAP

- ✓ Dialogue Oral / Dialogue Multimodal
- ✓ Langage Naturel / Langage Artificiel
- ✓ Situation Réelle / Environnement contrôlé
- ✓ Compréhension / Reconnaissance
- ✓ Parole Continue / Mots Isolés
- ✓ Système Adaptable / Système
- ✓ Multilocuteur / Monolocuteur
- ✓ ...

Le signal de Parole

- ◆ Les sons de parole sont produits par deux processus différents
 - Vibration des cordes vocales
 - ◆ **Source de voisement**
 - Turbulence créée par l'air
 - ◆ s'écoulant rapidement dans une constriction du conduit vocal
 - ◆ lors de relâchement d'une occlusion du conduit vocal
 - ◆ c'est une **Source de bruit**
- ◆ Et « une » modulation

Phonèmes

- ◆ Les phonèmes sont les élément sonores les plus brefs qui permettent de distinguer différent mots
- ◆ Exemples [p] [b]
 - pas / bas
 - paie / baie
 - pot / beau

Phonèmes du français

TABLEAU I. — *Les phonèmes du français*

Consonnes

[p] paie	[t] taie	[k] quai
[b] baie	[d] daïs	[g] gai
[m] mais	[n] nez	[ɲ] gagner
[f] fait	[s] sait	[ʃ] chez
[v] vais	[z] zéro	[ʒ] geai
[w] ouais	[y] huer	[j] yéyé
	[l] lait	[R] raie

Voyelles

[i] lit	[y] lu	[u] loup
[e] les	[ø] leu	[o] lot
[ɛ] lait	[œ] leur	[ɔ] lotte
[a] là	[ə] le	
[ɛ̃] lin	[ɑ̃] lent	[õ] long

Note : Les distinctions vocaliques [e]-[ɛ], [ø]-[œ] et [o]-[ɔ] ne sont pas faites dans tous les contextes et par tous les locuteurs du français. Par contre, certains locuteurs font aussi des distinctions entre patte et pâte, ([a]-[ɑ]) ainsi qu'entre brin et brun ([ɛ̃]-[œ̃]).

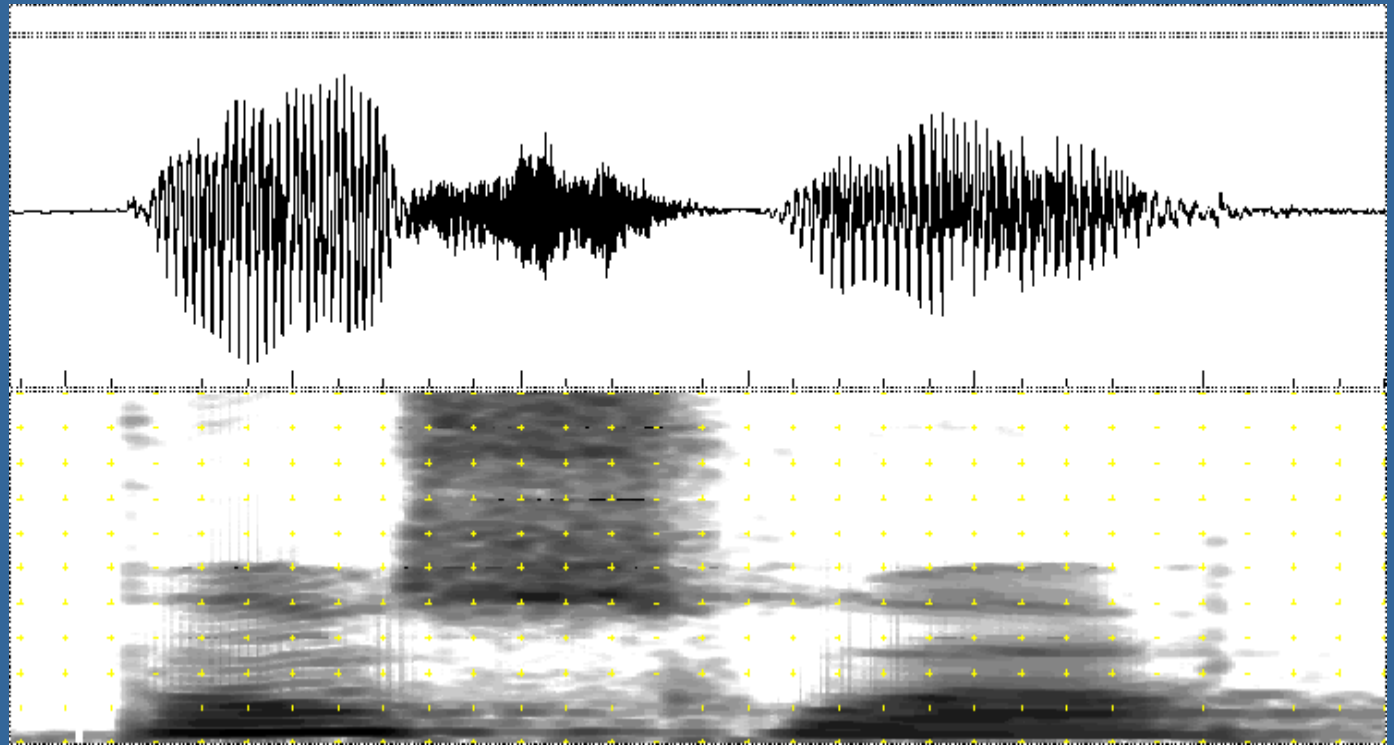
Phonèmes du français

TABLEAU II. — *Classification des phonèmes du français en traits distinctifs*

CONSONNES				Lieu d'articulation
Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales ←	
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[z]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[y]	[j]	
Liquides		[l]	[R]	
VOYELLES				
Orales	Antérieures		Postérieures	
	Non arrondies		Arrondies	
Fermées	[i]	[y]	[u]	
	[e]	[ø]	[o]	
	[ɛ]	[œ]	[ɔ]	
Ouvertes	[a]			
Nasales	Antérieures		Postérieures	
Fermées	[ɛ̃]		[õ]	
Ouvertes		[ã]		

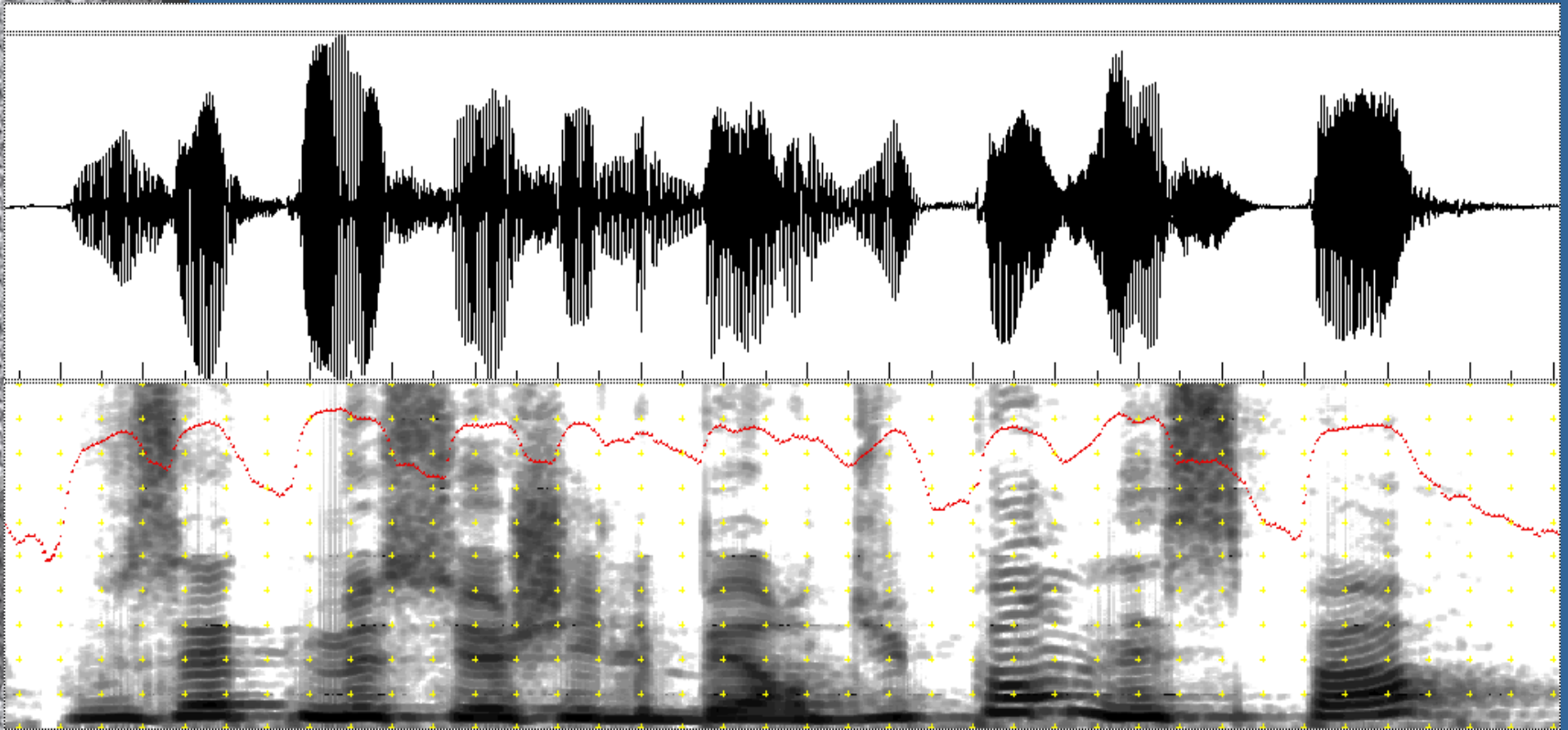
Signal de Parole

Représentation temps- fréquence



Bonsoir

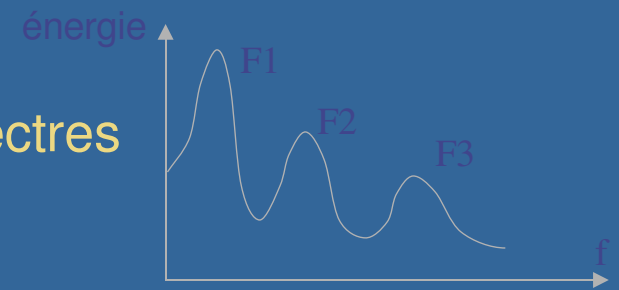
Signal de Parole



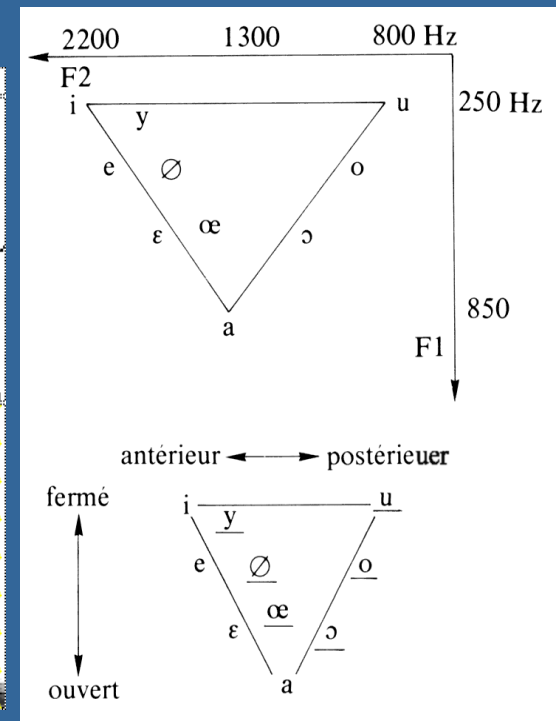
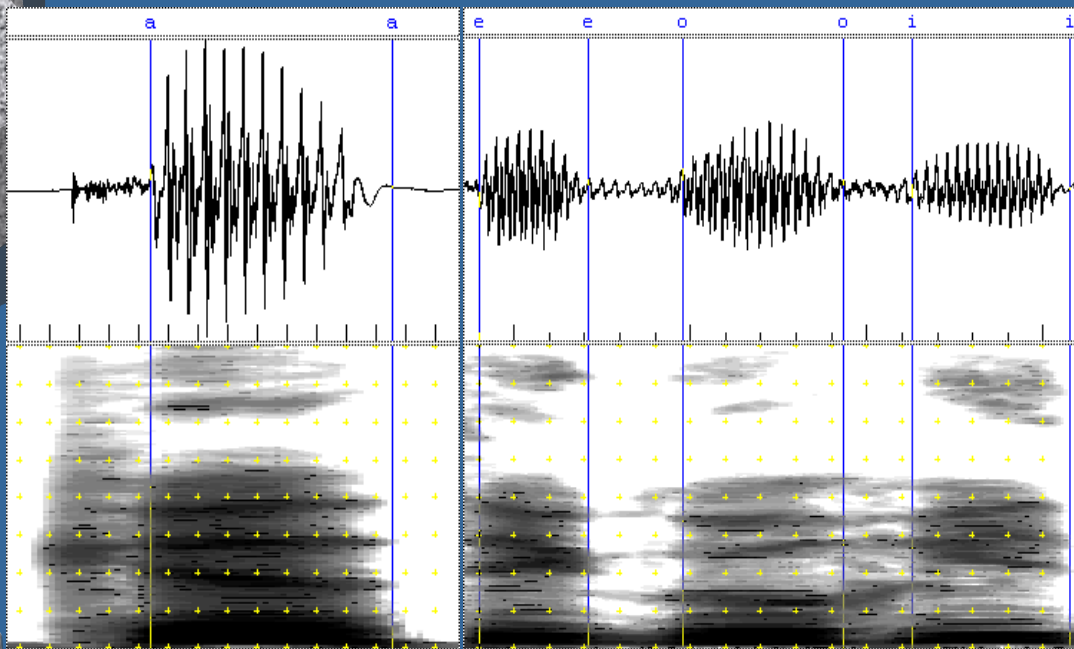
Vous êtes Monsieur Gilbert Dupont n'est-ce pas ?

Formants

spectres



- ◆ Fréquences de résonance du conduit vocal
- ◆ Triangle vocalique des voyelles





Fréquence fondamentale ou *f0*

- ◆ Vibration des cordes vocales
- ◆ Dépend essentiellement
 - de l'âge et du sexe du locuteur
 - du locuteur
- ◆ Valeurs standards
 - 100 à 150 Hz pour l'homme adulte
 - 140 à 240 Hz pour la femme adulte
- ◆ Mais une grande variété !
- ◆ peut présenter des variations considérables chez un même locuteur
 - selon le type de phrase prononcée
 - selon l'état émotif et l'attitude du locuteur

Prosodie

- ◆ Hauteur de la voix (*pitch* ou fréquence fondamentale)
- ◆ Intensité de la voix (énergie)
- ◆ Durées successives des segments syllabiques

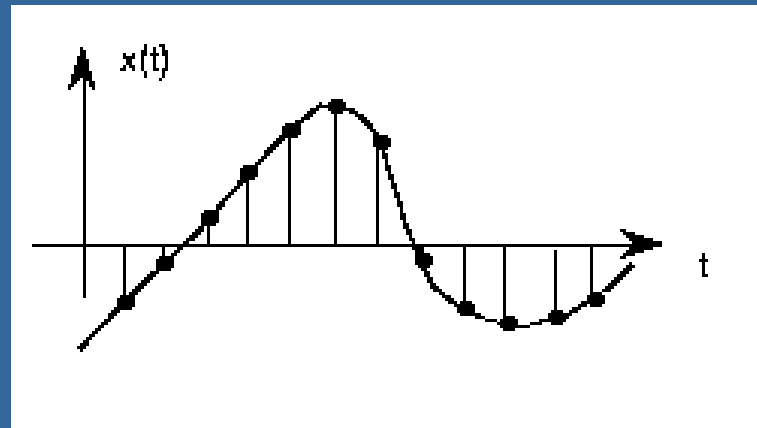


Intonation / Mélodie de la voix

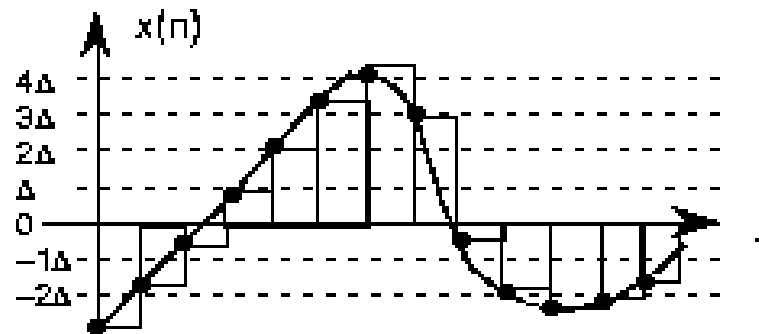
Conversion analogique-numérique

- ◆ Etape d'échantillonnage puis de quantification

échantillonnage



quantification



Energie d'un signal

- ◆ Energie d'un signal continu $s(t)$ sur l'intervalle de temps $[t_1, t_2]$

$$W_s(t_1, t_2) = \int_{t_1}^{t_2} s^2(t) dt$$

- ◆ Energie d'un signal discret $s(n)$ sur l'intervalle $[n_1, n_2]$

$$W_s(n_1, n_2) = \sum_{n=n_1}^{n_2} s^2(n)$$

Rapport signal/bruit

- ◆ La qualité d'un signal est souvent représentée par le **Rapport Signal/Bruit** ou **RSB** (*SNR* en Anglais)
- ◆ Pour $x(t)=s(t)+n(t)$

$$SNR = \frac{W_s}{W_n}$$

$$SNR_{dB} = 10 \log_{10} SNR$$



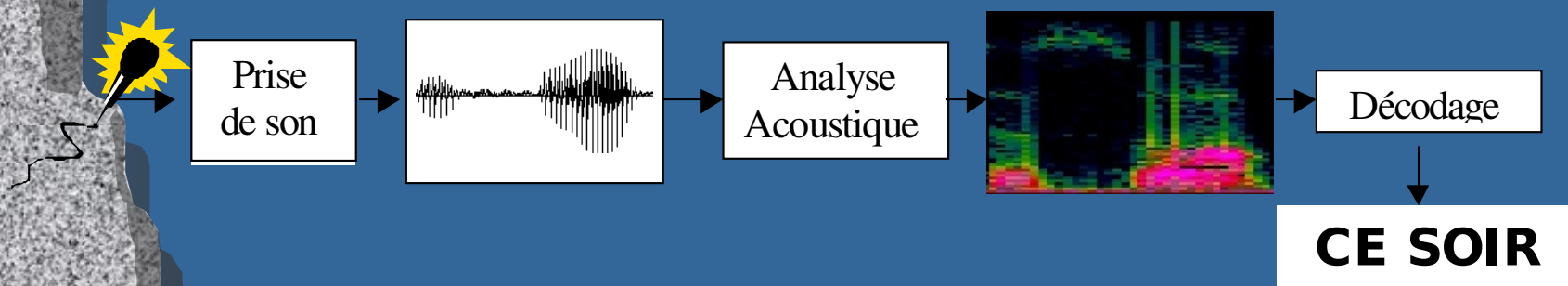
Reconnaissance de la parole: Historique

- ✓ Années 70 : méthodes à base de connaissances, **décodage acoustico-phonétique**
- ✓ Fin 70 : Reconnaissance de mots isolés, programmation dynamique
- ✓ 1980 : Modèles de Markovs cachés
- ✓ 1990 : Parole continue, Grands vocabulaires, adaptations

Les Systèmes de R.A.P.

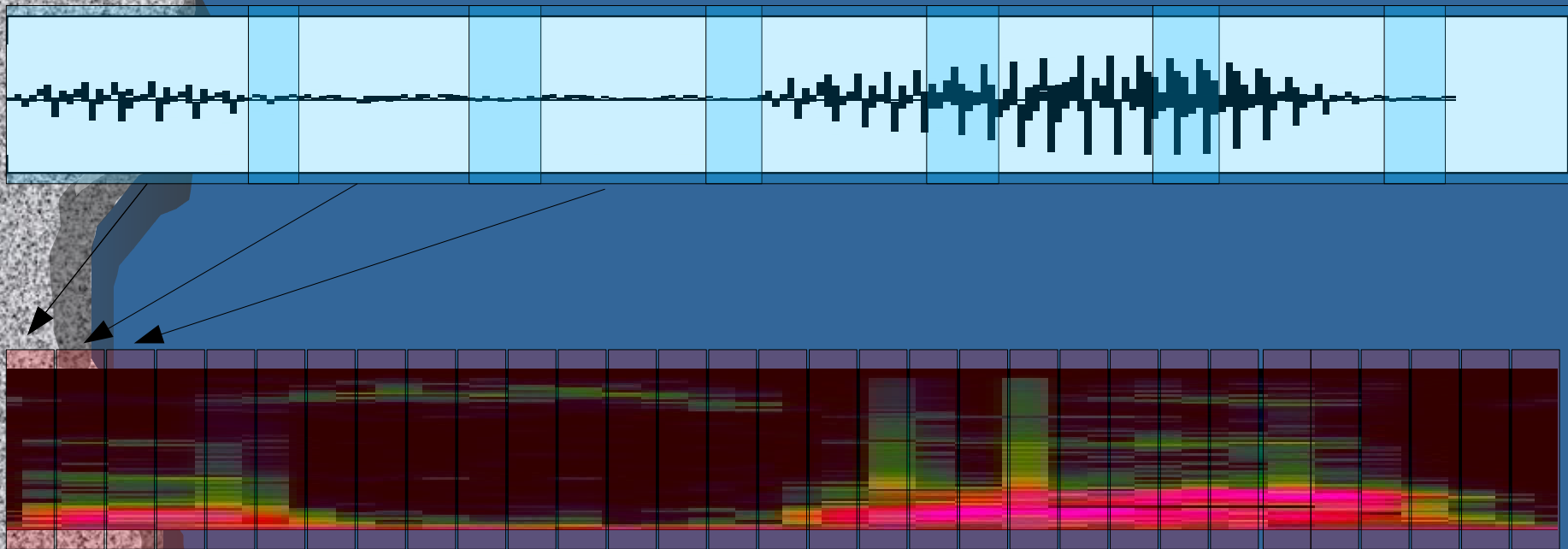
Chaîne de traitement :

- (1) **Analyseur Acoustique** (paramétrisation)
 - Vecteurs acoustiques toutes les 10 ou 20 ms
- (2) Moteur de reconnaissance
 - Détermine les mots reconnus à partir des vecteurs issus de (1)



Paramétrisation

- Analyse sur une fenêtre glissante
 - Ordre de grandeur : 30 ms
 - Recouvrement



Paramétrisation

- Généralités :

- Filtre passe bas (8khzt)

Préaccentuation (diminuer la dynamique du spectre) :

- Élimination de la composante continue
 - Pas d'information utile

Paramétrisation

- Analyse temps fréquence (spectro):
 - Transformée de Fourier à court terme
 - Convolué avec une fenetre qui évite les effets de bord
 - Energie dans chaque bande de fréquence
 - Calcul rapide
 - ... jamais utilisé directement

Paramétrisation

- LPC (*Linear Predictive Coding*)
 - *Modèle autoregressif (AR)*
 - *Principe :*
 - *Éliminer la redondance temporel du signal*
 - *Filtre AR :*

$$s(n) = \sum_{i=1}^P a_i s(n-i)$$

Paramétrisation

- *Erreur de prédiction*

$$e(n) = x(n) - s(n) = x(n) - \sum_{i=1}^P a_i x(n-i)$$

erreur de prédiction

échantillon observé

prédiction

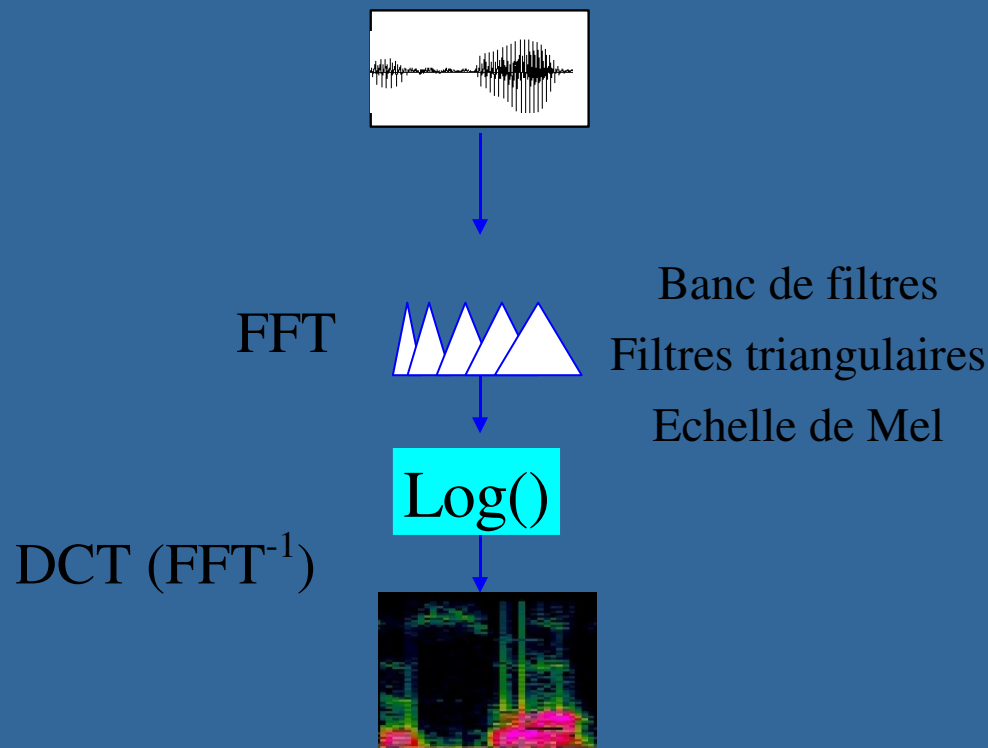
Coefficients
d'autocorrelation

Paramétrisation

- PLP (Perceptually-based Linear Prediction)
 - Inspiré des modèles de perception
 - Approximation de la densité spectrale à résolution variable (échelle de Bark)
 - Préaccentuation :
 - L'intensité perçue dépend de la fréquence
 - Préaccentuation basée sur des abaques

Paramétrisation

- MFCC (*Mel Frequency Cepstrum Coefficients*)
- *La plus fréquemment utilisée*



Paramétrisation (MFCC)

source

conduit

$$s(t) = e(t) * h(t)$$

log(fft)

$$\log(S(f)) = \log(E(f)) + \log(H(f))$$

$$s'(cef) = e'(cef) * h'(cef)$$

Paramétrisation (MFCC)

Autres types de transformation:

- *Traits acoustiques : formants, dpz, etc.*

Problèmes :

- augmenter le robustesse
- Réduire la dimension du problème
 - Choix des coefficients
 - Méthode d'analyse de données
 - ACP : analyse en composantes principales
 - LDA : analyse discriminante

Les systèmes à base de règles

- Principe :
 - Utilisation de connaissances explicites
 - Règles formulées par des experts
- Problèmes :
 - Performances faibles
 - Collecte de la connaissance



DTW

(Dynamic Time Warping)

- Principe :
 - Chaque mot est modélisé par une réalisation (un exemple de référence)
 - Calcul d'une distance de l'observation aux références
 - Le mot reconnu est celui dont la référence est le plus proche de l'observation

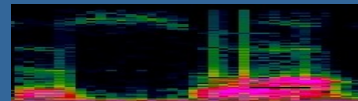
DTW

(Dynamic Time Warping)

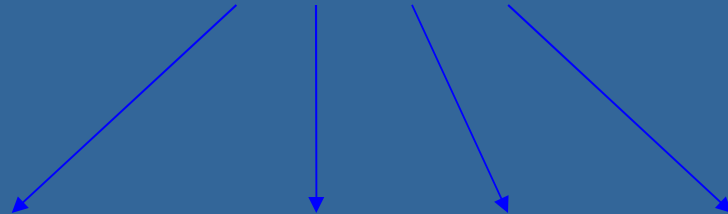
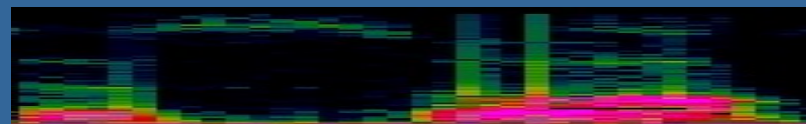
Problème : quelle distance utiliser ?

Les séquences sont de longueurs variables

observation



référence



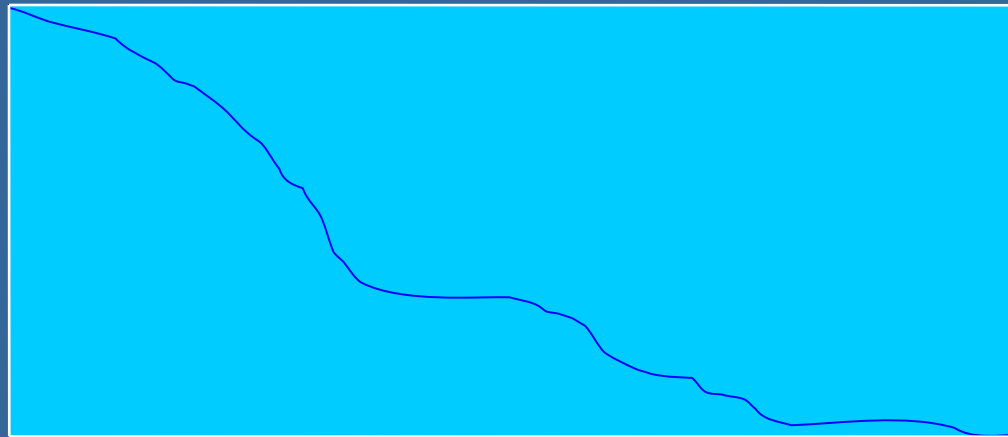
DTW

(Dynamic Time Warping)

✓ Distance : coût du chemin de déformation minimale

observation

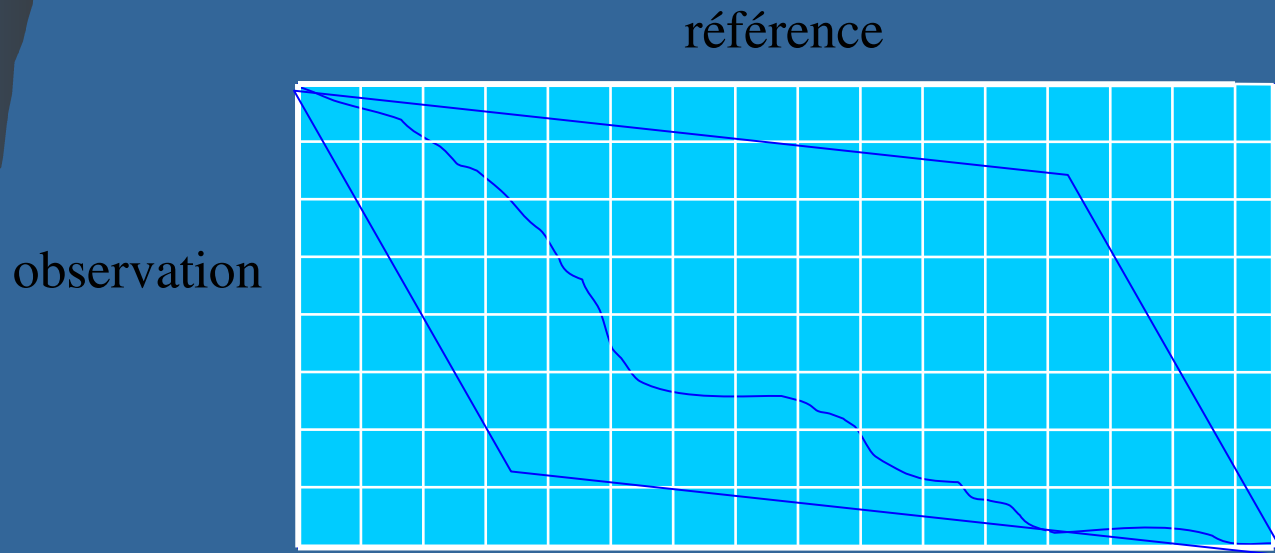
référence



DTW

(Dynamic Time Warping)

- ✓ Algorithme d'alignement dynamique
 - ✓ Avec contraintes de déformation



DTW

(Dynamic Time Warping)

✓ Coût d'un chemin :

$$d(\Phi_x, \Phi_y) = \sum_{n=1}^L D(X_{\phi_x(n)}, Y_{\phi_y(n)}) + \sum_{n=1}^{L-1} a[\Phi_x(n+1) - \Phi_x(n), \Phi_y(n+1) - \Phi_y(n)]$$

Distance (quadratique ?) trames/trames

Coût des transitions

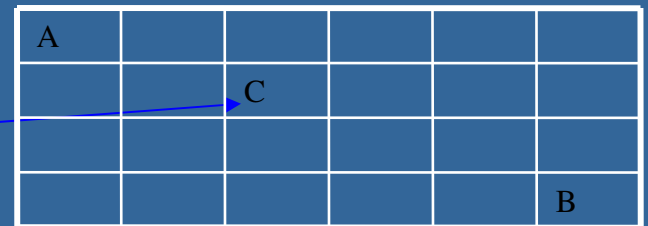
DTW

(Dynamic Time Warping)

✓ Algorithme de parcours :

- ✓ Principe : le meilleurs chemin allant de A à B en passant par C est :
 - ✓ le meilleurs chemin de A à C
 - ✓ ...suivi du meilleur de C à B

Coût du chemin optimal de A à C



DTW

(*Dynamic Time Warping*)

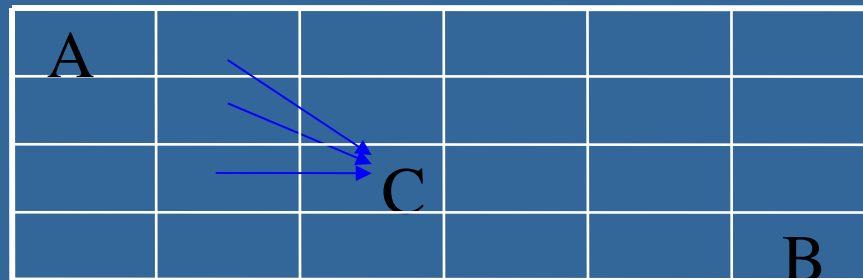
Pour chaque colonne c ,

Pour chaque ligne l ,

Pour chaque ligne lp ,

Si $V(c,l) > V(c-1,lp) + a[l,l-lp] + D(Xc,Yl)$ alors

$$V(c,l) = V(c-1,lp) + a[l,l-lp] + D(Xc,Yl)$$



Complexité : $L \times L \times C$



DTW

(Dynamic Time Warping)

- ✓ Avantages :

- ✓ Rapidité de paramétrage du système
- ✓ Rapidité de décodage

- ✓ ...mais

- ✓ Mots isolés !
- ✓ Faible robustesse au bruit
- ✓ Mono-locuteur
- ✓ Une référence par mot : petit vocabulaire
- ✓ Pas d'exploitation de l'information linguistique



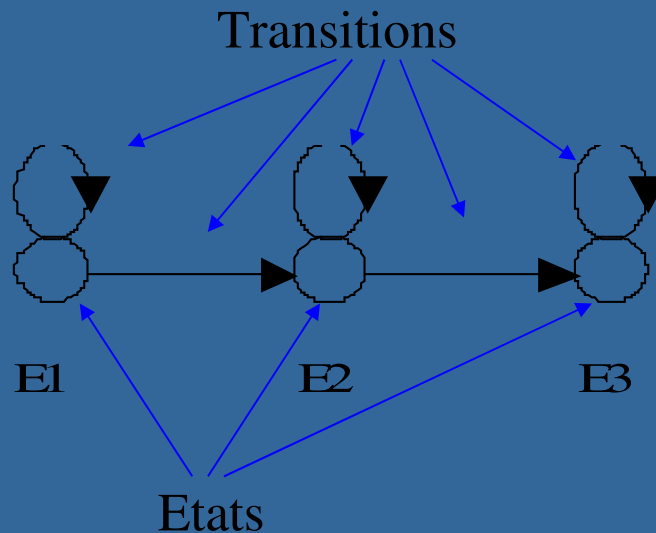
DTW

(Dynamic Time Warping)

- ✓ Améliorations :
 - ✓ Heuristiques sur le parcours
 - ✓ Choix des références
 - ✓ Moyennes, multiples, etc.
 - ✓ Distances et coûts des transitions peuvent être affinées

Modèles de Markov

Une unité acoustique (ex : phonème, mot) :
un modèle



Modèles de Markov

- ✓ Etats : une fonction de densité de probabilité : $\text{Prob}(X_k/E_i)$
 - ✓ Probabilité de produire l'événement X_k sur l'état E_i ,
 - ✓ Modélisant les formes rencontrées.
- ✓ Transitions : probabilité de transition d'un état à un autre : $P(E_{i+1}/E_i)$
 - ✓ Contraint l'ordre temporel dans lequel les formes doivent être observées

Modèles de Markov

- ✓ Les états :
 - ✓ Codent des « moments » des réalisations acoustiques
 - ✓ Estimateurs de probabilités par combinaison de fonctions élémentaires
 - ✓ Généralement, les estimateurs sont des mixtures de gaussiennes (*GMM, Gaussian Mixture Model*)
 - ✓ Les GMM permettent d'estimer $P(X_i|E_i)$

Les GMMs

- ✓ Une Mixture de gaussiennes approche n'importe quelle fonction continue
- ✓ ... plus ou moins bien

$$f(x) = \sum_{n=0}^N p_n \mathcal{H}(x_t)$$

Loi normale

Poids de la pdf



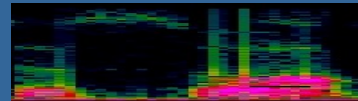


Estimation des probabilités acoustiques

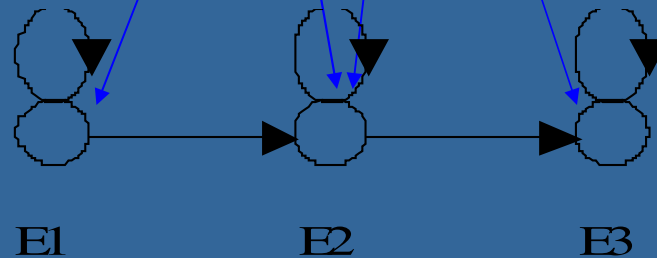
- ✓ Estimation de la vraisemblance d'une séquence d'observations sachant un modèle :
 - ✓ Alignement optimal de la forme acoustique sur la chaîne de Markov : chemin de probabilité maximale
 - ✓ Algorithme de Viterbi : alignement dynamique

Algorithme de Viterbi

observation



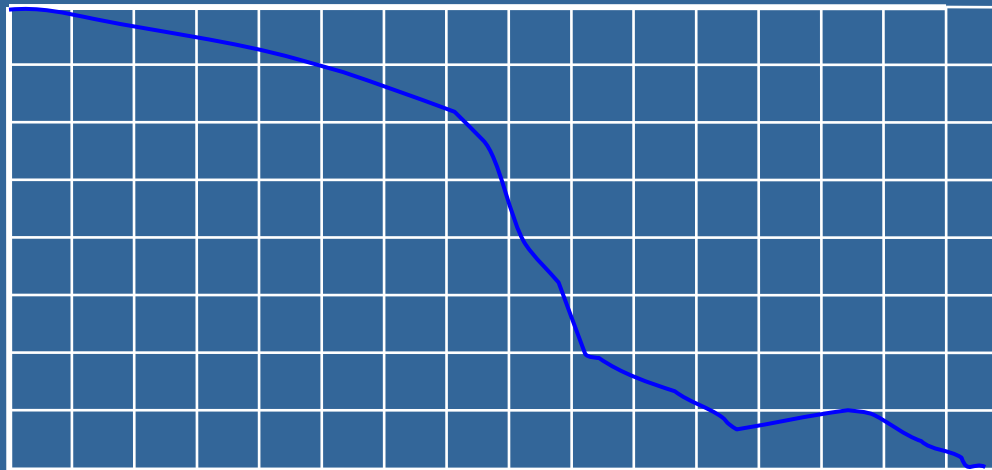
modèle



Algorithme de Viterbi

observations

états



Algorithme de Viterbi

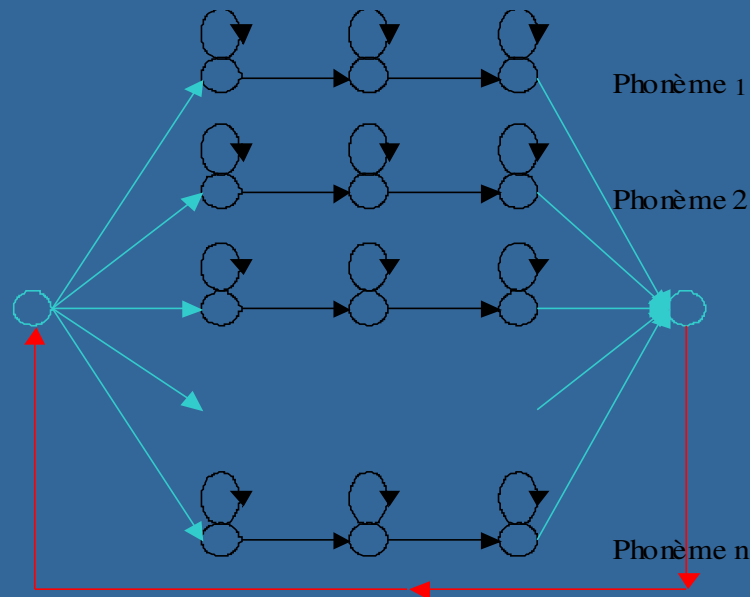
- Même principe que la DTW
 - Distance entre trames :
 - vraisemblance d'un vecteur d'observation sachant un état
 - Coût de transition :
 - probabilité de transition

Unités acoustiques

- ✓ Phonèmes :
 - ✓ Environ 50 unités
- ✓ Diphones, triphones
 - ✓ Modèles contextuels
 - ✓ Problème de complexité
 - ✓ D'apprentissage
 - ✓ Partage de paramètres : états, gaussiennes partagés entre plusieurs modèles

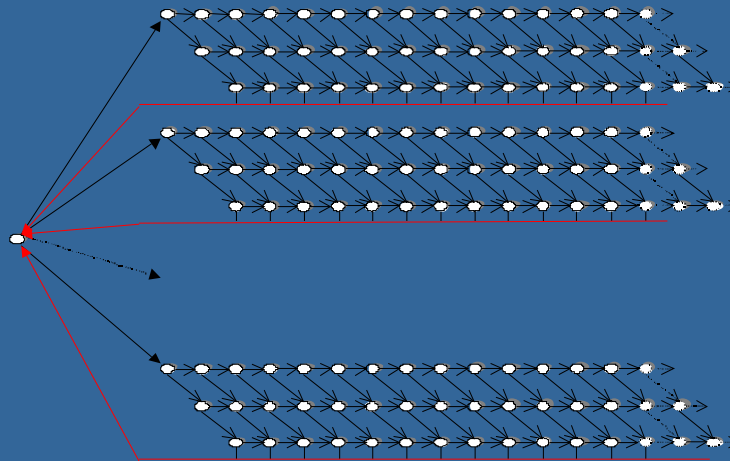
Décodage Acoustico Phonétique

- ✓ Principe : du signal à la suite phonétique



Décodeur mots isolés avec des HMMs

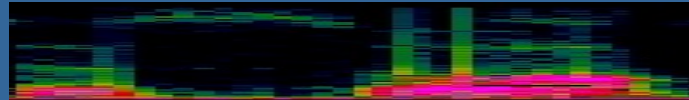
- ✓ Codage des mots :
 - ✓ un modèle par mot
 - ✓ mots isolés, petit vocabulaire



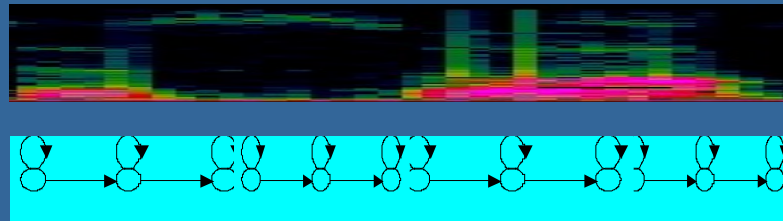
Apprentissage des HMM

- ✓ Estimation des GMM
 - ✓ Principe :
 - ✓ Apprentissage supervisé
 - ✓ Basé sur un corpus d'apprentissage
 - ✓ Corpus : ensemble de données audio étiquetées

Apprentissage des HMM



Salut tout le monde



Salut tout le monde

Apprentissage des HMM

- ✓ Estimation des GMM : processus itératif segmentation/estimation
 - ✓ Segmentation :
 - ✓ DAP contraint par la séquence de mots solutions
 - ✓ DAP semi-contraint par les variantes phonétiques possibles
 - ✓ Chaque état doit être associé à l'ensemble des trames qu'il émet

Apprentissage des HMM

- ✓ Estimation
 - ✓ Ajustement des paramètres maximisant la vraisemblance
 - ✓ Algorithme de Baum-Welch ou approximation EM

Apprentissage des HMM

- ✓ Difficultés :
 - ✓ Quantité de données (diphones, triphones, etc)
 - ✓ Liée à la complexité des modèles
 - ✓ Précision des modèles
 - ✓ Représentativité des données
 - ✓ Conditions apprentissage/test

Apprentissage des HMM : EM

- ✓ Expectation-Maximisation
 - ✓ algo itératif
 - étape 1 : estimation des paramètres
 - étape 2 : maximisation de la vraisemblance

Apprentissage des HMM : EM

✓ Formules de mises à jour :

- Vraisemblances : $l_i(x) = w_i * N_i$

- Probabilités : $P_i(X) = \frac{l_i(X)}{\sum l_i(X)}$

- Ré-estimation des poids :

$$w'_i = \frac{\sum_X P_i(X)}{\sum_X \sum_i P_i(X)}$$

Apprentissage des HMM : EM

- ✓ Formules de mises à jour :
 - Ré-estimation des moyennes :

$$m'_i = \frac{\sum_X P_i(X) \cdot X}{\sum_X \sum_i P_i(X)}$$

Apprentissage des HMM : EM

- ✓ Formules de mises à jour :
 - Ré-estimation des variances :

$$v'_i = \frac{1.0}{\sum_X P_i(X) - 1} \sum_X P_i(X) * (X - m'_i)^2$$

Apprentissage des HMM : EM

- ✓ Itérations EM : ré-estimation des paramètres jusqu'à convergence
 - typiquement, 8-16 itération
- ✓ Initialisation :
 - K-Means : clustering des trames
 - split de gaussiennes : 1 gaussienne, divisée et perturbée

Apprentissage des HMM : EM

- ✓ Processus itératif alignement-estimation :
 - pour un modèle donné:
 - alignement du corpus
 - » indexation des trames émises par un état donné
 - maximisation de la vraisemblance :
 - » EM, itérations



Partage des paramètres

- ✓ Objectif :
 - ✓ Réduire la complexité des modèles
 - ✓ Améliorer leur qualité en augmentant la quantité de données d'apprentissage
 - ✓ Augmenter le nombre de triphones
- ✓ Principe :
 - ✓ Partage d'états
 - ✓ Partage de Pdf

Partage des paramètres

- ✓ Partage des états :
 - ✓ Classification hiérarchique des états
 - ✓ À posteriori :
 - ✓ Regroupement d'états proches
 - ✓ A priori :
 - ✓ arbre de classification
 - ✓ Questions linguistiques
 - ✓ Mixte : génération automatique des questions

Partage des paramètres

- Partage des états par arbre de classification :
 - ✓ choix à priori d'un jeu de questions
 - question relative à la nature et au contexte du triphone
 - ex: liquide à droite, fricative, voisée,
« aa ii uu »
 - ✓ Bootstrap : un modèle acoustique non-contextuel (M_0)
 - *topologie fixée à priori*
 - ✓ 1 arbre à estimer pour chaque état de M_0

Partage des paramètres

- Construction de l'arbre :
 - à chaque noeud, évaluation de chaque question disponible :
 - 1 question bi-partitionne le corpus
 - estimation d'un modèle pour chaque partie
 - calcul du gain de vraisemblance
 - stop : trop peu de données sur les feuilles, faible gain en vraisemblance, plus de questions !

Partage des paramètres

- Choix des triphones pour une ensemble d'arbre donné :
 - pour chaque triphone, recherche de la séquence d'état
 - regroupement des triphones partageant la mme séquence d'états
 - construction d'une table des correspondance

Partage des paramètres

- Méthode ascendante :
 - ✓ sélection à priori des triphones
 - à partir de leur fréquence
 - ✓ estimation de “petits” modèles pour chacun d'eux (typiquement 4 gaussiennes/état)
 - ✓ regroupement des GMM les plus proches

Adaptation des modèles acoustiques

- ✓ Motivation
 - ✓ Adapter les HMMs au contexte de la tâche
 - ✓ Locuteur, bruit, canal, ...
 - ✓ En disposant de *relativement* peu de données
- ✓ Principe
 - ✓ maximiser la vraisemblance
 - ✓ ... en préservant la généricité des modèles

Adaptation des modèles acoustiques

- ✓ supervisée : suppose qu'on dispose d'un corpus d'adaptation annoté
- ✓ non-supervisée :
 - pas de corpus fourni
 - une première passe de décodage : transcription imparfaite, utilisée comme cible
 - l'ensemble de l'information impliquée dans le décodage est utilisée pour l'adaptation acoustique

Adaptation des modèles acoustiques

✓ MAP

- ✓ maximisation de la vraisemblance a posteriori

$$\lambda_{map} = \operatorname{argmax}_{\lambda} P(\lambda/O) = \operatorname{argmax}_{\lambda} P(O/\lambda) * P(\lambda)$$

- Problème : quantité de données

Adaptation des modèles acoustiques

- ✓ MLLR :
 - ✓ méthode généralement utilisée en non-supervisée
 - ✓ régression linéaire maximisant la vraisemblance
 - transformation globale sur des classes de modèles
 - ex: translation de tous les modèles, par phone, etc..
- ✓ TRÈS SOUVENT, EN SECONDE PASSE

Lexiques

- ✓ Ensemble des mots reconnus par le système
- ✓ En RAP moyen et grand vocabulaire :
 - ✓ Un mot est représenté par un ensemble de séquences phonétiques
 - ✓ Chaque séquence correspond à une séquence de HMM

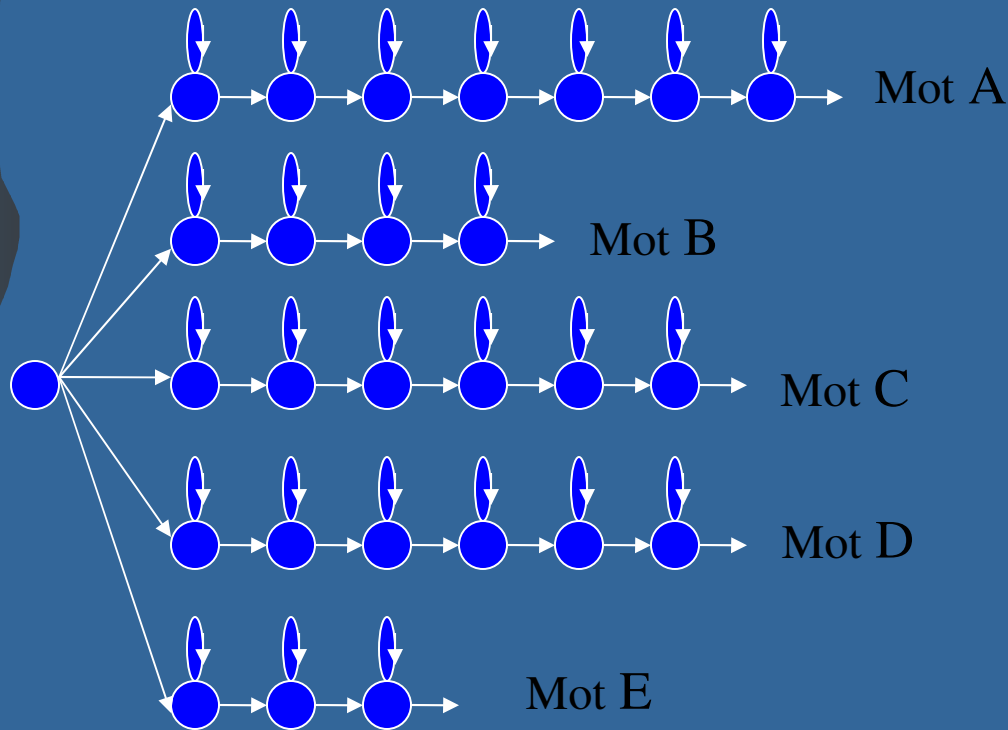
Lexiques

- ✓ Problème : phonétisation
 - ✓ Multiple :
 - ✓ Liaisons (*Amis/Les amis*)
 - ✓ Accents (*pain (pp in | pp un)*)
 - ✓ Langues (*jakson, karayan, poggio*)
 - ✓ Automatisation difficile
 - ✓ Acronymes, etc

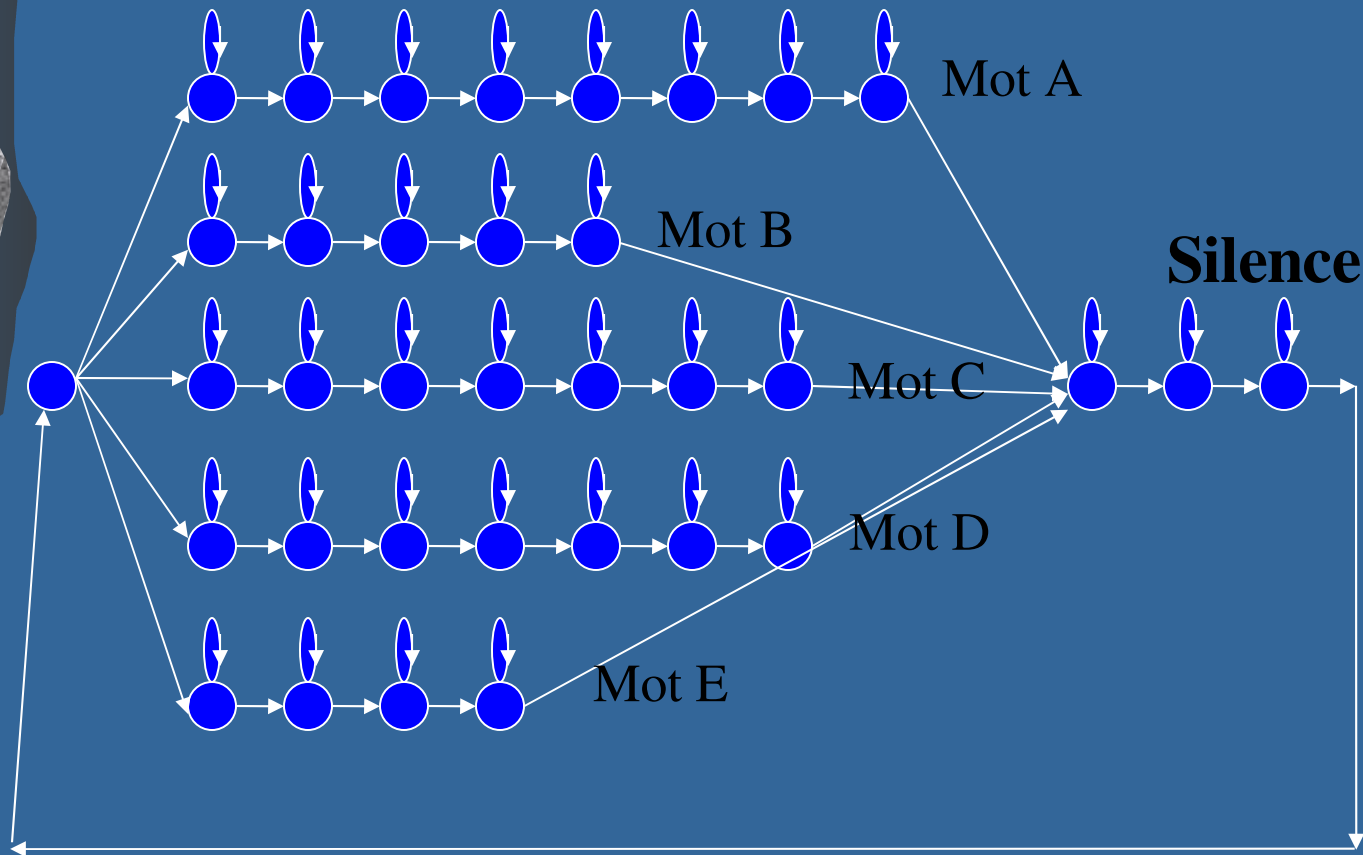
Lexiques

- ✓ La complexité du décodage dépend de la taille du lexique
- ✓ Tailles classiques dans les campagnes d'évaluation : 5k-65k mots
- ✓ Mots hors vocabulaire (vocabulaire ouvert/fermé)
- ✓ Codage en graphe de HMMs

Représentation des mots du lexique

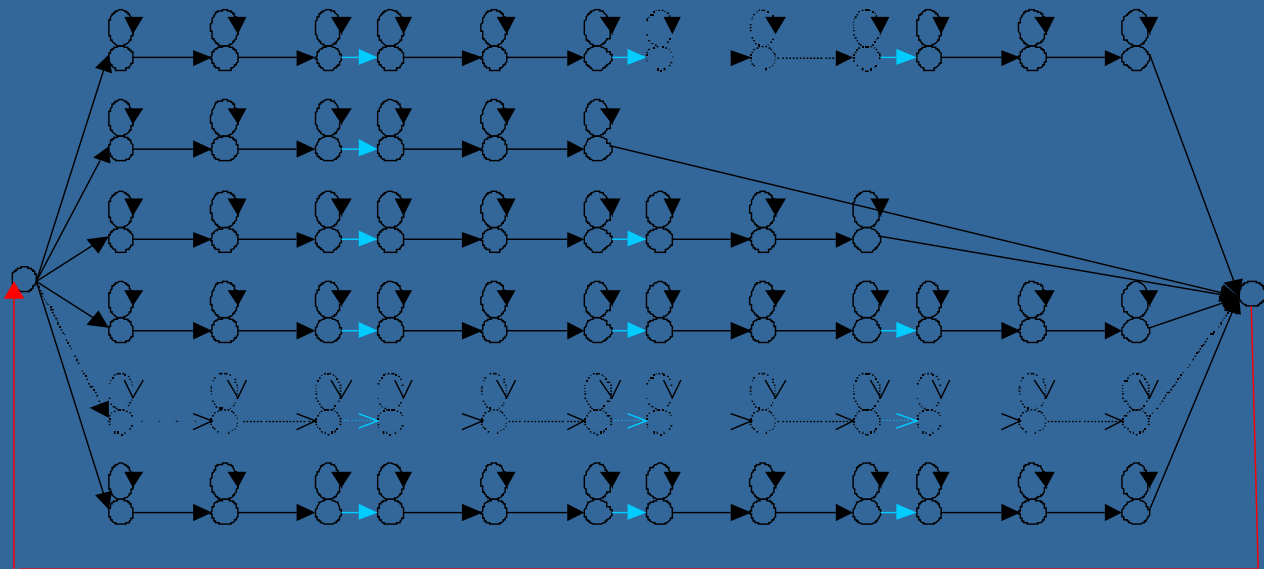


Représentation Globale (mots isolés, enchaînés)

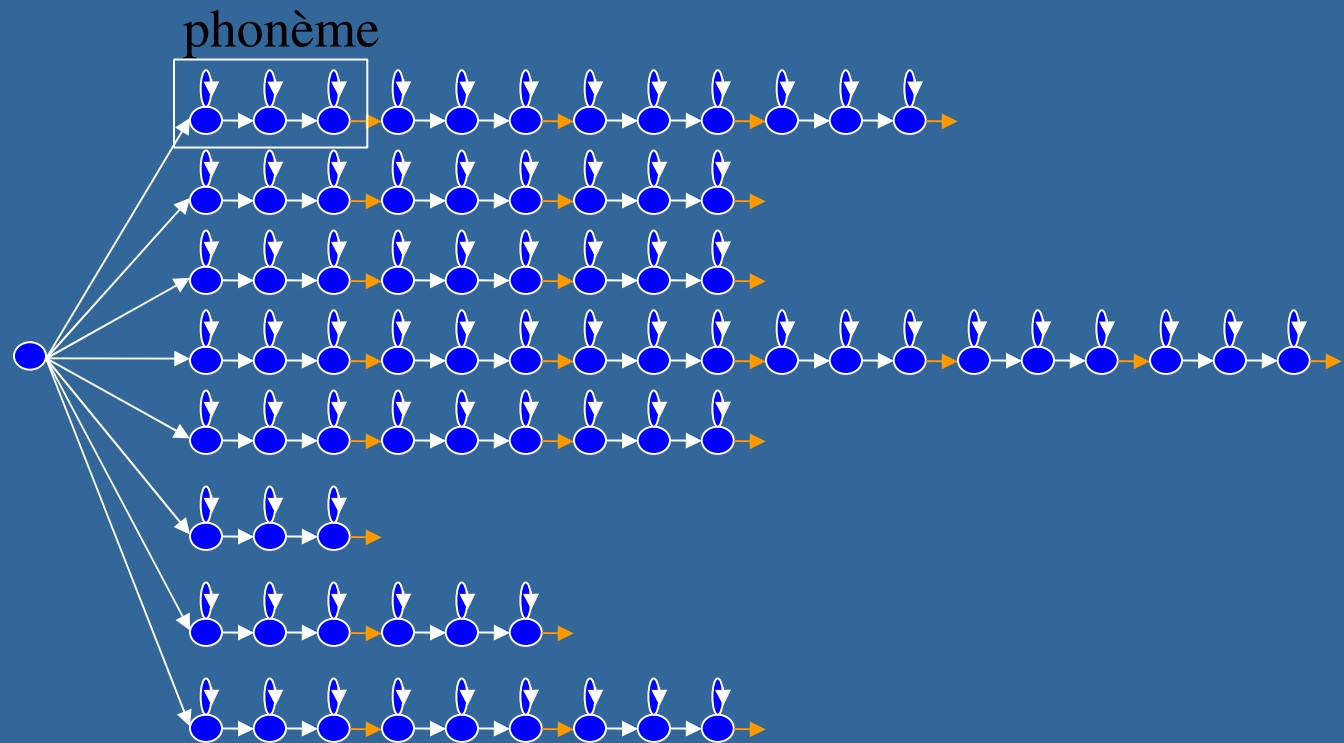


Autres représentations (grand vocabulaire)

- Un mot est codé comme une suite de machines de Markov



Représentation phonémique du lexique



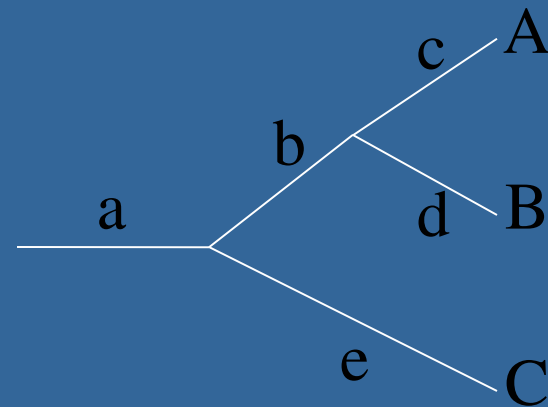
Représentation en Arbre du graphe

- Construire à partir des listes de phonèmes un arbre a tête commune de tous les mots :

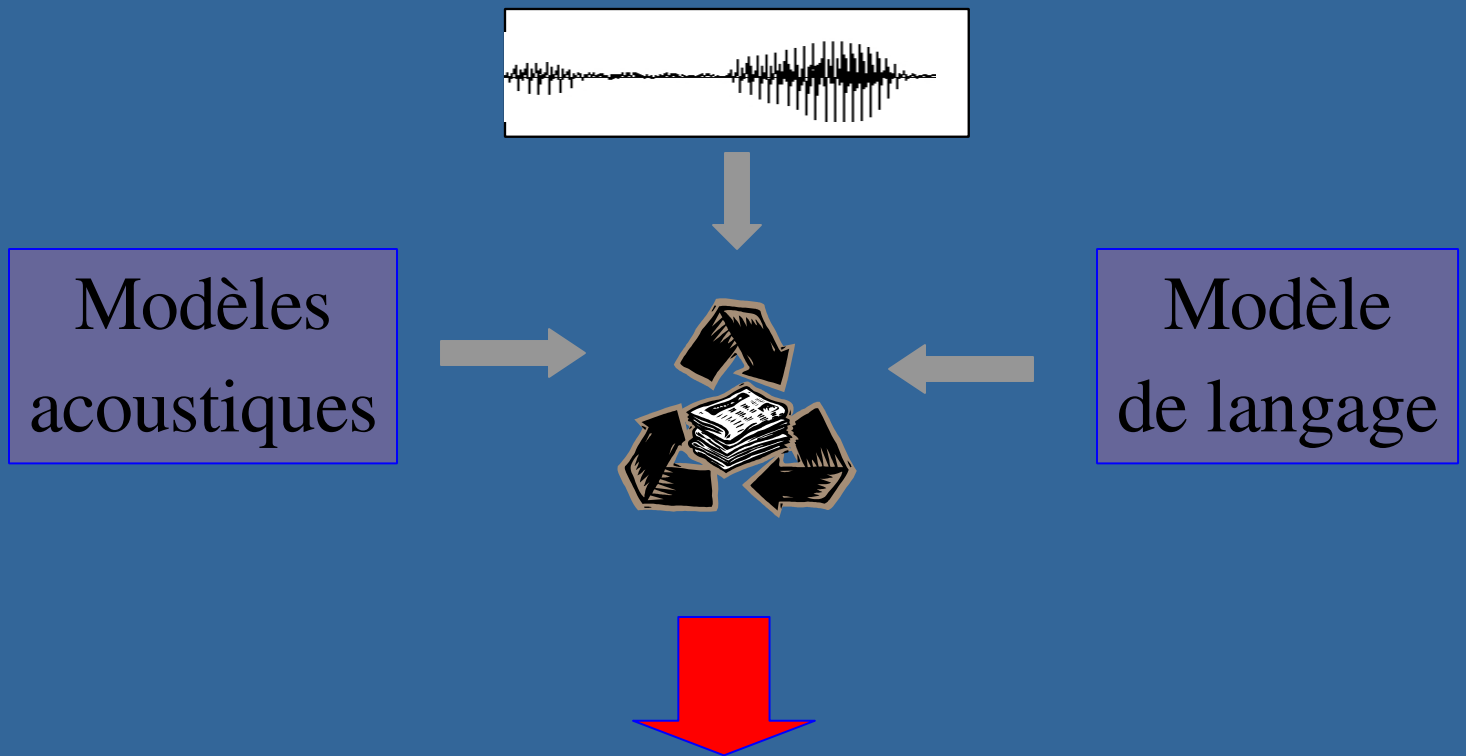
- Exemple :

— lexique

- $A = a \ b \ c$
- $B = a \ b \ d$
- $C = a \ e$



Le décodage



« *Salut tout le monde* »

Modèles de Langages

Des modèles statistiques

$$P(M_1^n) = \prod_{i=1}^n P(m_i / h)$$

Bi-grammes

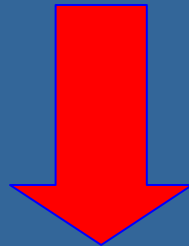
$$P(m_i / m_1 \cdots m_{i-1}) \approx P(m_i / m_{i-1})$$

Tri-grammes

$$P(m_i / m_1 \cdots m_{i-1}) \approx P(m_i / m_{i-2}, m_{i-1})$$

Modèles de langage

- ✓ Problèmes :
 - ✓ Estimation
 - ✓ Bigrammes/trigrammes non-observés
 - ✓ Technique du repli
 - ✓ Représentativité des corpus :
 - ✓ Couverture / précision
 - ✓ Acquisition des corpus



- ✓ Adaptation/sélection des modèles

Modèles de langage

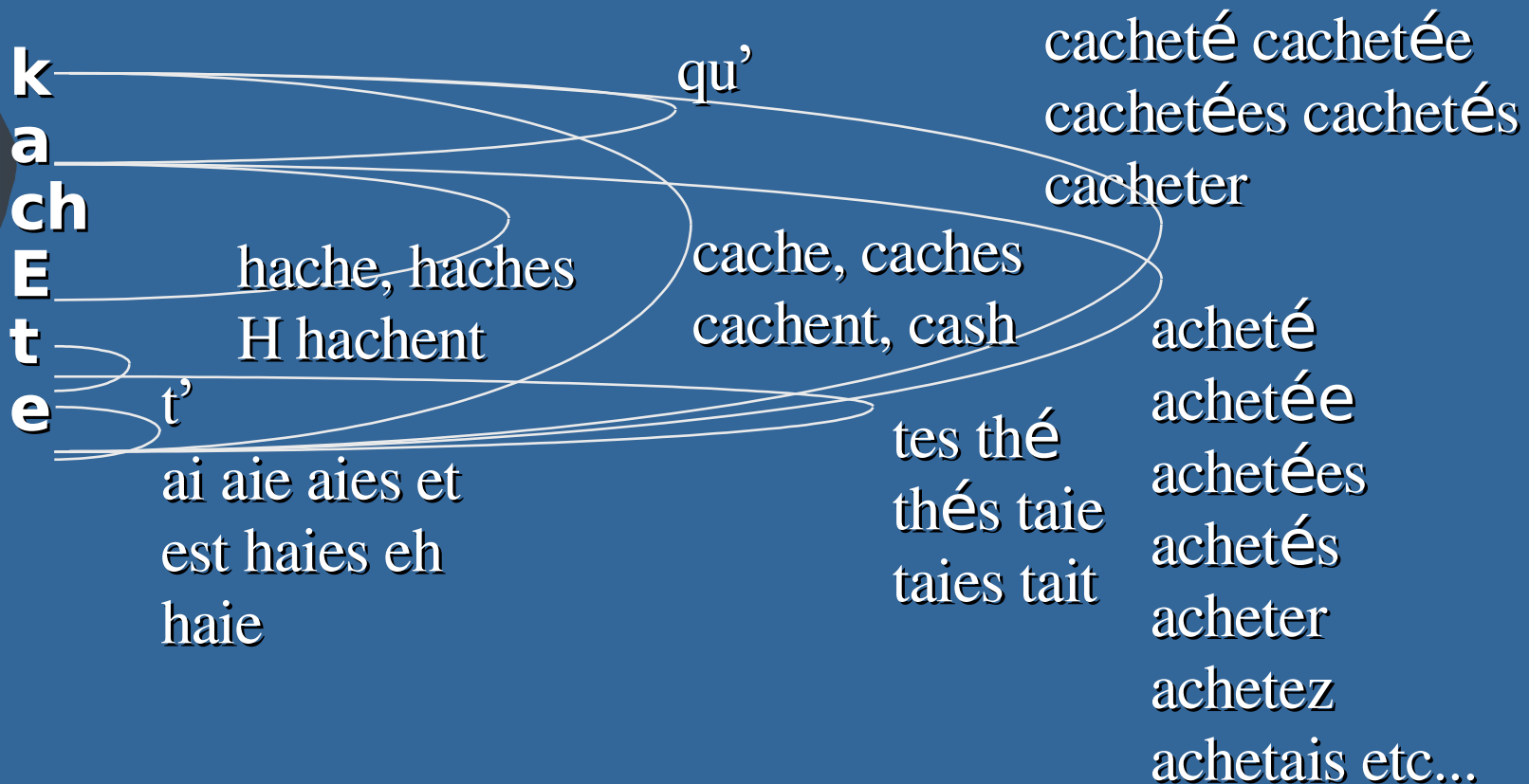
- ✓ Modèles à syntaxe fixe
 - ✓ Ensemble de règles exprimées en CFG (*Context free grammar*)
 - ✓ Chemin dans le graphe des règles : phrase du langage
 - ✓ Problèmes :
 - ✓ impossibilité de représenter toutes les séquences de mots du langage
 - ✓ Cadre non-probabiliste (pas d'estimation de la vraisemblance d'une séquence)

Le décodage

- Difficultés :
 - Exploitation/combinaisons des sources d'information
 - Exploration de l'espace des hypothèses
 - Construction du graphe
 - Stockage
 - **Complexité du parcours**

Le décodage

Difficultés : pas de frontières entre les mots

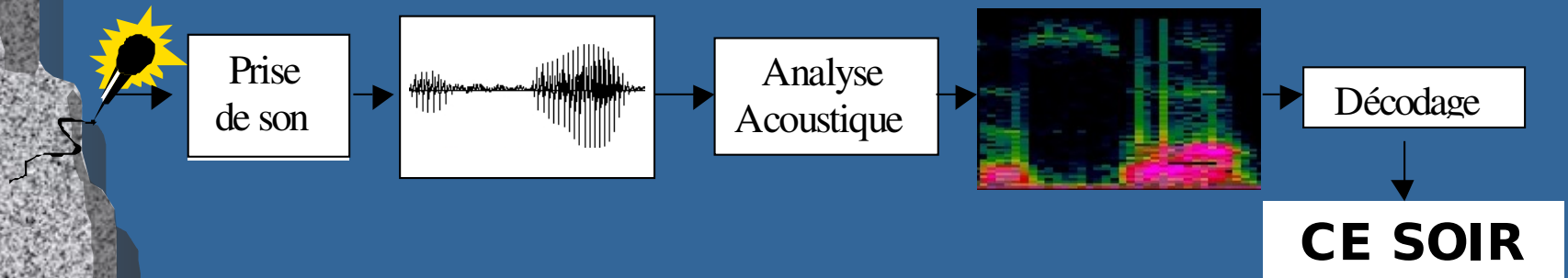


Le décodage

Approche Statistique

$$P(W|X) = P(X|W).P(W)/\sum P(X)$$

- ✓ $P(X/w)$: probabilité d'émission de l'observation X pour une phrase donnée w
- ✓ $P(w)$: probabilité à priori d'une séquence de mots w



Le décodage

Approche Statistique

Facteur d'échelle

$$W = \underset{w}{\operatorname{argmax}} P(X|W).P(W)$$

Proba acoustique

Proba linguistique



Algorithmes de décodage

- 2 catégories :
 - *stack decoder* : décodeur à piles
 - Pile d'hypothèses intermédiaires
 - Programmation dynamique
 - Parcours de treillis

Les décodeurs à pile

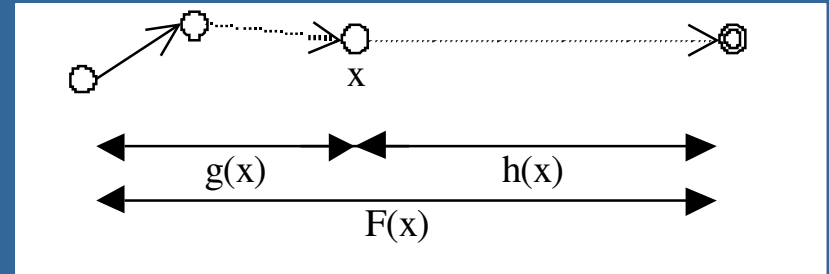
- Principe
 - Le système maintient une pile d'hypothèses partielles
 - Les hypothèses sont ordonnées par leurs vraisemblances
 - Les meilleures hypothèses sont prolongées, et insérées dans le pile

L'algorithme A*

- Décodeur à pile
- Algorithme classique de recherche du meilleur chemin dans un graphe
- Algorithme asynchrone
- L'ordre d'exploration des noeuds est déterminé par la fonction $F(n)$ qui représente une estimation du coût du meilleur chemin passant par n .

L'algorithme A*

$$F(x) = g(x) + h(x)$$



- $g(n)$: coût du chemin allant du début du graphe à n,
- $h(n)$: l'estimation du coût du chemin allant de n à la fin du graphe,

Coût du chemin parcouru : coût acoustique + coût linguistique



L'algorithme *Two-level*

- Algorithme d'alignement dynamique
- Principe :
 - Pour chaque segment du message, évaluation des meilleurs formes de référence (niveau 1)
 - Alignement optimal des solutions issues du niveau 1 (niveau 2)



L'algorithme *Level-Building*

- Optimisation du two-level:
 - Seules les hypothèses prolongeant des hypothèses partielles « acceptables » sont évaluées



L'algorithme *beam-search*

- One-pass :
 - Algorithme synchrone
 - Toutes les hypothèses sont évaluées en parallèle
- Amélioration beam-search (recherche en faisceau)
 - Les hypothèses partielles trop peu vraisemblables sont écartées
 - Algorithme très utilisé dans les systèmes actuels



L'algorithme *token-passing* (*passage de jeton*)

- Formulation unifiée de l'algorithme de Viterbi
- Principe :
 - Des jetons valués sont propagés dans le graphe d'hypothèses
 - Pour chaque nœud, le meilleur jeton est attribué au nœud



Décodages multipasses

- Motivation :
 - Réduction de la complexité
 - 1 passe : 1 filtrage de l'ensemble des hypothèses
- Adaptation non supervisées :
 - Un premier décodage est utilisé pour l'adaptation du système

Mise en oeuvre d'un SRAP

- Toolkits public ou semis-publics :
- HTK (université de Cambridge)
- ISIP : Université du Mississippi
- Sphinx : CMU
- LIA : SPEERAL (Toolkit GPL, moteur libre pour enseignement/recherche)

Mise en oeuvre d'un SRAP

- Difficultés :
 - identifier la tâche :
 - les contraintes
 - Les performances attendues
 - Réglage « fin » du système
 - Paramètres acoustiques
 - Modèles de langage
 - Vitesse de décodage



Performances

- Exprimées en terme de Taux d'erreur Mot (WER)
- Broadcast News : 10-25%
- Cours/conférences : 25-40%
- Réunions, conversations : 60%

Perspectives

- Paramétrisations (LDA, VTLN)
- Modélisation : Apprentissage discriminant
- Linguistique : mises à jours des lexiques, traitement des mots hors vocabulaire
- Décodage : combinaison de systèmes



Perspectives

- Combinaison de systèmes :
 - bas niveau : acoustique
 - cross-adaptation
 - combinaison des one-best/ réseaux de confusion
- Utilisabilité :
 - réduire les coûts d'apprentissage
 - corpus acoustiques
 - collecte des corpus texte, sélection du vocabulaire, etc.

Perspectives

- Nouvelles modélisations acoustiques
 - les HMM reposent sur des hypothèses fortes – et fausses
 - combinaison de systèmes pour le décodage
 - multi-linguisme
 - interfaces TALN