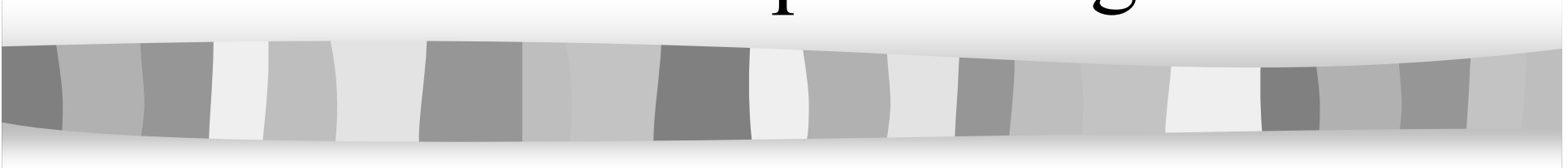


# 1. Parole et technologies vocales (6h)

- **Introduction au traitement numérique du signal**
- **Du signal de parole aux technologies vocales**
- **Modélisation stochastique d'objets sonores (parole et sons)**

# Introduction au traitement numérique du signal





# Bibliographie

- Cours ENSIMAG de Jim Crowley
  - <http://www-prima.imag.fr/Prima/Homepages/jlc/Courses/Courses.html>
- *DSP First, A Multimedia Approach*, J.H. McClellan, R. W. Schafer, M.A. Yoder
- *Traitement numérique des signaux*, M. Kunt, Presses Polytechniques Romandes
- *Théorie et traitement des signaux*, F. De Coulon, Editions de l'Ecole Polytechnique Fédérale de Lausanne



# Pourquoi du Traitement du Signal ?

- Aspect Multimédia, savoir manipuler / traiter :
  - sons
  - images
  - vidéos
- Aspect Réseau, savoir transmettre :
  - de l'information audio, vidéo ou autre
  - nécessité du codage / décodage de l'information
- Ex. :
  - transmission de parole via Internet (VoIP)
  - transmission de données via le réseau GSM



# Quelques définitions

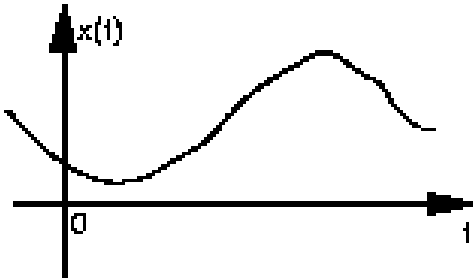
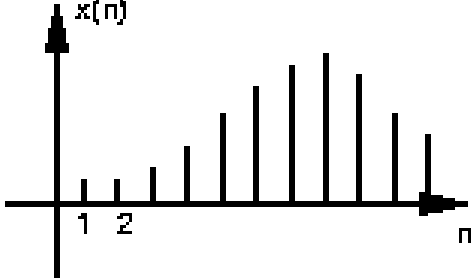
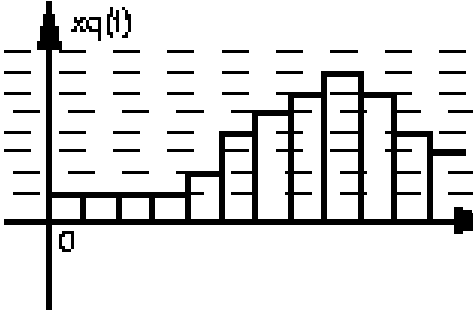
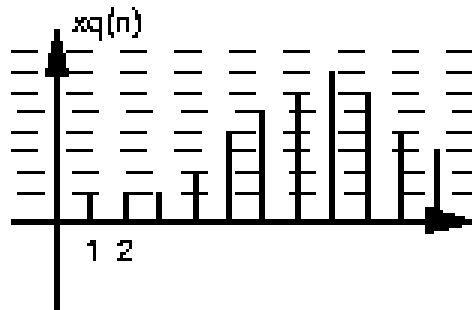
- Signal : représentation physique de l'information ;
- Bruit : tout phénomène perturbateur gênant l'interprétation d'un signal
- Traitement du signal : théorie permettant d'effectuer une description (une modélisation) et une analyse des signaux et des systèmes porteurs d'information



# Exemple de Signaux

- D'un point de vue analytique, le signal est une fonction d'une variable réelle (en général le temps  $t$ )
- Signal audio :  $x(t)$
- Signal video N&B :  $p(x,y,t)$
- Signal video couleur :  $p(c,x,y,t)$ 
  - composante couleur  $c=\{r,v,b\}$

# Représentation analogique et numérique

	TEMPS CONTINU	TEMPS DISCRET
AMPLITUDE CONTINUE	 <p>A graph showing a continuous signal <math>x(t)</math> plotted against time <math>t</math>. The vertical axis is labeled <math>x(t)</math> and the horizontal axis is labeled <math>t</math>. The origin is marked with 0. The signal is a smooth, continuous curve that starts at a low value, rises to a peak, and then falls.</p>	 <p>A graph showing a discrete signal <math>x(n)</math> plotted against time <math>n</math>. The vertical axis is labeled <math>x(n)</math> and the horizontal axis is labeled <math>n</math>. The origin is marked with 0. The signal consists of vertical bars at discrete time intervals, with the first two bars labeled 1 and 2. The bars vary in height, forming a discrete approximation of the continuous signal.</p>
AMPLITUDE DISCRETE	 <p>A graph showing a discrete signal <math>x_q(t)</math> plotted against time <math>t</math>. The vertical axis is labeled <math>x_q(t)</math> and the horizontal axis is labeled <math>t</math>. The origin is marked with 0. The signal consists of horizontal bars at discrete time intervals, with the first two bars labeled 1 and 2. The bars vary in height, forming a discrete approximation of the continuous signal.</p>	 <p>A graph showing a discrete signal <math>x_q(n)</math> plotted against time <math>n</math>. The vertical axis is labeled <math>x_q(n)</math> and the horizontal axis is labeled <math>n</math>. The origin is marked with 0. The signal consists of vertical bars at discrete time intervals, with the first two bars labeled 1 and 2. The bars vary in height, forming a discrete approximation of the continuous signal.</p>



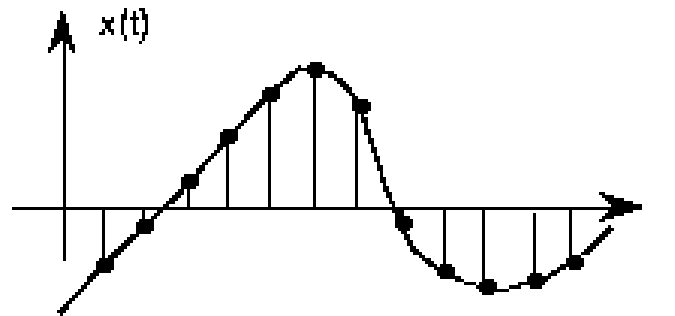
# Traitement numérique

- Avantages : les moyens informatiques actuels permettent le traitement des signaux sous forme numérique (traitements plus rapides)
- Problème : le passage en numérique conduit à une perte d'information
- Nécessité :
  - Convertisseurs analogique-numérique (A/N)
  - Convertisseurs numérique-analogique (N/A)

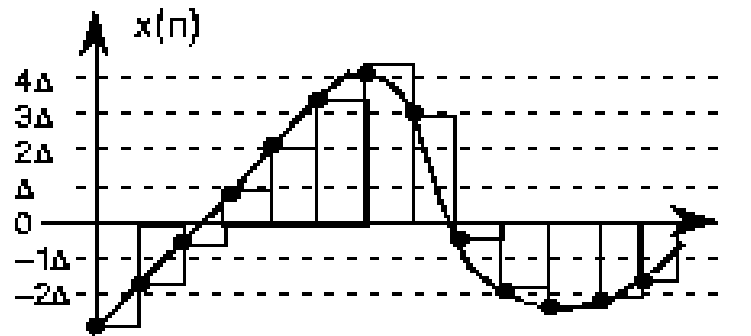
# Conversion analogique-numérique

- 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  $x(n)$  sur l 'intervalle  $[n_1, n_2]$

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

# Autres grandeurs

**Puissance moyenne  
sur [t1,t2]**

$$P_x(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} x^2(t) dt$$

**Energie totale**

$$W_x = \int_{-\infty}^{+\infty} s^2(t) dt$$

**Puissance moyenne  
totale**

$$P_x = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t) dt$$

# Rapport signal/bruit

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

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

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



# Rapport signal/bruit

- On peut aussi faire une approximation du RSB en estimant un rapport entre l'amplitude du signal et l'amplitude du bruit

$$SNR_{dB} = 20 \log_{10} \frac{\textit{Amplitude}_s}{\textit{Amplitude}_n}$$

# Rapport signal/bruit

- Si  $E(\text{signal})=E(\text{bruit})$   $\longrightarrow$  SNR=0dB
- Si  $E(\text{signal})=2E(\text{bruit})$   $\longrightarrow$  SNR=3dB
- Si  $E(\text{signal})=10E(\text{bruit})$   $\longrightarrow$  SNR=10dB
- Si  $E(\text{signal})=100E(\text{bruit})$   $\longrightarrow$  SNR=20dB
- Si  $E(\text{signal})=1000E(\text{bruit})$   $\longrightarrow$  SNR=30dB

$$E(\text{signal}) = 10^N E(\text{bruit}) \Rightarrow \text{SNR} = N * 10\text{dB}$$

# Exemples de signaux bruités

RSB=40dB (Signal propre) 

RSB=26dB 

RSB=14dB 

RSB=8dB 

RSB=0dB 



# Transformée de Fourier - TF

- Instrument de base de la théorie du signal
- Représentation spectrale des signaux
- Exprime la répartition en fréquence de l'amplitude et de la phase de l'énergie d'un signal

# Rappels

$$e^{j\omega t} = \cos(\omega t) + j \sin(\omega t)$$

$$e = \lim_{n \rightarrow \infty} \left\{ \left( 1 + \frac{1}{n} \right)^n \right\} = 2.7182818284$$

$$j = \sqrt{-1}$$

$$e^{j\omega t} + e^{-j\omega t} = 2 \cos(\omega t)$$

$$e^{j\omega t} - e^{-j\omega t} = 2j \sin(\omega t)$$

# TF d'un signal continu

- Soit  $x(t)$  un signal complexe
- La TF est une fonction complexe de la variable réelle  $\omega = 2\pi f$  définie par :

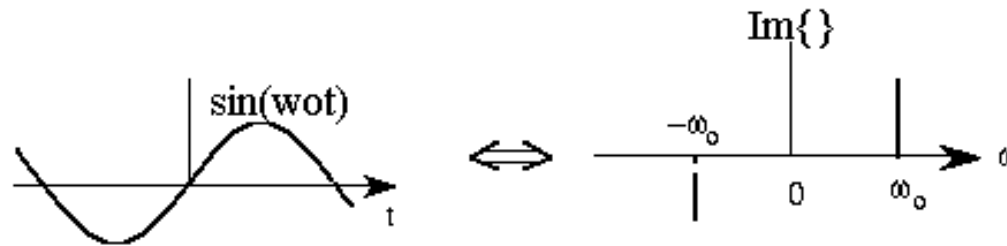
$$F\{x(t)\} = X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t} dt$$

- La transformée inverse est donnée par :

$$x(t) = F^{-1}\{X(\omega)\} = \int_{-\infty}^{+\infty} X(\omega)e^{+j\omega t} d\omega$$

# TF du Sinus

$$\sin(\omega_0 t) \stackrel{TF}{\leftrightarrow} \frac{1}{2j} [\delta(\omega - \omega_0) - \delta(\omega + \omega_0)]$$



La transformée d'un sinus de fréquence  $\omega_0$  est une somme de 2 impulsions dans la partie imaginaire

# Représentation temps-fréquence

- Une partition musicale est un exemple de représentation temps-fréquence



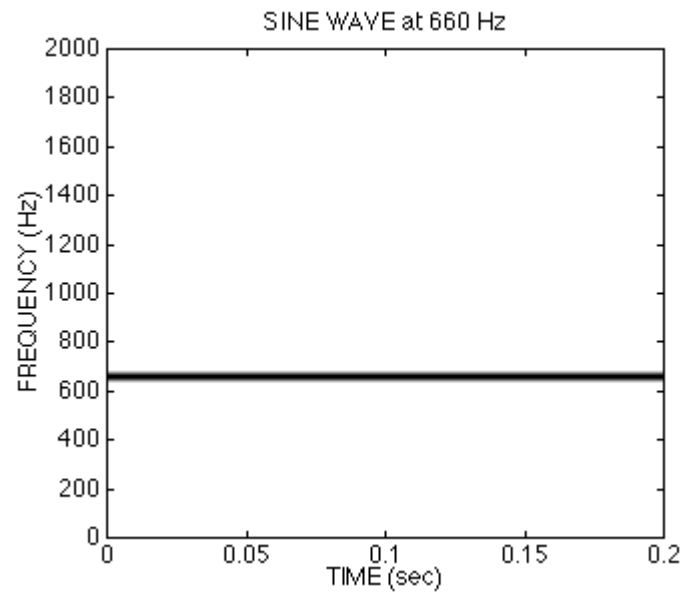
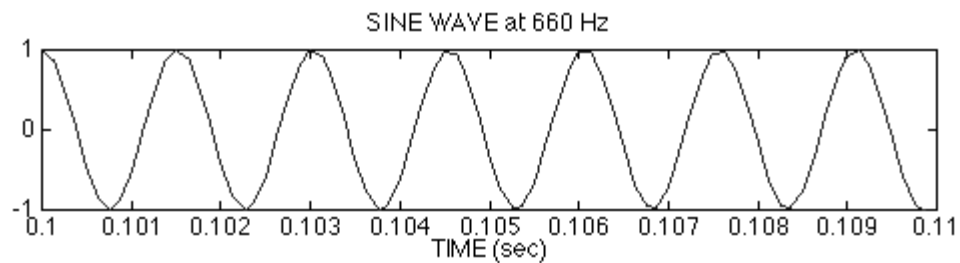


# Spectrogramme temps-fréquence

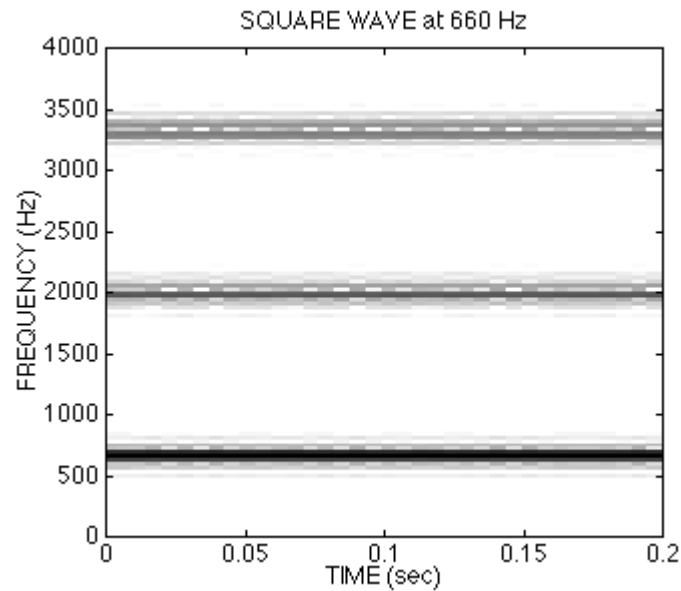
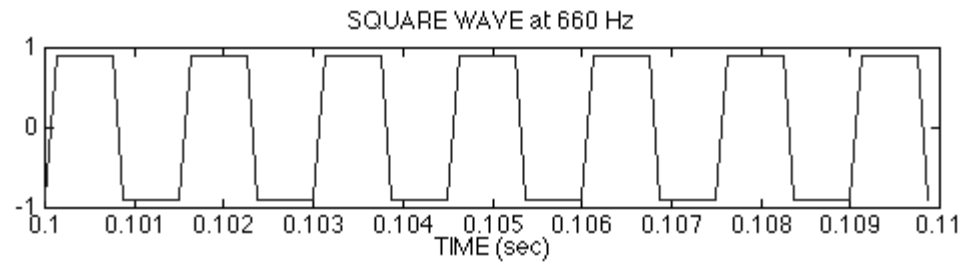
$$S_x(t, f) = \left| \int_{-\infty}^{+\infty} x(s) h^*(s - t) e^{-i2\pi fs} ds \right|^2$$

Représentation du module du spectre du signal (donc de l'énergie) sur une fenêtre glissante temporelle  $h$

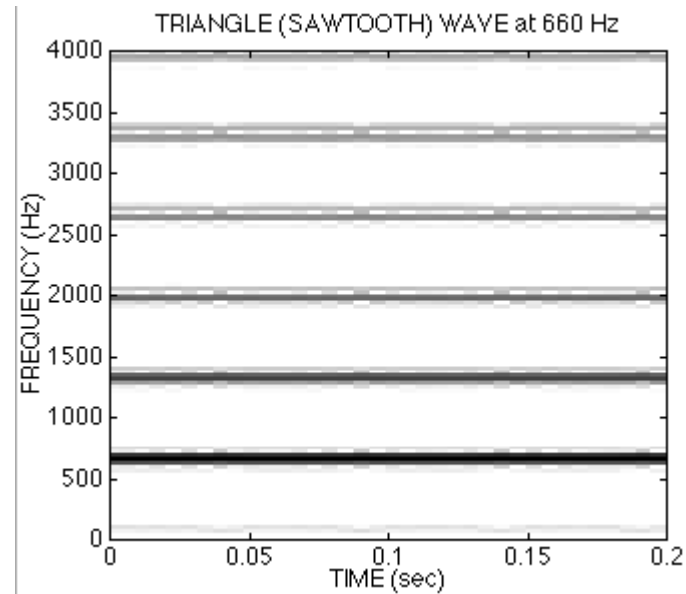
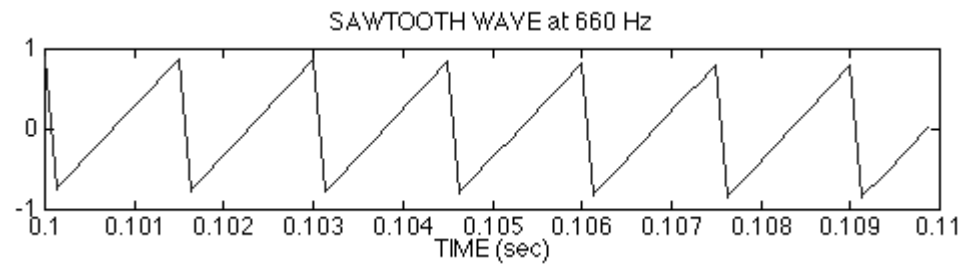
# Cas d'une sinusoïde



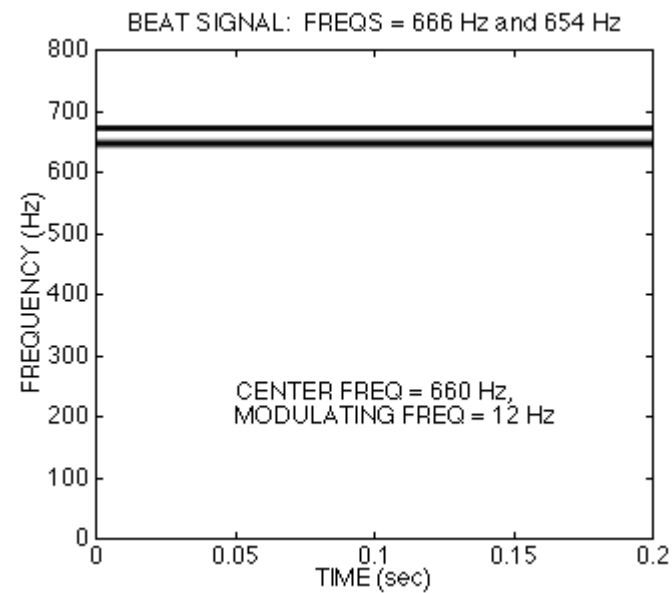
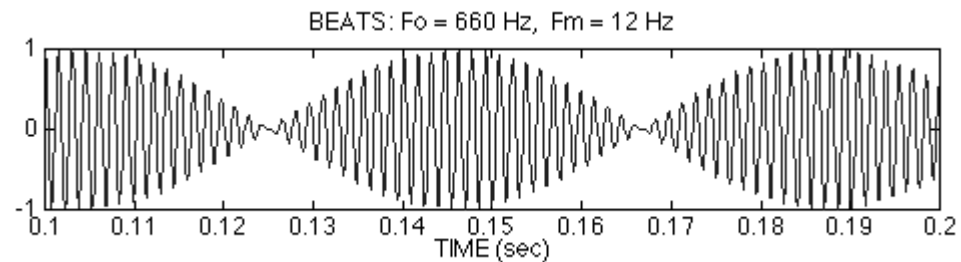
# Cas de signaux en créneaux



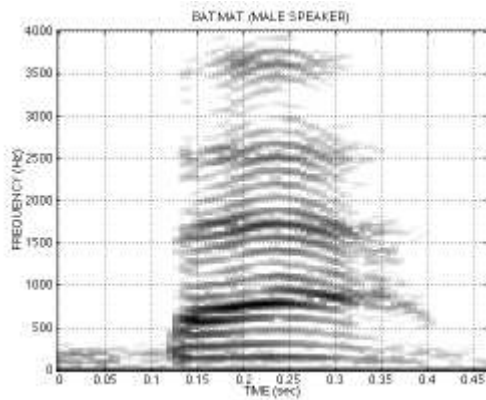
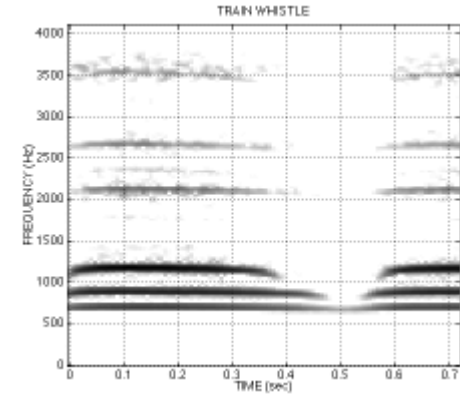
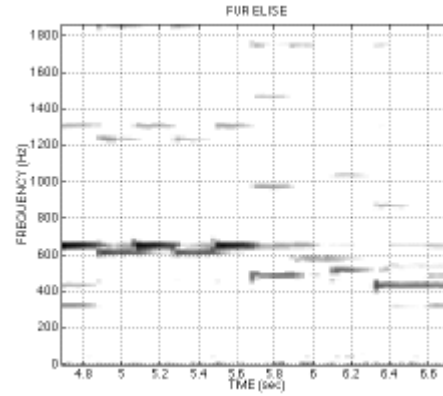
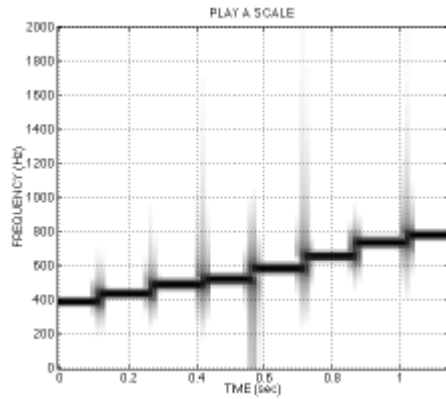
# Cas de signaux en dent de scie



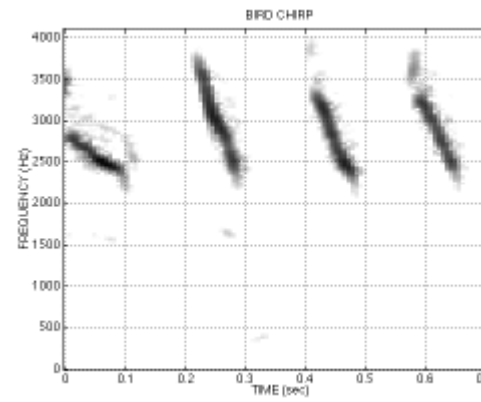
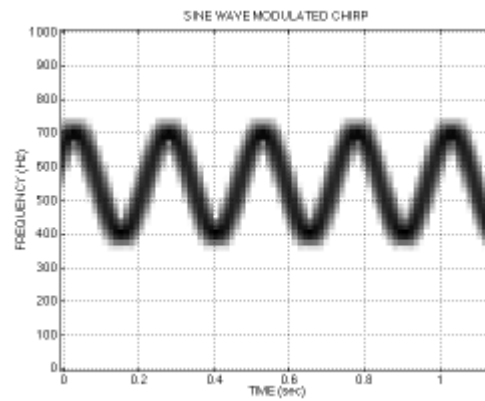
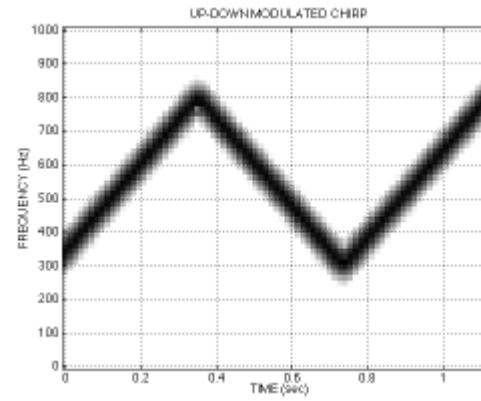
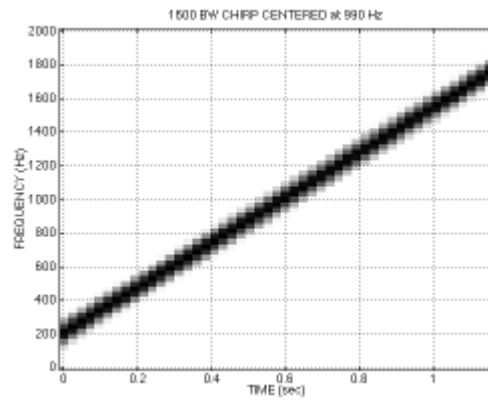
# Cas de signaux de battements



# Autres exemples

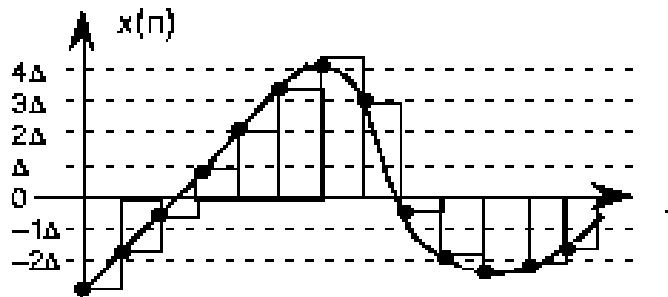


# Chirps



# Représentation numérique des échantillons

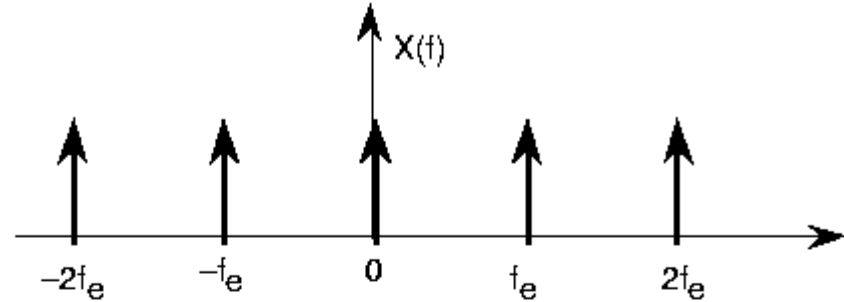
- Quantification et codage



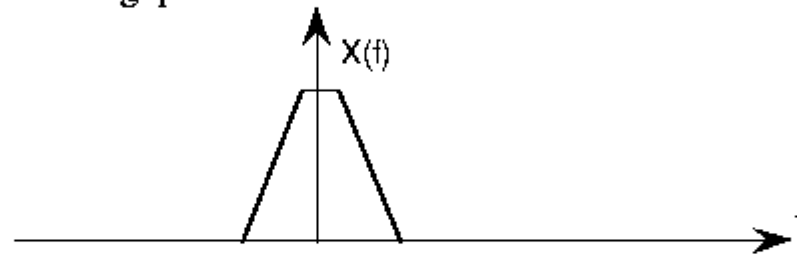
- Echantillonnage + Quantification = Conversion Analogique/Numérique A/N (*Analog to Digital A/D*)
- Distorsion (perte d'information) systématique introduite par la conversion A/N

# TF d'un signal échantillonné :

Spectre d'un échantillonneur idéal :

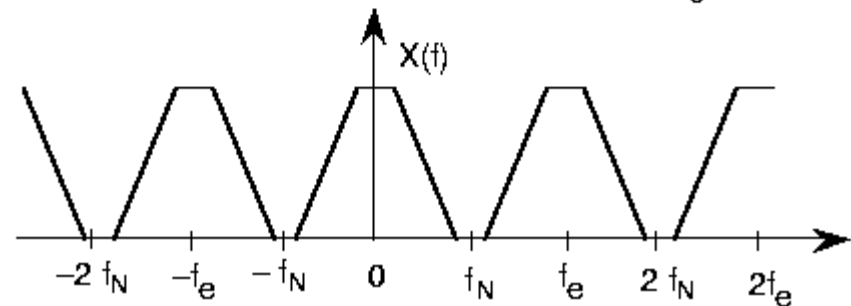


Spectre du signal analogique :



Spectre du signal après échantillonnage (idéalisé) :

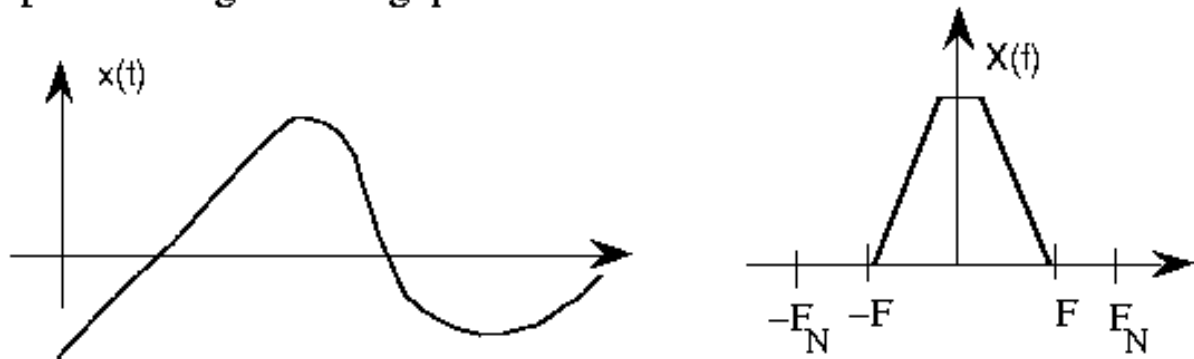
Replié autour du fréquence de "Nyquist".  $f_N = \frac{f_e}{2} = \frac{1}{2T_e}$



# Théorème de Shannon

- Un signal analogique  $x(t)$  ayant une largeur de bande finie limitée à  $2F$  hz, ne peut être reconstitué exactement à partir de ses échantillons  $x(n\Delta t)$  que si ceux-ci ont été prélevés avec une période  $T_e = \frac{1}{f_e} \leq \frac{1}{2F}$

Spectre du signal analogique :



# Théorème de Shannon

- Pour que la répétition périodique du spectre ne modifie pas le motif répété, il faut et suffit que la fréquence d'échantillonnage soit supérieure ou égale à 2 fois la fréquence maximum  $F$  du signal

$$F \leq f_N = \frac{f_e}{2}$$

# Exemples

- Signaux quantifiés sur 8 bits (256 valeurs possibles) et échantillonnés à 8kHz
    - ➔ débit binaire = 64kbit/s
  - Signaux quantifiés sur 16 bits (65536 valeurs possibles) et échantillonnés à 16kHz
    - ➔ débit binaire = 256kbit/s
- Nécessité de codage (MPEG, GSM, G723)  
des signaux pour la transmission sur un réseau




# Exemple de signaux codés (1)

- MPEG I, Layer II,

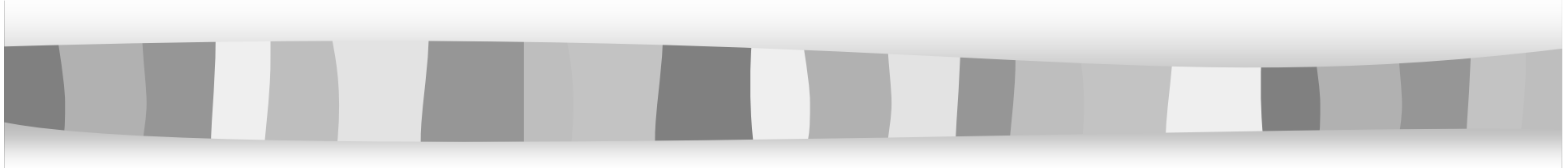
Bit rate	64 kb/s	32 kb/s	24 kb/s	16 kb/s	8 kb/s
----------	---------	---------	---------	---------	--------



# Exemple de signaux codés (2)

- GSM 
- G723 (vidéoconférence) 
- Même signal non codé 

# Du signal de parole aux technologies vocales...





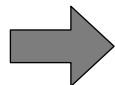
# La parole, une modalité parmi d'autres ?

- Du point de vue de l'utilisateur
  - En entrée
    - *Visuel*
    - **Auditif**
    - Olfactif
    - Gustatif
  - En sortie
    - **Oral**
    - Gestuel

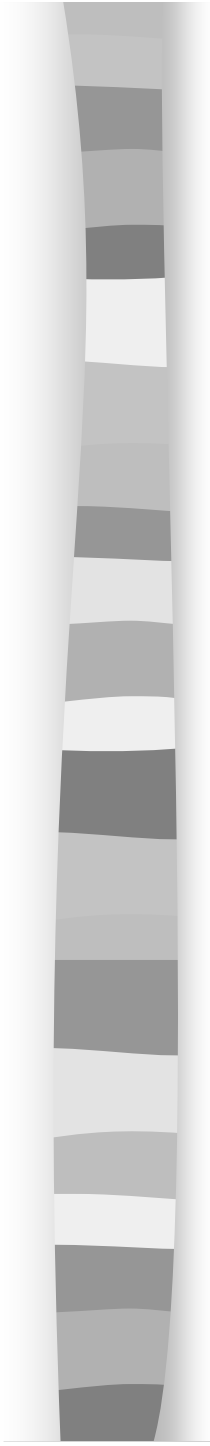
# Utilisation de la parole dans une interface : aspect ergonomique

## ■ Mode de communication

- naturel
- rapide et concis
  - l'opérateur n'a pas à réfléchir comment déclencher une série d'actions pour obtenir le résultat souhaité
- mains libres
  - les mains peuvent être utilisées pour des actes plus en lien avec la tâche
- permet au regard de rester fixé sur l'endroit de l'écran où se déroule l'action demandée
  - va-et-vient des yeux nettement diminué
- aide aux personnes handicapées



Confort et souplesse d'utilisation



# Contraintes engendrées par la parole (1)

- Situations exceptionnelles (d 'alarme)
  - dégradation des performances
- Apprentissage nécessaire de la machine
  - peut être long et contraignant (connaissez-vous IBM ViaVoice?)
- Contraintes linguistiques
  - taille du vocabulaire que le système pourra traiter/reconnaître
  - structure syntaxique des phrases (langage artificiel ? naturel ?)
  - mode d 'élocution de l 'opérateur (mots isolés,

# Contraintes engendrées par la parole (2)

## ■ Facteurs humains

- casque + micro => gêne de l'utilisateur
- peur de se voir rejeté par la machine
- confidentialité de l'utilisation de l'interface en public

– peur d'être ridicule en public

**réaction « normale » face à une technologie nouvelle**



# Buts à atteindre

- Satisfaction de l'utilisateur
- Qualité
  - temps de réponse du système
  - facilité et sûreté de fonctionnement
    - prévoir un mode de remplacement si le bruit devient important (robustesse & plasticité)
- Coûts des systèmes
  - l'utilisation de la parole doit avoir une influence limitée sur les coûts



# Ecueils à éviter

- danger du « tout vocal »
  - la parole n'est qu'une modalité supplémentaire
- la parole ne doit pas compliquer le système
  - attention à l'ergonomie de l'interface
- ne pas énerver l'utilisateur
  - robustesse nécessaire



# Les Technologies de la Parole dans les Services Télécom

- Compression de parole pour réseaux de télécommunication (*wireless, IP,...*)
- Synthèse, reconnaissance et compréhension pour les systèmes de dialogues (accès à l'information...)
- Authentification de l'utilisateur par la voix (reconnaissance du locuteur) pour accès confidentiel à l'information et sécurité



# Différents types de communication

- Communication Homme/Homme Médiatisée :
  - audio et video-conférence
  - apprentissage à distance
  - traduction assistée
- Communication Homme/Machine
  - recherche d'informations
  - e-commerce
  - interfaces à commande vocale



# Traitement de la Parole

- La parole se distingue des autres sons par ses caractéristiques acoustiques
- 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



# Phonèmes

- La fonction principale des sons dans une langue est d'établir des distinctions entre les unités de signification
- 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] dais	[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

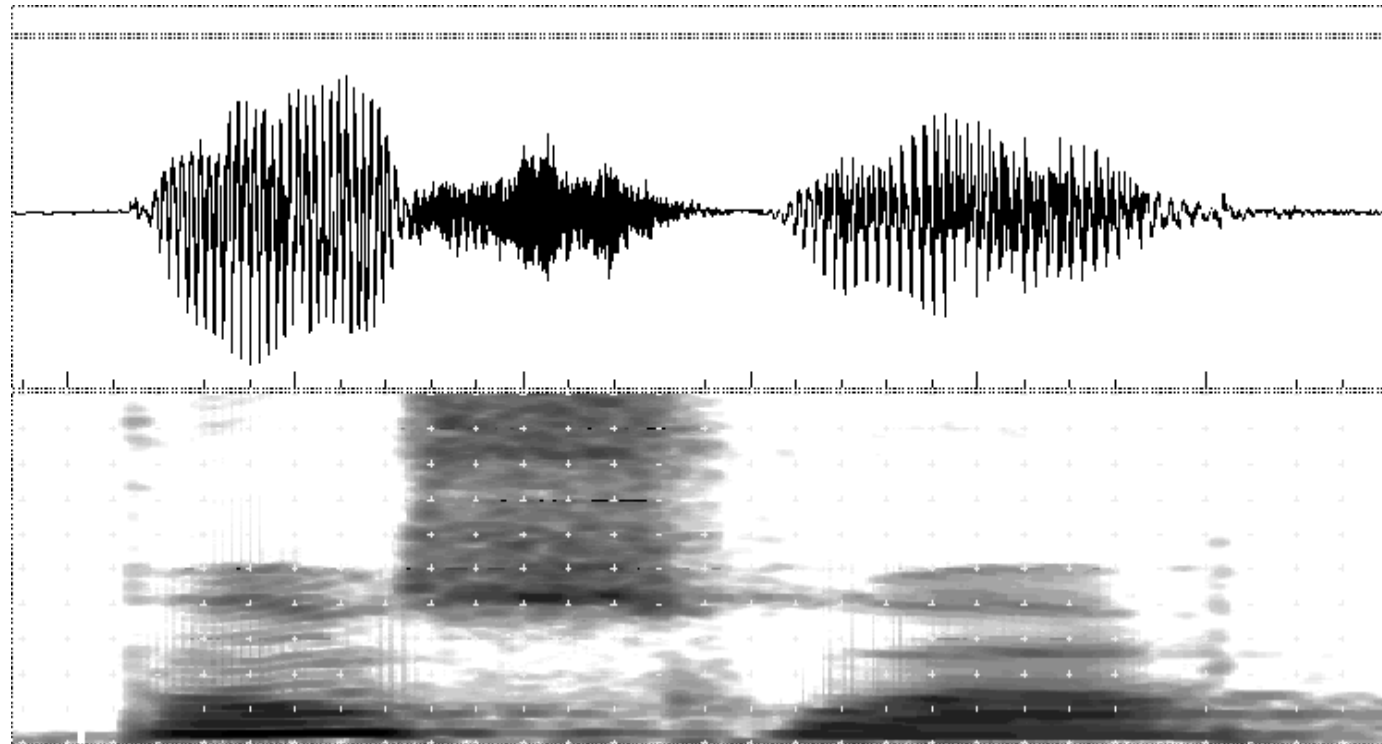
*Note* : Les distinctions vocaliques [e]-[ɛ], [ø]-[œ] et [o]-[ɔ] ne sont pas faites dans tous les contextes et par tous les locuteurs du français. Par contre, certains locuteurs font aussi des distinctions entre patte et pâte, ([ɑ]-[ɑ]) ainsi qu'entre brin et brun ([ɛ̃]-[œ̃]).

# Phonèmes du français

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

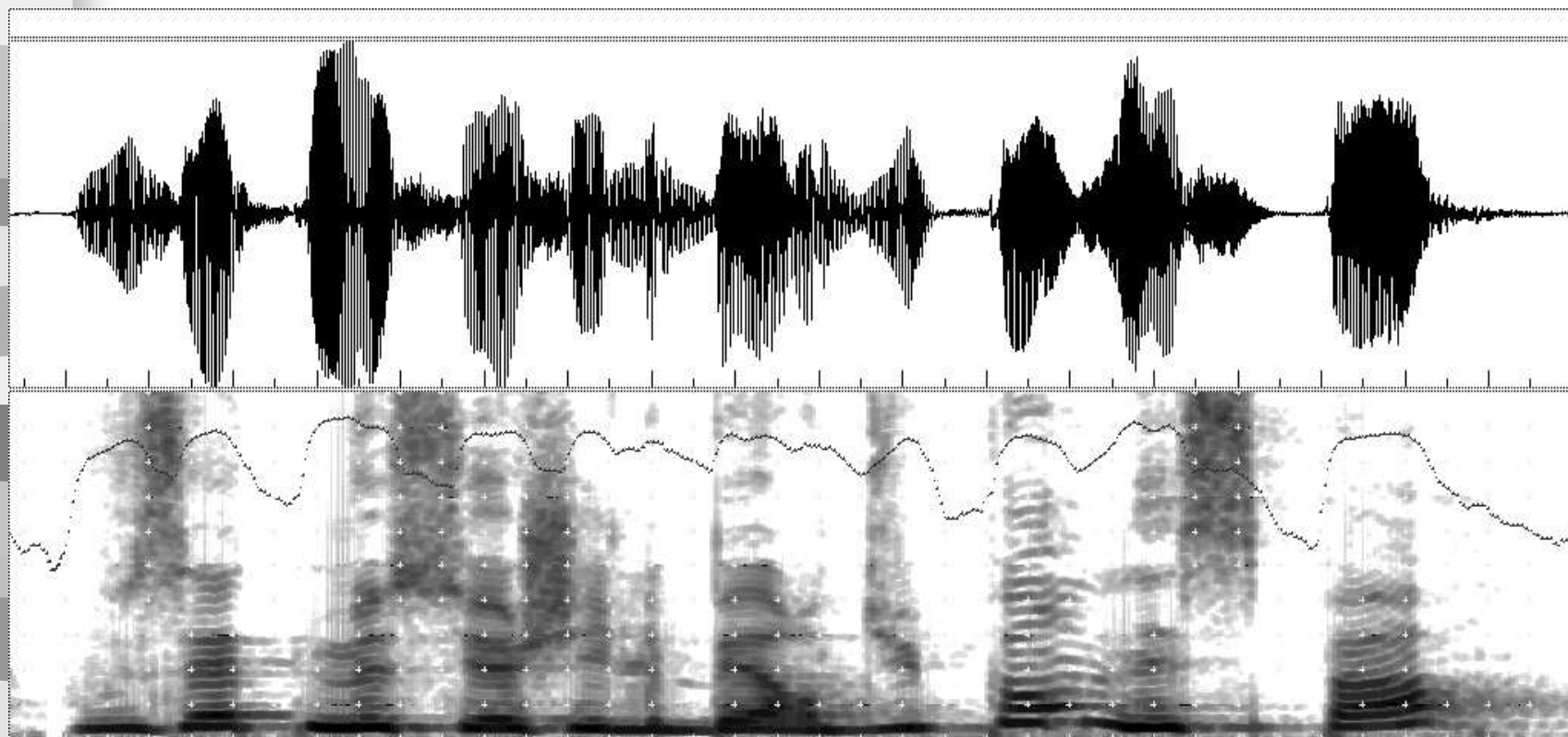
CONSONNES				Lieu d'articulation
Mode d'articulation	Labiales	Dentales	Vélo-palatales	←
↓				
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ŋ]	
Fricatives				
non voisées	[f]	[s]	[z]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[y]	[j]	
Liquides		[l]	[R]	
VOYELLES				
Orales	Antérieures		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



Bonsoir

# Signal de Parole



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

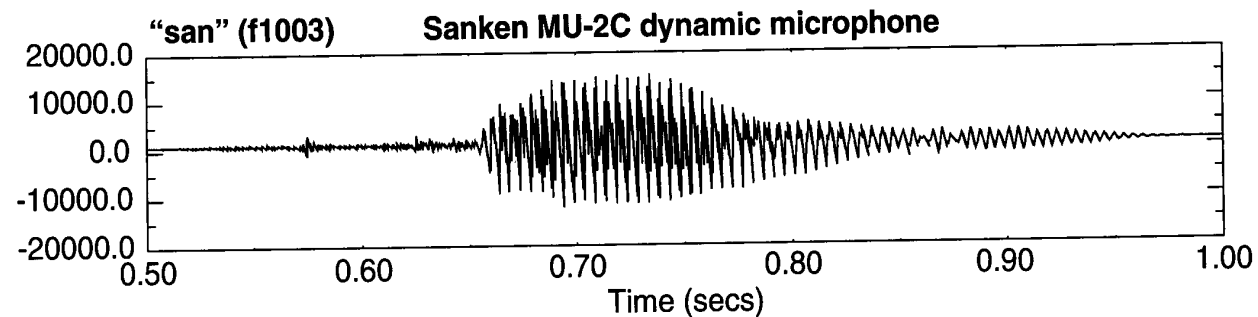
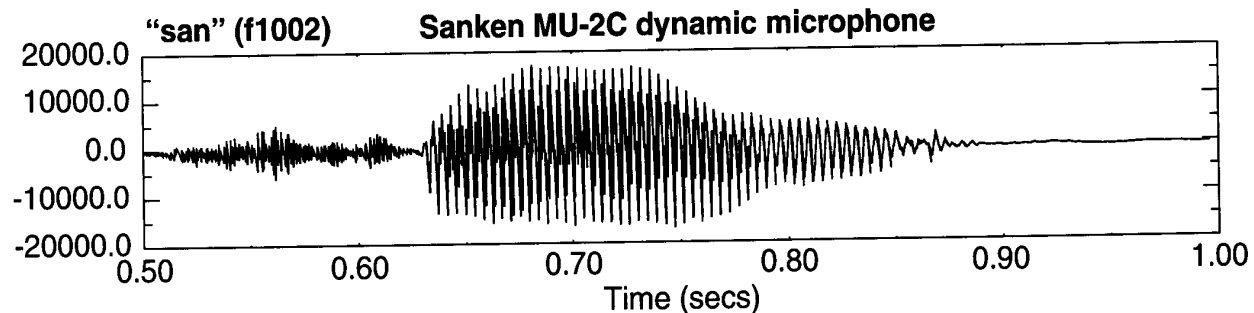


# Variabilité de la parole

- La première caractéristique d'un signal de parole est sa variabilité
  - une personne ne prononce jamais deux fois le même son de la même façon
  - deux personnes ne prononcent pas le même son de la même façon
- pourtant, ce son est parfaitement reconnu et compris par un auditeur humain !!

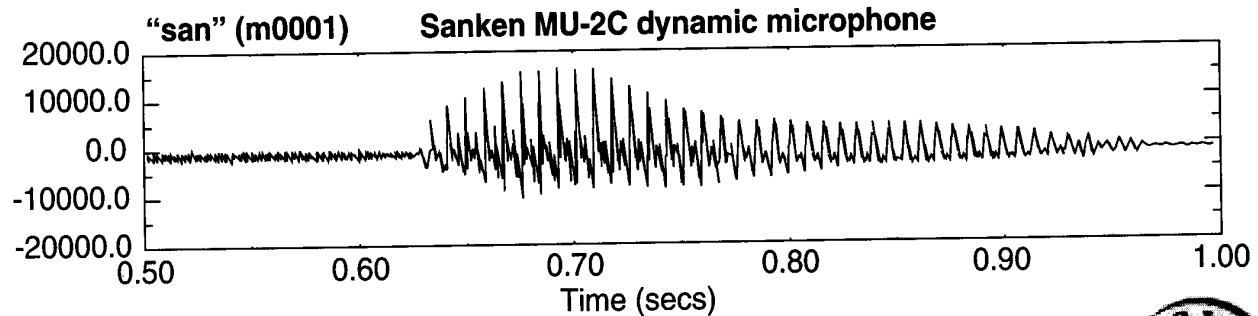
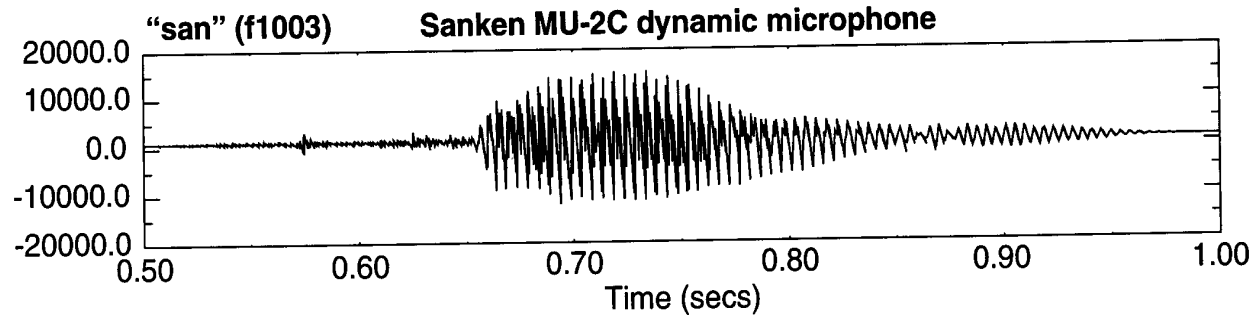
# Variabilité du locuteur

Même sons, même personne,  
mêmes conditions d'enregistrement



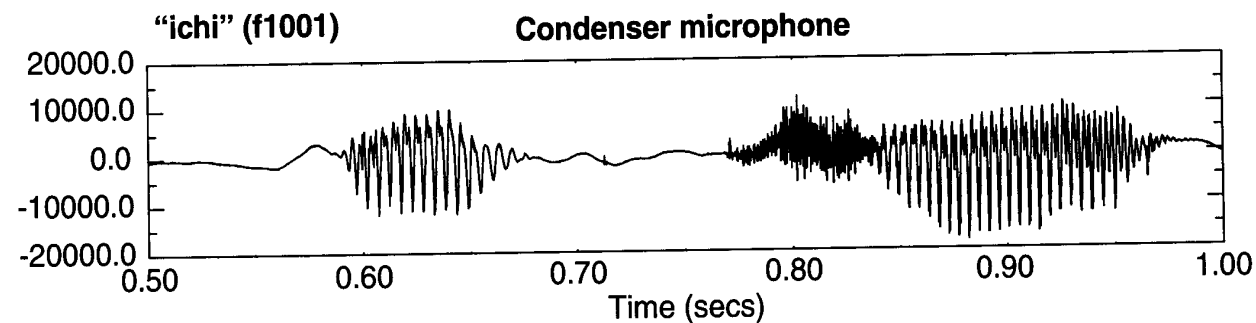
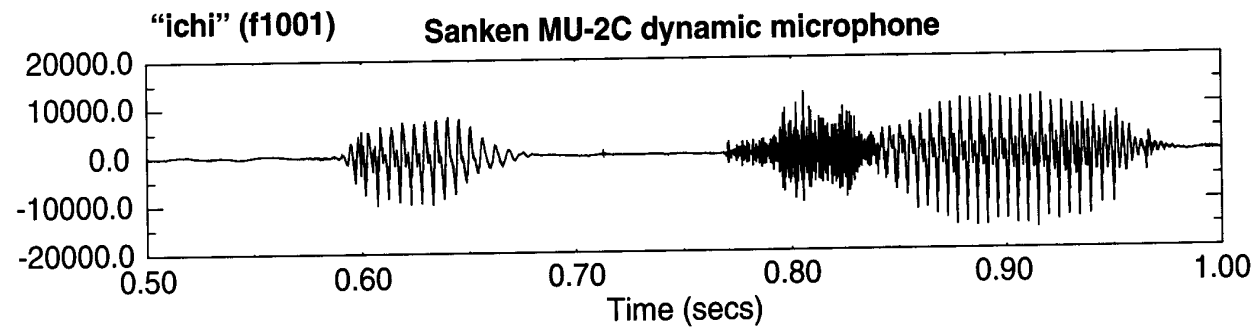
# Variabilité due à la différence des locuteurs

Même son, mêmes conditions d'enregistrement  
deux locuteurs différents



# Variabilité due à l'enregistrement

Même son, même locuteur, deux microphones différents

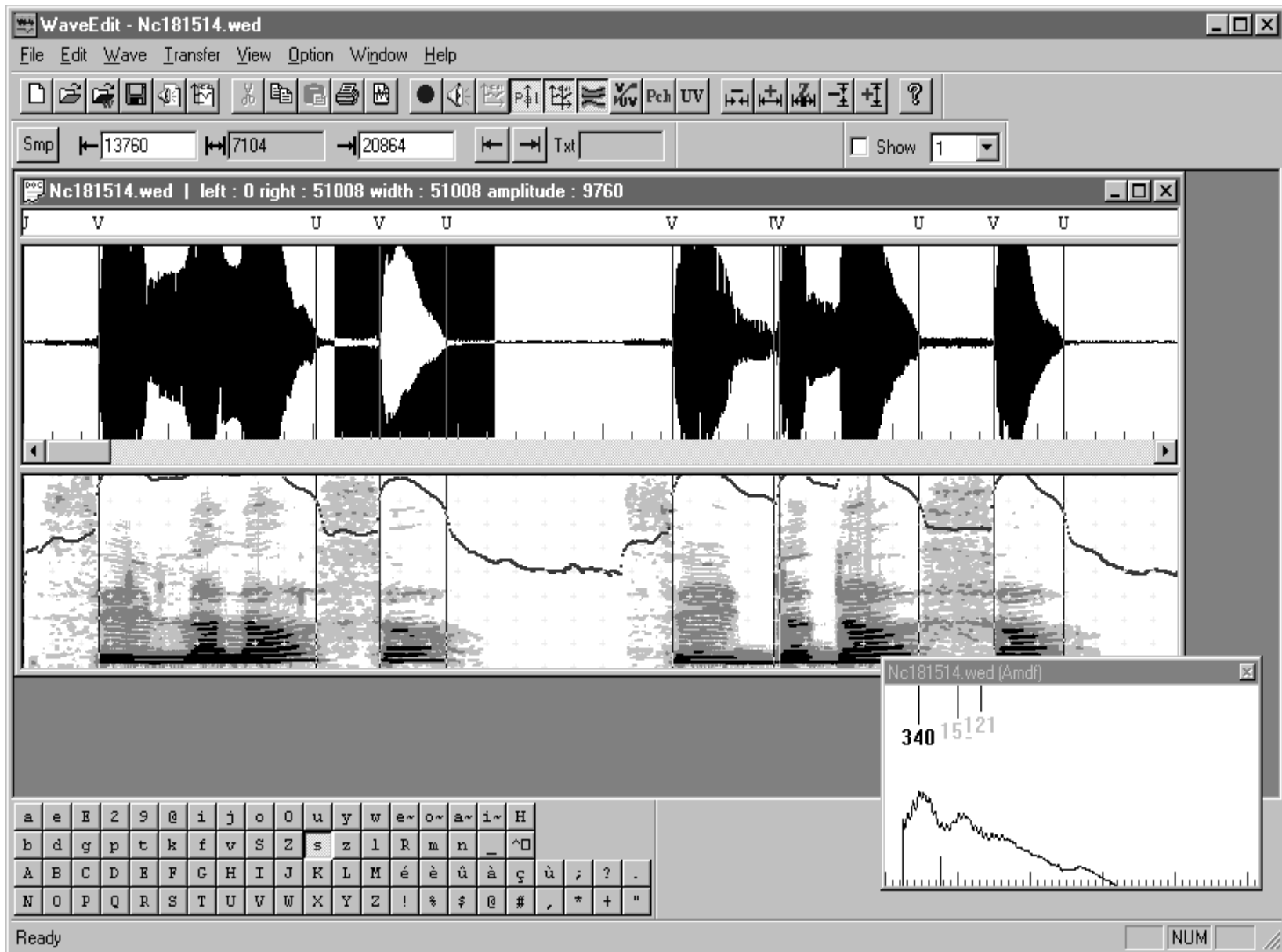


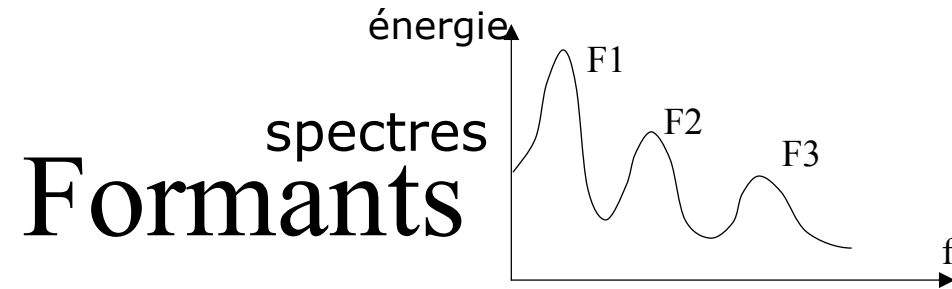


# Pourquoi analyser la parole ?

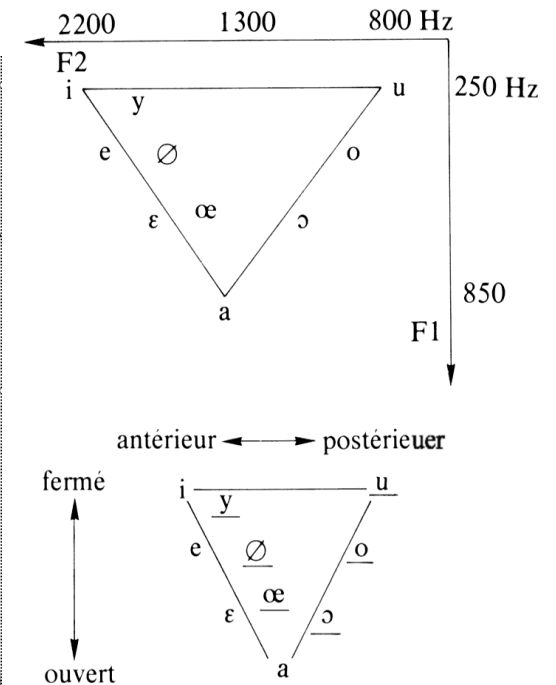
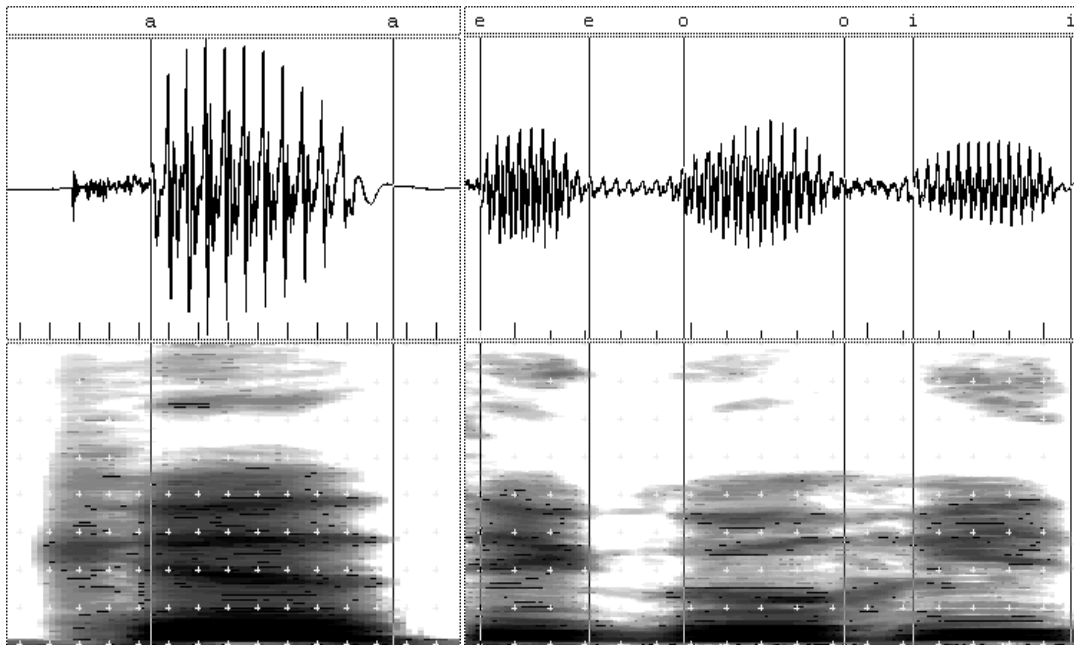
- Pour étudier et comprendre les phénomènes physiques mis en jeu
  - être un peu moins ignorants...
  - comprendre aussi les dysfonctionnements (handicapés du langage)
  - être capable d'utiliser ces connaissances pour l'apprentissage des langues étrangères
- Pour reproduire la parole sous forme artificielle
  - synthèse de la parole (synthèse à formants)
  - modélisation de l'appareil de production
- Pour déterminer des mesures de caractérisation pouvant être utilisées pour le codage ou par les moteurs de reconnaissance de parole
  - caractéristiques spectrales (LPC, MFCC, fréquence fondamentale, etc...)

# Outils d'analyse de la parole

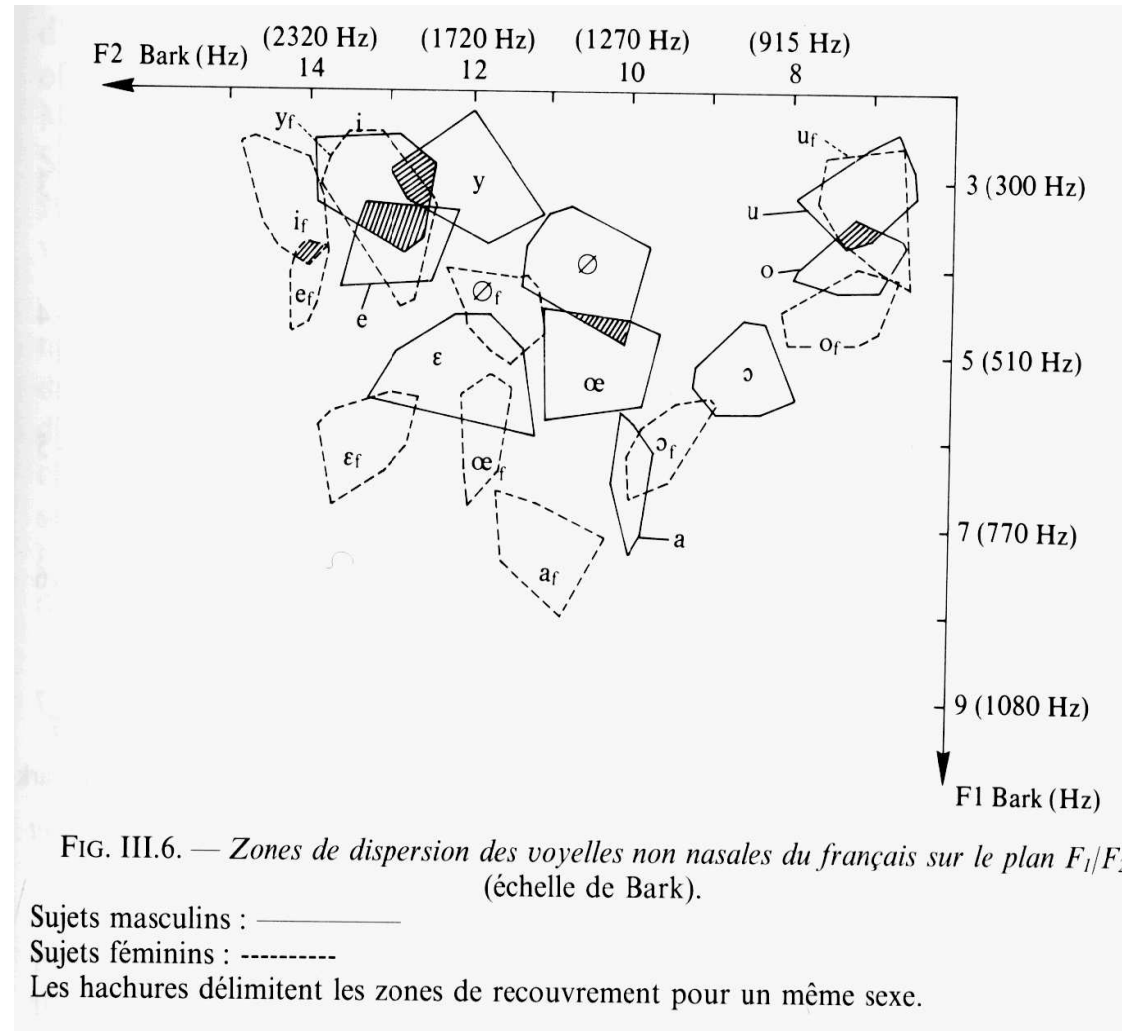




- Fréquences de résonance du conduit vocal
- Triangle vocalique des voyelles



# Formants : ellipses de dispersion



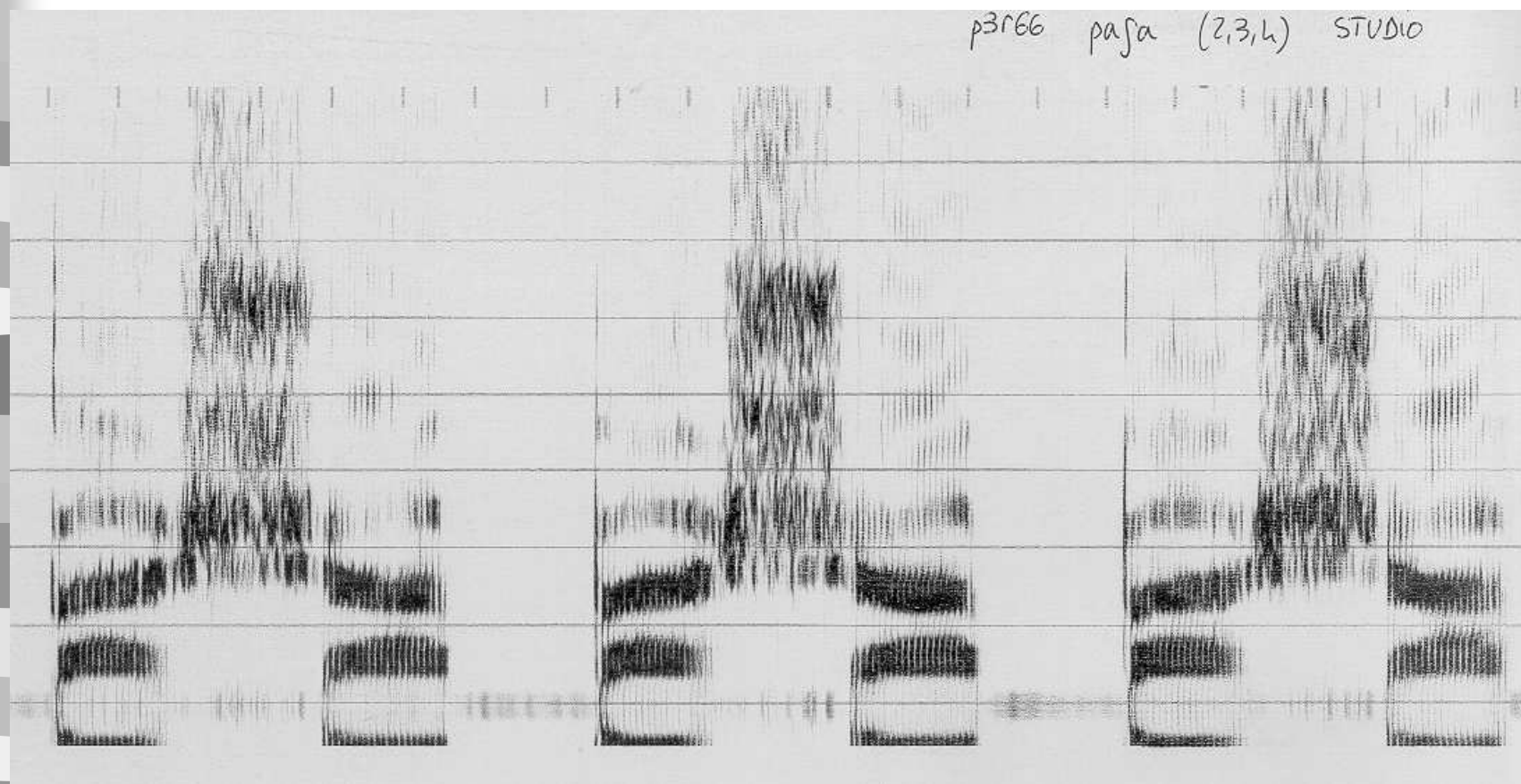
Dispersion due à la grande variabilité de la parole et des locuteurs

# Formants : quelques valeurs

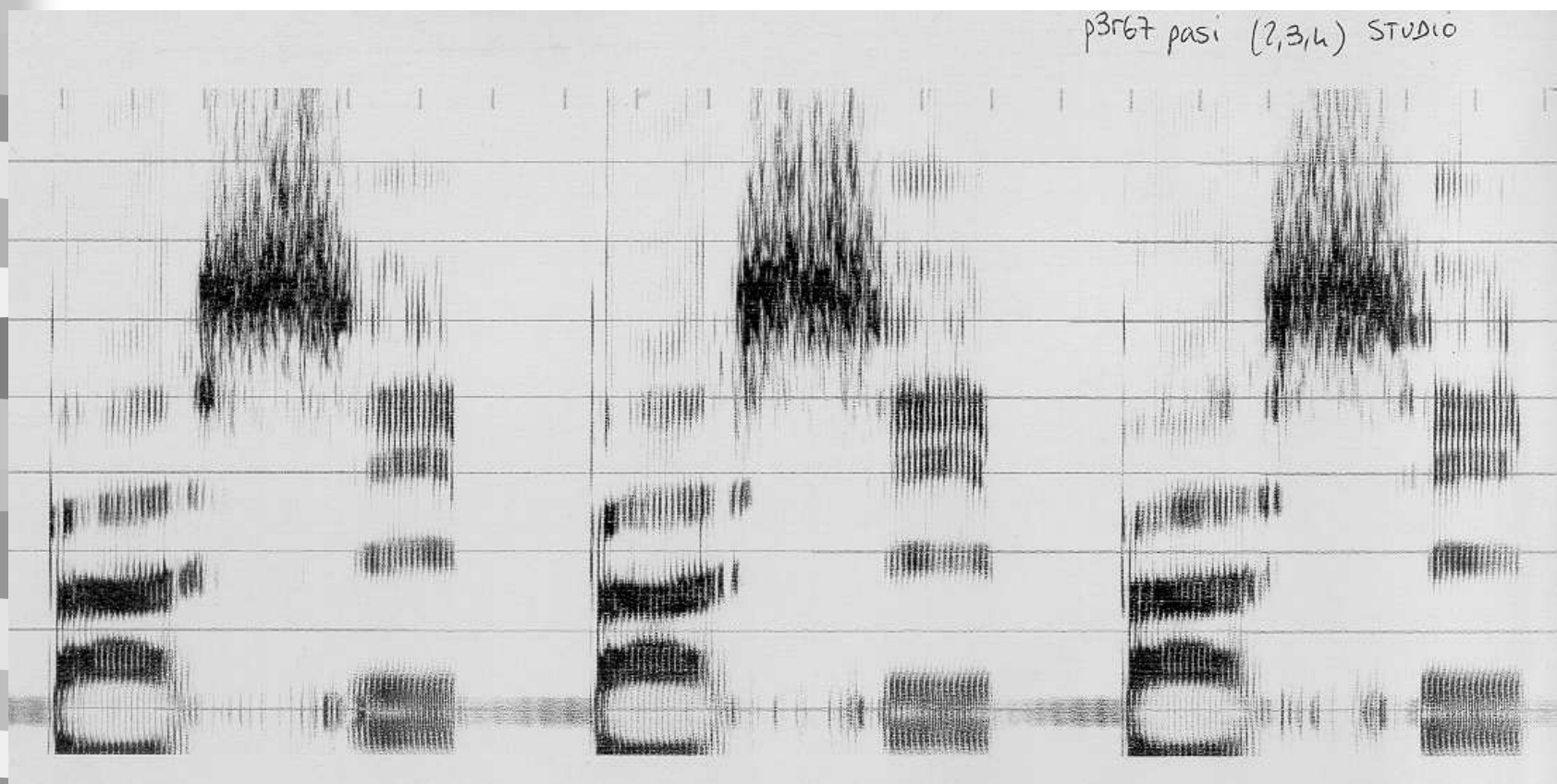
TABLEAU III.1. — Valeur en Hz des 4 premières fréquences formantiques. 10 sujets masculins ; 9 sujets féminins — Analyse LPC. Med. : valeur de la médiane ;  $\sigma$  : écart-type ; Ki : coefficient d'écart entre formants masculins et féminins :  $(Ki = F1_m / F1_f)$ . Corpus : [pV] ou [pVR], 2 répétitions par sujet.

voyelles	Sujets masculins					Sujets féminins			
		$F_1$	$F_2$	$F_3$	$F_4$	$F_1$	$F_2$	$F_3$	$F_4$
i	Med.	308	2064	2976	3407	306	2456	3389	3966
	$\sigma$	34	134	147	208	42	111	68	169
	Ki	0,99	1,19	1,14	1,16				
e	Med.	365	1961	2644	3362	417	2351	3128	4161
	$\sigma$	31	119	107	155	31	52	115	121
	Ki	1,14	1,20	1,18	1,24				
ɛ	Med.	530	1718	2558	3300	660	2080	2954	4231
	$\sigma$	49	132	103	221	46	108	156	210
	Ki	1,25	1,21	1,15	1,28				
a	Med.	684	1256	2503	3262	788	1503	2737	4143
	$\sigma$	47	32	131	155	51	86	174	192
	Ki	1,15	1,20	1,09	1,27				
ɔ	Med.	531	998	2399	3278	634	1180	2690	3950
	$\sigma$	39	60	116	155	48	59	198	201
	Ki	1,19	1,18	1,12	1,21				
o	Med.	383	793	2283	3256	461	855	2756	3805
	$\sigma$	22	63	126	161	38	73	240	183
	Ki	1,20	1,08	1,21	1,17				
u	Med.	315	764	2027	3118	311	804	2485	3550
	$\sigma$	43	59	136	172	43	53	284	197
	Ki	0,99	1,05	1,23	1,14				
y	Med.	300	1750	2120	3145	305	2046	2535	3570
	$\sigma$	37	121	182	141	68	124	139	216
	Ki	1,02	1,17	1,20	1,14				
ø	Med.	381	1417	2235	3215	469	1605	2581	4005
	$\sigma$	44	106	113	201	36	90	148	168
	Ki	1,23	1,13	1,15	1,25				
œ	Med.	517	1391	2379	3353	647	1690	2753	4038
	$\sigma$	42	94	91	149	58	47	155	202
	Ki	1,25	1,21	1,16	1,20				

# Spectrogramme - 1

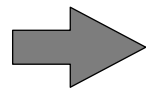


# Spectrogramme (bande large) - 2



# Prosodie

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



**Intonation / Mélodie de la voix**



# Fréquence fondamentale ou $f_0$

- Résulte de la vibration des cordes vocales
- Se traduit par l'évolution de la fréquence laryngienne réalisée à partir du signal de parole
- Dépend essentiellement de l'âge et du sexe du locuteur
  - 100 à 150 Hz pour l'homme adulte
  - 140 à 240 Hz pour la femme adulte
  - peut présenter des variations considérables chez un même locuteur
    - selon le type de phrase prononcée
    - selon l'état émotif et l'attitude du locuteur



# Compléments bibliographiques sur l'analyse de parole

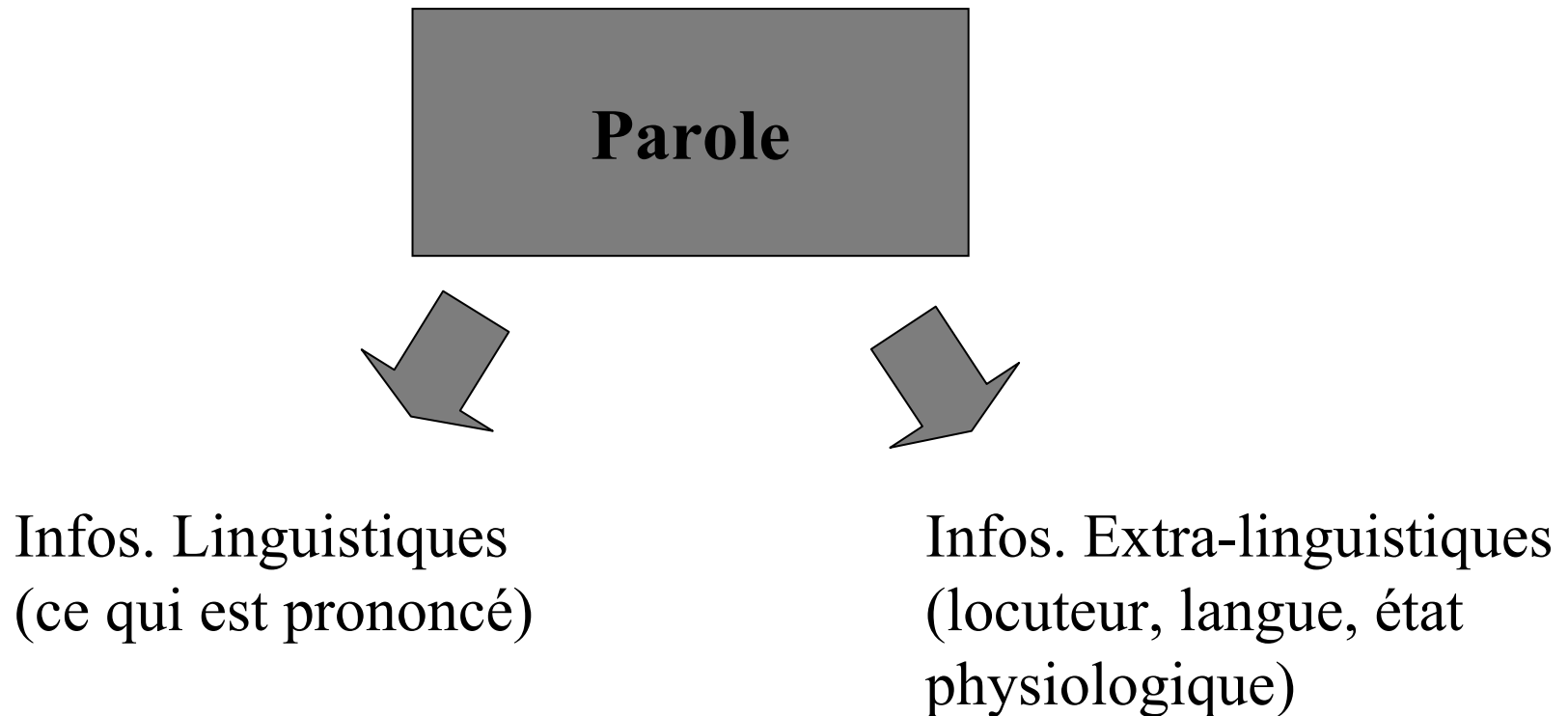
- CALLIOPE La parole et son traitement automatique
  - 1989, Masson, CENT, ENST
- FANT G. Acoustic theory of speech production
  - Mouton, The Hague (1960)
- R. BOITE, H. BOURLARD, T. DUTOIT, J. HANCQ, H. LEICH Traitement de la parole
  - Presses Polytechniques et Universitaires Romandes - Collection Electricité (2000)
- J.P. HATON & al., « La parole : du signal à son interprétation », Dunod, 2006.



# Illustration pratique

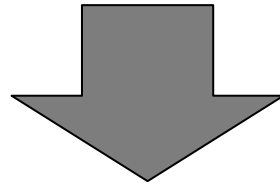
- <http://www-clips.imag.fr/geod/User/laurent.besacier/NEW-TPs/TP-CL/tp7.html>

# La parole, une source d'informations

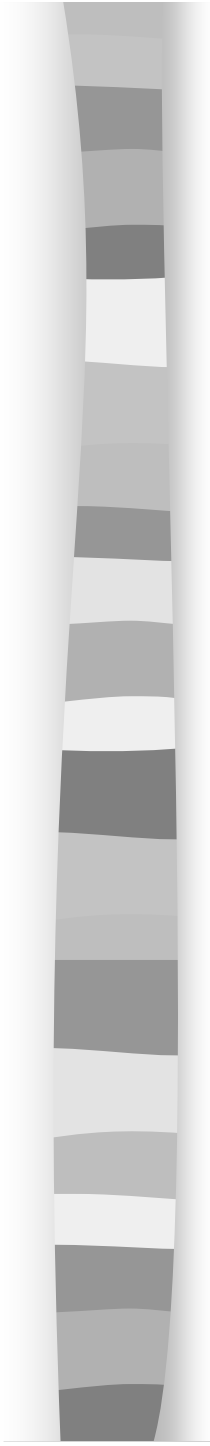


# Informations linguistiques...

- Ce qui est prononcé par le locuteur....

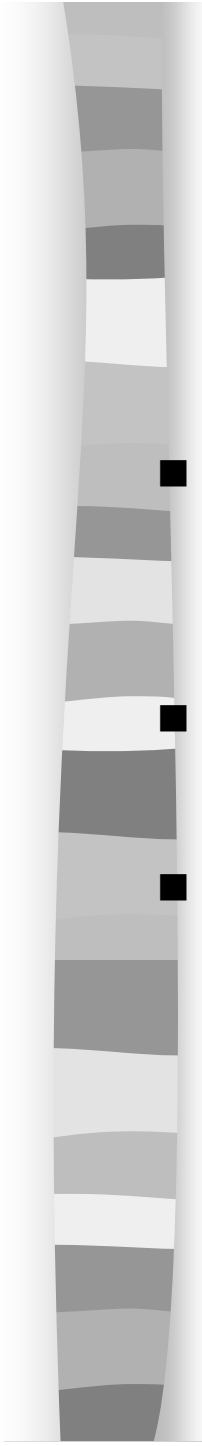


**Reconnaissance Automatique de la Parole (RAP)**



# Différents niveaux de difficulté (1)

- **Nombre de locuteurs** : systèmes mono locuteurs ...jusqu'à multi locuteurs
- **Nombre de mots du dictionnaire** : systèmes petit vocabulaire ... jusqu'à grand vocabulaire
- **Canaux de transmission** : « en direct », via le téléphone, réseaux mobiles...



# Différents niveaux de difficulté (2)

- **Environnement acoustique** : calme (utopique), normal (bureau, pièce isolée), bruité (hall de gare, rue), extrême (cockpit d'avion...)
- **Style de parole** : digits, mots isolés, mots enchaînés, parole continue (lue, spontanée)
- Une seule personne à la fois ou conversation (« cocktail party »)



# Différents domaines d 'application (1)

- **Services télécom** (ou serveurs vocaux) : renseignements par commande vocale via le téléphone (séances cinéma, annuaires, renseignements)
- **Commerce électronique** (E-commerce)
- **Services Web** : remplissage de formulaires par la voix...
- **Bornes vocales** (sur site) : renseignements touristiques, achat tickets transport (SNCF)



# Différents domaines d 'application (2)

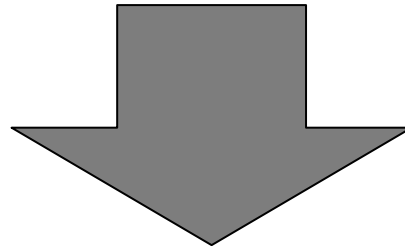
- **Bureautique** : logiciels de dictée vocale, pilotage de l 'environnement, dictée de messages électroniques
- **Enseignement des langues** : évaluation de la prononciation
- **Transports** : commande vocale pour l 'aide au pilotage



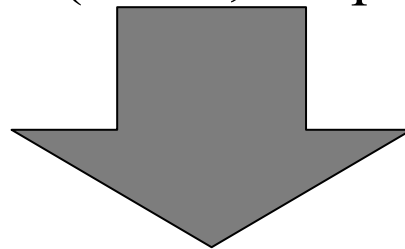
# Différents domaines d 'application (3)

- **Aide aux handicapés** : saisie de données à la voix, commandes vocales (ouverture porte, contrôle des équipements au domicile)
- **Archivage & recherche d 'informations** : transcription automatique de documents radio ou télédiffusés, recherche d 'informations dans des BD audiovisuelles

# Différents domaines d 'application (4)



- Différentes tâches (mots isolés, parole continue...)
- Différents environnements (calme, bruité...)
- Différents canaux (direct, téléphone...)



**Systemes de reconnaissance automatique de parole  
dépendants de l 'application considérée**



# Quelques systèmes commerciaux

- IBM Via Voice
- Dragon Naturally Speaking
- Phillips
- Lernhout & Hauspie
- ATT
- Boites à outils : Sphinx (cf TP)



# La reconnaissance automatique de la parole

- **Les meilleurs systèmes obtiennent\***
  - ~10-12% de taux d'erreur de mots pour l'anglais sur des documents de journaux télévisés ou des enregistrements du parlement européen
  - ~20% de taux d'erreur de mots pour l'anglais sur des conversations téléphoniques
- **Progrès réguliers**  
Voir évaluations DARPA & NIST ...

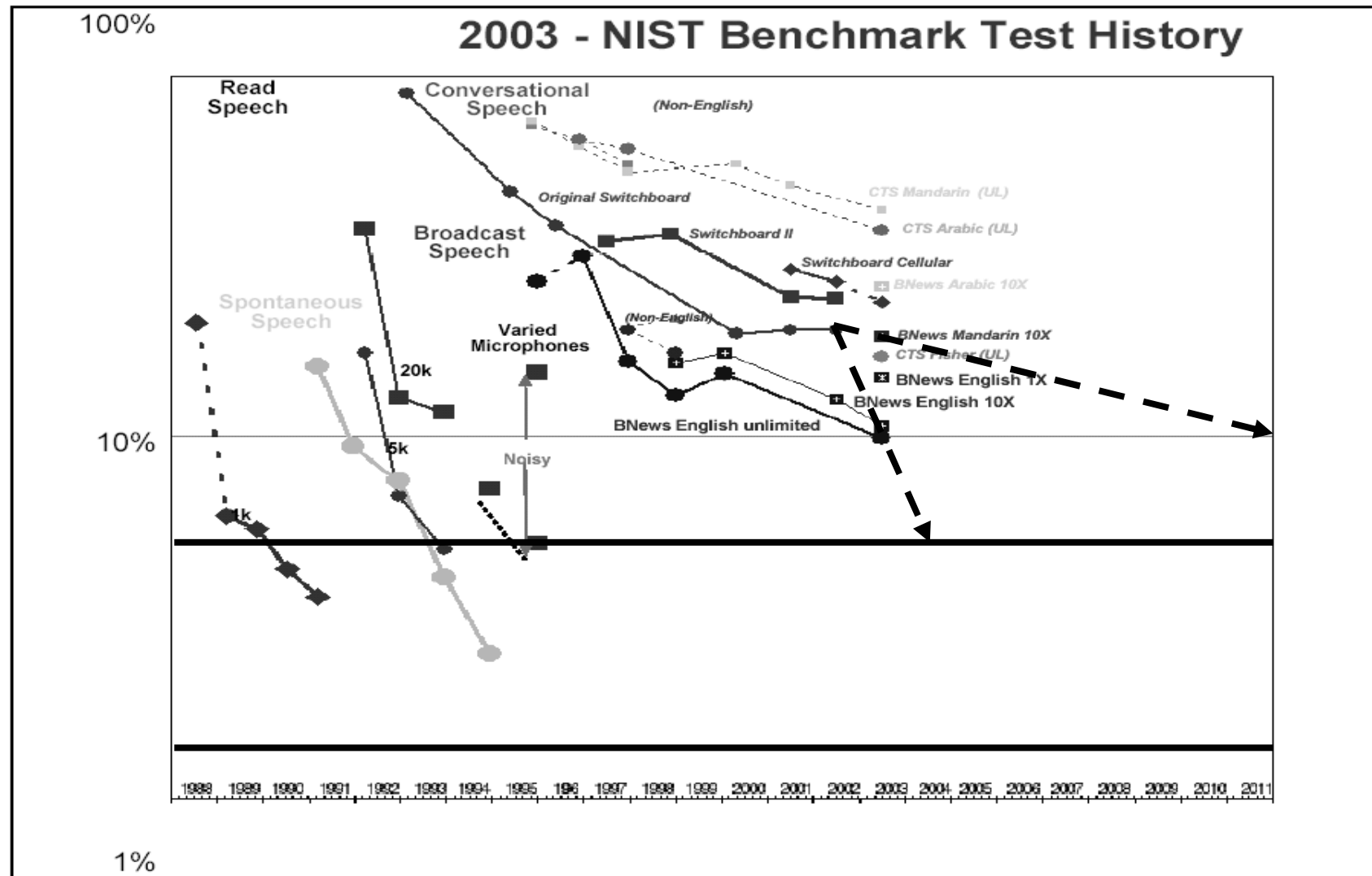
\*sources: projets TCSTAR & GALE

# Une idée des performances...pour l'anglais

<i>Task</i>	<i>Condition</i>	<i>Word Error</i>
<b>Dictation</b>	read speech, close-talking mic.	3-4% (humans 1%)
	read speech, noisy (SNR 15dB)	10%
	spontaneous dictation	14%
	read speech, non-native	20%
<b>Found audio</b>	TV & radio news broadcasts	10-15% (humans 4%)
	TV documentaries	20-30%
	Telephone conversations	20-30% (humans 4%)
	Lectures (close mic)	20%
	Lectures (distant mic)	50%
	EPPS	8%

[Gauvain, 2007]

# La reconnaissance automatique de la parole





# Informations **extra** linguistiques...

- Identité du locuteur
- Langue parlée (c 'est quand même une info linguistique, mais bon....)
- Etat physiologique et psychologique du locuteur (stress, maladie...)



# Reconnaissance Automatique du Locuteur (RAL)

- Caractéristiques biométriques
  - Empreinte digitale
  - Empreinte génétique
  - Visage
  - **Voix** (serrure vocale, transactions à distance ...)
  - Signature
  - Ecriture



# Différentes tâches

- **Identification du locuteur** : par quel locuteur de la base de donnée le signal de parole a-t-il été prononcé ? (réponse 1 parmi N, *matching*)
- **Vérification du locuteur** : le locuteur est-il bien celui qu'il prétend être ? (réponse binaire, *acceptation* ou *rejet*)



# Différents niveaux de difficulté

- **Dépendance au texte** : systèmes dépendants du texte, systèmes à textes « promptés », systèmes indépendants du texte
- **Canaux de transmission et environnement** (cf RAP)
- **Variabilité...**



# Variabilité

- **variabilité due au locuteur**
  - émotion, fatigue, stress
- **conditions d'enregistrement variables**
  - microphone, bruit ambiant
- **conditions de transmission variables**
  - canal téléphonique
- **nouveaux problèmes**
  - Mobiles : codage, bruit évoluant au cours du temps



# Différents domaines d'application

- **Applications sur site** : serrures vocales, cabines bancaires en libre service
- **Applications liées aux télécommunications** : accès à un service de transactions bancaires, ou à des données confidentielles
- **Applications judiciaires** (« *forensic* ») : recherche de suspects, expertises vocales
- **Indexation multimédia** : indexation / loc.

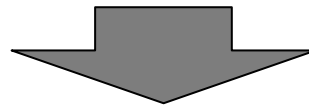


# Reconnaissance Automatique des Langues

- Prise en compte de l'aspect multilingue des serveurs vocaux
- Diriger un appel prononcé dans une langue quelconque vers l'opérateur ou le service automatique correspondant
- Proche de la RAL car identification d'un *groupe* de locuteurs qui parlent la même langue

# Autres infos. extra linguistiques

- **Etat psychologique du locuteur** : stress (boîtes noires d'avions), émotion
- **Etat pathologique du locuteur** : pathologies vocales, fatigue (intéressant à détecter pour les métiers « à risque »...mais il existe de bien meilleurs capteurs!!)



**Difficile car besoin dans des conditions extrêmes...**

# Indexation et recherche (BD audio)

## ■ Quelques chiffres

- 45 ans d'archives télé. signal → 300 000 h de
- 60 ans d'archives radio signal → 400 000 h de

## ■ Problèmes

- stockage des informations
- **accessibilité** aux données



# Différentes clés d'indexation

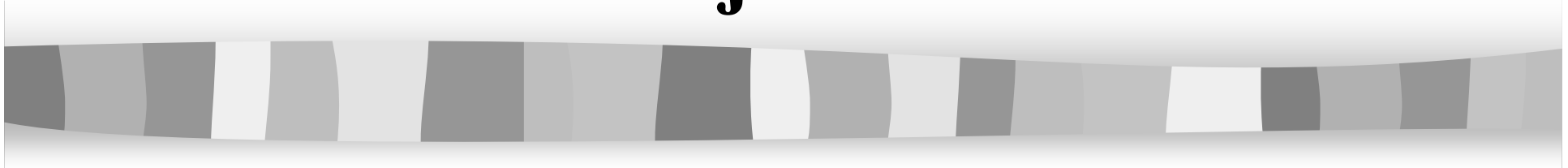
- Mots (*word spotting*)
  - recherche de documents dans lesquels un mot clé est prononcé
- Thème (*topic detection*)
  - archivage (ou recherche) de documents par thèmes
- Locuteur (*speaker segmentation / indexation*)
  - Savoir qui parle et quand
  - L'identité des locuteurs peut être inconnue au départ



# Exemples d 'application de l 'indexation

- Archivage automatique de programmes radio-diffusés par thème
- Trouver tous les discours prononcés par telle personnalité politique (calcul du temps de parole lors d 'une campagne électorale)
- Recherche des messages par locuteur (ou par thème) sur un répondeur téléphonique
- Transcription automatique de documents audio.

# **Modélisation stochastique d'objets sonores**



# Modélisation probabiliste

$$\hat{P}(Y|X)$$

Séquence d'observations acoustiques  
(vecteurs multi-dimensionnels)

- *tranches (trames) de signal*
- *coefficients de bancs de filtres*
- *coefficients cepstraux*
- *composantes principales temps-  
fréquence*
- *...*

hypothèse de classe  
ou d'objet sonore

- *type de son (parole / musique / ...)*
- *locuteur / langue / canal*
- *phonème / syllabe / mot*
- *événement sonore (jingle)*
- *passé / futur d'une rupture*
- *...*

→ Approche générique

# Bayes

- $x$  : observation (signal)
- $c_i$  : classe à reconnaître

$$c^* = \arg \max_i p(c_i / x) = \arg \max_i \frac{p(x / c_i).P(c_i)}{p(x)} = \arg \max_i p(x / c_i).P(c_i)$$

- Reconnaissance automatique de la parole

$$w^* = \arg \max_i \frac{p(x / w_i).P(w_i)}{p(x)} = \arg \max_i p(x / w_i).P(w_i)$$

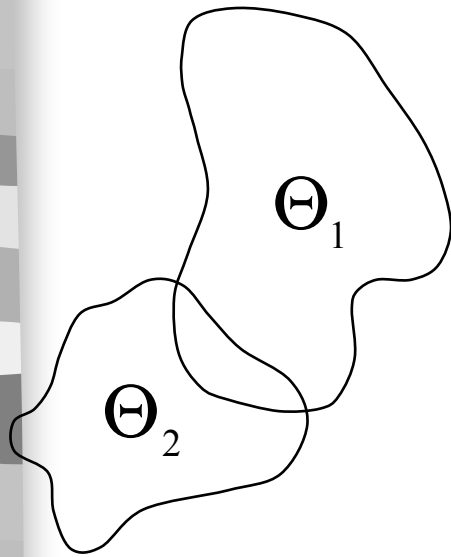
Modèle acoustique ↑  
Modèle de langage ↓

- Traduction automatique statistique

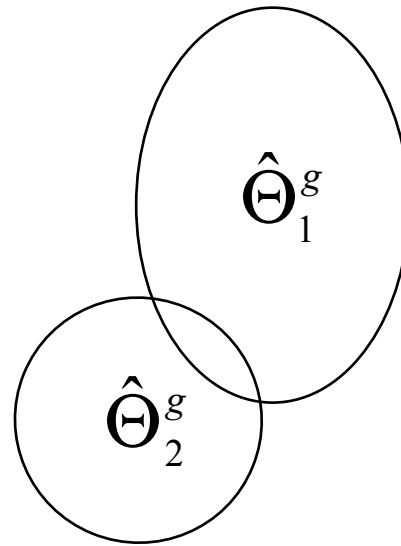
$$e^* = \arg \max_i \frac{p(f / e_i).P(e_i)}{p(f)} = \arg \max_i p(f / e_i).P(e_i)$$

Modèle de traduction ↑  
Modèle de langage ↓

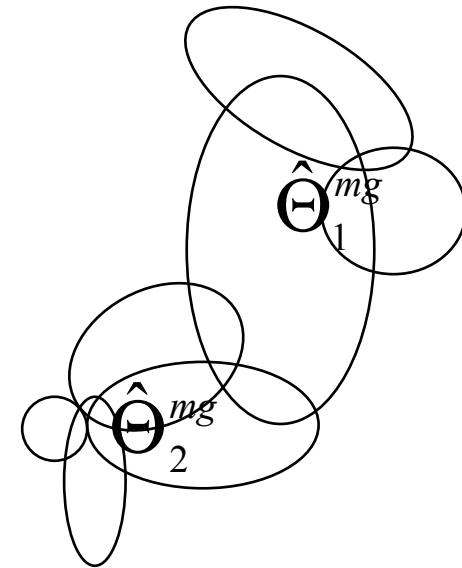
# Gaussiennes



Distribution réelle



Modèle gaussien



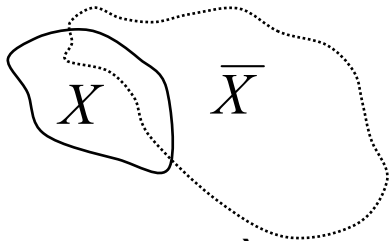
Modèle multigaussien  
(GMM)

# Automates

- Traitement / Modélisation de processus séquentiels
- Structures séquentielles complexes décomposées en segments élémentaires stationnaires
- Chaque segment : fonction déterministe ou stochastique
- Permet de décrire des modèles de langues, de lexiques, de phonèmes...
- Exemple : Modèles de Markov Cachés (HMMs)
  - 2 processus stochastiques :
    - Séquence d'états du HMM (structure séquentielle des données)
    - Probabilité d'émission de l'état (caractéristiques locales des données)
    - Exemple : modèle HMM gauche-droit de phonème avec distributions multigaussiennes

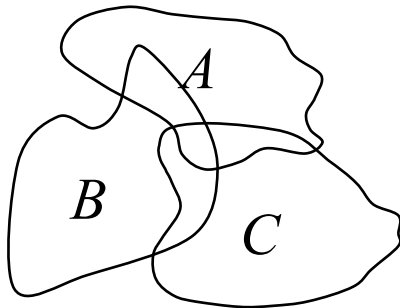
# Types de problèmes traités

## Détection



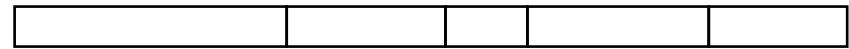
→ Tests d'hypothèses binaires

## Classification



→ Maximum A Posteriori

## Segmentation



→ Détection de ruptures

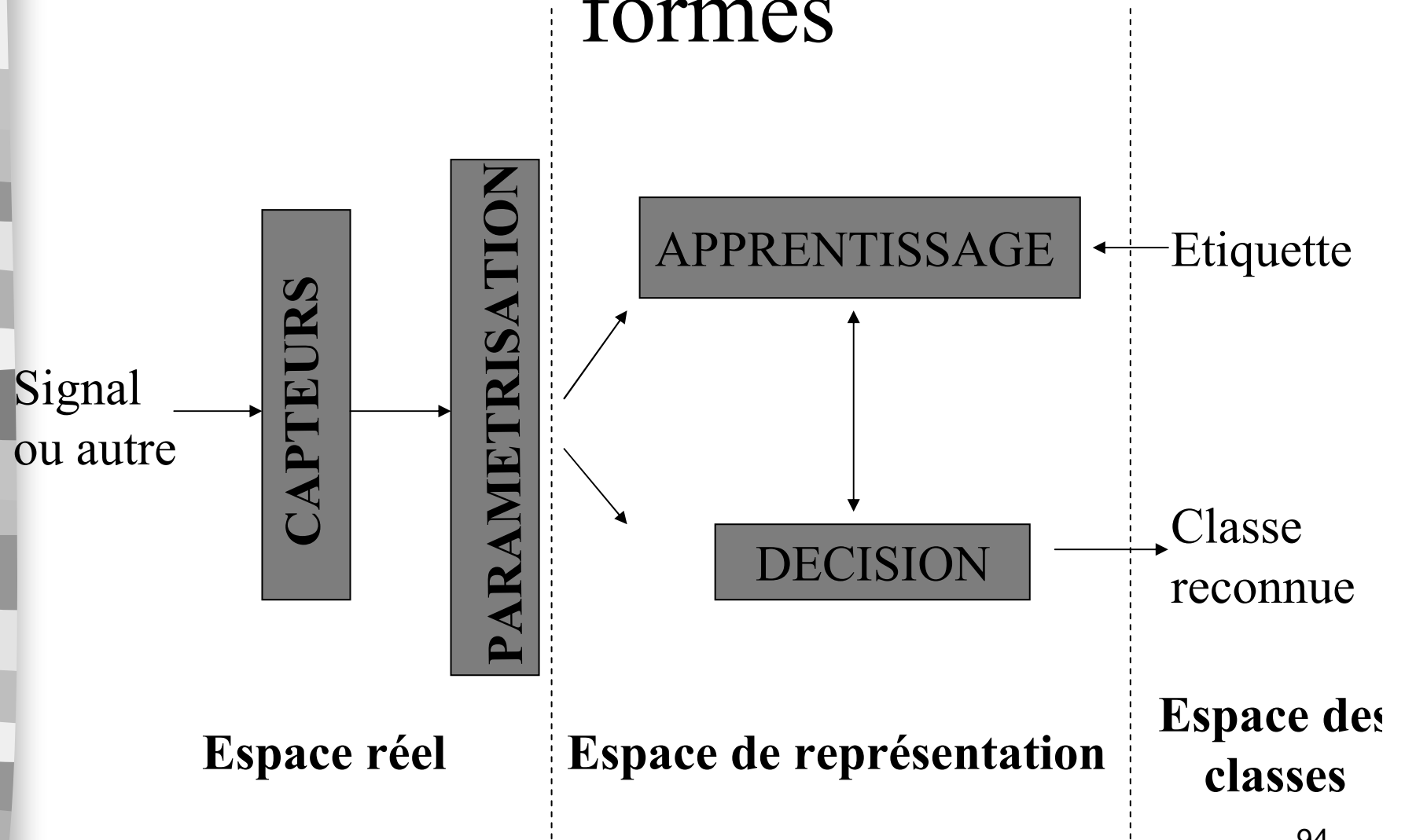
## Décodage



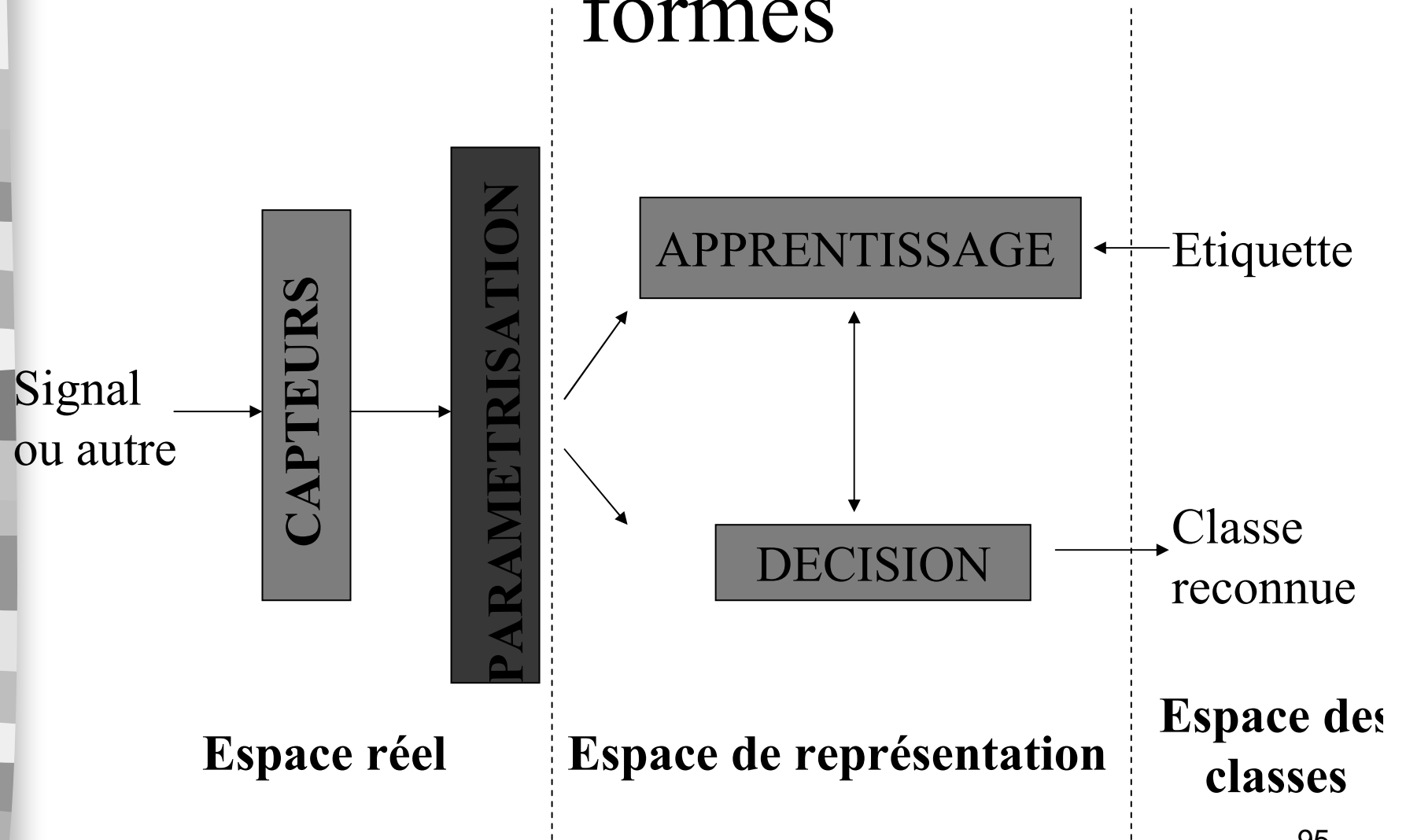
→ Recherche de séquences d'états

\*Transparent emprunté à F. Bimbot

# Base de reconnaissance des formes



# Base de reconnaissance des formes





# Paramétrisation en parole...

- Espace de représentation = paramètres acoustiques :
  - LPC (Linear Predictive Coefficients)
  - MFCC (Mel Frequency Cepstral Coefficients)
  - FilterBanks

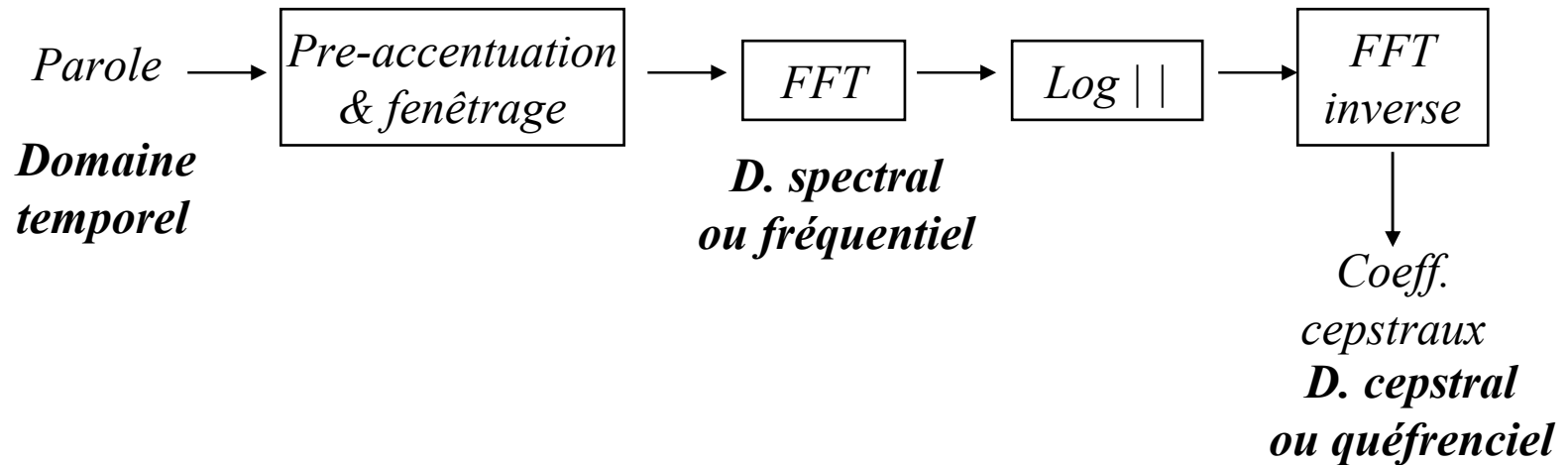


# Paramétrisation du signal de parole

- Essentiellement pour la reconnaissance :
  - analyse spectrale
  - analyse cepstrale
  - prédiction linéaire
- On utilise aussi :
  - informations prosodiques (fréquence fondamentale, durée)

# Paramétrage acoustique (1)

- Coefficients banc de filtres : énergie du signal dans différentes bandes de fréquence
- Coefficients cepstraux : séparer (par déconvolution) source et conduit vocal



# Paramétrage acoustique (2)

- Méthode LPC (Linear Predictive Coding) : fondée sur les connaissances en production de la parole
  - on prédit l'échantillon suivant, comme étant une somme pondérée des échantillons précédents

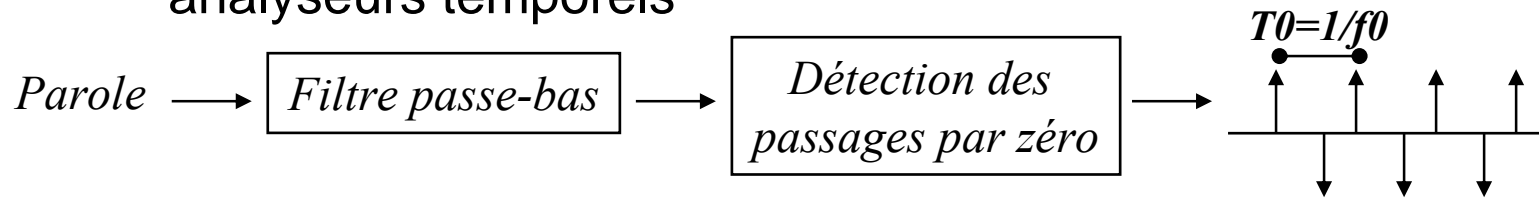
- $p$  est l'ordre du modèle  $\hat{s}_n = \sum_{i=1}^p a_i s_{n-i}$
- les  $a_i$  = coefficients de prédiction linéaire
- différentes méthodes d'estimation de ces coeff.

# Paramétrage acoustique (3)

- Détection de la fréquence fondamentale (pitch ou  $f_0$ )

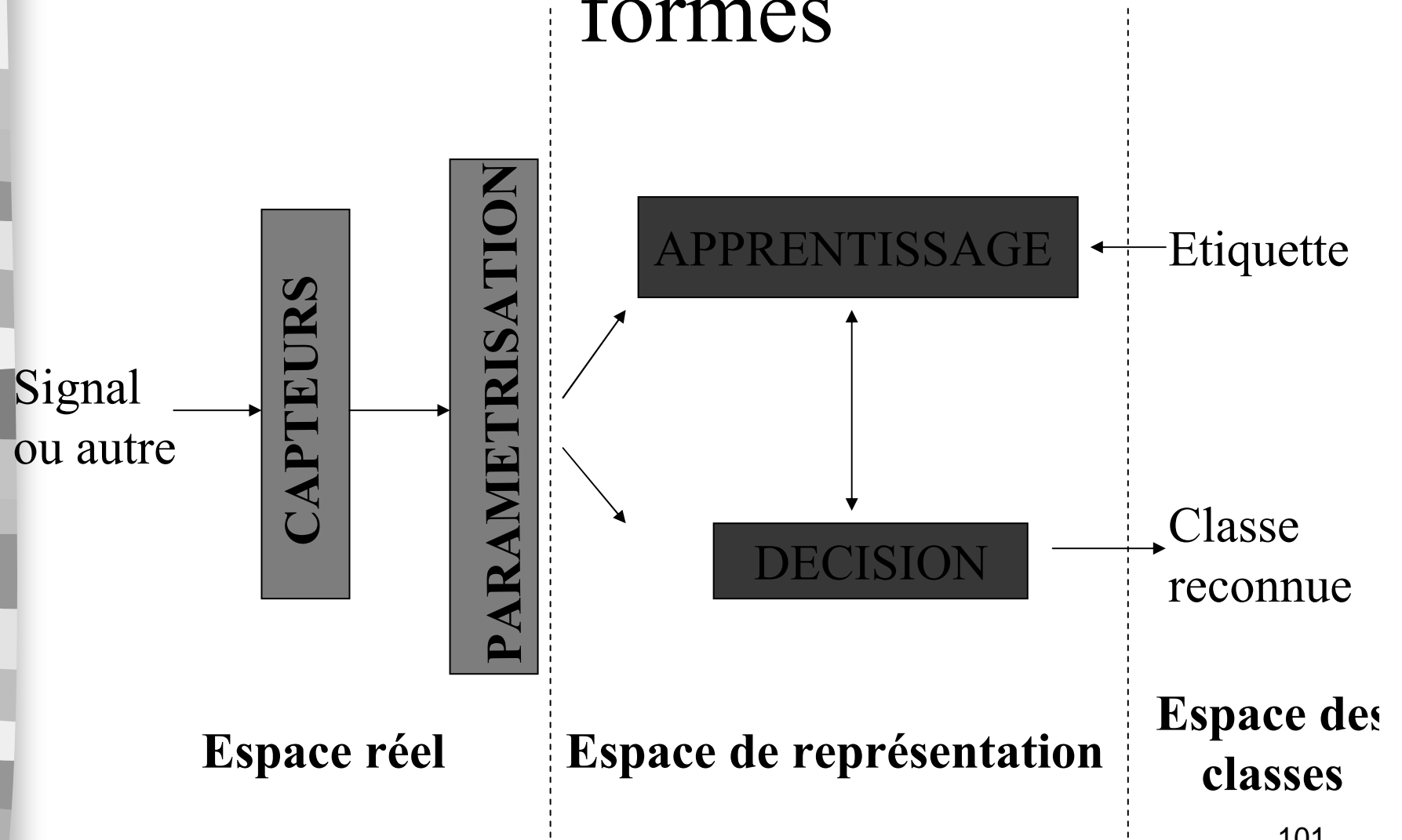
:

- analyseurs temporels



- analyseurs spectraux : pitch déterminé à partir des harmoniques du signal de parole
  - revient à calculer le PGCD des maxima du spectre d'amplitude
- problème du pitch : grande variabilité, estimation fine difficile...

# Base de reconnaissance des formes





# Approche statistique

- Toute forme est représentée par un **vecteur** et une **métrique** est définie pour comparer les formes entre elles
- 2 phases :
  - **apprentissage** : caractériser statistiquement les classes dans l'espace de représentation
  - **décision** : recherche, pour une forme donnée, la classe qui maximise la probabilité d'appartenance parmi l'ensemble des classes



# Différents modes d'apprentissage (1)

- **Supervisé** : l'algorithme utilise *a priori* la connaissance de la classe d'appartenance de chaque vecteur d'apprentissage
- **Non supervisé** : l'algo. considère l'ensemble des vecteurs d'apprentissage et propose une classification de ces vecteurs dans l'espace de représentation  
partition de l'espace en classes définies *a posteriori*



# Différents modes d'apprentissage (2)

- **Discriminant** : on cherche à maximiser le taux de bonne reconnaissance (=minimiser le taux d'erreur)
  - pas forcément généralisable sur d'autres données
- **Non discriminant** : on cherche à maximiser la probabilité d'appartenance à une classe donnée
  - pas forcément optimal au sens du taux de reconnaissance...

# Théorie bayésienne de la décision

- Approche qui minimise le *risque global d'erreur* du système
- Soient  $w_i$   $i=1, K$ , les classes possibles pour une observation  $x$  inconnue, la classe reconnue est  $w^*(x) = \arg \max_{i=1..K} P(w_i / x)$
- Avec la probabilité *a posteriori* :

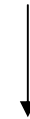
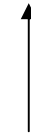
$$P(w_i / x) = \frac{p(x / w_i) \cdot P(w_i)}{p(x)}$$

# La formule...en parole...

- $x$  : suite de vecteurs acoustiques
- $w_i$  : mot ou phonème

$$w^* = \arg \max_i \frac{p(x / w_i) \cdot P(w_i)}{p(x)} = \arg \max_i p(x / w_i) \cdot P(w_i)$$

Modèle acoustique



Modèle de langage

# La formule...en parole...

- $x$  : suite de vecteurs acoustiques
- $w_i$  : mot ou phonème

$$w^* = \arg \max_i \frac{p(x / w_i) \cdot P(w_i)}{p(x)} = \arg \max_i p(x / w_i) \cdot P(w_i)$$

Modèle acoustique

Modèle de langage

# Reconnaissance de parole

- Ce qui différencie les méthodes, c'est l'estimation de  
!!  $p(x / w_i)$
- Plusieurs méthodes d'apprentissage suivant les hypothèses + ou - fortes que l'on fait
- **Méthodes paramétriques** : supposent une forme de loi connue (souvent gaussienne)
- **Méthodes non paramétriques** : pas d'hypothèse sur la forme des lois

$$p(x / w_i)$$



# Modélisation d'une classe acoustique

- **Modélisation statistique**
- **Modèle paramétrique ou non de  $p(x/w_i)$**
- **Principe de la vraisemblance**
  - **Distribution réelle (modèle non paramétrique)**
  - **Données manquantes – modèle paramétrique**

# Modélisation mono-gaussienne

$$l_t = G(y_t / X) = \frac{1}{(2\pi)^{p/2} (\det X)^{1/2}} e^{-\frac{1}{2}(y_t - \bar{x})^T X^{-1} (y_t - \bar{x})}$$

avec  $\bar{x} = \frac{1}{M} \sum_{t=1}^M x_t$  et  $X = \frac{1}{M} \sum_{t=1}^M (x_t - \bar{x})(x_t - \bar{x})^T$



# Modélisation multi-gaussienne (GMMs)

$$l_t = p(x_t | \lambda_s) = \sum_{i=1}^N p_i^s b_i^s(x_t)$$

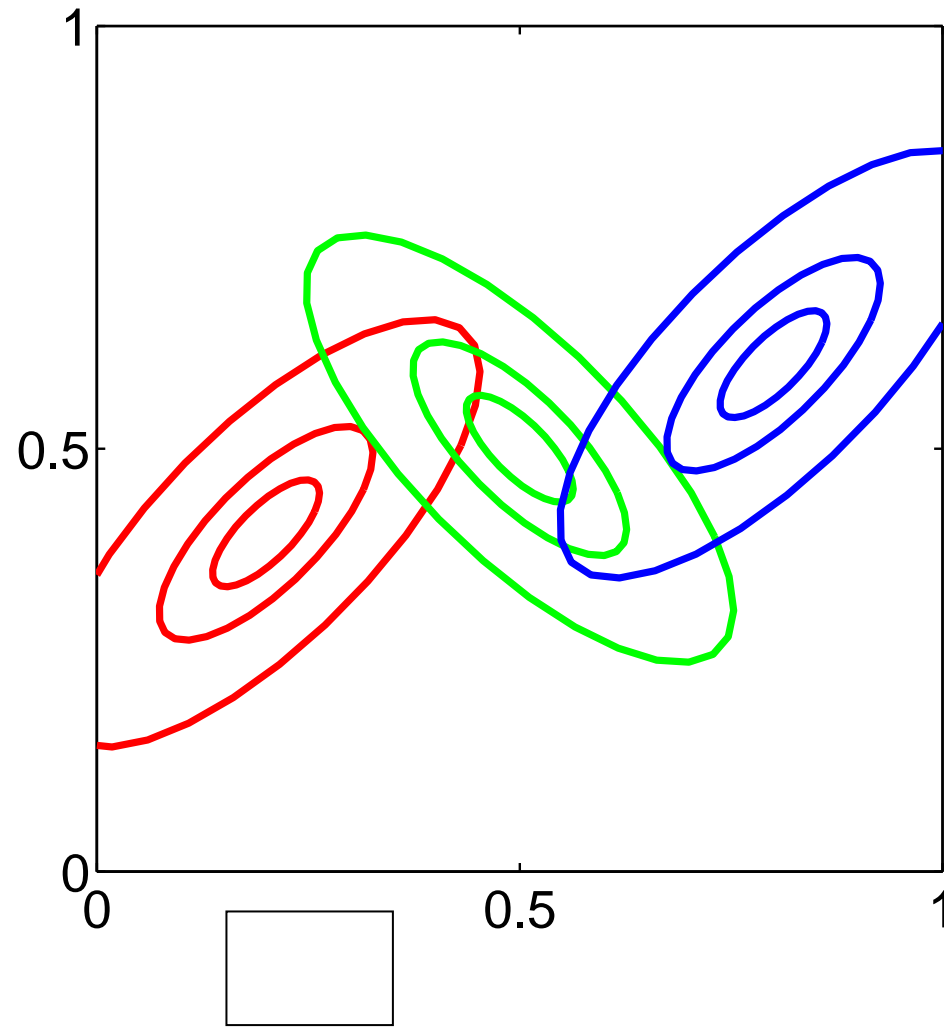
- Les paramètres du modèle de la classe  $s$  sont alors :  $\lambda_s = (p_i^s, \mu_i^s, \Sigma_i^s)_{1 \leq i \leq N}$
- Algorithme EM
- Algorithme MAP



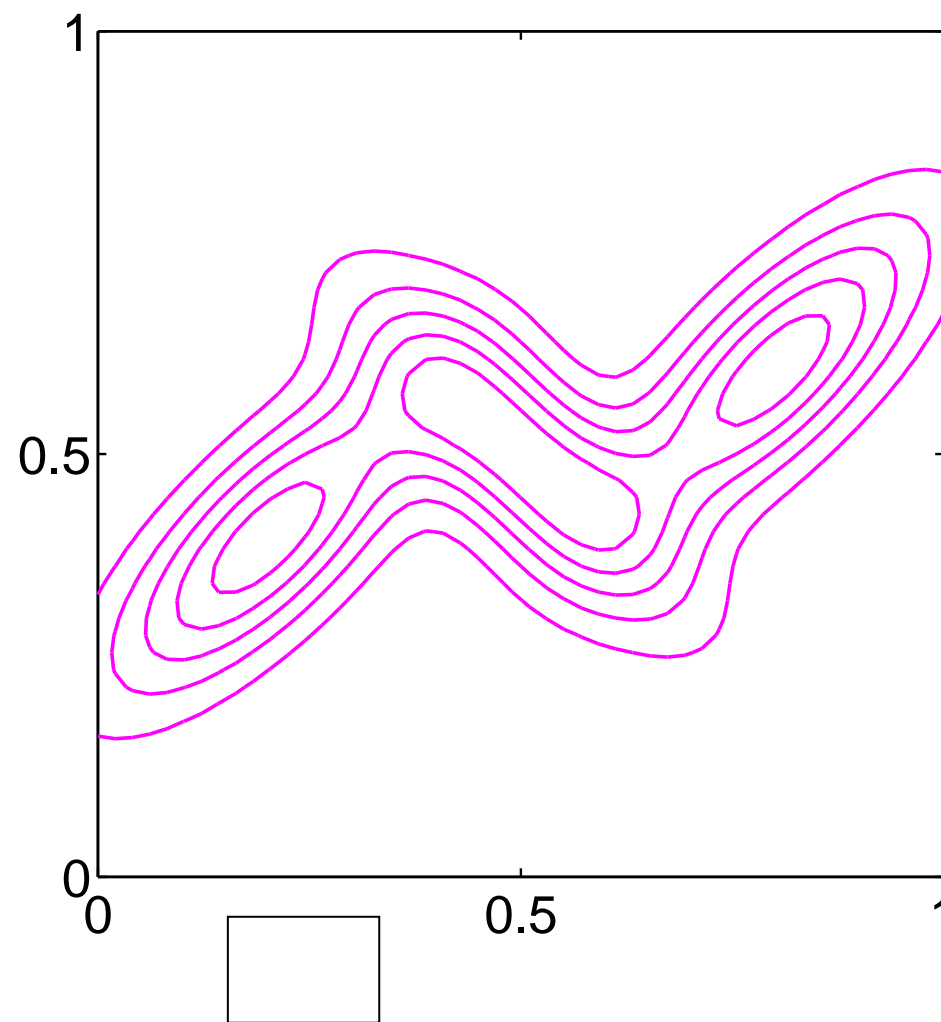
# Modélisation multi-gaussienne (GMMs)

- Un GMM peut être vu comme un HMM ergodique avec des probabilités de transitions entre états égales, et une probabilité d'émission gaussienne pour chaque état.

# Exemple: Melange de 3 Gaussiennes



# Contours de la distribution





# Problème

- Comment calculer les paramètres d'un modèle GMM à partir d'observations.
  - expectation-maximization (EM) algorithm



# Algorithme EM (1)

- Algo. itératif
  - Initialiser les paramètres (aléatoire, clustering)
  - Enchaîner les étapes suivantes:
    1. E-step: calculer l'appartenance supposée de chaque point pour chaque classe
    2. M-step: mettre à jour les valeurs des paramètres maximisant la vraisemblance des données (ML), connaissant les appartenances des paramètres juste calculées

# Algorithme EM (2)

- Maximisation de la log-vraisemblance (dérivée nulle)

$$-\sum_{n=1}^N \frac{\pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\underbrace{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\gamma_j(\mathbf{x}_n)}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j) = 0$$

- Donne

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^N \gamma_j(\mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^N \gamma_j(\mathbf{x}_n)}$$

Moyenne pondérée des données

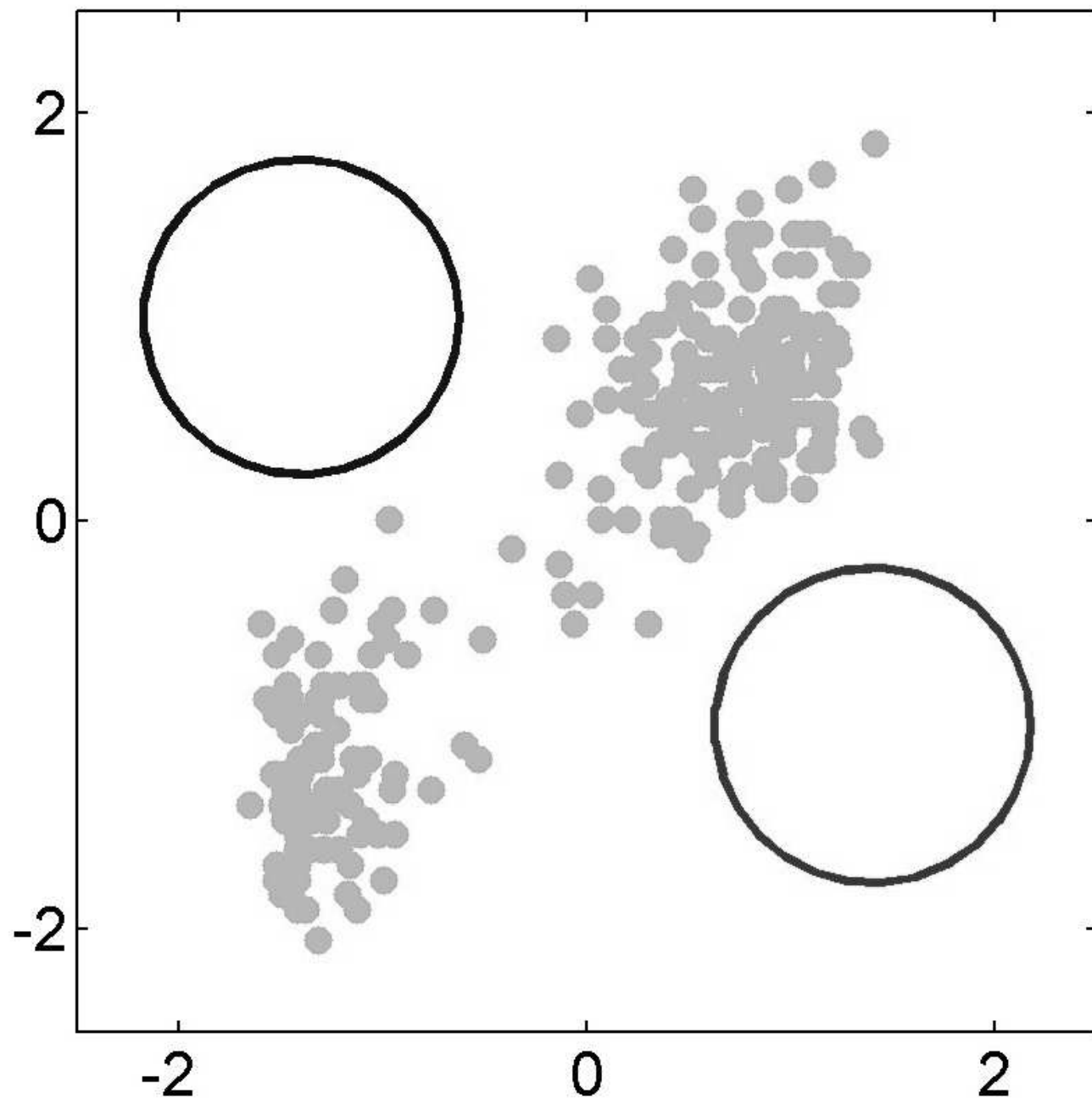
# Algorithme EM (3)

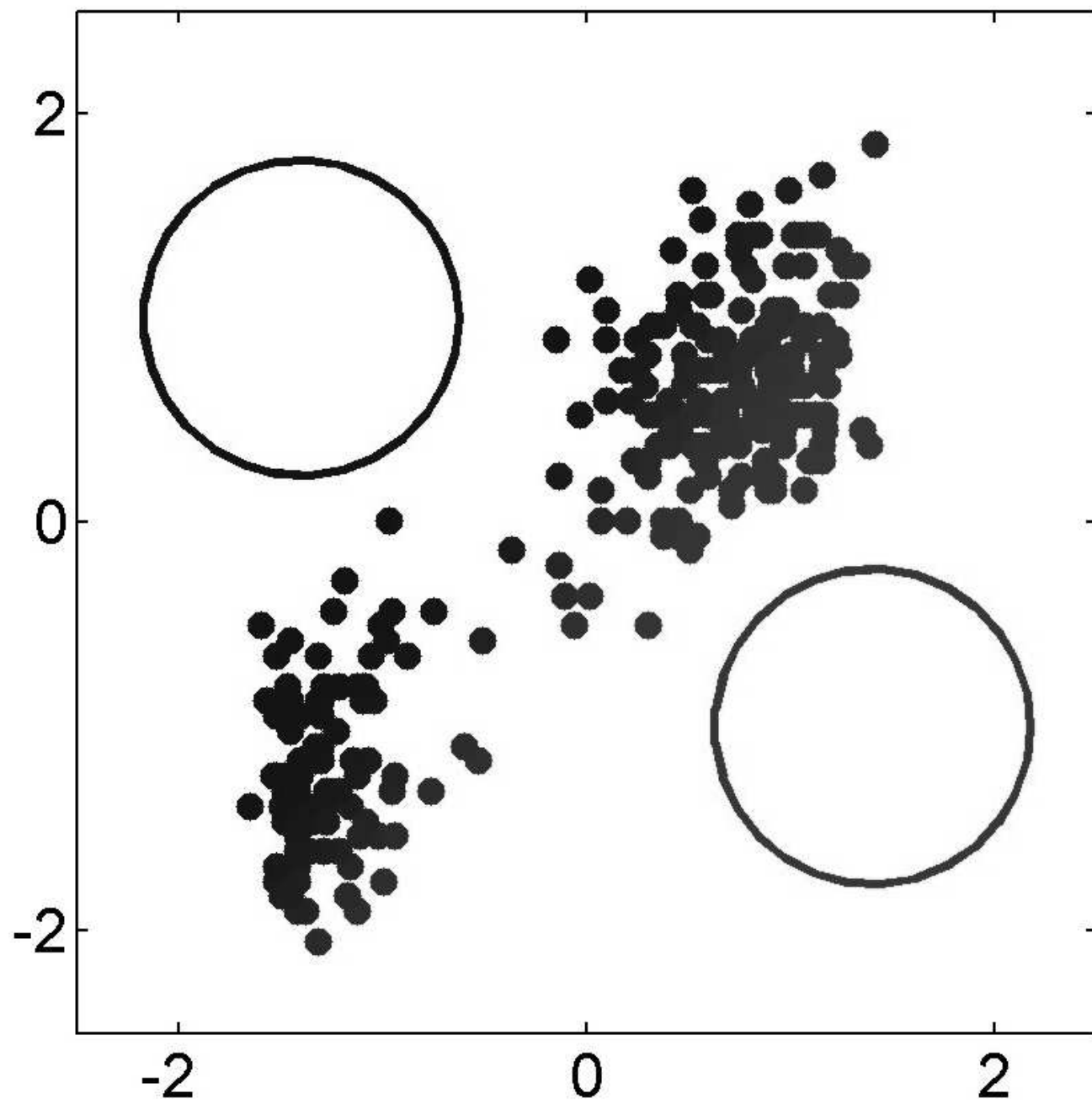
- Pareil pour les covariances

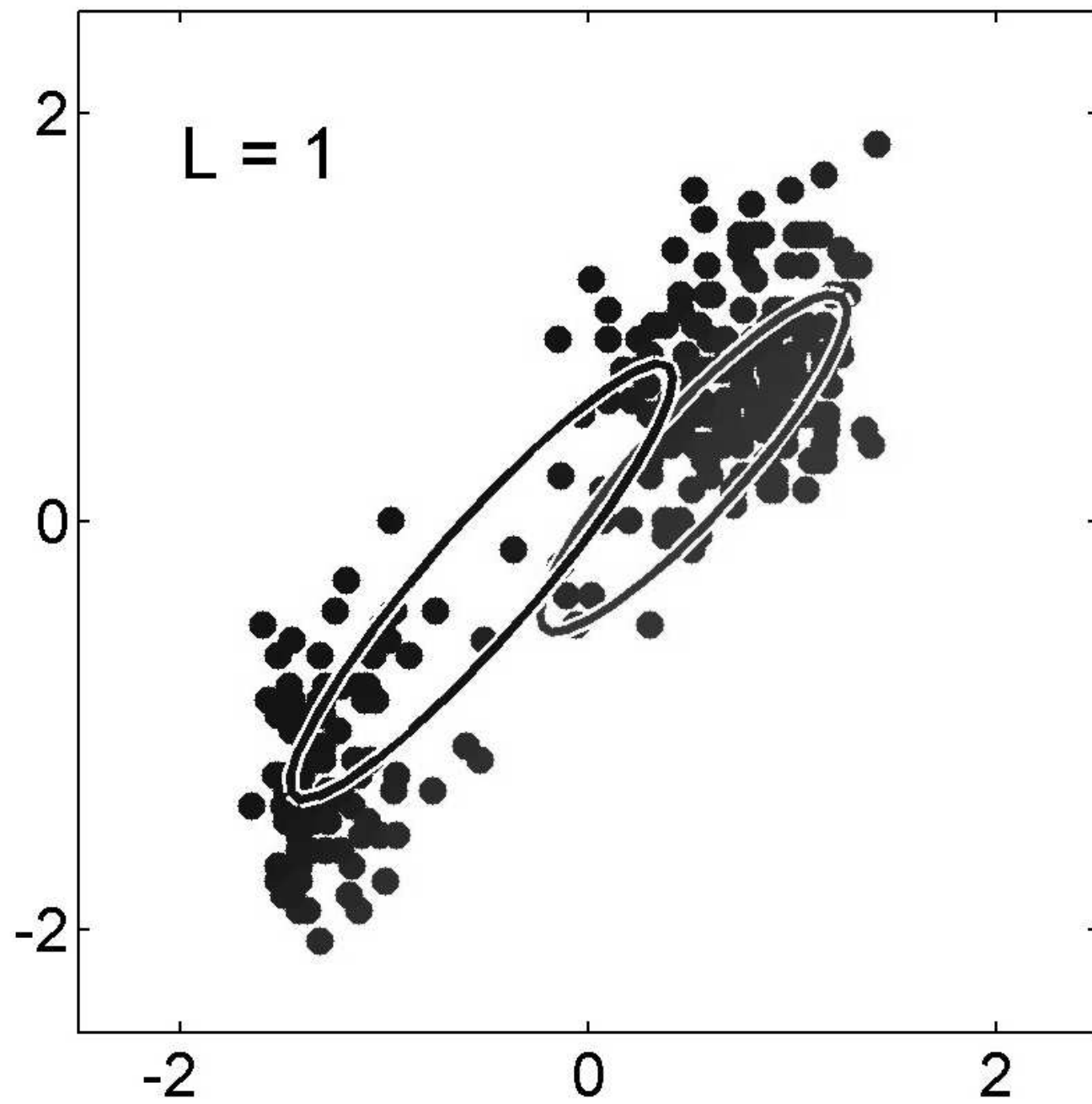
$$\Sigma_j = \frac{\sum_{n=1}^N \gamma_j(\mathbf{x}_n) (\mathbf{x}_n - \boldsymbol{\mu}_j) (\mathbf{x}_n - \boldsymbol{\mu}_j)^\top}{\sum_{n=1}^N \gamma_j(\mathbf{x}_n)}$$

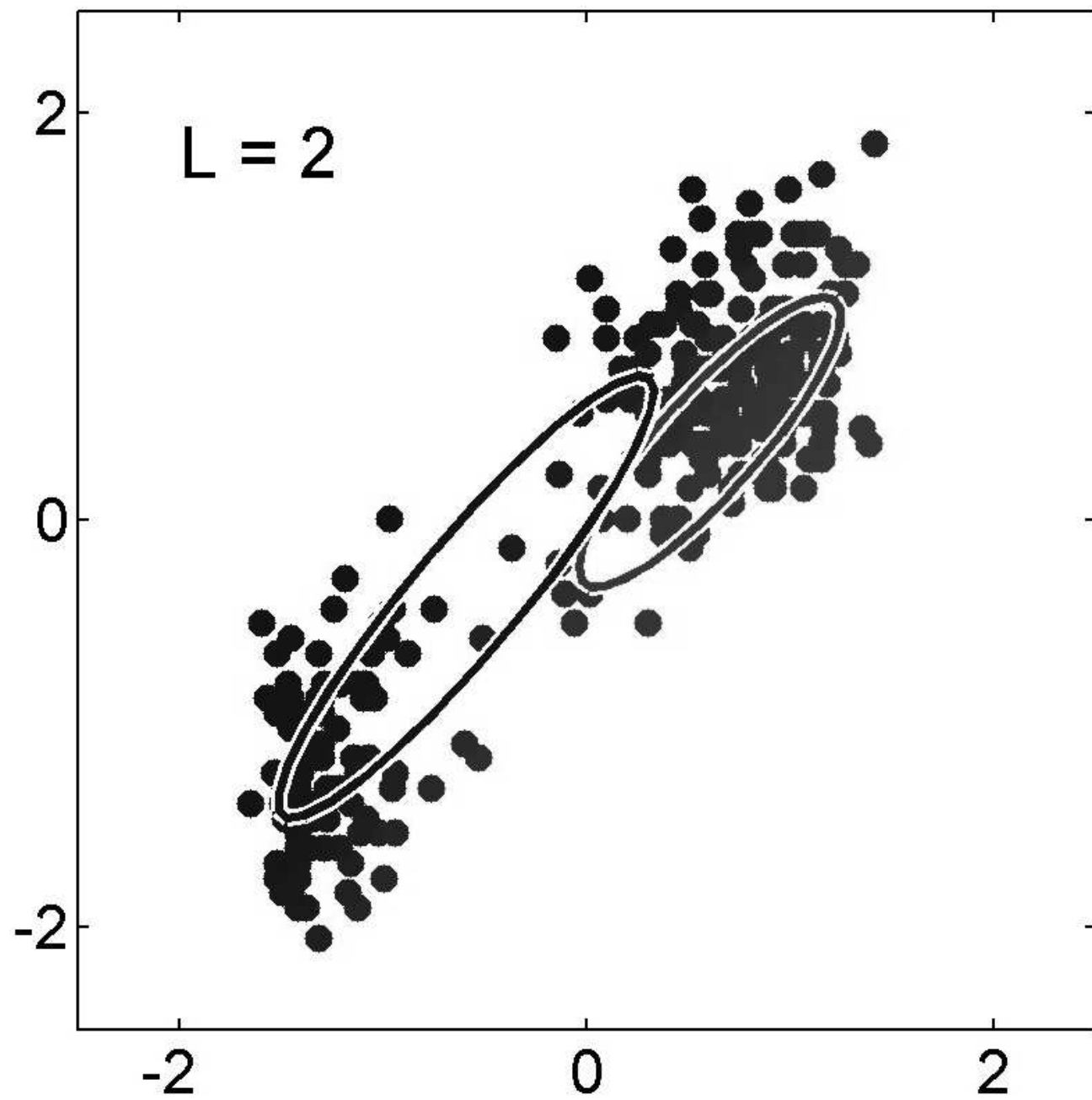
- | .

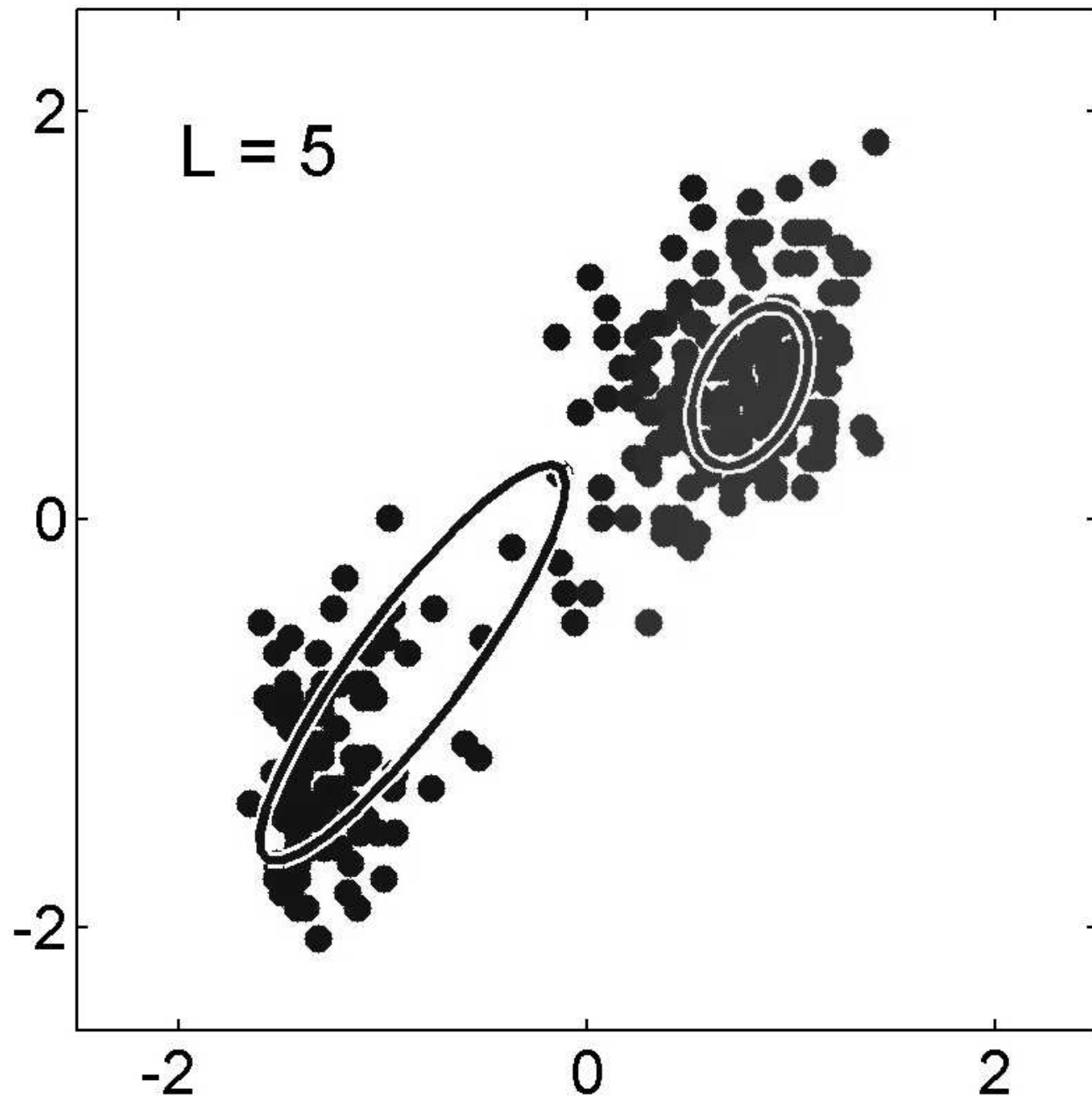
$$\pi_j = \frac{1}{N} \sum_{n=1}^N \gamma_j(\mathbf{x}_n)$$

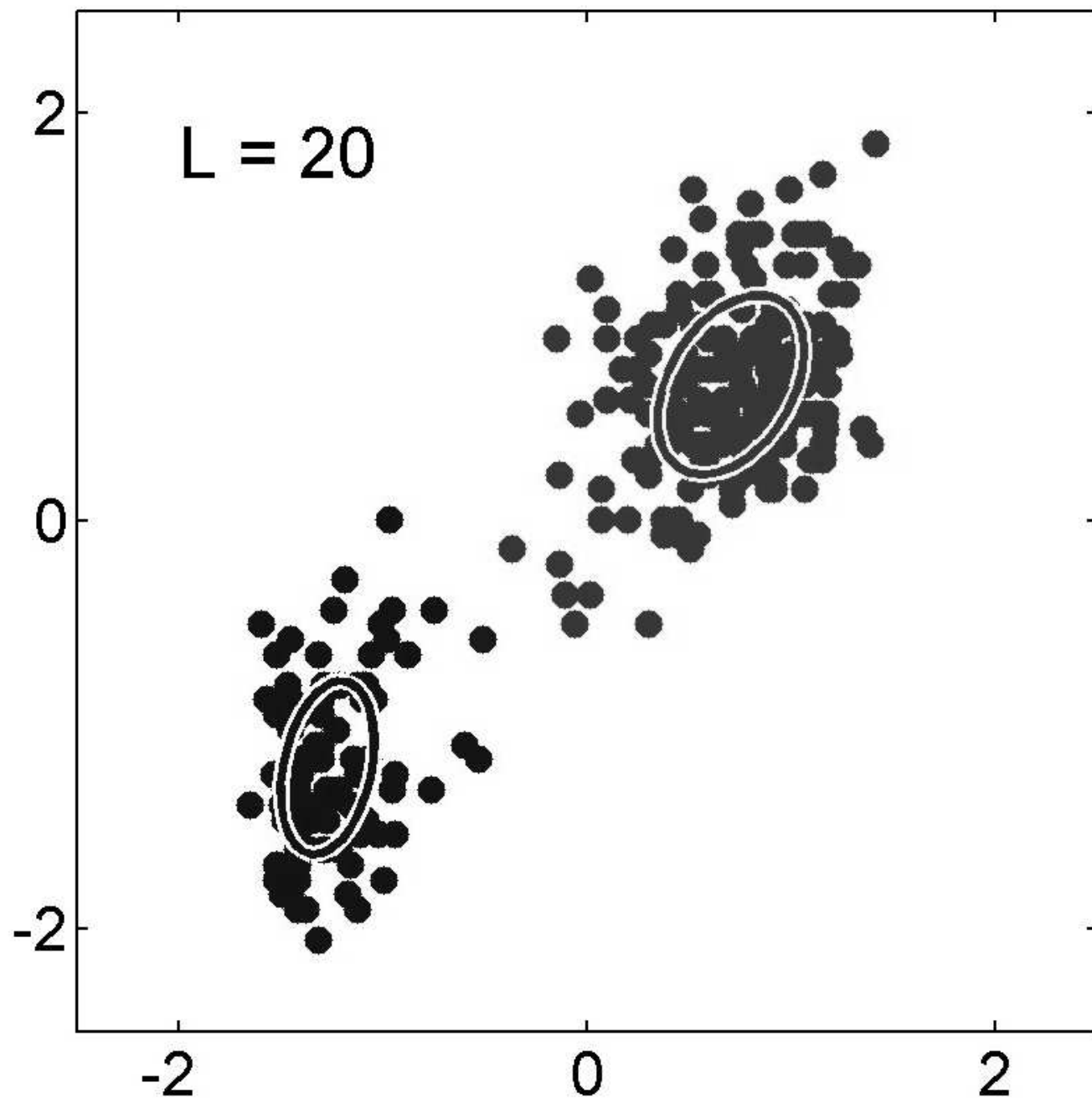














# Modèles de Markov Cachés (HMMs)

[Intro aux HMMs : \(en anglais\)](#)

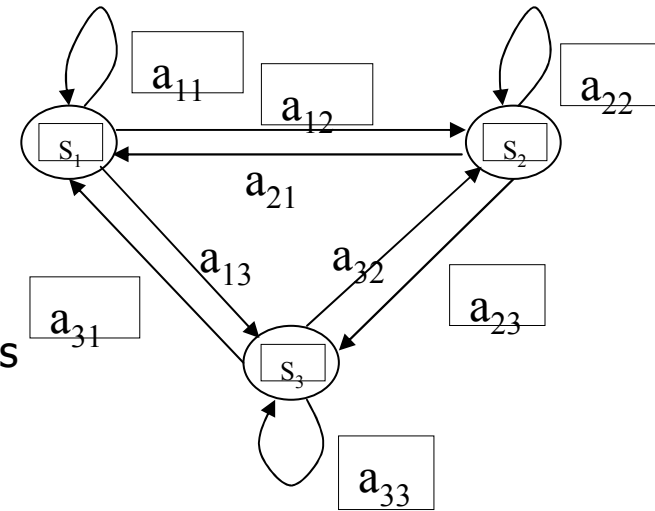
# Modèle de Markov Caché (HMM)

- Un HMM est caractérisée par les paramètres :
  - N, nombre d'états dans le modèle,  $S=\{S_1, S_2, \dots, S_N\}$
  - M, nombre de symboles d'observations pour un état,  $V=\{v_1, v_2, \dots, v_M\}$
  - à chaque état on associe une distribution de probabilités
    - Probabilité de transition  $A=\{a_{ij}\}$ .
    - Probabilité d'observation du symbole k dans l'état j :  $b_{jk}$
    - Probabilité d'états initiaux  $\pi=\{\pi_i\} 1 \forall i \forall N$ .
  
- Si l'ensemble des symboles d'observations V est fini alors on parle de HMM discret (si V infini alors HMM continu).

# HMM pour la reconnaissance de la parole

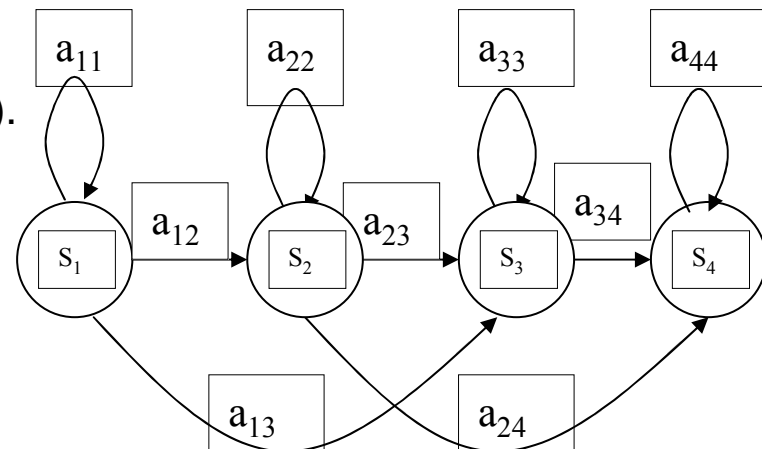
## HMM exemple général

calcul des probabilités de tous les états et de toutes les transitions



Exemple HMM ergodique

- La parole est un phénomène dont les propriétés changent dans le temps
  - Elle peut être modélisée par des HMM gauche-droit (exemple: modèle de Bakis).
- Les propriétés fondamentales du modèle gauche-droit :
  - $a_{ij} = 0$  avec  $j < i$
  - $(\pi_i = 0$  avec  $i \neq 1)$  et  $(\pi_1 = 1$  avec  $i = 1)$
  - $a_{NN} = 1$



Exemple de HMM gauche-droite

# Trois problèmes basiques pour HMM

- A partir d'une suite d'observations  $O$  et d'un HMM  $\lambda$ 
  - comment calculer les probabilités  $P(O|\lambda)$  ?
    - $P(O|\lambda)$  la probabilité de la suite d'observations  $O$  pour le modèle  $\lambda$
  - La solution à ce problème d'**évaluation** est l'algorithme **Forward-Pass**
- A partir de la suite d'observations  $O$  et d'un modèle HMM  $\lambda$ 
  - comment choisir une suite d'états  $Q$  pour maximiser la probabilité  $P(Q|O, \lambda)$  ?
    - $P(Q|O, \lambda)$  la probabilité de la suite  $Q$  pour la suite d'observations  $O$  et modèle  $\lambda$
  - La solution à ce problème de **décodage** est l'algorithme **Viterbi**
- A partir de la suite d'observations  $O$  et d'un modèle  $\lambda$ 
  - comment ajuster les paramètres du modèle pour maximiser la probabilité  $P(O|\lambda)$ ? Ce problème est la phase d'**apprentissage** des paramètres d'un modèle.
  - Algorithme de Baum-Welch, algorithme **EM** (expectation-maximization)

# Algorithme Forward pass (1)

- Cet algorithme est utilisé pour calculer la probabilité d'une séquence de T observations :

$$Y^{(k)} = y_{k_1}, \dots, y_{k_T}$$

- où chaque  $y$  est une des observations. Les probabilités intermédiaires ( $\alpha$ 's) sont calculées récursivement en calculant d'abord, pour chaque état  $j$  et pour  $t=1$ .

$$\alpha_1(j) = \pi(j) \cdot b_{jk_1}$$

# Algorithme Forward pass (2)

- Ensuite, pour  $t=2$  à  $T$ , les probabilités partielles sont calculées pour chaque état:

$$\alpha_{t+1}(j) = \sum_{i=1}^n (\alpha_t(i) a_{ij}) b_{jk_t}$$

- correspondant à la somme des probabilités de chaque chemin menant à l'état considéré (valeurs calculées à l'étape précédente ... récursion), multipliée par la probabilité de l'observation dans l'état considéré
- Finalement, la somme de toutes les probabilités partielles au temps  $T$ , donne la probabilité de l'observation étant donné le modèle HMM

$$Pr(Y^{(k)}) = \sum_{j=1}^n \alpha_T(j) \lambda$$

# Algorithme de Viterbi (1)

- Cet algorithme est utilisé pour calculer la séquence d'états la plus probable  $\mathbf{X}_i = (X_{i_1}, X_{i_2}, \dots, X_{i_T})$  à partir d'une séquence de T observations

$$Y^{(k)} = y_{k_1}, \dots, y_{k_T}$$

- Les probabilités intermédiaires ( $\delta$ 's) sont calculées récursivement en calculant d'abord, pour chaque état  $i$  et pour  $t=1$ .

$$\delta_1(i) = \pi(i)b_{ik_1}$$

# Algorithme Viterbi (2)

- Ensuite, pour  $t=2 \dots T$  et  $i=1 \dots n$ , on calcule

$$\delta_t(i) = \max_j (\delta_{t-1}(j) a_{ji} b_{ik_t})$$

$$\phi_t(i) = \operatorname{argmax}_j (\delta_{t-1}(j) a_{ji})$$

- Alors,  $i_t = \operatorname{argmax}_i (\delta_T(i))$  correspond à l'état le plus probable à temps  $t=T$
- Il reste à revenir en arrière (« *backtracking* ») pour calculer le chemin le plus probable. On calcule pour  $t=T-1 \dots 1$

$$i_t = \phi_{t+1}(i_{t+1})$$

# Algorithme de Viterbi (3)

- $\hat{i}_1, \hat{i}_2, \dots, \hat{i}_T$ , est la séquence d'états cachés la plus probable, pour l'observation et le HMM donné.
- *Remarque* : le calcul de  $\hat{i}_t$  est similaire au calcul de  $\alpha_t$  pour l'algorithme *forward*. La seule différence est que la **somme** de l'algorithme *forward* est remplacée par le **max** dans l'algorithme de *Viterbi*.
- Pour *Viterbi*, on s'intéresse donc au chemin le plus probable, plutôt qu'à la probabilité totale.

# Algorithme Baum-Welch

- Le jeu optimal de paramètres  $\hat{\lambda}$  du HMM est obtenu de manière itérative par cet algo.
- Les formules de réestimation de l'algo. permettent d'obtenir à partir des param.  $\lambda_i$  de l'itération  $i$ , les nouvelles valeurs qui améliorent la vraisemblance des observations  $X$  du corpus d'apprentissage  
 $\lambda_{i+1} = F(\lambda_i)$

$$\prod_X P(X / \lambda_{i+1}) \geq \prod_X P(X / \lambda_i)$$



# Modèles acoustiques de phonèmes

- Généralement, les unités acoustiques modélisées sont de phonèmes, plutôt que des mots
  - Exemple : 46 modèles de phonèmes pour le français



# Modèles Dépendants vs. Indépendants du contexte

- Indépendants : on modélise chaque unité indépendamment les unes des autres
- Dépendants : on modélise différemment les unités suivant les unités de droite et de gauche
- Modèles triphones : on modélise un phonème en tenant compte du phonème de gauche et du phonème de droite



# Problématique

- Constitution d'un modèle acoustique : apprentissage nécessite beaucoup de données
- Modèle acoustique dépendant du contexte : il faut beaucoup d'exemples de chacun des contextes.
- Corpus de parole disponibles trop petits



# Coarticulation

- Pourquoi un modèle dépendant du contexte ?
- Observation : la prononciation d'un phonème dépend des phonèmes qui le précèdent et qui le suivent
- Phénomène de coarticulation



# Clustering

- Nombre de contextes important :  
42 phonèmes → 70 000 triphones
- Regroupement des contextes en classes  
→ clustering
- Construction d'un modèle par classe