Les descripteurs audio pour la parole

Master ATAL

Marie Tahon
MCF, dpt. Informatique

8 septembre 2018

PLAN DE LA SECTION ACTUELLE

INTRODUCTION

Le traitement automatique de la parole consiste à réaliser des opérations sur un signal sonore de parole afin d'en extraire des informations de haut-niveau:

- qui parle ? $\rightarrow$ identification du locuteur
- qu'est-ce qui a été dit ? $\rightarrow$ reconnaissance des mots
- dans quelle langue ?
- dans quel état psychologique $\rightarrow$ affective computing, détection des émotions.

Descripteurs Audio

$\downarrow$

Modèles statistiques

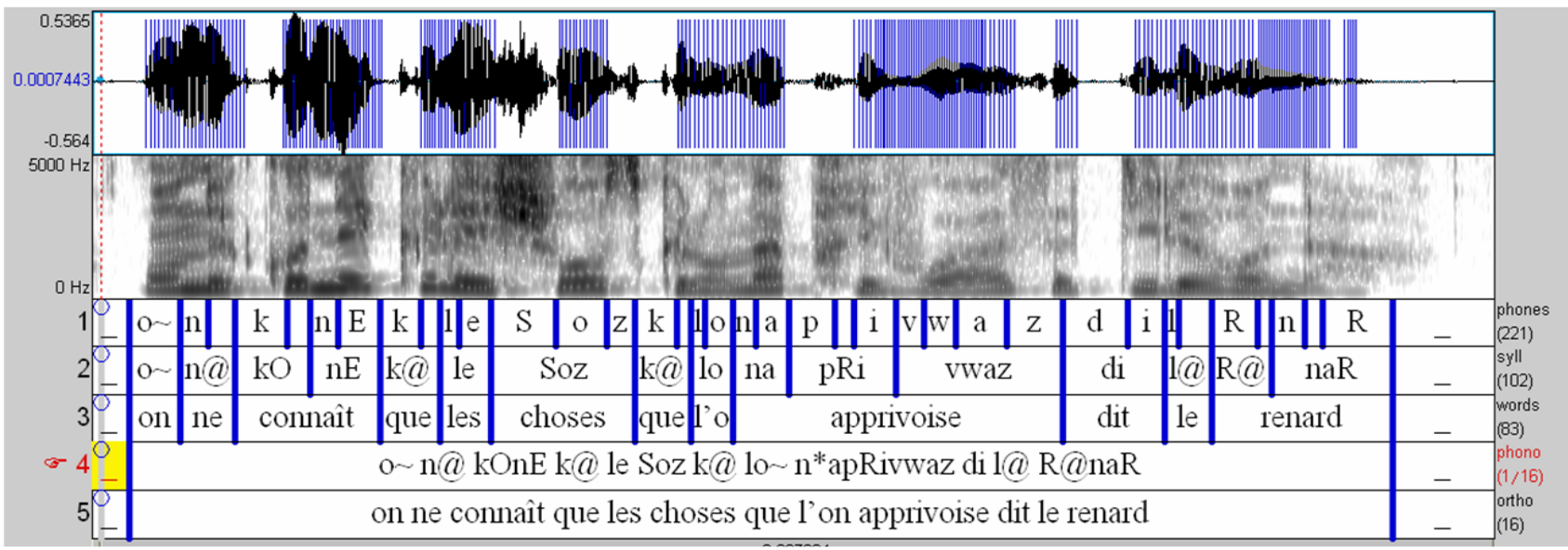
$\downarrow$

Information haut-niveau

INTRODUCTION

Complexité de l'oral par rapport à l'écrit:

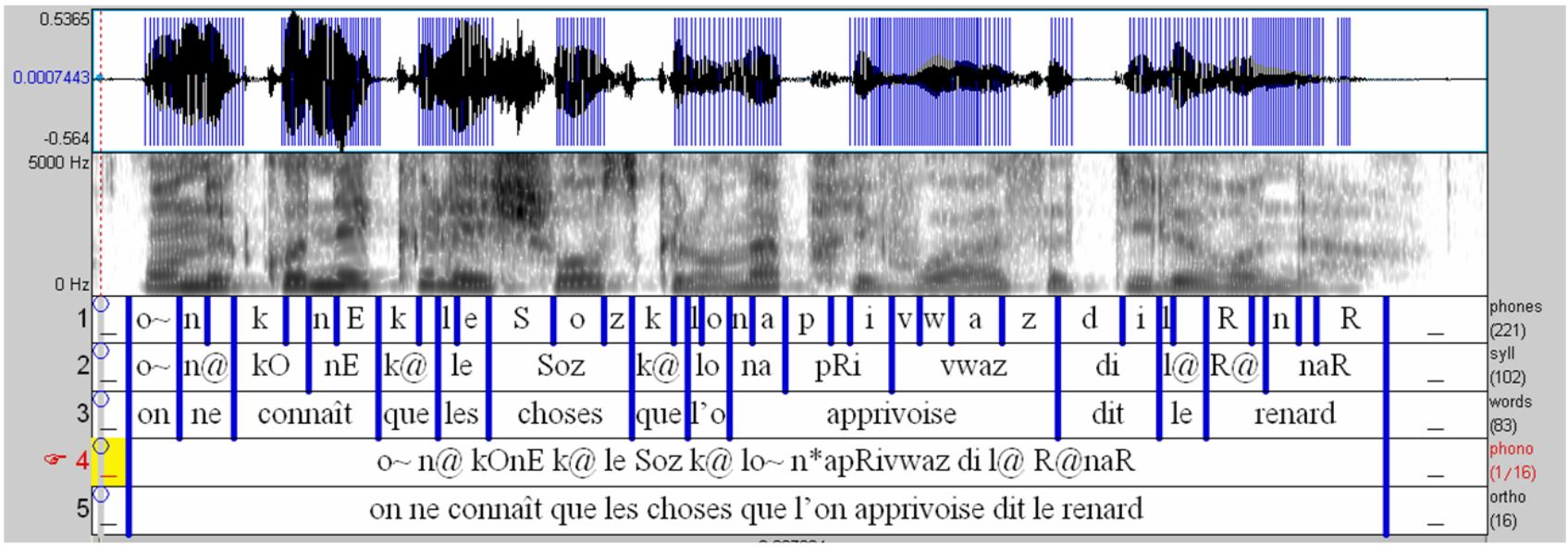
- un signal continu, coarticulé → il faut le segmenter
- existence de distorsions temporelles → ex: débit de parole variable
- présence de variabilités inter-locuteur (expression), intra-locuteur (timbre), conditions acoustiques.
- aspects non-verbaux → timbre, qualité vocale, prosodie, disfluences, ...



SEGMENTATION

Le choix de la taille de fenêtre détermine le type d'information que l'on peu extraire du signal.

- fenêtre glissante de taille fixe → niveau spectral
- segmentation sur les groupes de souffle (pause > 300 ms) → niveau prosodique
- segmentation en phonèmes → niveau formantique



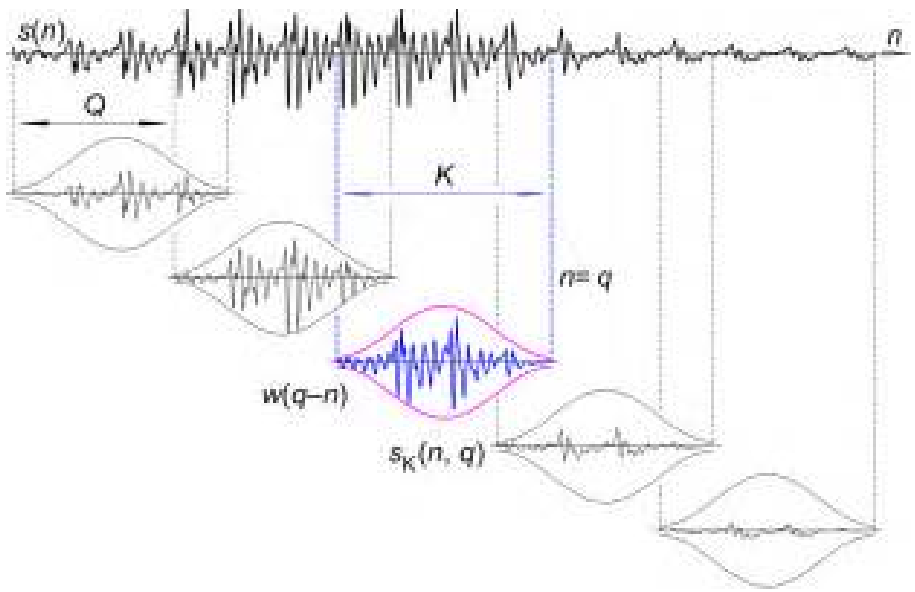
La segmentation en phonèmes/mots reste une tâche difficile.

PLAN DE LA SECTION ACTUELLE

SEGMENTATION D'UN FLUX AUDIO

Le signal de parole est un flux de paramètres. On ne sait pas a priori où segmenter $\rightarrow$ segmentation de taille fixe.

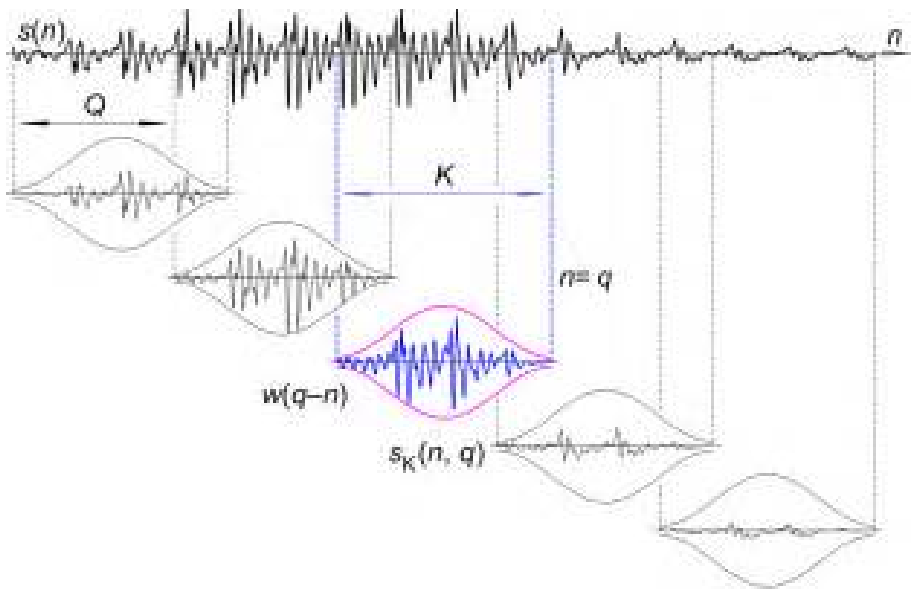
- fenêtre temporelle (Hamming, Hanning, rectangulaire, ...)
- taille de la fenêtre $K \simeq 30$ ms
- pas $Q \simeq 10$ ms
- overlap $\dfrac{K-Q}{K} \times 100 \simeq 33\%$



LE VECTEUR ACOUSTIQUE

À chaque pas (Q) on associe un vecteur de k paramètres acoustiques extraits de la fenêtre (cf spectrogramme) statique et dynamique (les Δ et $\Delta\Delta$)

temps	Q	2Q	3Q	...	nQ
p_1	a_{11}	a_{12}	a_{13}	...	a_{1n}
p_2	a_{21}	a_{22}	a_{23}	...	a_{2n}
...					
p_k	a_{q1}	a_{q2}	a_{q3}	...	a_{qn}
Δa_1	0	$a_{12} - a_{11}$	$a_{13} - a_{12}$	...	$a_{1n} - a_{1(n-1)}$
...					



LE VECTEUR ACOUSTIQUE

Les paramètres du vecteur acoustique peuvent être:

- des descripteurs de spectre:
  - le spectre à court terme (calcul de FFT), généralement 512 points ($\simeq$30 ms à $F_e = 16$ kHz).
  - les ondelettes (utilisées pour caractériser les signaux de parole)
  - les coefficients LPC (linear prediction coefficients) utilisés pour l'extraction des formants $\simeq$ 40 points
  - l'énergie par bande spectrale (Mel, Bark, Harmonique) entre 10 et 40 points.
  - les coefficients PLP (perceptual linear prediction) coefficients LPC obtenus sur une échelle perceptive de Bark.
- des descripteurs de cepstre:
  - les coeffcients cepstraux (MFCC) généralement 13 points
  - les coefficients LPCC (linear prediction cepstral coefficients) sont des LPC obtenus sur le cepstre

LE CEPSTRE

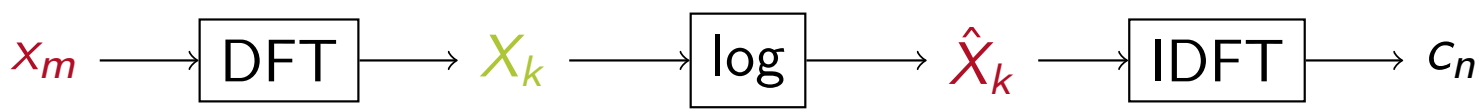
La représentation cepstrale a été conçue pour représenter les modèles source/filtre comme celui de la parole.

- DFT du signal source: $G(f)$
- DFT du filtre: $H(f)$
- On définit le cepstre réel du signal fréquentiel $X(f) = F(f) \times H(f)$, avec τ la quéfrence (homogène à un temps).

$$c(\tau) = FFT^{-1} \log |X(f)| = FFT^{-1} \log |G(f)| + FFT^{-1} \log |H(f)|$$

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X_k| e^{2j\pi kn/N}$$

$$= \frac{1}{N} \sum_{k=0}^{N-1} \log |\sum_{m=0}^{N-1} x_m e^{-2j\pi km/N}| e^{2j\pi kn/N}$$



LE CEPSTRE

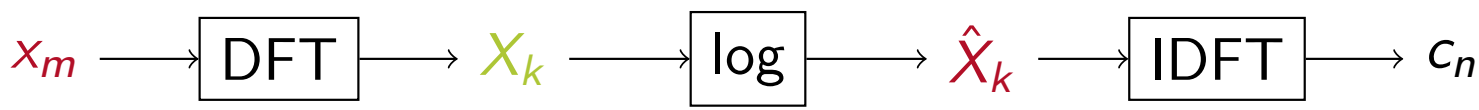
La représentation cepstrale a été conçue pour représenter les modèles source/filtre comme celui de la parole.

- DFT du signal source: $G(f)$
- DFT du filtre: $H(f)$
- On définit le cepstre réel du signal fréquentiel $X(f) = F(f) \times H(f)$, avec τ la quéfrence (homogène à un temps).

$$c(\tau) = FFT^{-1} \log |X(f)| = FFT^{-1} \log |G(f)| + FFT^{-1} \log |H(f)|$$

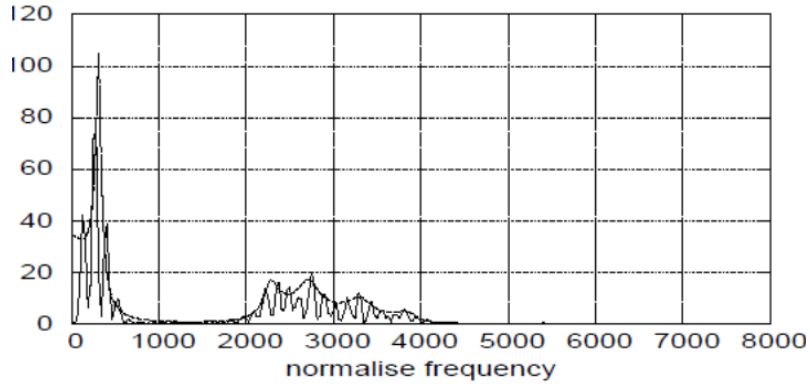
$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X_k| e^{2j\pi kn/N}$$

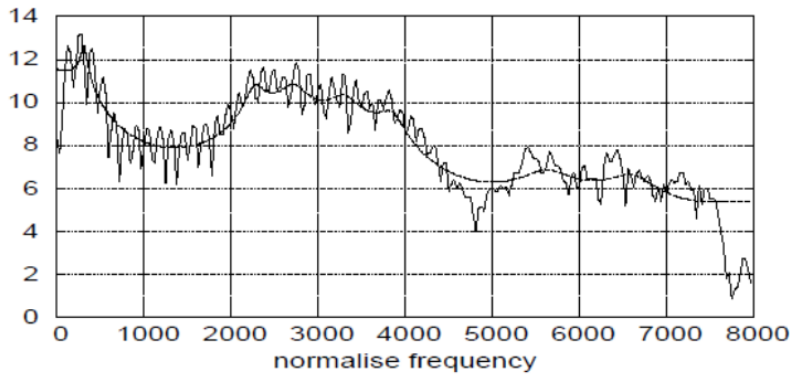
$$= \frac{1}{N} \sum_{k=0}^{N-1} \log |\sum_{m=0}^{N-1} x_m e^{-2j\pi km/N}| e^{2j\pi kn/N}$$

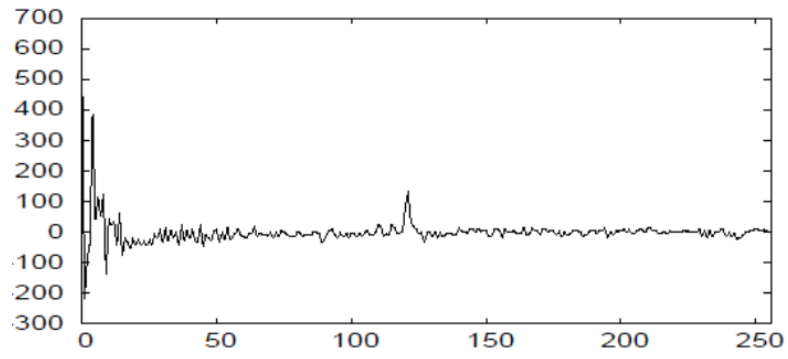


Qu'est-ce que N ?

LE CEPSTRE







[Taylor]

x_m (signal échantillonné)

↓

DFT

↓

X_k (spectre discret $f < 8\ kHz$)

↓

log

↓

$\hat{X}_k$ (spectre discret en dB)

↓

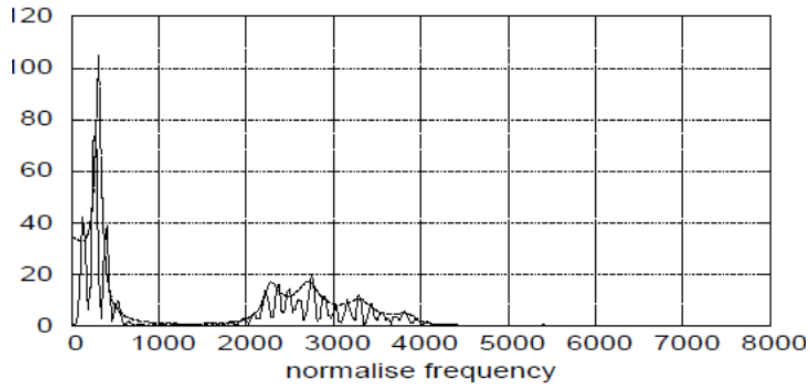
IDFT

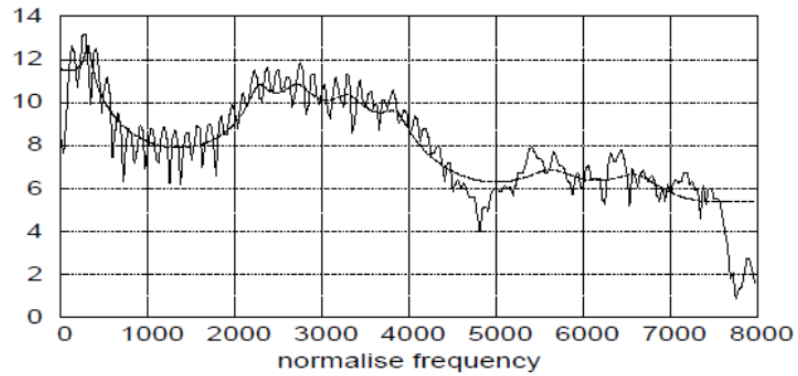
↓

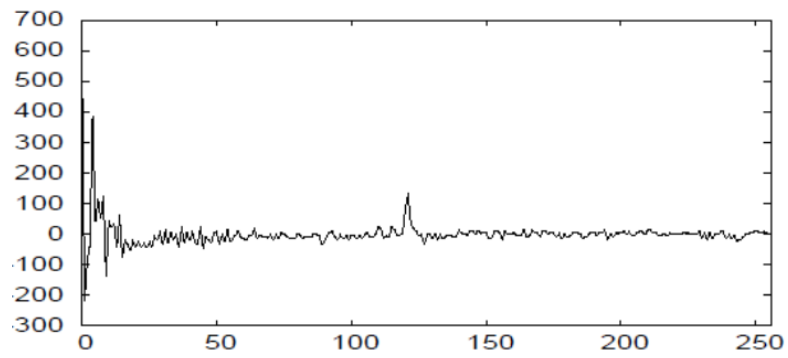
c_n (cepstre $\tau < 256$ ech.)

NB: F_e =?, Δt =?, pic à 120 ech.?

LE CEPSTRE







[Taylor]

x_m (signal échantillonné)

↓

DFT

↓

X_k (spectre discret $f < 8$ kHz)

↓

log

↓

$\hat{X}_k$ (spectre discret en dB)

↓

IDFT

↓

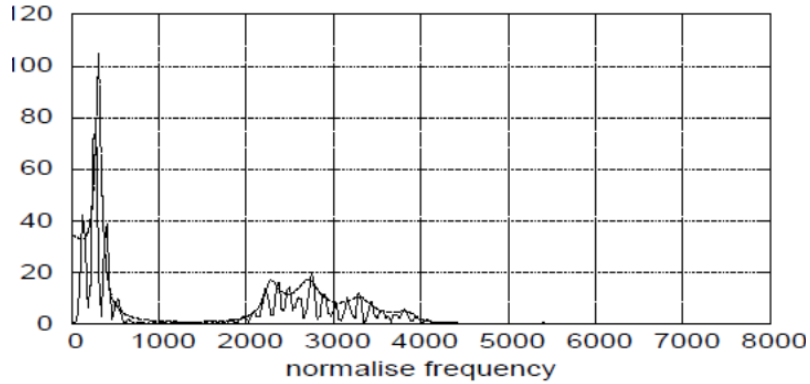
c_n (cepstre $\tau < 256$ ech.)

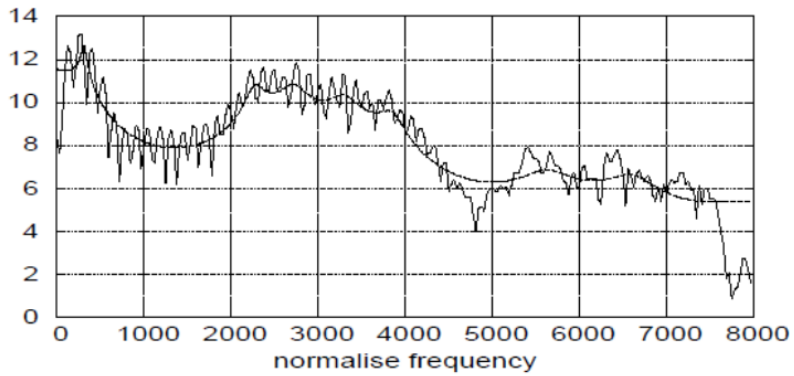
NB: $F_e = 2F_{max} = 16$ kHz,

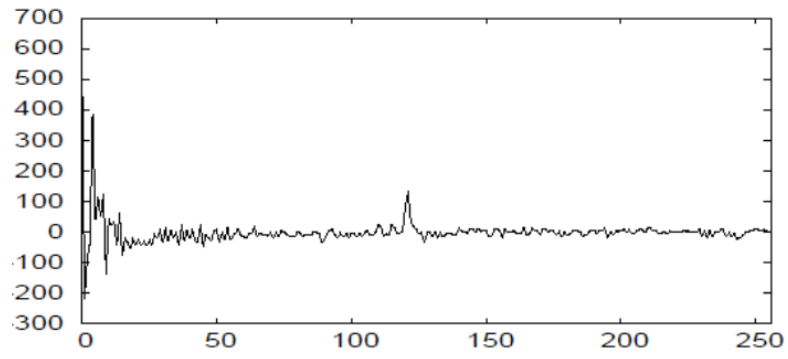
256 ech. $\rightarrow \Delta t = 512/F_e = 32$ ms,

$$F_0 = \frac{F_e}{2 \cdot 120} \simeq 67 \text{ Hz}$$

LE CEPSTRE







[Taylor]

x_m (signal échantillonné)

↓

DFT

↓

X_k (spectre discret $f < 8\ kHz$)

↓

log

↓

$\hat{X}_k$ (spectre discret en dB)

↓

IDFT

↓

c_n (cepstre $\tau < 256$ ech.)

Pas de phase

log $\rightarrow$ importance de la périodicité.

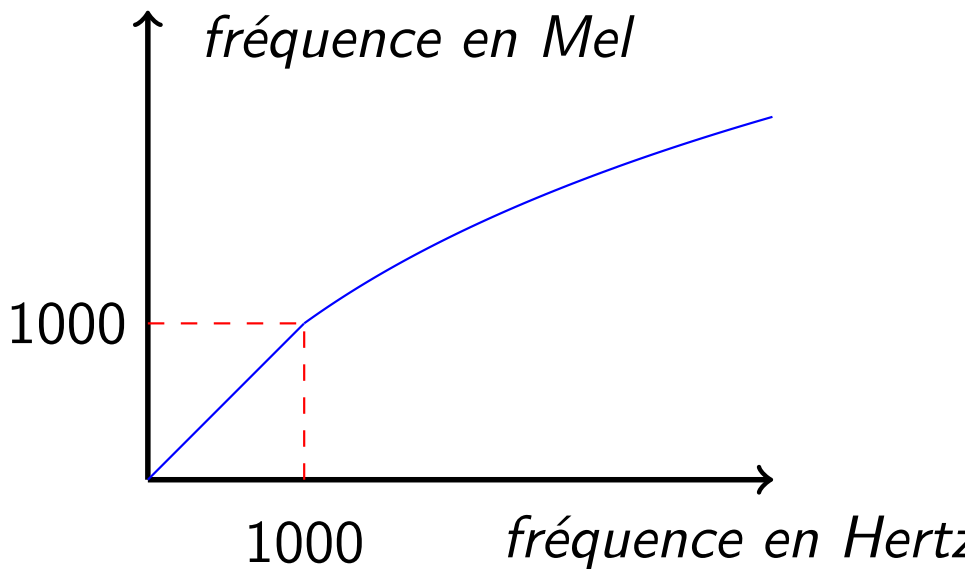
τ petit: conduit vocal, τ grand: source.

LES MFCCs

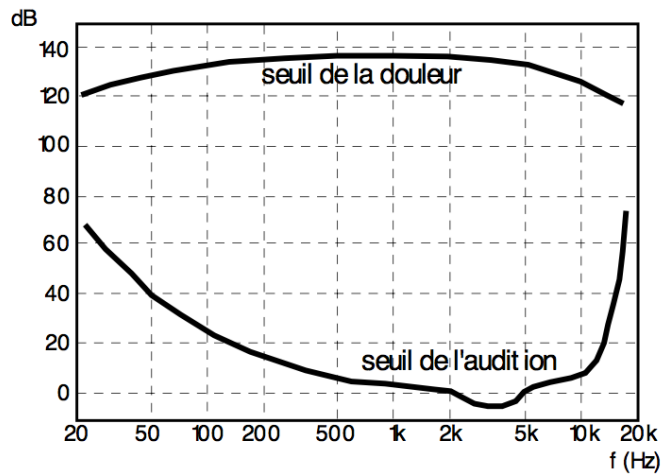
Echelle Mel:

- Approximation de la sensation psychologique de hauteur d'un son (sonie)
- Formules analytiques [Fant]

$$Mel(f) = \begin{cases} 1000 \log_2 \left(1 + \frac{f}{1000}\right) & f \geq 1000 \\ f & f < 1000 \end{cases}$$



Champ de l'audition humaine (sonie).



LES FILTRES DE MELS

- On définit R filtres triangulaires fréquentiels entre f_{min} et f_{max}.

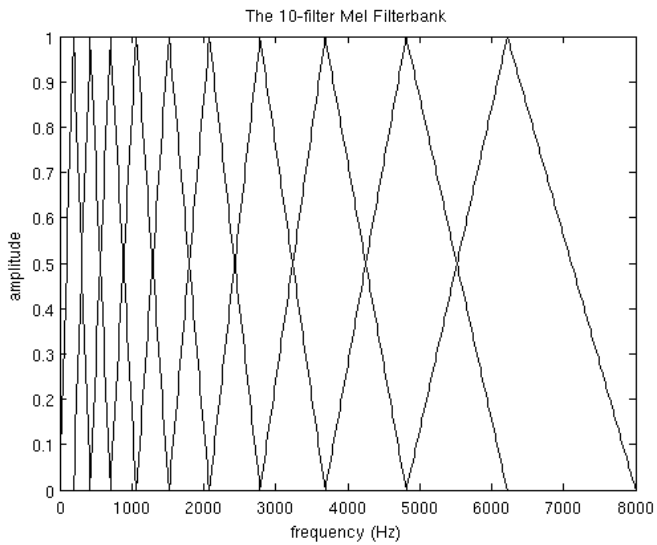
  $f_{min} = 300$ et $f_{max} = 4000$Hz et $R = 10$

- On détermine R fréquences centrales d'intervalle égaux sur une échelle de Mel:

  $Mel(f_{min}) = 401.25$ et $Mel(f_{max}) = 2834.99$.

  $mel = [401.25, 622.50, 843.75, 1065.00, 1286.25, 1507.50, 1728.74, 1949.99,$
  $2171.24, 2392.49, 2613.74, 2834.99]$

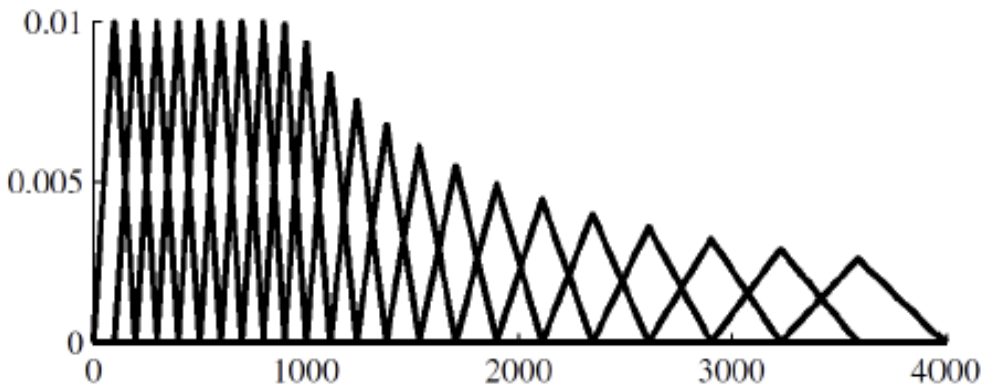
  $freq = [300, 517.33, 781.90, 1103.97, 1496.04, 1973.32, 2554.33,$
  $3261.62, 4122.63, 5170.76, 6446.70, 8000]$



LES FILTRES DE MELS

Implémentation classique

- $R = 22$ filtres fréquentiels
- $f_{min} = 20$
- $f_{max} = F_e/2 = 8000$ Hz.
- normalisation de l'aire des triangles à 1

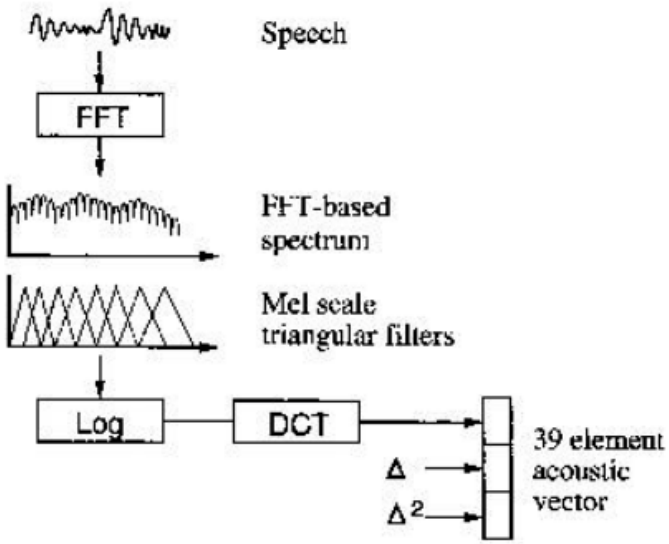


[Rabiner&Schaffer]

Attention: les fréquences centrales des filtres dépendent de F_e et de R !

LES MFCCs

Mel-Frequency Cepstral Coefficients: la paramétrisation la plus répandue.



Avec R filtres de Mel

$$S[k] = |FFT(s[n])|$$

$$\hat{S}_r \simeq \sum_{k=0}^{K} Mel_r[k] \times S[k]$$

$$MFCC_n = \frac{1}{R} \sum_{r=1}^{R} \log(\hat{S}_r) \cdot \cos\left(\frac{2\pi}{R}\left(r + \frac{1}{2}\right) n\right)$$

Typiquement pour la parole:

- $F_e = 8$ kHz, DFT sur 512 échantillons, $\Delta t = 30$ ms, fenêtre de Hamming.
- $R = 22$ filtres Mel
- $m \in [1, 13]$ ($\pm MFCC_0$)
- dérivées premières (Δ) et secondes ($\Delta\Delta$)
- vecteur acoustique: 39 paramètres

DCT vs DFT

DCT: Discrete Cosine Transform

$$X_{DCT}[n] = \sum_{n=0}^{N-1} x[n] \cdot \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right)$$

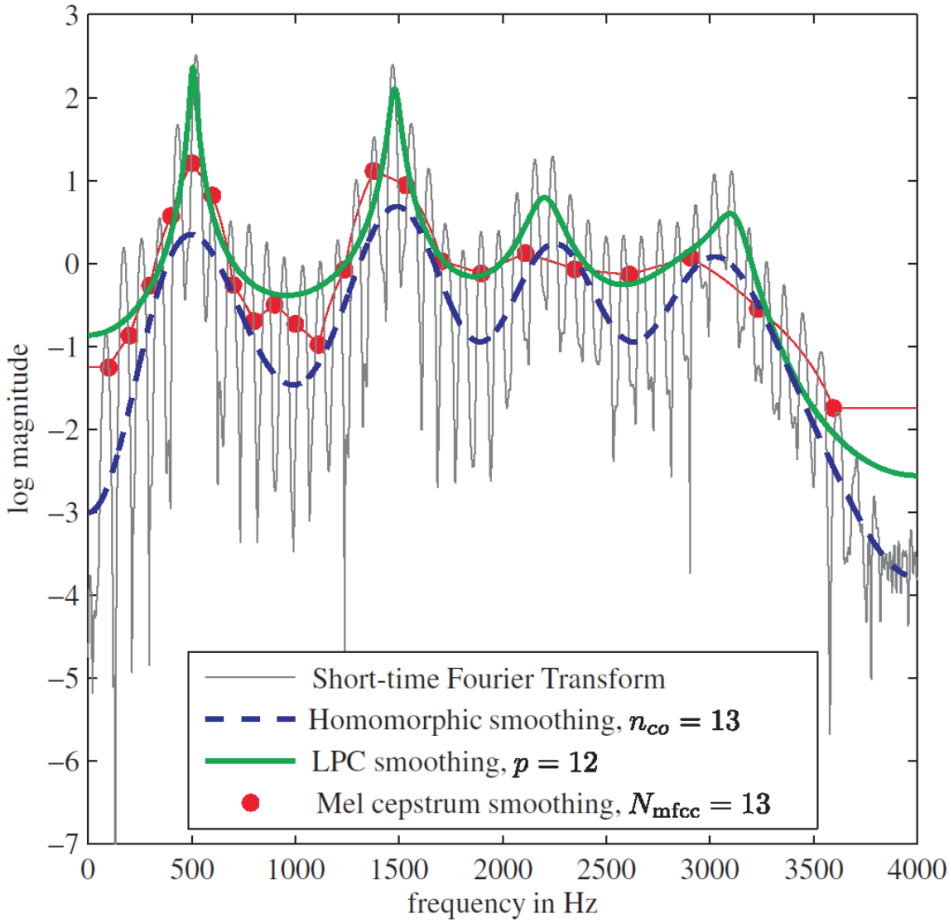
DFT: Discrete Fourrier Transform

$$X_{DFT}[n] = \sum_{n=0}^{N-1} x[n] \cdot \left(\cos\left(\frac{2\pi}{N}kn\right) + j\sin\left(\frac{2\pi}{N}kn\right)\right)$$

C'est comme si on avait $X_{DCT}[n] \simeq \Re\mathrm{e}(X_{DFT}[n])$

ENVELOPPE SPECTRALE

Les coefficients MFCCs permettent de reconstruire l'enveloppe spectrale:



[Rabiner&Schaffer]

APPLICATIONS

Avantages:

- permet de reconstruire l'enveloppe spectrale
- proche de la perception humaine (échelle log en amplitude et en fréquence).
- dissocie la source (excitation glottique: coefficients élevés) du filtre (conduit vocal: coefficient faibles)
- les coefficients sont décorrélés, cela fait un stockage minimum d'information (moyenne et écart-type) (comme une PCA)
- robuste au bruit de fond

Applications:

- reconnaissance automatique de la parole
- identification du locuteur
- affective computing

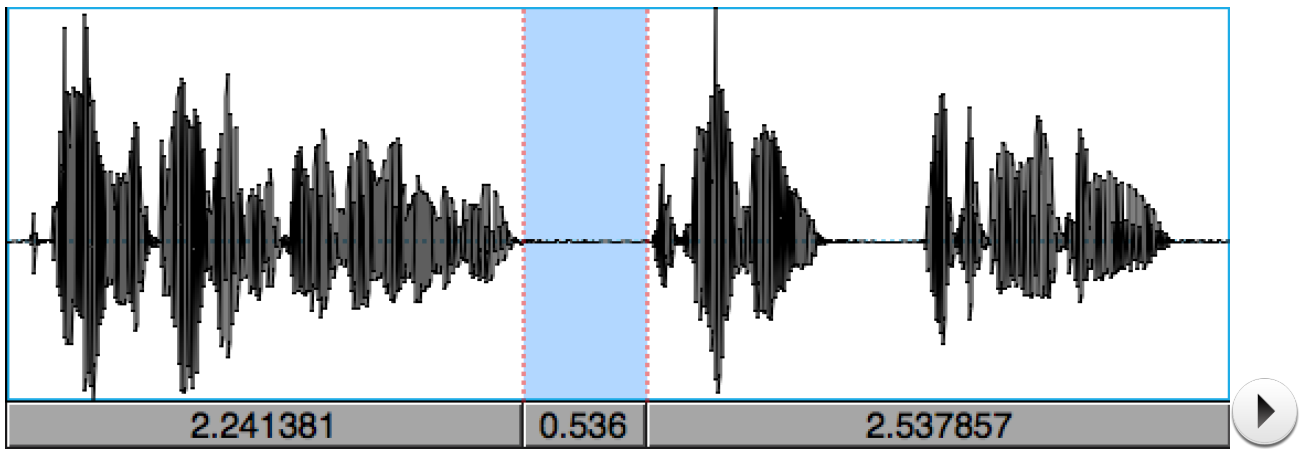
PLAN DE LA SECTION ACTUELLE

LA PROSODIE

- intonation ou fréquence fondamentale (F_0)
- énergie ou intensité
- rythme
- (timbre et qualité vocale)
- accentuation, proéminences

Segmentation prosodique

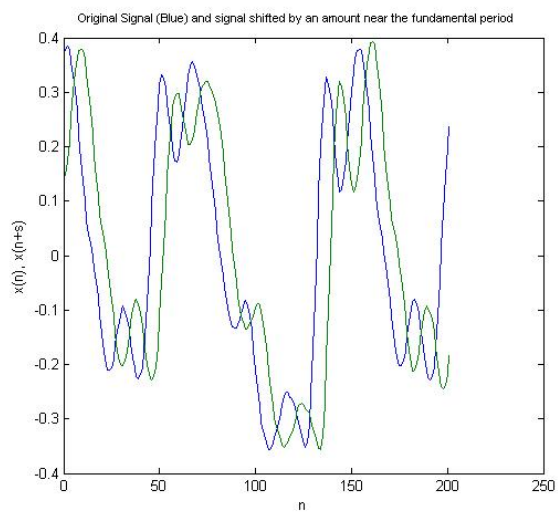
- groupe de souffle, délimiteurs = respirations, pauses $>$ 300 ms
- attention segment prosodique $\neq$ phrases



Alors je pleurais ce que je voyais si bien, et qui la veille, n'était pour moi que néant.

LA FRÉQUENCE FONDAMENTALE

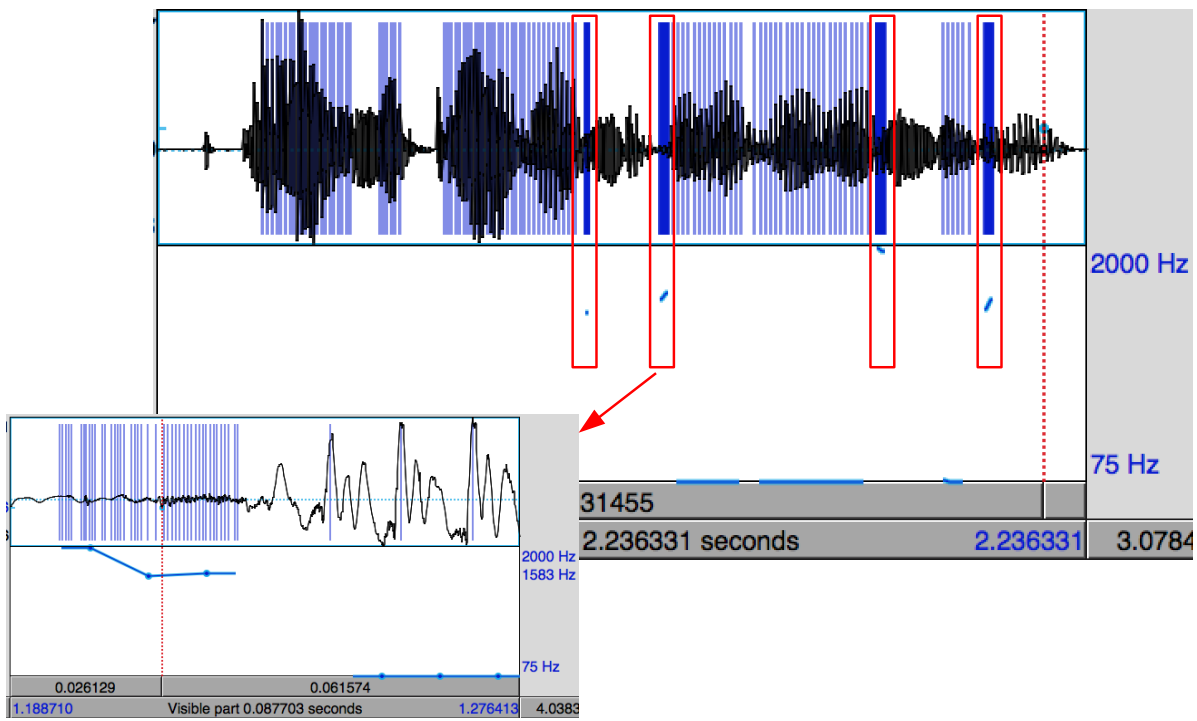
- correspond à la vibration des cordes vocales
- elle n'est pas définie pour des sons non-voisés (ce ne sont pas des signaux périodiques).
- extraite automatiquement à partir d'un signal:
  - méthode d'auto-corrélation (Praat, YIN),
  - fonction de différences moyennées (ASDF),
  - estimation de maxima de vraisemblance (Doval&Rodet),
  - algorithme de Viterbi,
  - estimation du cepstre.
- En parole la F_0 est estimée par pas de 10 ms.



LA FRÉQUENCE FONDAMENTALE

ATTENTION: tous les algorithmes d'extraction de F_0 sont susceptibles de faire des erreurs.

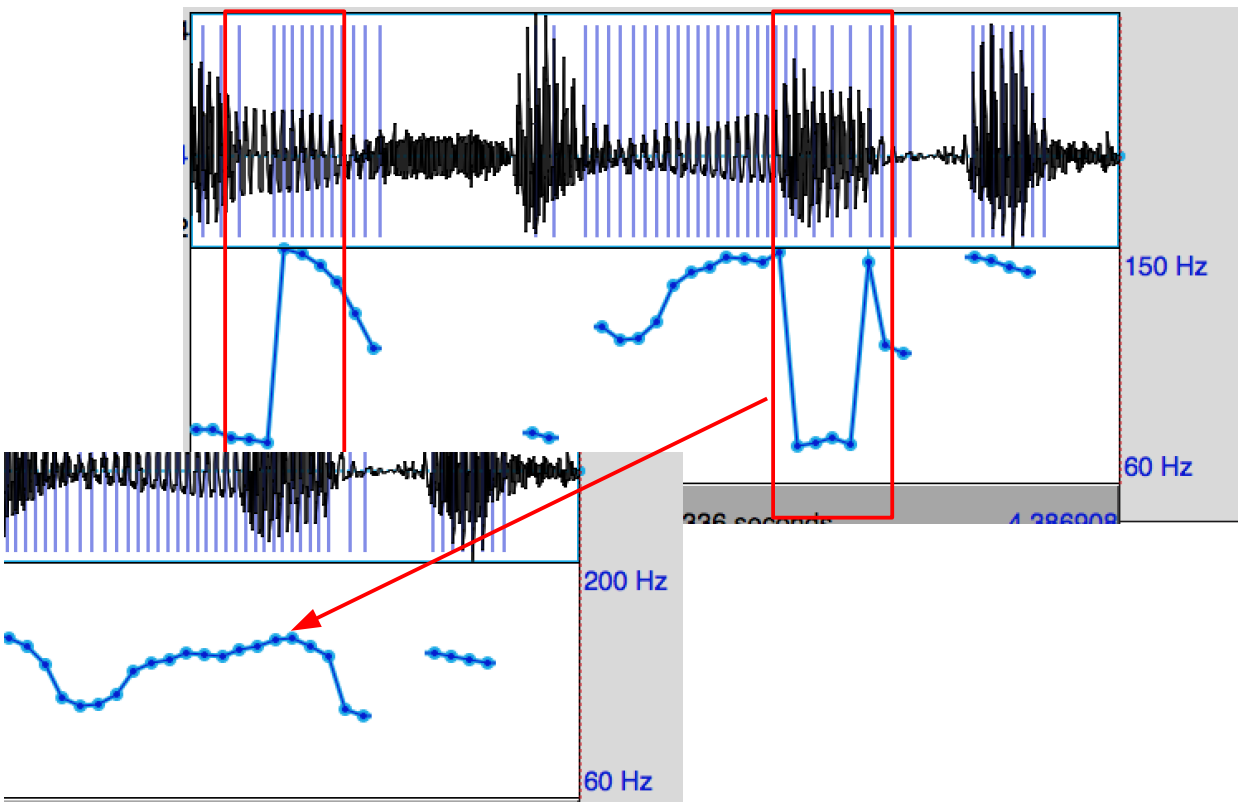
- erreur de voisement: vérifier le taux de voisement qui donne la confiance dans la détection de périodicité dans le signal (1: voisé, 0: non-voisé)
- ex: $F_0 > 500$ Hz très peu fréquent (surtout chez les adultes) sauf voix très expressive $\rightarrow$ erreur de voisement: détection d'une périodicité dans le bruit fricatif.



LA FRÉQUENCE FONDAMENTALE

ATTENTION: tous les algorithmes d'extraction de F_0 sont susceptibles de faire des erreurs.

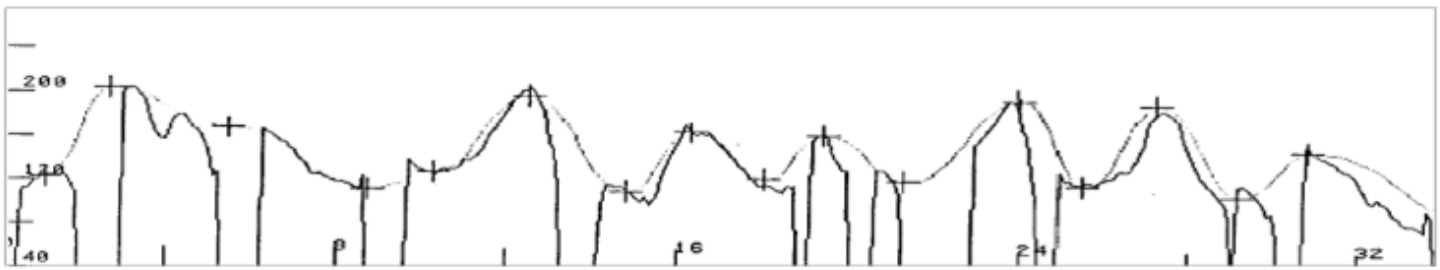
- saut d'octave: l'algorithme rate une période ($F_0/2$) ou bien accroche l'harmonique supérieure ($F_0 \times 2$)
- Saut d'octave entre 150 Hz et 75 Hz $\rightarrow$ adapter la résolution: $F_0 \in [60; 200]$ Hz.



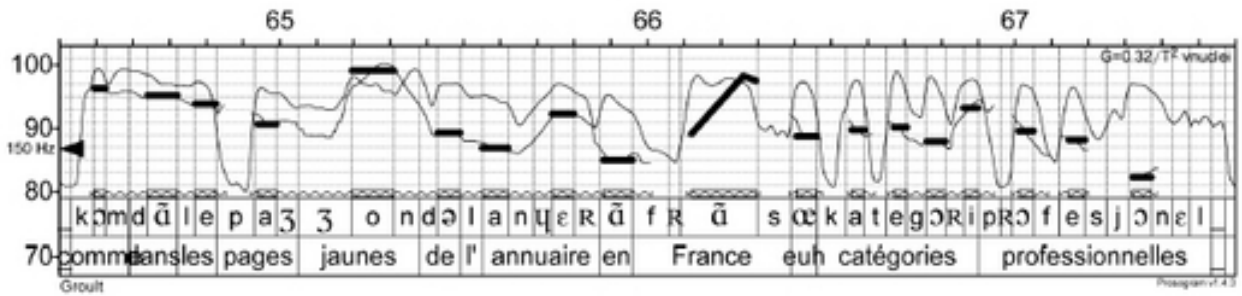
L'INTONATION

La stylisation de la courbe de F_0 permet de modéliser le contour intonatif.

MoMel: courbe continue interpôlée MoMel [Hirst&Espresser]



Prosogramme: alignement signal/phonèmes fondé sur un modèle de perception tonale [d'Alessandro&Mertens]



INTSINT: codage du contour avec des points cibles [Hirst]

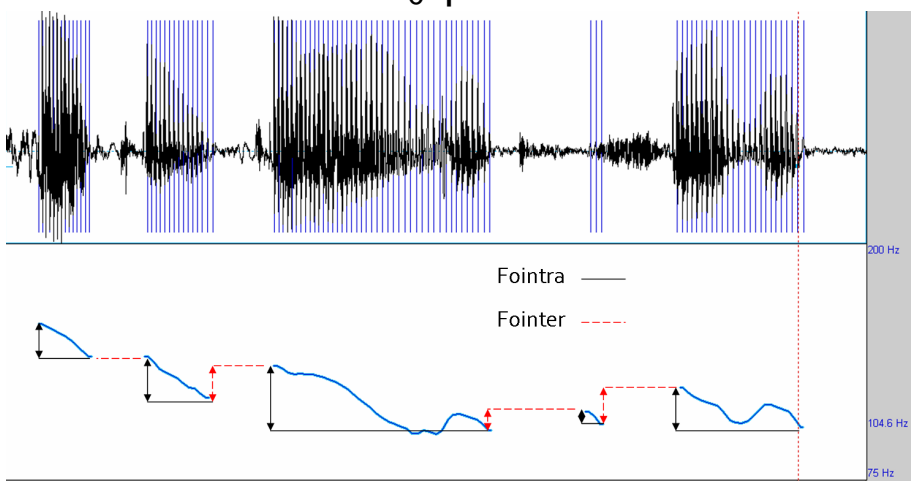
ASPECTS PERCEPTIFS

- Pour être plus proche de l'échelle de fréquence perçue (échelle musicale), on peut utiliser les semi-tons (avec une référence à 110 Hz (LA2):

$$F_{st} = 12 \log_2 \left(\frac{F_{Hz}}{110} \right)$$

- Ainsi on étudiera plutôt les rapports de fréquences (en Hz) plutôt que leur différences

Indices inter et intra F_0 pour l'étude des émotions



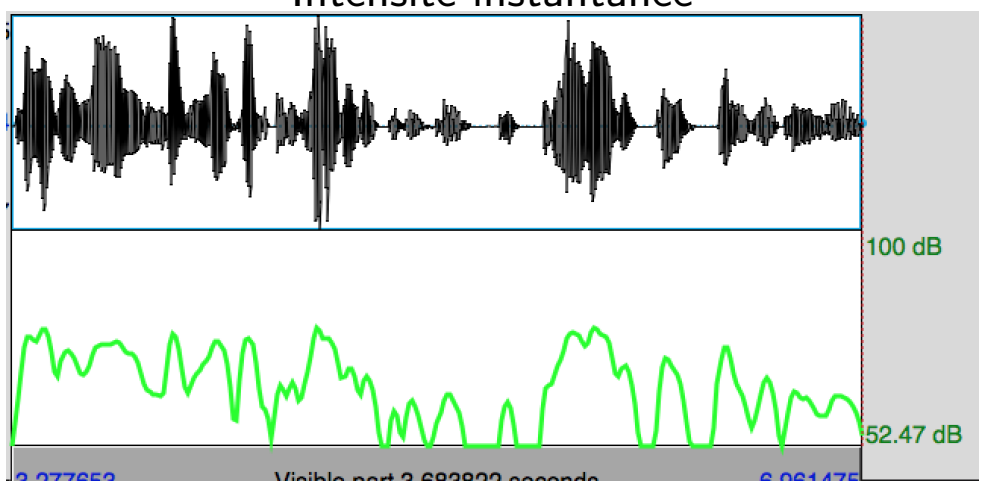
ENERGIE

- Energie du signal:

$$E = \int_{t_0}^{t_1} |x(t)^2| dt$$

- Intensité: $I = 20 \log_{10}(E)$
- En parole l'intensité est donnée en dB par pas de 10 ms
- L'intensité peut être modulée par un filtre perceptif (on entend moins fort les très hautes fréquences et les basses fréquences) c'est ce qu'on appelle loudness.

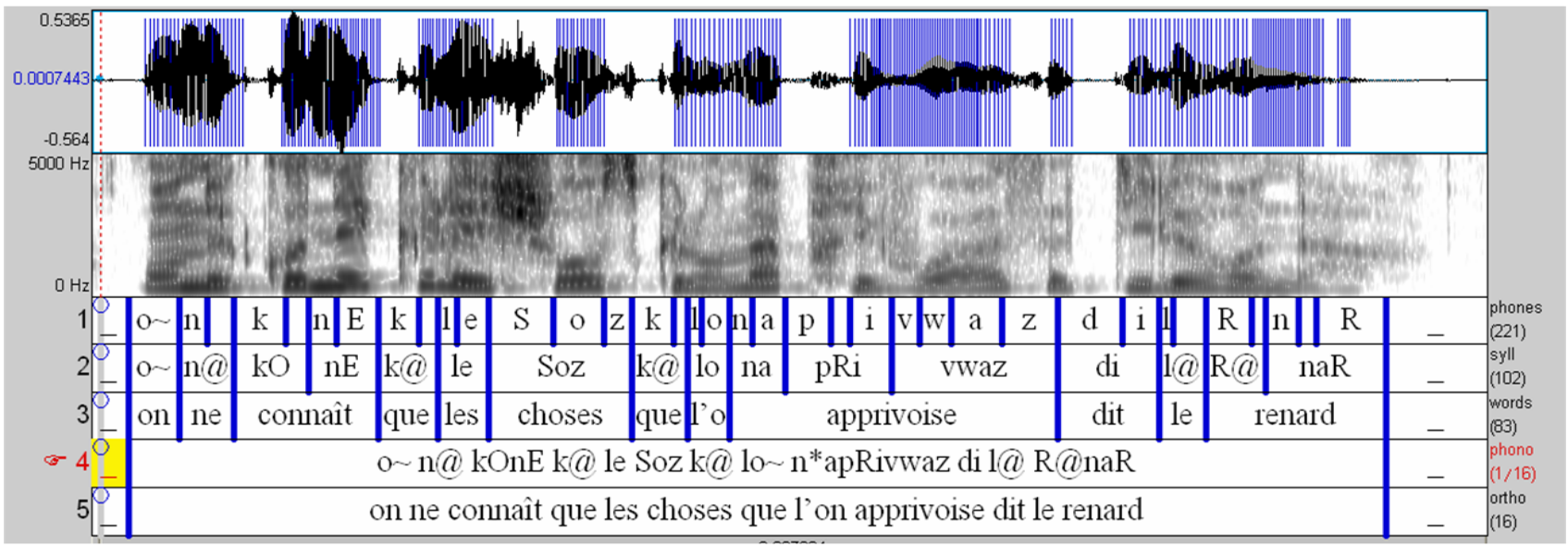
Intensité instantanée



RYTHME

Pas de consensus simple sur des mesures simples de rythme dans la parole:

- structure chaotique
- débit de voisement: $\frac{\text{Durée des parties voisées}}{\text{Durée totale}}$
- débit syllabique: $\frac{\text{Nombre de syllabes}}{\text{Durée totale}}$
- débit articulatoire: $\frac{\text{Nombre de syllabes}}{\text{Durée totale sans les pauses}}$
- $\rightarrow$ dépend fortement d'un alignement signal/phonème.



RYTHME

Autres mesures directement extraites du signal

- ondelettes
- beat detection: recherche de pulsations basses fréquences (valable en musique mais en parole ?)
- $\rightarrow$ dépend fortement de la qualité du signal d'origine.

RYTHME

Autres mesures étudiées pour l'identification d'une langue

- %V: proportion d'intervalles vocaliques
- ΔV: écart-type des intervalles vocaliques
- ΔC: écart-type des intervalles de consonnes

%V en fonction de ΔC [Grabe&Low]

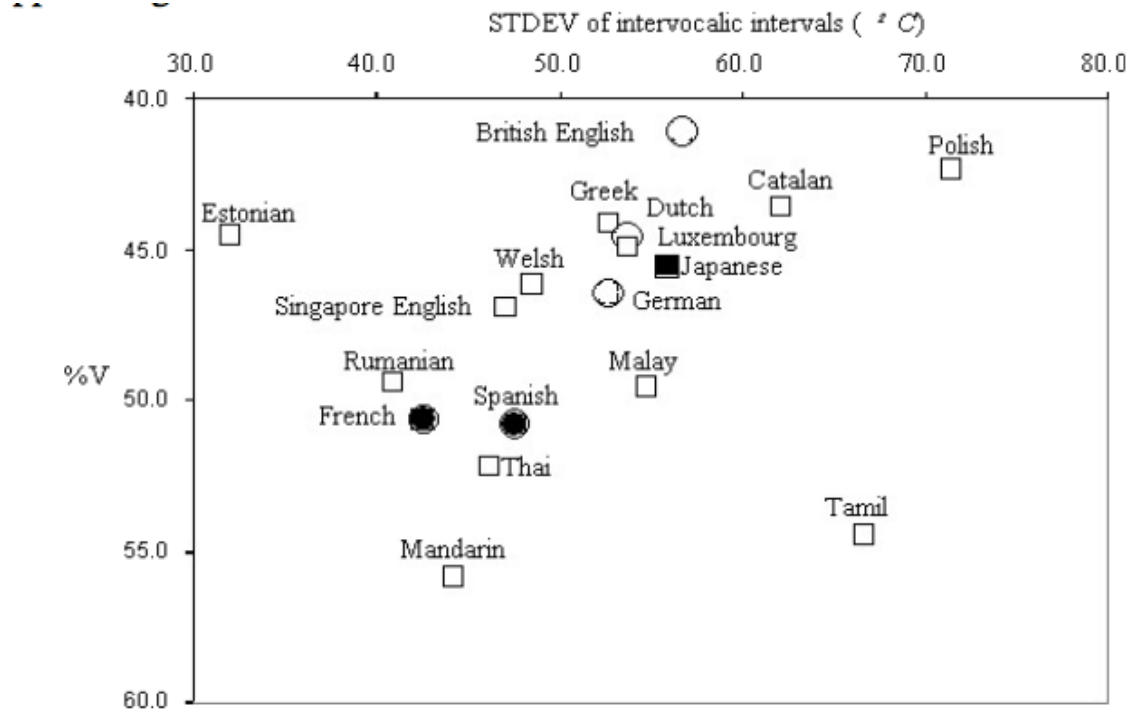


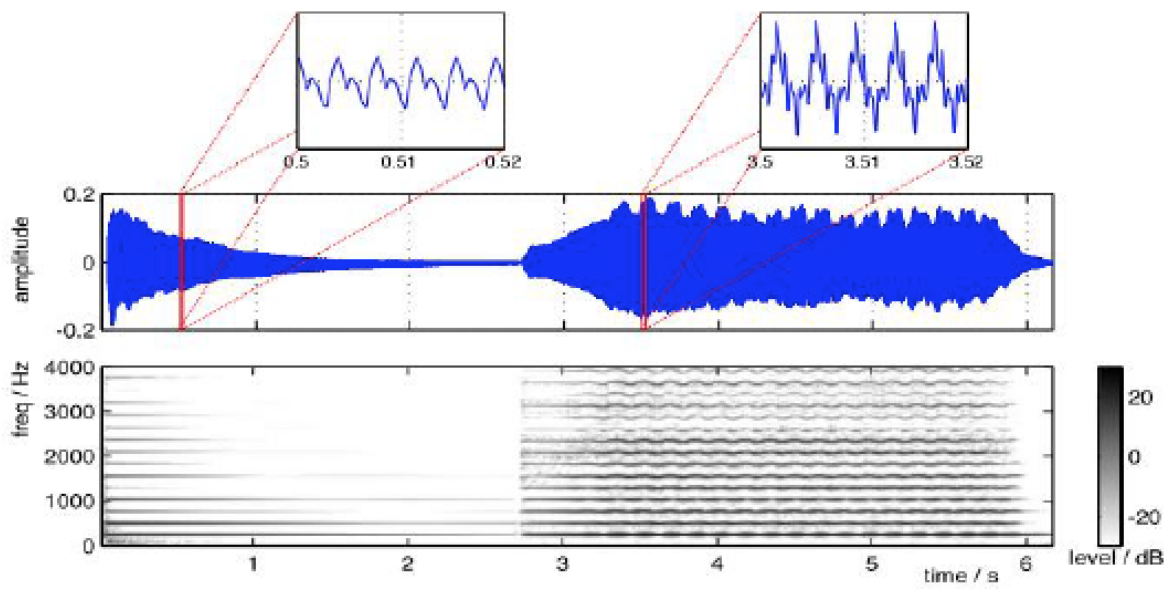
Figure 3. The measure *%V* is plotted on the y-axis, in reverse order. The standard deviation of intervocalic intervals **Δ** , is given on the x-axis.

TIMBRE

Le timbre se définit comme la répartition spectrale de l'énergie d'un son.

- on peut différencier les instruments de musique par leur timbre (et aussi leur attaque)
- l'humain a la faculté de changer son timbre de voix en modifiant la forme de son conduit vocal (acteur, expressivité)
- les humains se distinguent par leur timbre de voix

[Mueller]

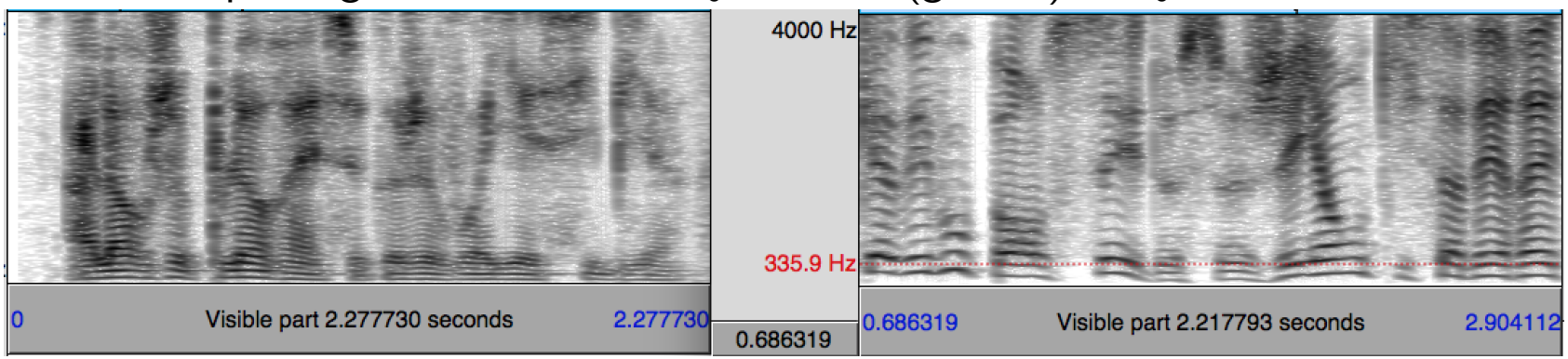


TIMBRE

Le timbre se définit comme la répartition spectrale de l'énergie d'un son.

- on peut différencier les instruments de musique par leur timbre (et aussi leur attaque)
- l'humain a la faculté de changer son timbre de voix en modifiant la forme de son conduit vocal (acteur, expressivité)
- les humains se distinguent par leur timbre de voix

spectrogramme d'homme $F_0 = 90$ Hz (gauche) et $F_0 = 136$ Hz



DESCRIPTEURS DE TIMBRE

Pour l'identification du locuteur, on utilise principalement les MFCCs

- 13 premiers coefficients (modélisation du conduit vocal)
- ajout du $MFCC_0$ pour avoir une information sur l'énergie
- éventuellement ajout de la F_0

Quel sens donner à un timbre donné ?

Timbre	sombre / clair
	détimbré / timbré
	sourd / brillant
	dureté
	nasillard
	présence de souffle sur la voix
Oscillations F_0	vibrato
	tremolo / tremor
	jitter
Mode de production	voix grinçante
	voix criée
	voix chuchotée

[Garnier & Abrilian]

DESCRIPTEURS DE TIMBRE

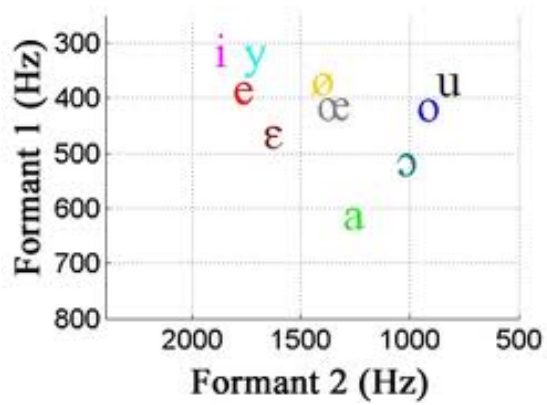
D'autres descripteurs de qualité vocale sont utilisés pour décrire la qualité vocale (très sensible au bruit de fond):

- jitter et shimmer: micro-variations de la F_0 ou énergie

$$J_N = \frac{N}{N-1} \frac{\sum_0^{N-1} T_0(k+1) - T_0(k)}{\sum_0^N T_0(k)}$$

- rapport harmonique sur bruit (HNR) $\rightarrow$ mesure de voisement
- coefficient de relaxation $\rightarrow$ mesure de relâchement des cordes vocales
- rugosité (modulation de la F_0)
- aire du triangle vocalique $\rightarrow$ mesure d'articulation

triangle vocalique



ACCENTUATION / PROÉMINENCES

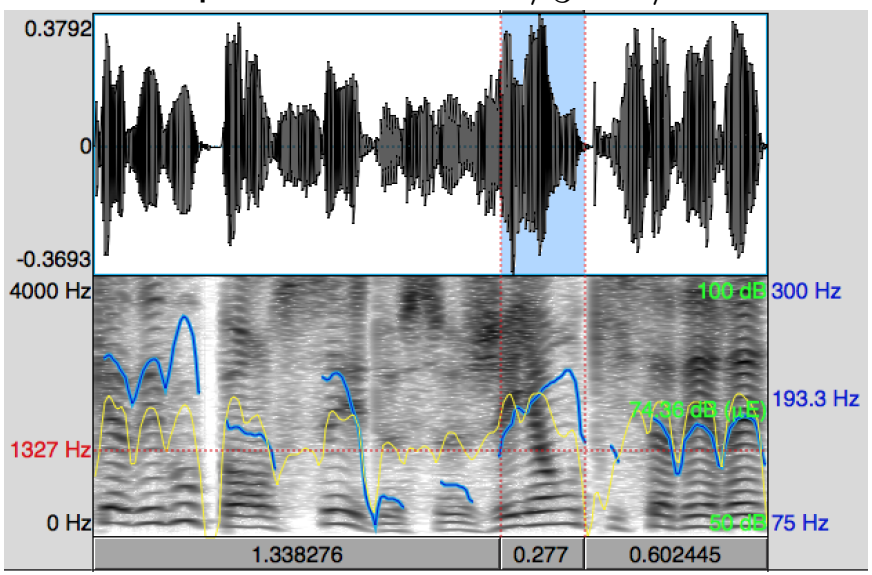
Accentuation:

- accentuation de certaines syllabes
- le système d'accent est spécifique à une langue
  - Français: accent pas nécessaire à la compréhension mais à l'expression
  - Anglais: accent nécessaire à la compréhension

Proéminence

- manifestation acoustique d'un accent autour d'un noyau syllabique.

proéminence sur /ʒ e ɑ̃/



- pic F_0
- forte intensité
- élongation (durée de la syllabe allongée)

VECTEUR PROSODIQUE

- Ex: pour la détection automatique des émotions dans la parole.
- Segment acoustique: segment de parole émotionnel correspondant à un groupe de souffle.
- 1 segment = 1 vecteur acoustique de 384 paramètres

LLD (16 × 2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F_0	kurtosis, skewness
(Δ) HNR	extremes: value, relative position, range
(Δ) MFCC1-12	linear regression: offset, slope, MSE

[Schuller]

APPLICATIONS

Avantages:

- permet de capturer des informations perceptives de haut-niveau.
- proche de la perception humaine.
- multiplicité des descripteurs

Inconvénients:

- très sensible au bruit de fond.
- les paramètres sont très corrélés entre eux.
- lesquels choisir ?

Applications

- détection des émotions,
- traitement du signal social,
- synthèse de la parole,
- ...

LES BOITES A OUTILS

Il existe plusieurs boîtes à outils qui extraient des descripteurs à partir du signal audio (parole et musique) moyennant le choix de quelques paramètres

- HTK : MFCCs principalement
- AUBIO, OpenSmile (C++), Yaafe: descripteurs spectraux et prosodiques
- librosa: librairie Python qui fait plein de choses
- SPRO4, YIN: extraction de la F0
- IrcamDescriptor
- etc...

Pour la visualisation des signaux:

- Praat
- Sonicvizualizer

PLAN DE LA SECTION ACTUELLE

LA PAROLE

Le language est:

- une faculté spécifiquement humaine et universelle
- un système de représentation régi par une grammaire

La langue est:

- une réalisation particulière de language
- constituée de règles et normées partagées par les membres d'une communauté.

La parole correspond à l'usage de la langue orale (par opposition à écrite).

LA PAROLE

Niveaux d'analyse:

Acoustic-phonétique	présence des sons d'une langue
phontactique	fréquence et enchaînement des sons
prosodique	intonation, rythme, accentuation
lexical	mots possible d'une langue
syntaxique	enchaînement possible des mots dans une langue
sémantique	sens de l'enchaînement des mots
pragmatique	informations relatives au contexte

PHONEMES

Résonateurs et formants:

- les formants sont les fréquences de résonances de la cavité bucale,
- leur valeur dépend du volume de la cavité et de ses ouvertures,
- les 3 premiers formants permettent de caractériser une voyelle
  - ex: /i/ $F_1 = 300$, $F_2 = 2200$, $F_3 = 3000$ Hz.

Paramètres en phonétique articulatoire:

- point d'articulation: position de la langue par rapport au palais
- Aperture: section du conduit vocal au point d'articulation
- Labialisation: forme des lèvres
- Nasalité: passage de l'air par le conduit nasal
- Latéralité: passage de l'air de part et d'autre de la langue

PHONEMES DU FRANÇAIS

- 36 phonèmes en français: 16 voyelles, 3 semi-consonnes (/j w ɥ/) + 17 consonnes
- Référence International Phonetic Alphabet

Voyelles orales

/i/	pie	/a/	patte
/e/	été	/ɑ/	pâte
/ɛ/	modèle	/o/	auditeur
/y/	puni	/ɔ/	porte
/ø/	deux	/u/	poux
/œ/	peur	/ə/	petite

Voyelles nasales

/ɑ̃/	an	/œ̃/	brun
/ɛ̃/	matin	/ɔ̃/	bon

PHONEMES DU FRANÇAIS

Plosives orales	labiales	alvéolaires	vélaires
sourdes	/p/: poids	/t/: toit	/k/: quoi
voisées	/b/: bois	/d/: doigt	/g/: goût

Occlusives	labiale	alvéolaire	palatales	
nasales	/m/: mon	/n/: nous	/ɲ/: agneau	/ŋ/:smoking

Fricatives	dentales	alvéolaires	post-alvéolaires
sourdes	/f/: feu	/s/: soir	/ʃ/: poche
voisées	/v/: voix	/z/: zéro	/ʒ/: jeu

Liquides	/l/: long	/ʁ/: rond

Semi-voyelles	/w/: oui	/j/: fille	/ɥ/: lui

INTERNATIONAL PHONETIC ALPHABET

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

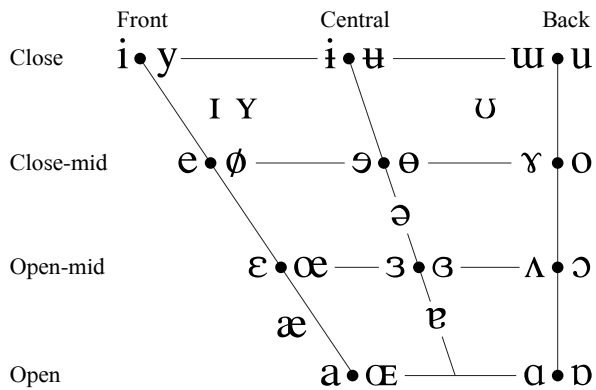
CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	’ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
ɕ ʑ Alveolo-palatal fricatives
w Voiced labial-velar approximant
ɺ Voiced alveolar lateral flap

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

INTERNATIONAL PHONETIC ALPHABET

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͡s k͡p

Open

Where symbols appear in pairs, the one to the right represents a rounded vowel.

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝ (ɪ̝ = voiced alveolar fricative)			
◌̩	Syllabic	n̩	◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)			
◌̯	Non-syllabic	e̯	◌̙	Advanced Tongue Root	e̙			
◌˞	Rhoticity	ɚ a˞	◌̘	Retracted Tongue Root	e̘			

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən

ˌ Secondary stress

ː Long eː

ˑ Half-long eˑ

◌̆ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
é̋ or ˥	Extra high		ě or ˩˥	Rising	
é ˦	High		ê ˥˩	Falling	
ē ˧	Mid		e᷄ ˦˥	High rising	
è ˨	Low		e᷅ ˩˧	Low rising	
ȅ ˩	Extra low		e᷈ ˩˥˩	Rising-falling	
↓	Downstep		↗	Global rise	
↑	Upstep		↘	Global fall	

Typefaces: Doulos SIL (metatext); Doulos SIL, IPA Kiel, IPA LS Uni (symbols)

COARTICULATION ET ASSIMILATION

La coarticulation:

- effet d'inertie articulatoire $\rightarrow$ minimisation de l'effort articulatoire.
- modification de la réalisation acoustique en fonction du contexte phonétique

L'assimilation:

- élision du /ə/ (schwa) en élocution rapide
  - ex: petite fille /p t i t ʃ i j ə/ au lieu de /p ə t i t ə ʃ i j ə/
- dévoisement des fricatives sonores si la consonnes suivante est sourde
  - ex: médecin /m e t s ɛ̃/ au lieu de /m e d ə s ɛ̃/
- voisement des plosives et fricatives sourdes si la consonnes suivante est voisée
  - ex: pâquebot /p a g b o/ au lieu de /p a k a b o/

APPLICATIONS

Avantages:

- un système international
- modélisation symbolique
- existence de beaucoup d'étude en linguistique pour faire le lien entre phonétique, linguistique, syntaxe et sémantique.

Inconvénients:

- difficulté pour faire le lien entre le symbole et sa réalisation acoustique (le phone)
- outils d'alignement (signal / chaîne phonétique) sensible au bruit et au locuteur

Applications:

- construction de modèles de language
- synthèse de parole