Laurent.Besacier@imag.fr

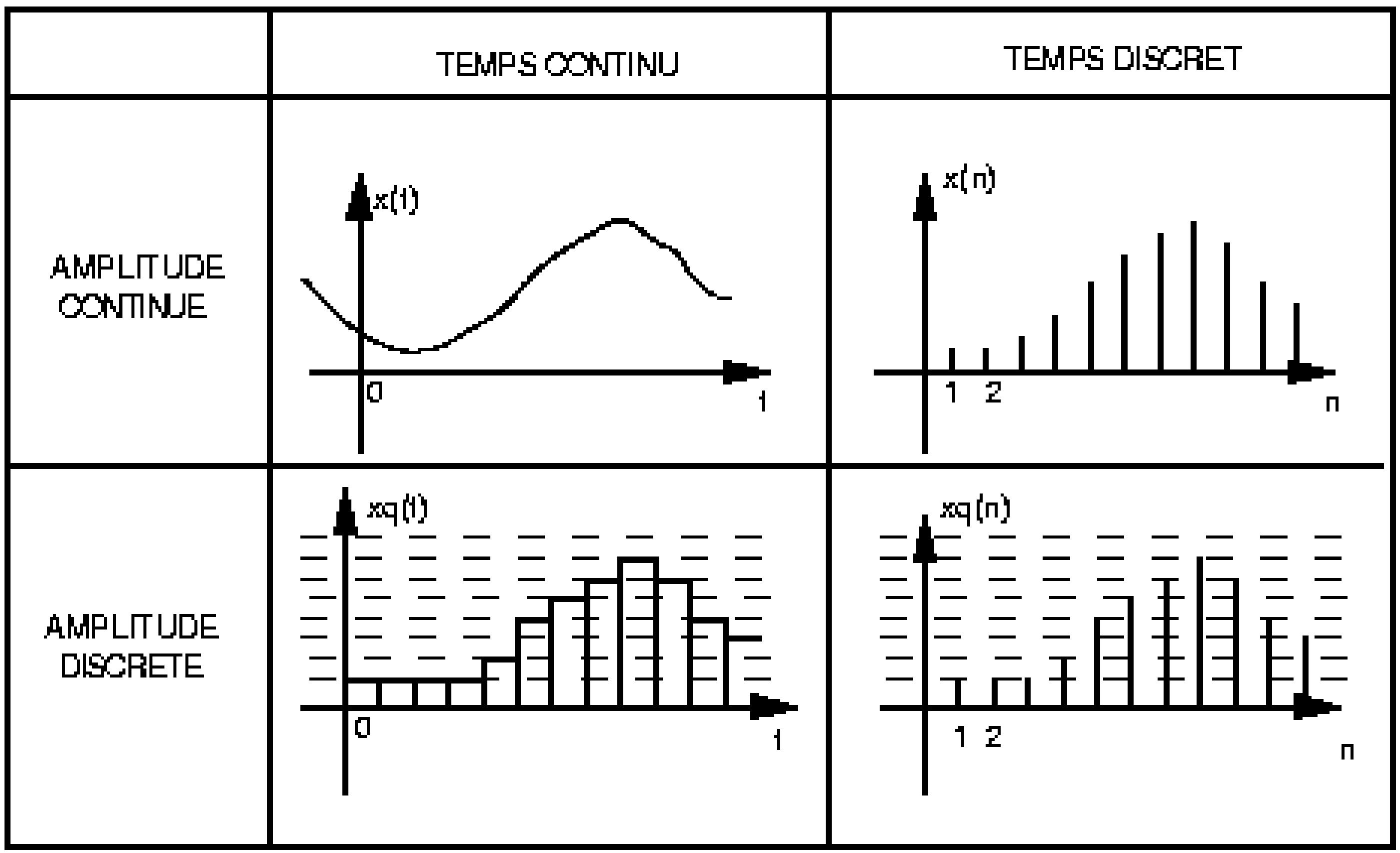
- 2. Speech and voice technologies
  - DSP reminder
  - The speech signal
  - Speech Technologies
    - Overview
    - Modelling (parameters, models)

Digital Signal Processing Reminder

Bibliography

- 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

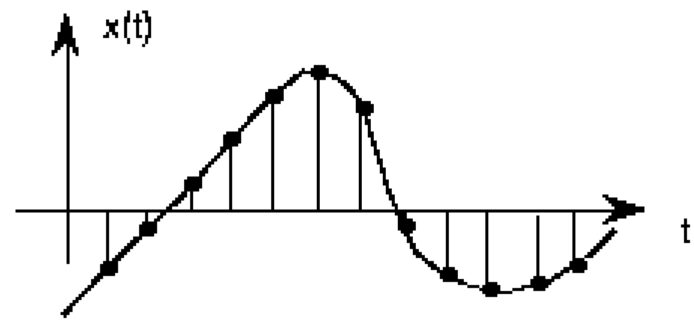
Digital / Analogic Signals



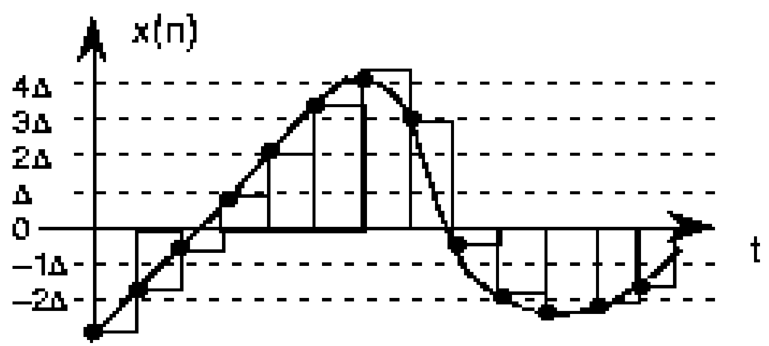
Analog-Digital Conversion

- **Sampling** and **Quantification**

sampling



quantification



Energy of a signal

- *Continuous signal s(t) on* [t1,t2]

$$W_s(t_1,t_2) = \int_{t1}^{t2} s^2(t)dt$$

- Discrete signal *x(n)* on [n1,n2]

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

SNR : Signal-to-Noise Ratio

- For $x(t)=s(t)+n(t)$

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

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

$$SNR_{dB} = 20\log_{10} \frac{Amplitude_s}{Amplitude_n}$$

Fourier Transform - TF

- Spectral representation of signals
- Core mathematical tool in DSP

TF for continuous signals

- x(t) signal
- TF is a function of variable $\omega = 2\pi f$ defined by :

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

- Inverse transform

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

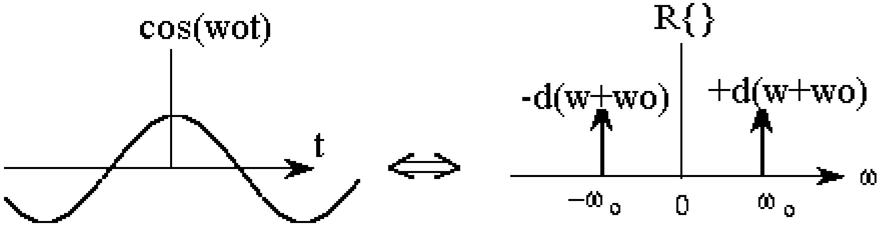
(2-D)

$$\mathcal{F}\{g(x,y)\} = G(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)e^{-j(ux+vy)}dxdy$$

$$g(x,y) = \mathcal{F}^{-1}\{G(u,v)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} G(u,v)e^{j(ux+vy)}dudv$$

TF of a periodic signal (cos)

$$\cos(w_0 t) \overset{TF}{\longleftrightarrow} \frac{1}{2}\left[\delta(w - w_0) + \delta(w + w_0)\right]$$

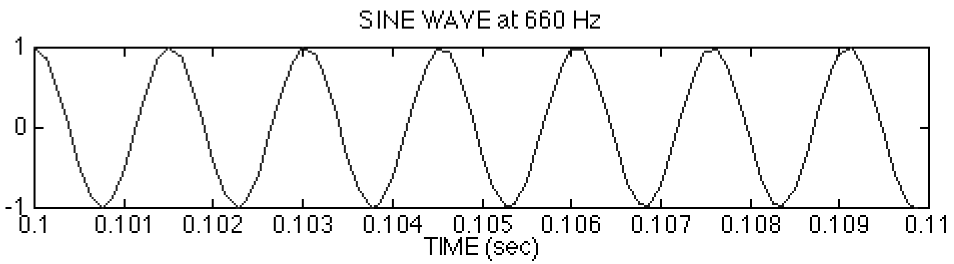


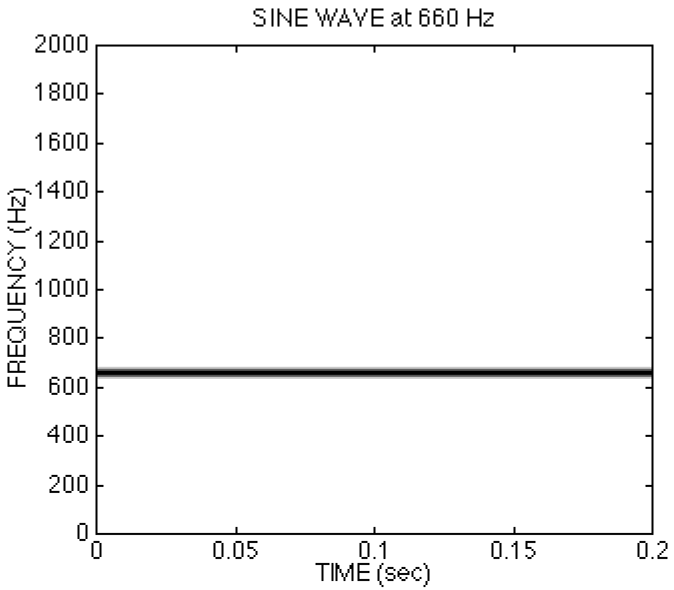
Time-frequency representation

Time-frequency representation Spectrogram

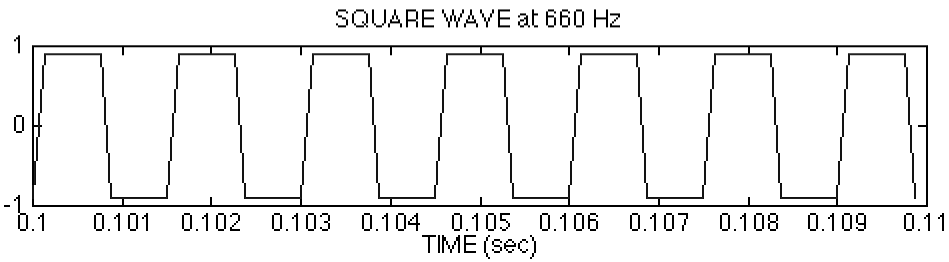
$$S_x(t,f)=\left|\int_{-\infty}^{+\infty} x(s)h^*(s-t)e^{-i2\pi fs}\,ds\right|^2$$

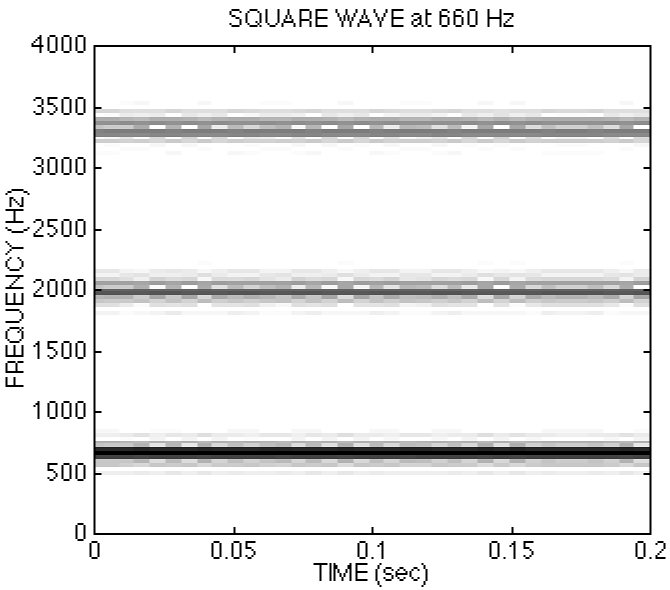
Case of a sinus



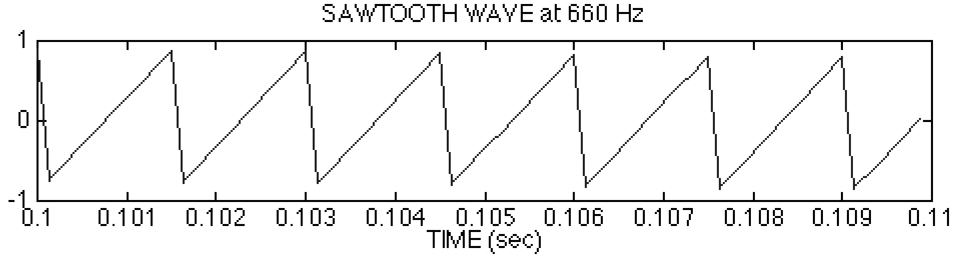


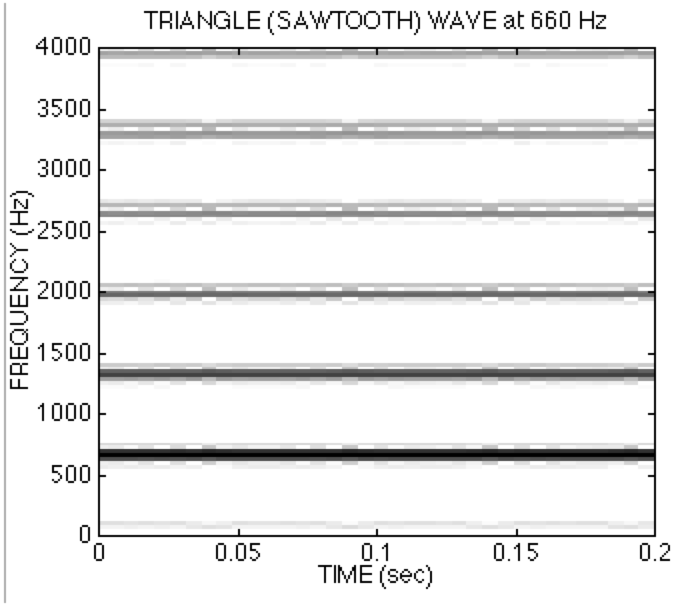
Case of square signal



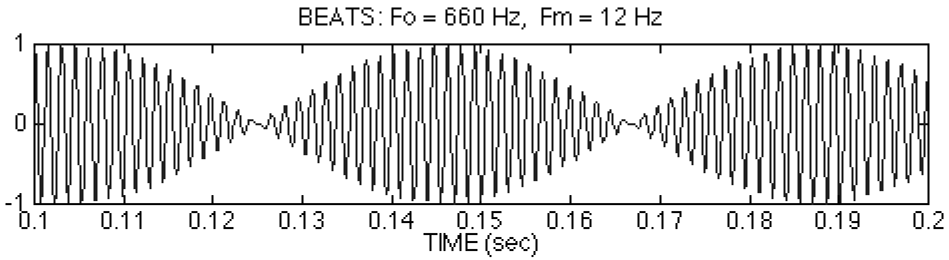


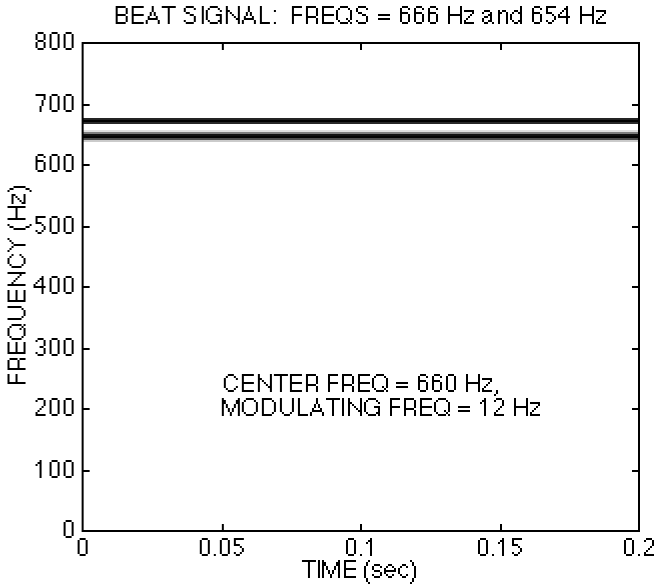
Sawtooth signal



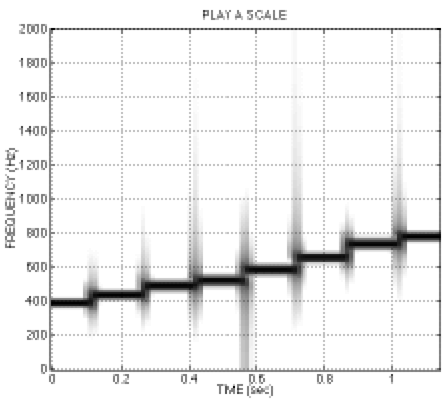


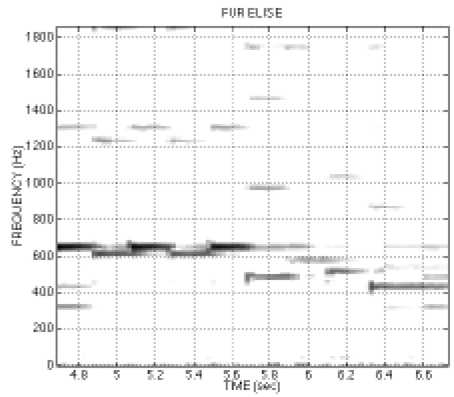
Beat Signals

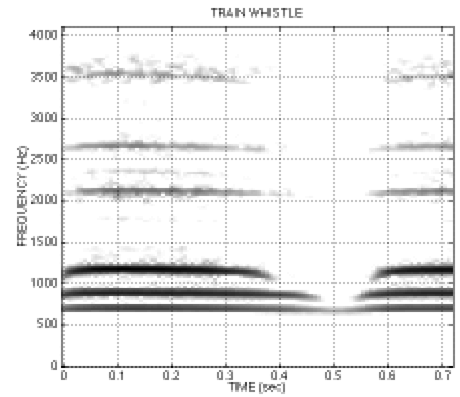


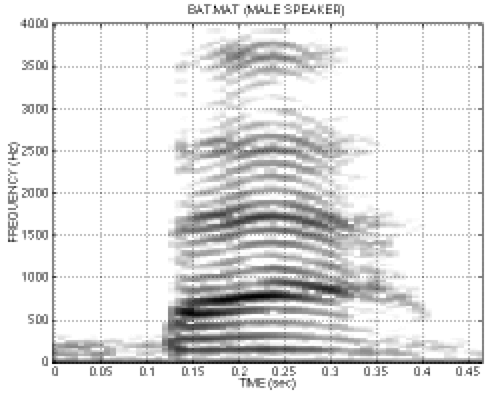


Other examples

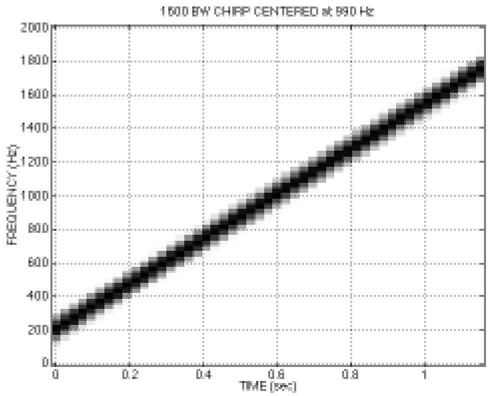


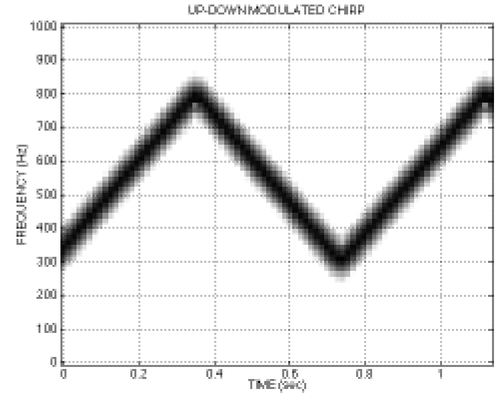


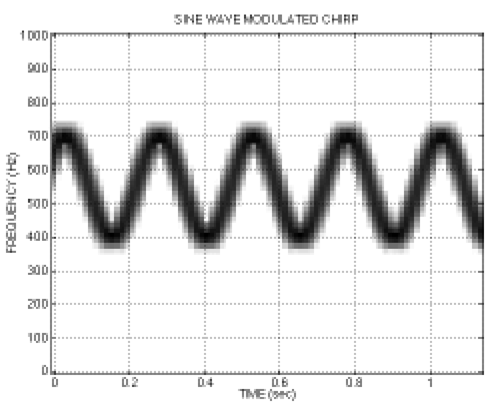


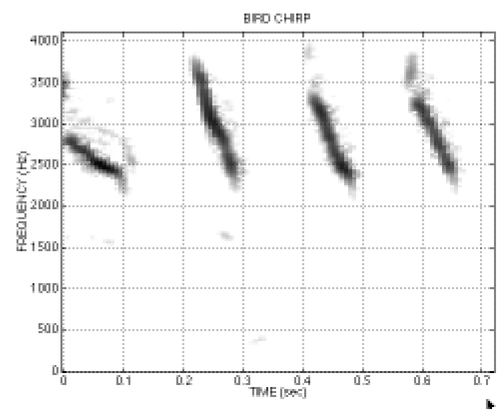


Chirps


1600 BW CHIRP CENTERED at 990 Hz
FREQUENCY (Hz)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
0.2
0.4
0.6
0.8
1
TIME (sec)


UP-DOWN MODULATED CHIRP
FREQUENCY (Hz)
1000
900
800
700
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
TIME (sec)


SINE WAVE MODULATED CHIRP
FREQUENCY (Hz)
1000
900
800
700
600
500
400
300
200
100
0
0
0.2
0.4
0.6
0.8
1
TIME (sec)


BIRD CHIRP
FREQUENCY (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (sec)

The Speech Signal

Speech : one modality among others ?

- From the user point of vue
  - As ouput
    - *Visual (2D,3D)*
    - **Sound/Speech**
    - …
  - As input
    - **Spoken command**
    - Gesture
    - …

Speech in HMI : ergonomic aspects

- Communication Mean
  - natural
  - fast and concise
  - Hands free
  - Gaze can be used for other task
  - Helpful for disabled people

But…

- Critical Situations
  - degradation of performances
- Training of the Machine Needed
  - Might be long, constraints (do you know IBM ViaVoice?)
- Linguistic constraints
  - Vocabulary size
  - Syntactic structure
  - Mode (isolated words, connected words, natural spoken language, disfluencies)
- Human factors
  - micro + headset => acceptability of the user
  - Privacy issues

Goals

- Satisfaction of the user
- Quality
  - System delay
  - Robustness
    - Need to have plan B (or C, or D…)
- Cost
  - Use of speech must have a limited impact on the costs

Avoid…

- Full speech interface
  - It is just another modality
- Complicated systems
  - ergonomy
- Annoy user
  - Robustness needed

Speech technologies

- Speech compression / coding (*wireless*, *IP*,…)
- Speech synthesis / recognition for dialog systems (information access…)
- Speaker authentication (voice biometrics)

- More details will be given in my last course…

Speech sounds : phonemes

- Establish distinctive units of meaning
- Phonemes are the shortest sound units in speech that allow to distinguish different words
- Examples [p] [b]
  - pas / bas
  - paie / baie
  - pot / beau

French phonemes

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	[ɥ] 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	[ɔ̃] 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]).

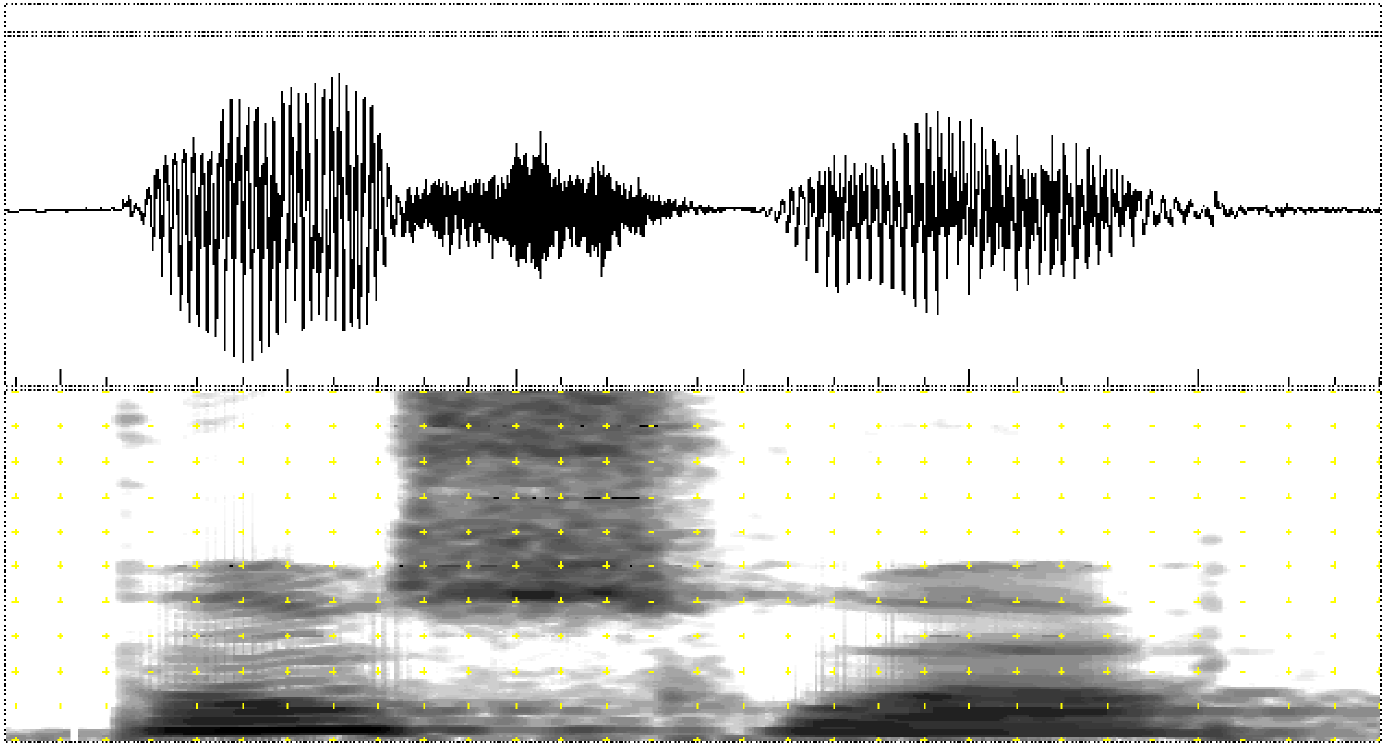
French phonemes

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

CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales	← Lieu d'articulation
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]	[ɥ]	[j]	
Liquides		[l]	[R]	

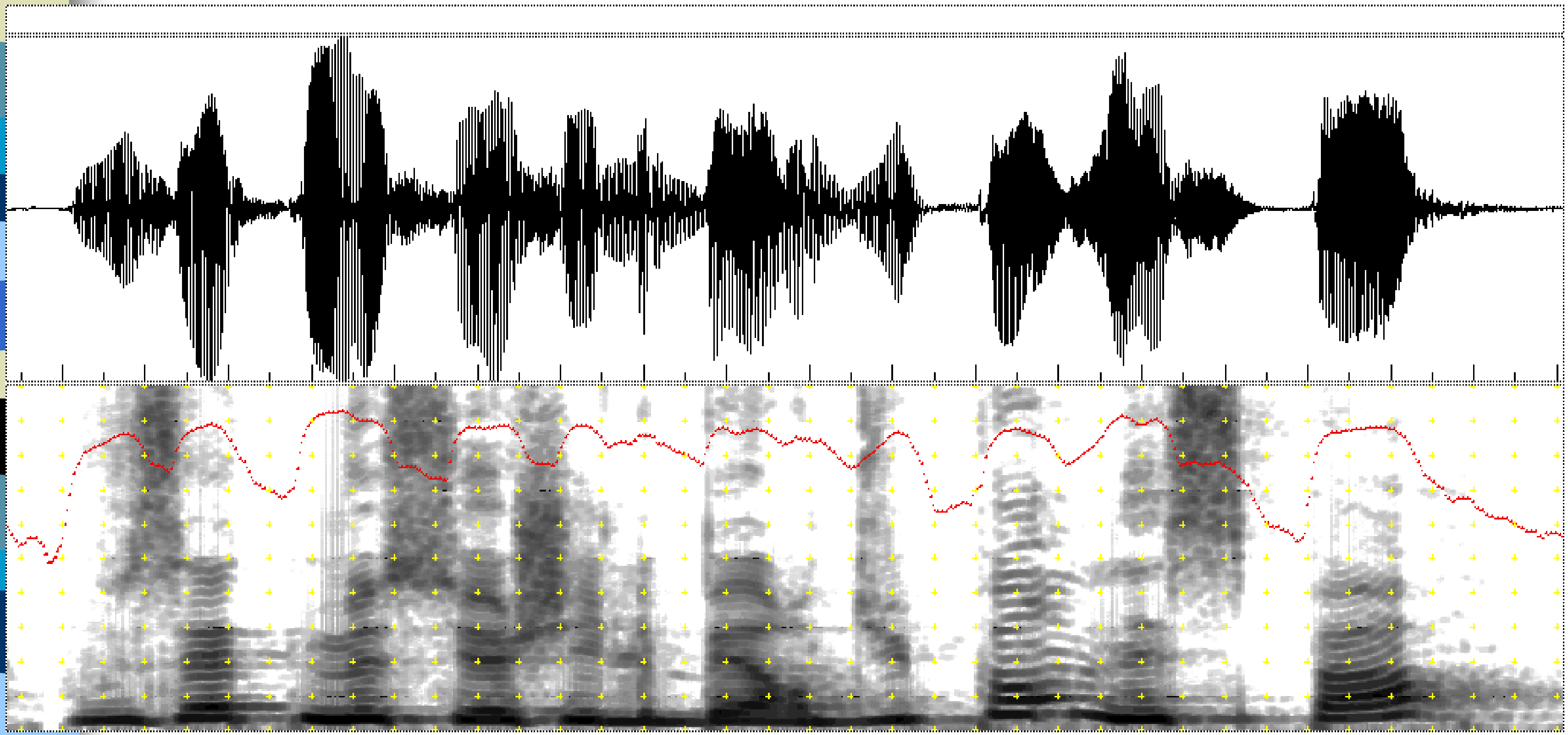
VOYELLES			
Orales	Antérieures		Postérieures
	Non arrondies	Arrondies	
Fermées	[i]	[y]	[u]
	[e]	[ø]	[o]
	[ɛ]	[œ]	[ɔ]
Ouvertes	[a]		
Nasales	Antérieures		Postérieures
Fermées	[ɛ̃]		[õ]
Ouvertes		[ã]	

Speech signal



Bonsoir

Speech signal



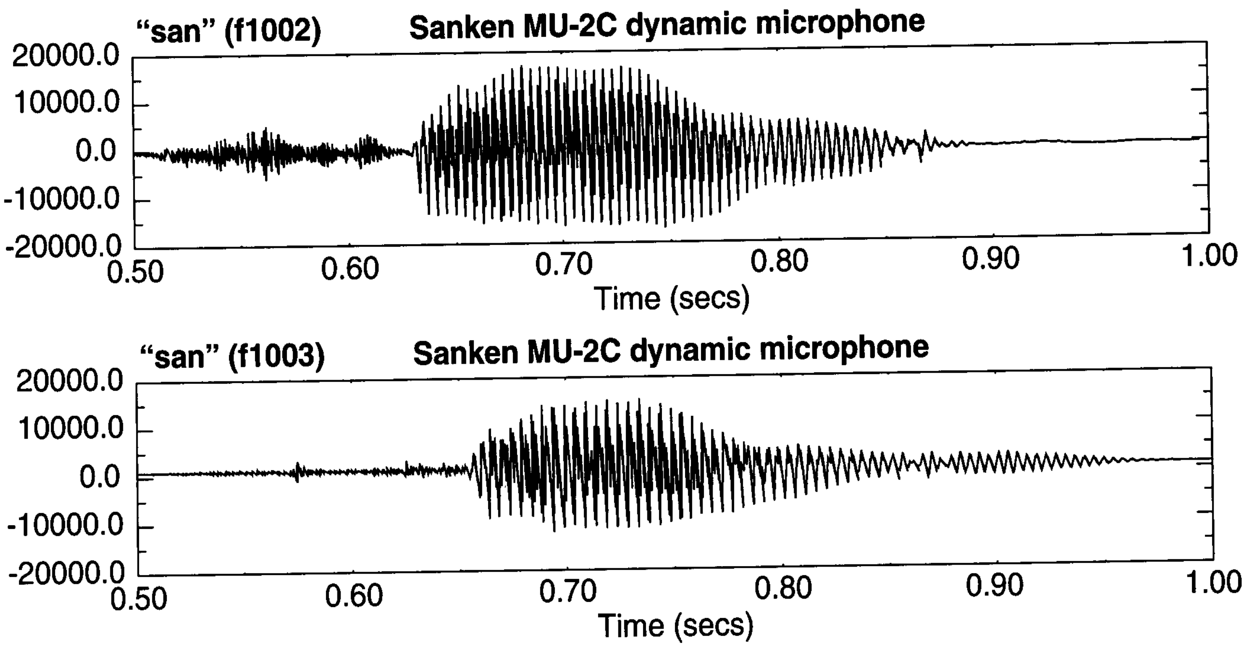
Vous êtes Monsieur Gilbert Dupont n’est-ce pas ?

Variability

- First caractersitic of speech signals is variability
  - A speaker nevers pronounces two times the exact same sound
  - Two speakers do not pronounce the same sound the same way
- However, this sound is perfectly recognized by a human listener

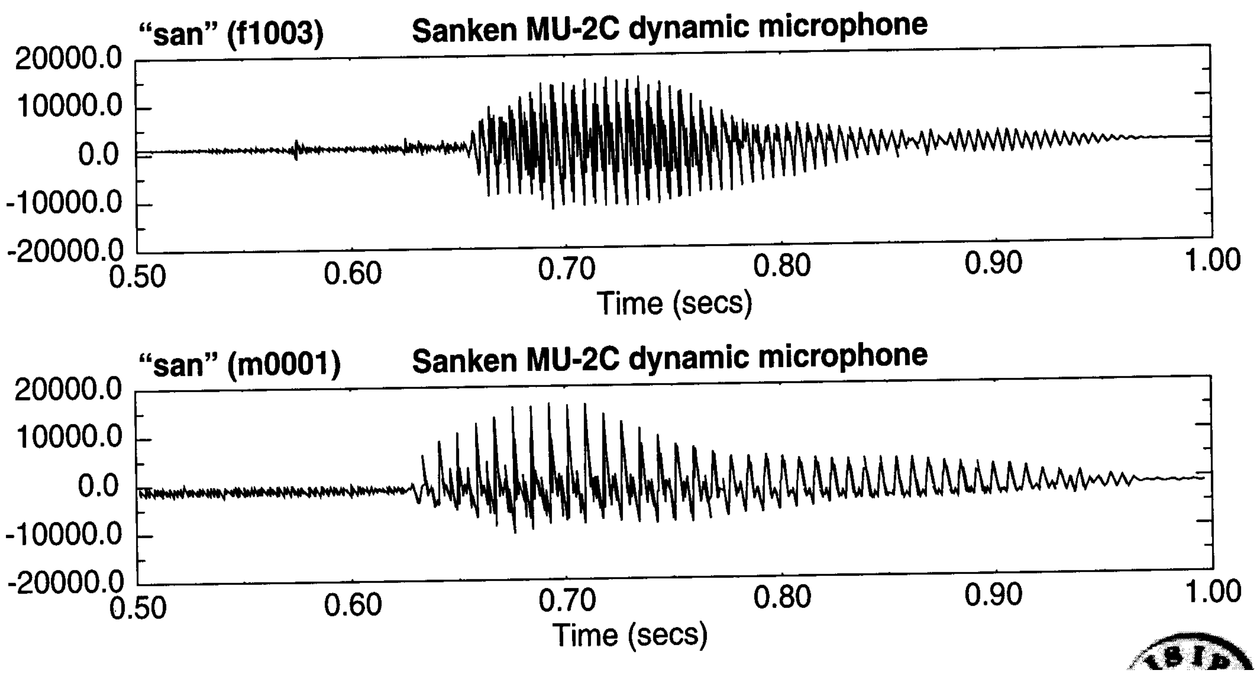
Intra-Speaker variability

Same sounds, same person,
same recording conditions



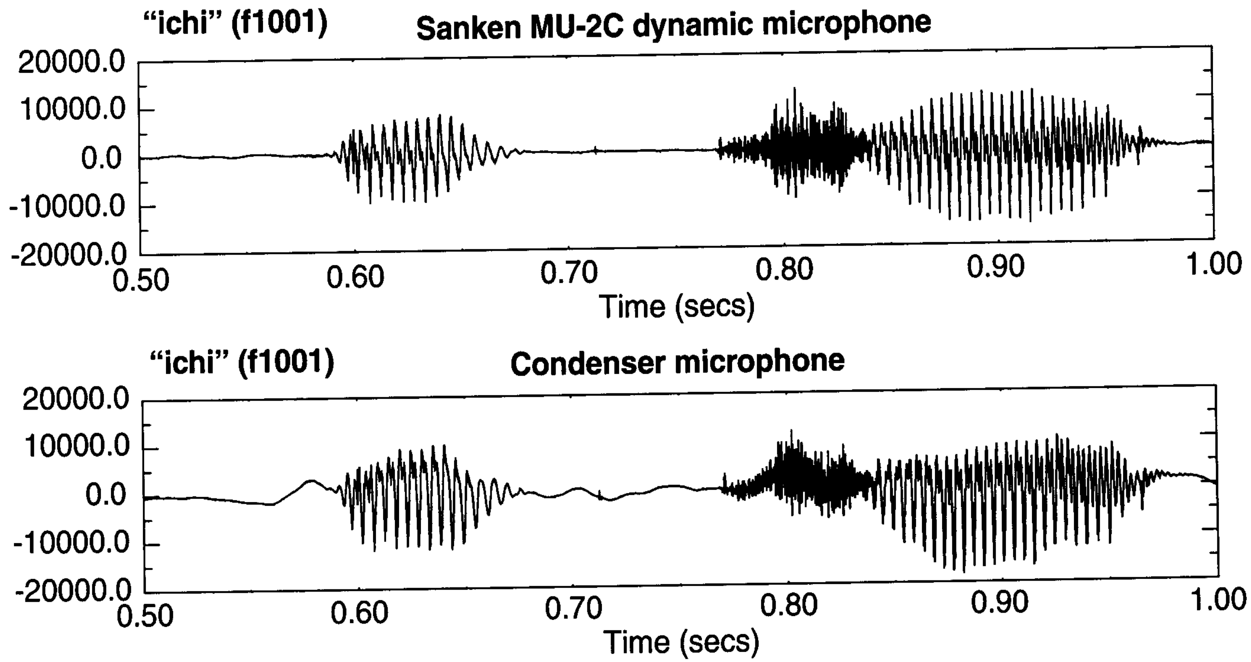
Inter-speaker variability

Same sounds, different persons,
same recording conditions

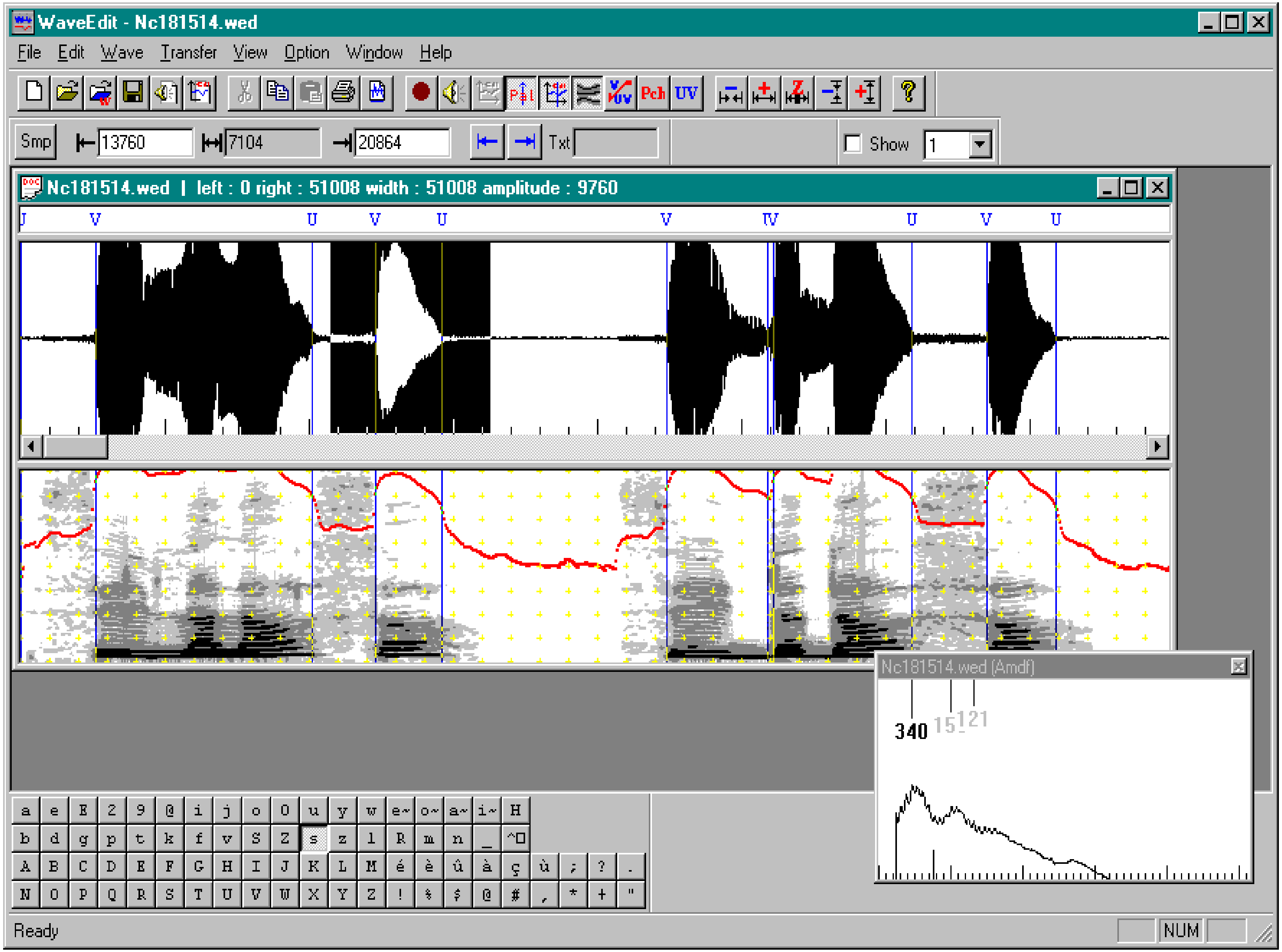


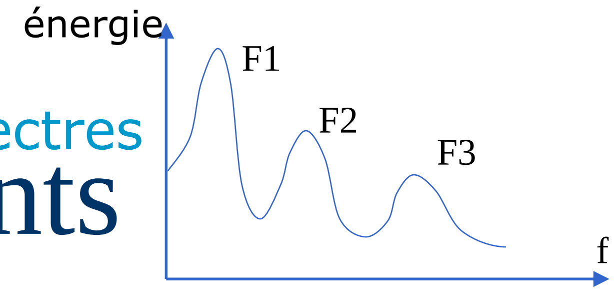
Variability due to recording conditions

Same sounds, same person,
different microphones



Tools for speech analysis

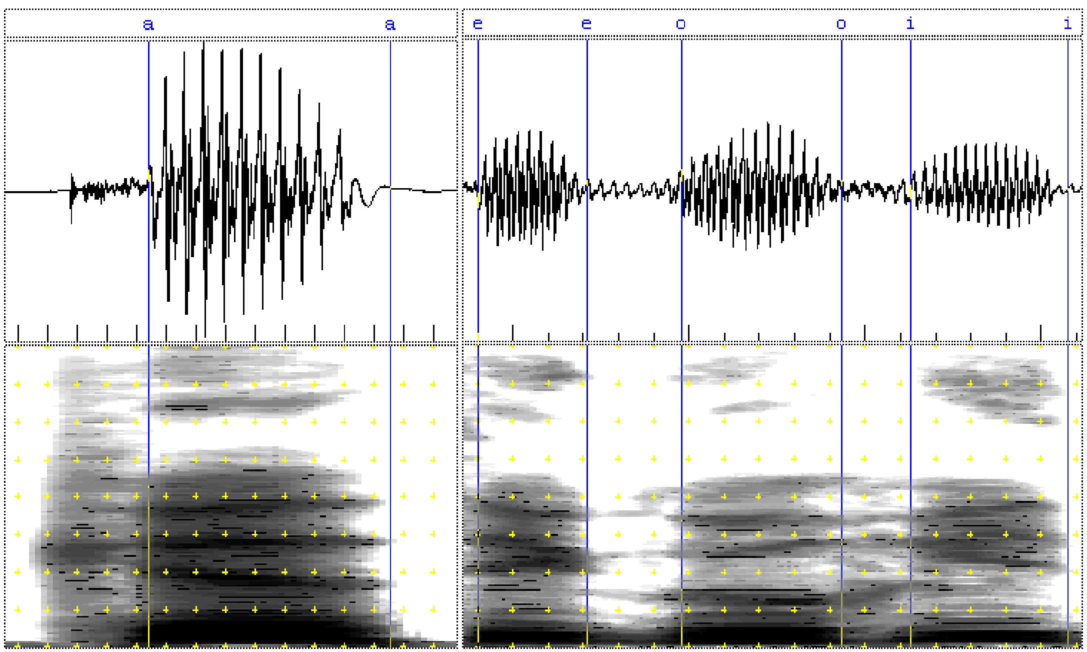


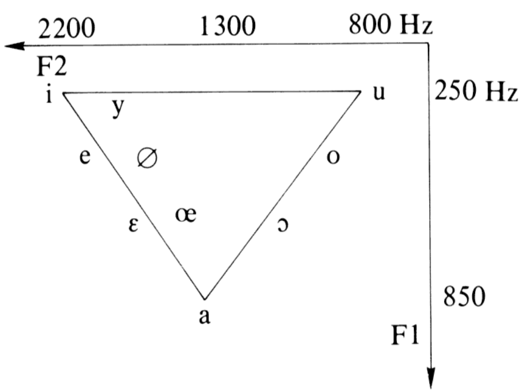


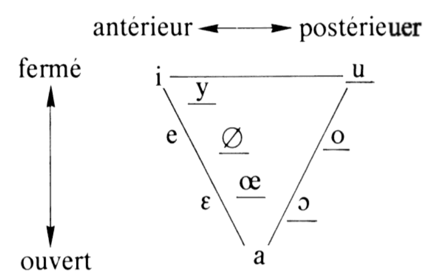
spectres

Formants

- The spectral peaks of the sound spectrum |P(f)|' of the voice
- Acoustic resonance of the human vocal tract
- Vocal triangle for vowels







Formants

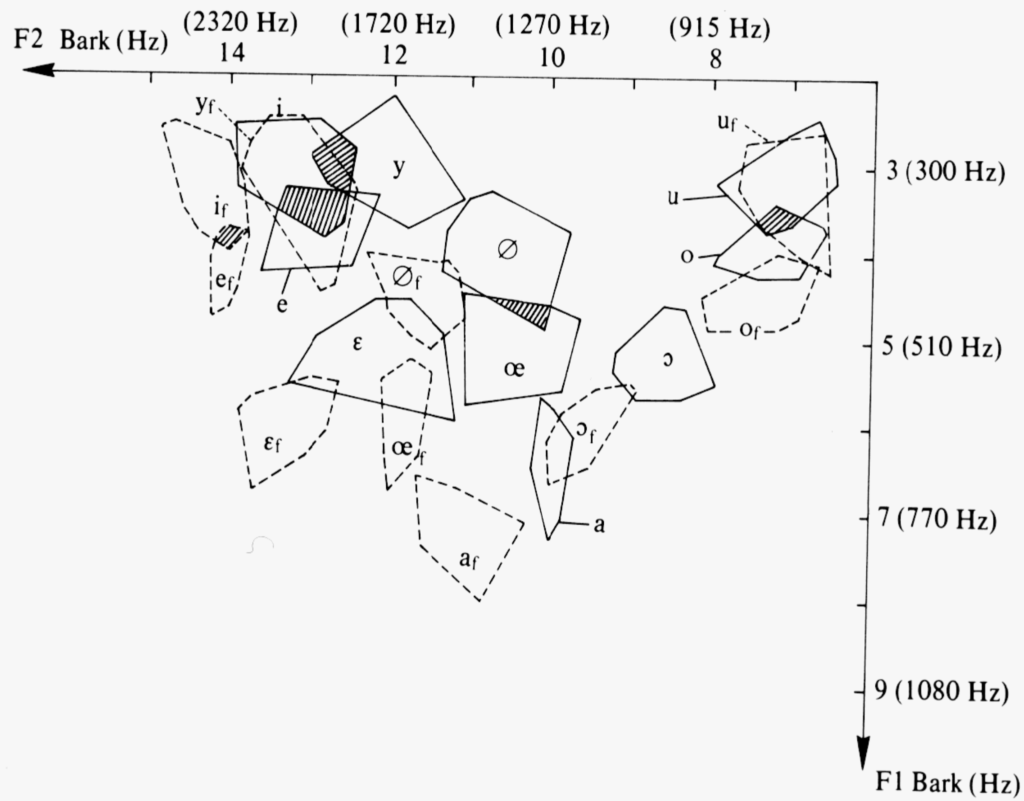


FIG. III.6. — *Zones de dispersion des voyelles non nasales du français sur le plan F_1/F_2* (échelle de Bark).

Sujets masculins : ————

Sujets féminins : ----------

Les hachures délimitent les zones de recouvrement pour un même sexe.

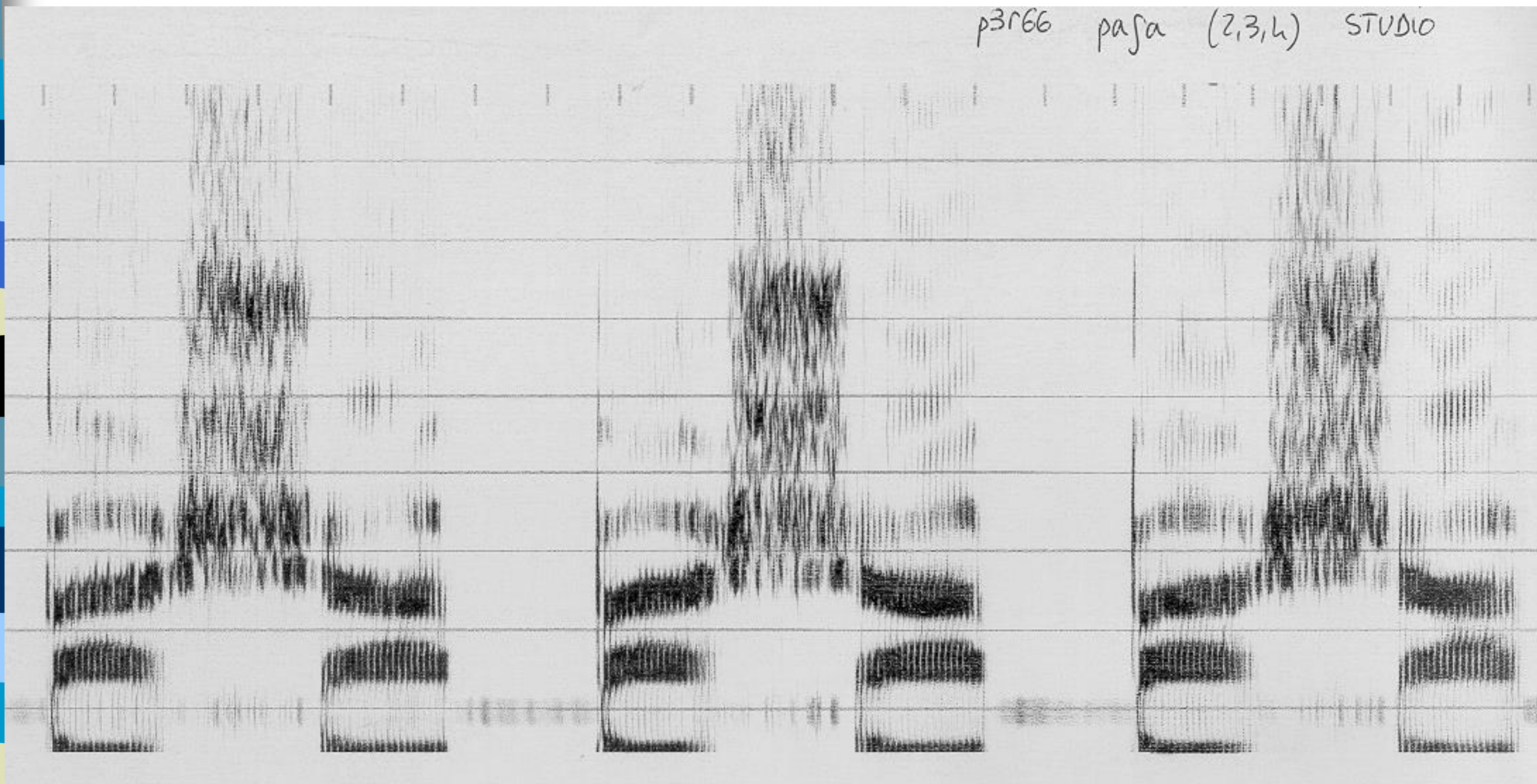
Variability of speech and speakers

Formants : values for french vowels

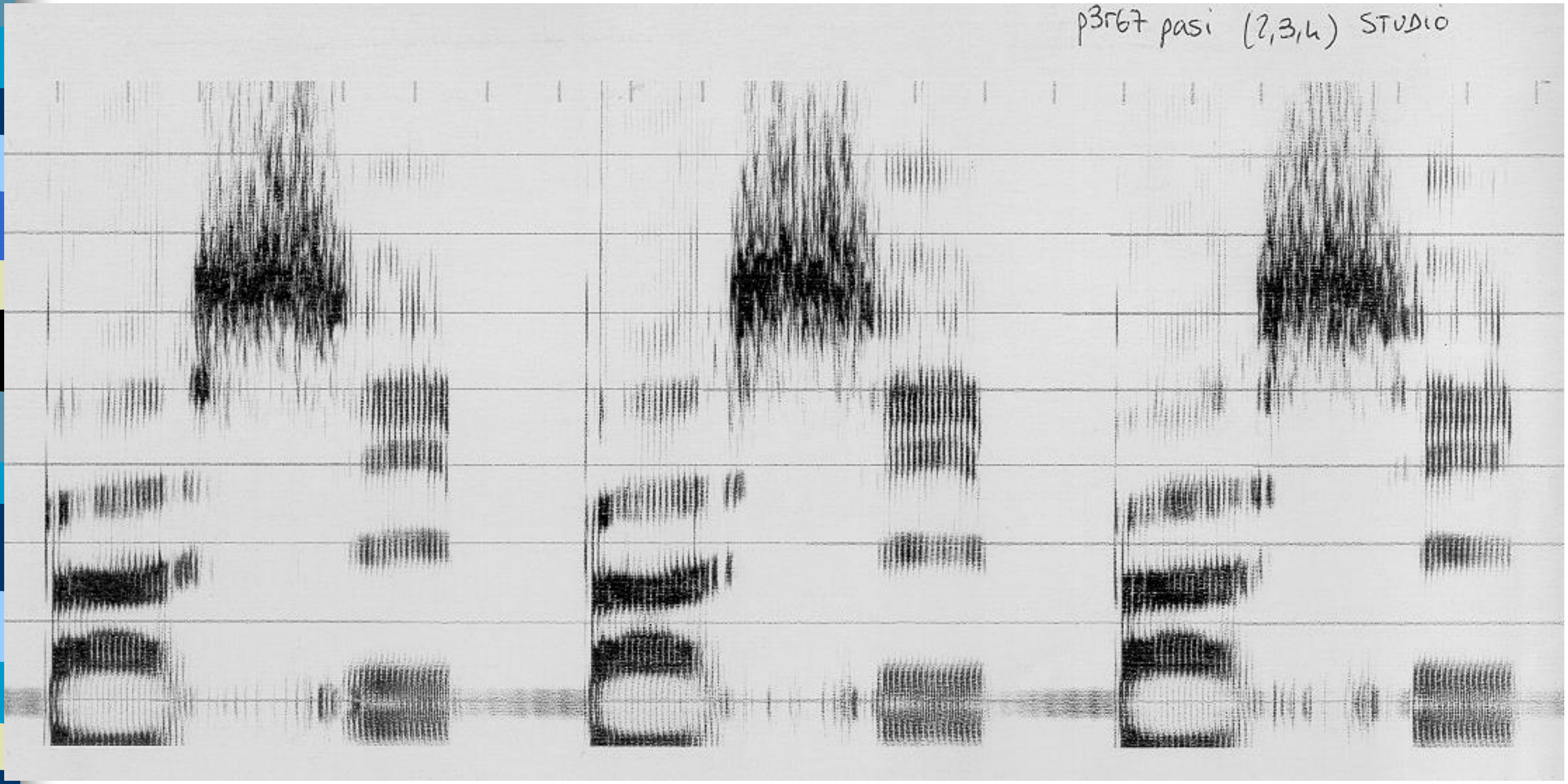
TABLEAU III.1. — *Valeur en Hz des 4 premières fréquences formantiques. 10 sujets masculins ; 9 sujets féminins — Analyse LPC. Med. : valeur de la médiane ; σ : écart-type ; Ki : coefficient d'écart entre formants masculins et féminins : (Ki = Fi♀/Fi♂). Corpus : [pV] ou [pVR], 2 répétitions par sujet.*

voyelles		Sujets masculins F_1	F_2	F_3	F_4		Sujets féminins F_1	F_2	F_3	F_4
i	Med.	308	2064	2976	3407	Hz	306	2456	3389	3966
	σ	34	134	147	208	Hz	42	111	68	169
	Ki	0,99	1,19	1,14	1,16	%				
e	Med	365	1961	2644	3362		417	2351	3128	4161
	σ	31	119	107	155		31	52	115	121
	Ki	1,14	1,20	1,18	1,24					
ε	Med	530	1718	2558	3300		660	2080	2954	4231
	σ	49	132	103	221		46	108	156	210
	Ki	1,25	1,21	1,15	1,28					
ɑ	Med	684	1256	2503	3262		788	1503	2737	4143
	σ	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
	σ	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
	σ	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
	σ	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
	σ	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
	σ	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
	σ	42	94	91	149		58	47	155	202
	Ki	1,25	1,21	1,16	1,20					

Spectrogram



Spectrogram (large band) - 2



Prosody

- Pitch or fundamental frequency
- Voice energy
- Syllable / phoneme duration

→ **Intonation / Voice « melody »**

Fundamental frequency *f0*

- Vibration of the vocal cords
- Depnds on speaker age and sex
  - 100 à 150 Hz for adult male speaker
  - 140 à 240 Hz for adult female speaker

  - Can have huge variations for a single speaker
    - Depending on the type of sentence uttered
    - Depending on the emotional / affect of the speaker

Practical lab (optional)

- Speech Analysis Tools
- http://www-clips.imag.fr/geod/User/laurent.besacier/NEW-TPs/TP-CL/tp7.html

Speech technologies

Overview

Modelling (parameters, stochastic models HMMs)

Speech, a source of informations

Speech

Linguistic informations
(what is uttered)

Extra-linguistic info.
(speaker, language, speaker state)

Linguistic Informations

- What is uttered by the speaker….

Automatic Speech Recognition (ASR)

Different levels of difficulty

- **Number of speakers** : systems mono-speakers …until multi-speakers
- **Vocabulary size**
- **Transmission channel** : «direct mic. », téléphone, mobile phone, VoIP

Different levels of difficulty

- **Acoustic Environment :** quiet**,** normal (officeroom), noisy (train station, street), extreme (plane cockpit)
- **Speaking style** : digits, isolated words, connected words, continuous speech (read, spontaneous)
- 1 person or conversation

Applications

- **Services** (vocal servers)
- **Vocal terminals** (on site)
- **Transports** (vocal commands help, command for navigation system, EVV)
- **Language learning**
- **Dictation**
- **Information retrieval**
- **Control / vocal commands**

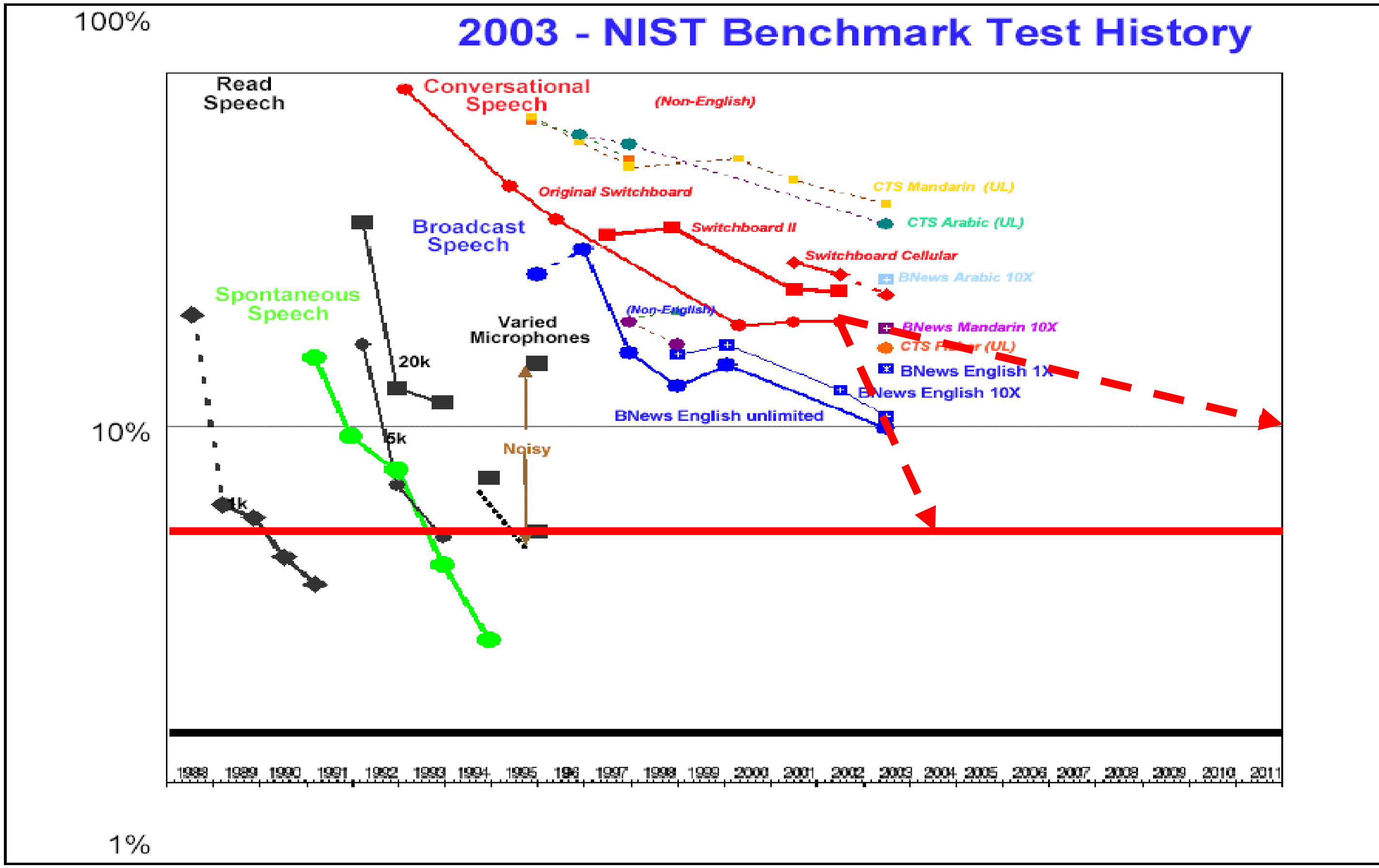
Where we are…

- Best systems achieve*
  - ~10-12% WER for English on European Parliament Speeches or Broadcast News Data !
  - ~20% WER for English on broadcast or telephone conversations
- Large Improvements over the years

  See DARPA & NIST evaluations…

*sources: TCSTAR & GALE projects

Where we are…



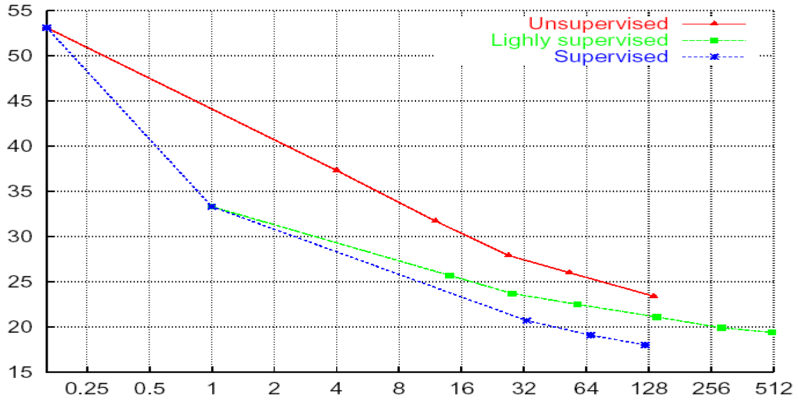
2003 - NIST Benchmark Test History
100%
10%
1%
Read Speech
Conversational Speech
(Non-English)
Original Switchboard
CTS Mandarin (UL)
CTS Arabic (UL)
Broadcast Speech
Switchboard II
Switchboard Cellular
BNews Arabic 10X
Spontaneous Speech
(Non-English)
BNews Mandarin 10X
Varied Microphones
20k
BNews English 1X
BNews English 10X
BNews English unlimited
5k
Noisy
1k
1988 1989 1990 1991 1992 1993 1994 1995 196 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011

Where we are…

- Improvements over the last 15 years mostly due to…
  - Better modeling : discriminant approaches (MMI,MPE), tying (mixtures, states)
  - Adaptation techniques (MAP,MLLR,VTLN)
  - Computational power : for multipass decoding and multiengine approaches (ROVER)
  - And last but not least…

Where we are…

- **More Data !!**
- **"*There's no data like more data*"**, Robert L. Mercer



From LIMSI, Lamel (2002)

Training (hrs)	141	297	602	843
WER(%)	17.2	15.4	14.7	14.5

From RT03 (BBN)

Where we go…

- Evolution of the domain
  - ‘Simple’ Transcription → Rich Transcription
  - Controlled Audio Stream → Continuous Audio Stream
  - One sensor → Multiple sensors
  - Monolingual → Multilingual
  - Audio only → Multimodal
- Increasing difficulty of the tasks

Dictation → ***Broadcast news Transcription*** → ***Meetings Smart rooms***

Human versus machine

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

Source: R. Lippmann, Speech Communication, 1997

Limits and open issues

- Rich Transcription
  - Mark speaker turns, disfluencies, …
- Continuous Audio Flaw
  - Need for sentence breaks, punctuation, …
- Multiple sensors
- Multilingual
  - Portability to new languages, non native speakers
- Multimodal
  - Multiple data streams, asynchronism

Statistical modelling

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

Sequence of acoustic observations

- *Signal frames*
- *Filterbank coefficients*
- *Cepstral coefficients*
- *Time-frequency principal components*
- *…*

Sound object (or class) hypothesis

- *Sound type (speech / music / …*
- *speaker / language / channel*
- *phone / syllable / word*
- *Sound event (jingle)*
- *Past or future of a break (ex: speaker change)*
- *…*

→ Generic Approach

Speech parameters

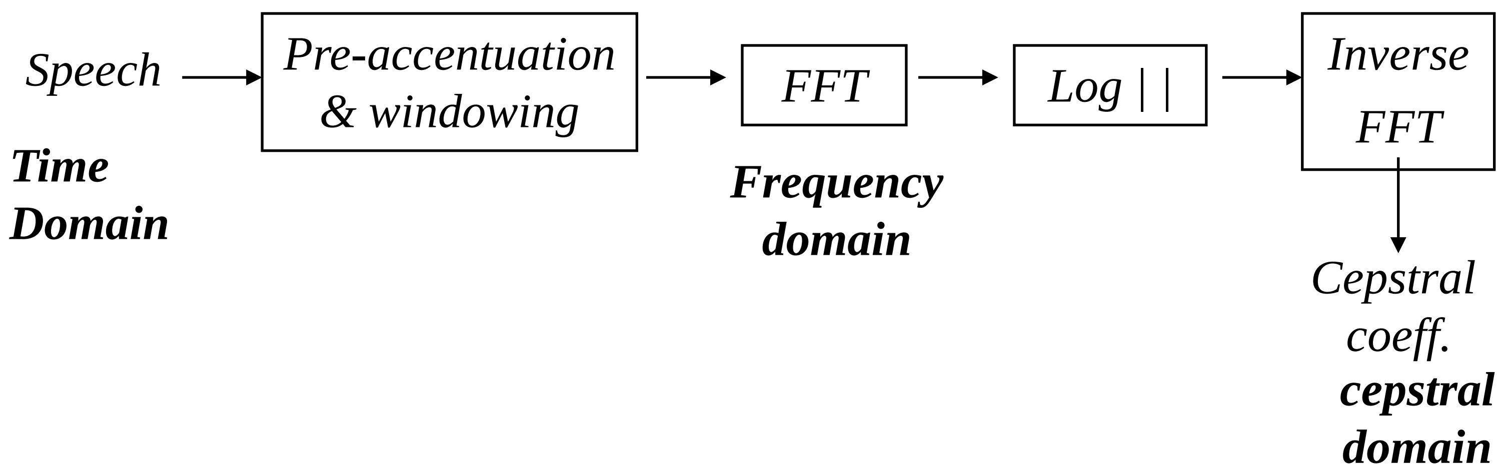
- Acoustic parameters extracted from speech
  - LPC (Linear Predictive Coefficients)
  - MFCC (Mel Frequency Cepstral Coefficients)
  - FilterBanks

Speech parameters

- Mostly for automatic speech recognition and speech compression
  - Spectral analysis
  - Cepstral analysis
  - Linear prediction
- Also used
  - Prosodic information (fundamental frequency, energy features, duration)

Acoustic parameters

- Filterbank coefficients : signal energy in different frequency bands
- Cepstral coefficients



Acoustic parameters

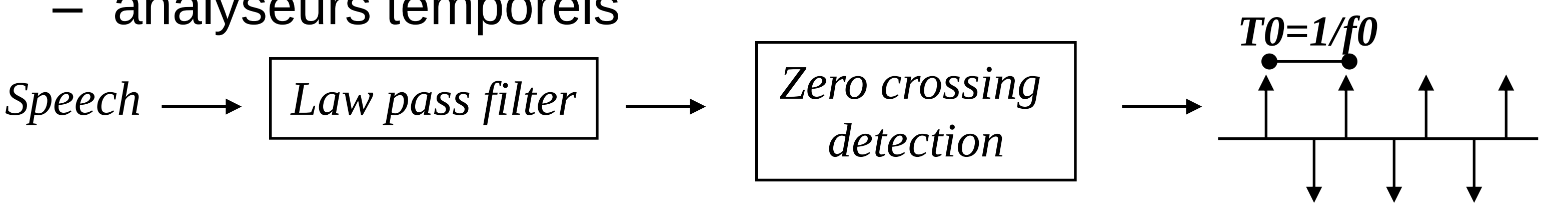
- LPC (Linear Predictive Coding)
  - A sample is predicted as a weighted sum of preceding samples

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i}$$

  - p is the model order
  - a_i = linear prediction coefficients
  - different methods to predict this coeff. (levinson-durbin algo.)

Acoustic parameters

- Fundamental frequency (pitch or *f0*) :
  - analyseurs temporels



  - Problem with pitch : large variability, fine estimation is difficult...

Statistical modelling

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

Sequence of acoustic observations

- *Signal frames*
- *Filterbank coefficients*
- *Cepstral coefficients*
- *Time-frequency principal components*
- *...*

Sound object (or class) hypothesis

- *Sound type (speech / music / ...*
- *speaker / language / channel*
- *phone / syllable / word*
- *Sound event (jingle)*
- *Past or future of a break (ex: speaker change)*
- *...*

→ Generic Approach

Bayes

- x : observation (signal)
- ci : class to be recognized

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

- Automatic Speech Recognition (ASR)

Acoustic model

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

Language model

Bayes

- x : observation (signal)
- ci : class to be recognized

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

- Automatic Speech Recognition (ASR)

$$w^* = \arg\max_i \frac{p(x / w_i).P(w_i)}{p(x)} = \arg\max_i \overbrace{p(x / w_i)}^{\text{Acoustic model}}.\underbrace{P(w_i)}_{\text{Language model}}$$

- Statistical Machine Translation (SMT)

$$e^* = \arg\max_i \frac{p(f / e_i).P(e_i)}{p(f)} = \arg\max_i \overbrace{p(f / e_i)}^{\text{Translation model}}.\underbrace{P(e_i)}_{\text{Language model}}$$

Phone (Acoustic) Models

- Generally, the acoustic units modeled are phonemes rather than words
  - Exemple : ~40 phone models for french
- To calculate p(x/w_i) an acoustic model, as well as a pronunciation dictionary are needed

Context Dependent vs. Context Independent Models

- Independent : each unit is modeled independtly of the others
- Dependent : different models for a same phone unit according to the left-right context
- triphones : only nearest left and right phonemes are considered

- =>due to coarticulation
- =>Problem : corpora never big enough to estimate robust models

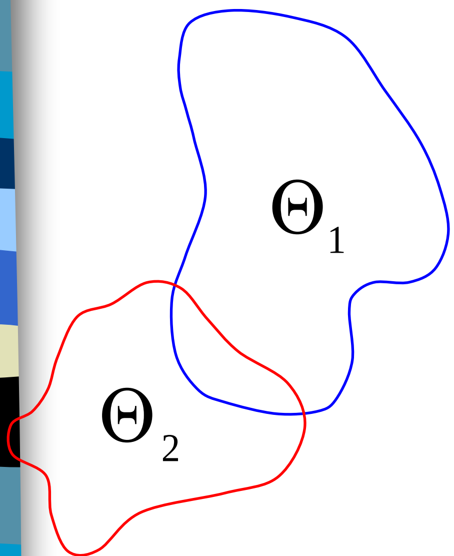
What are those models ?

Many possibilities but we'll talk only of…

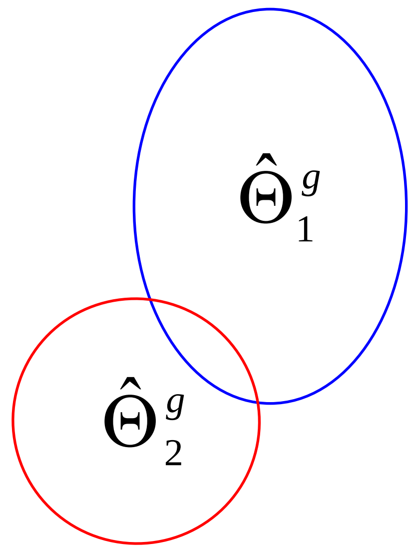
What are those models ?

… Hidden Markov Models with Gaussian Distributions

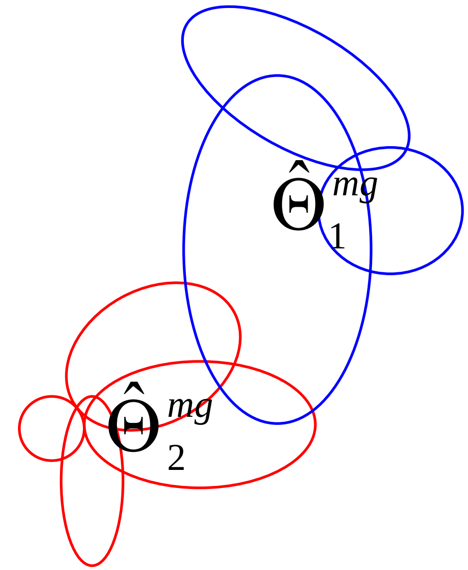
Gaussians



Real distribution



Gaussian model



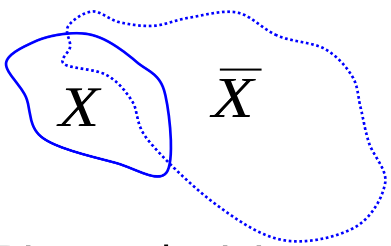
Gaussian mixture model (GMM)

Automata

- For sequence processing
- Complex sequential patterns decomposed into piecewise stationary segments
- Each segment : deterministic or stochastic function
- Can describe grammar, lexicon, phone models…
- Example : Hidden Markov Models (HMMs)
  - 2 concurrent stochastic processes :
    - Sequence of HMM states (sequential structure of the data)
    - State output processes (local characteristics of the data)
    - Example : left-right HMM phone model with gaussian mixture output distributions

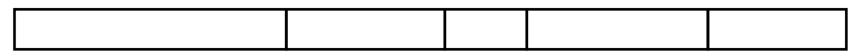
Different problems

Detection



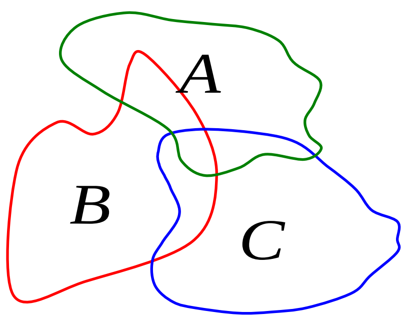
→ Binary decision tests

Segmentation



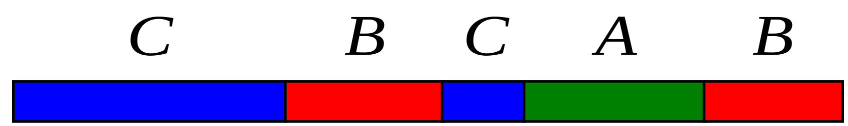
→ Change point detection

Clustering



→ Maximum A Posteriori

Decoding



→ State sequence search

Hidden Markov Models (HMMs)

Intro to [HMMs](http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html)

http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html

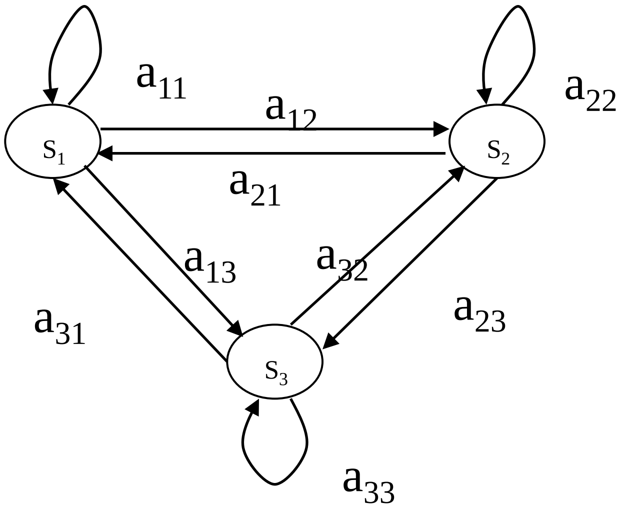
Hidden Markov Models (HMMs)

A HMM is defined by :

- N, number of states in the model, $S=\{S_1,S_2,...S_N\}$
- M, number of output (emission) symbols per state, $V=\{v_1, v_2...v_M\}$
- Propability distribution are defined
  - Transition probabilities $A=\{a_{ij}\}$.
  - Emission probabilitiy of symbol k in state j : b_{jk}
  - Initial state probabilites $\pi=\{\pi_i\}$ 1 $\forall$ i $\forall$ N.

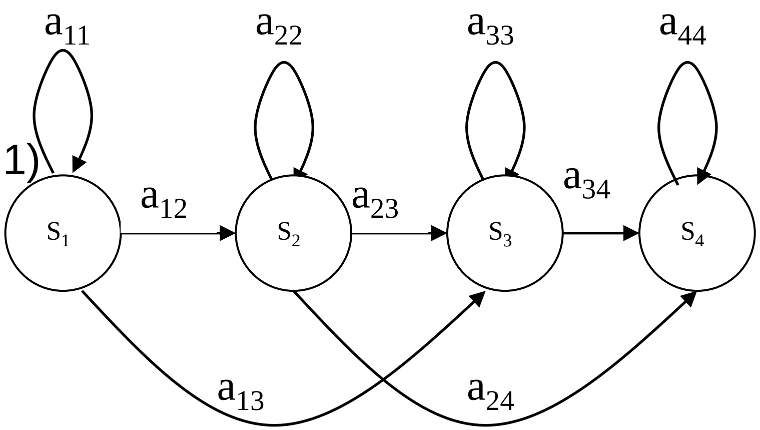
- If the set of emission symbols V is finite, the HMM is called discrete (if V is infinite, then the HMM is continuous).

HMM for speech recognition



Ergodic HMM

- Temporal aspect of speech
  - Use of left-right HMMs (Bakis model).
- Left-right HMM properties
  - aij = 0 when j < i
  - (πi = 0 when i $\neq$ 1) and (πi = 1 when i = 1)
  - aNN = 1



Left-right HMM

Three fundamental problems of HMMs

- Given observations O and HMM λ
  - How to calculate $P(O|\lambda)$?
  - The solution to this problem called **evaluation** is the algorithm ***Forward-Pass***
- Given observations O and HMM λ
  - How to choose the most probable state sequence Q that maximizes $P(Q|O, \lambda)$?
  - The solution to this problem called decoding is the algorithm ***Viterbi***
- Given observations O and HMM λ
  - How to adjust (train) the parameters of the model to maximize $P(O|\lambda)$? This is the **training** of the model parameters.
  - Algorithm Baum-Welch, algorithm **EM** (expectation-maximization)

Algorithm Forward pass (1)

- Sequence of T observations :

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

- Partial probabilities (α's) are calculated iteratively

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

Algorithm Forward pass (2)

- Then for t=2 .. T, : $$\alpha_{t+1}(j) = \sum_{i=1}^{n} \left(\alpha_t(i) a_{ij} \right) b_{jk_t}$$
- Corresponding to the sum of the probabilities of each path leading to the considered state (j) multiplied by the emission probability in the considered state
- Finally, the sum of all the partial probabilities at time T, gives the probability of the observation given the HMM model

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

Algorithm Viterbi (1)

- Calculate most probable state sequence

$$\mathbf{X}_i = (X_{i_1}, X_{i_2}, \ldots, X_{i_T})$$

- From T observations

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

- Partial probabilities (δ's) are calculated iteratively

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

Algorithm Viterbi (2)

- Then for *t=2 .. T* and *i=1 .. n*, one calculate

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

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

- Then, $i_t = argmax(\delta_T(i))$ corresponds to the most probable state at time *t=T*
- Finally, « *backtracking* ») is necessary to calculate the most probable path

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

FIN