UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN – Mohamed BOUDIAF
Faculté des mathématiques et d'informatique- Département d'informatique

2018-2019

# Cours de Traitement Automatique de la Parole

1ière année Master IAA

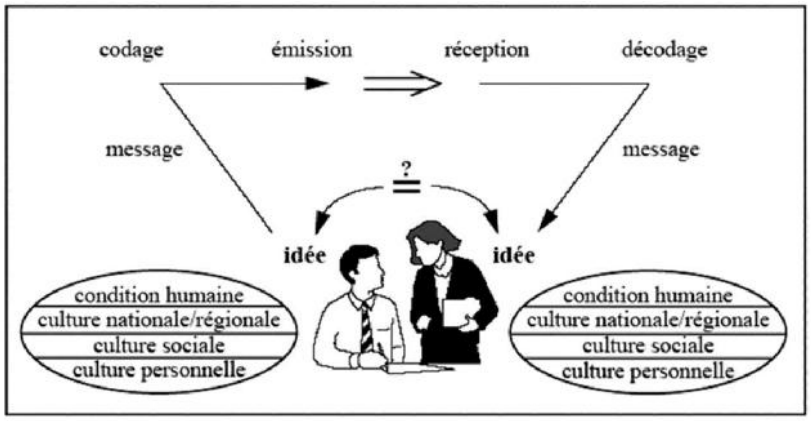

Mr. BENDAHMANE Abderrahmane

# Plan du cours

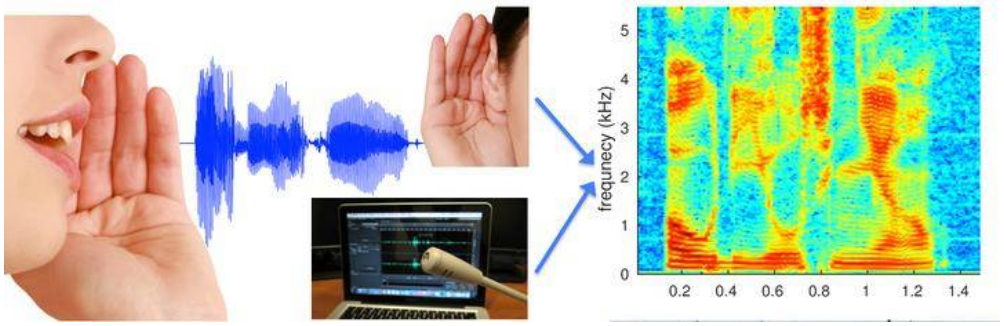
# I Introduction au traitement automatique de la parole

## I.1 Introduction

La parole est un moyen de communication qui permet d'exprimer sa pensée via des sons émis par les organes phonateurs.




Le Traitement Automatique de la Parole (TAP) c'est le faite de manipuler des signaux de parole sur machine.




## I.2 Domaines d'application

Le TAP s'applique dans plusieurs domaines tel que :

**Traduction automatique**
à partir de la parole et synthèse du résultat.

**Commande vocale**
de véhicules, jeux vidéo, robots...

**Télécommunications**
Compression de la parole, cryptage, transmission.

**Dictée vocale et synthèse**
Speech to text, Text to speech afin de simplifier l'envoi et la lecture des messages.

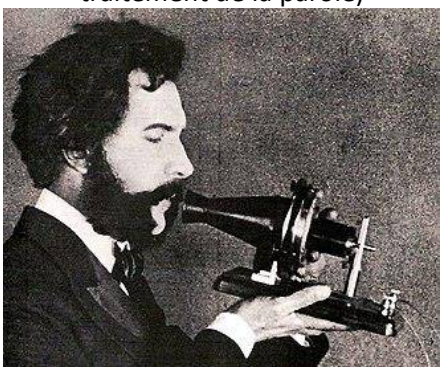
**Aide aux personnes avec un handicap**
(ex : synthèse vocale pour les malvoyants).

## I.3 Evènements historiques

Voici quelques évènements historiques qui ont favorisés les recherches en TAP.

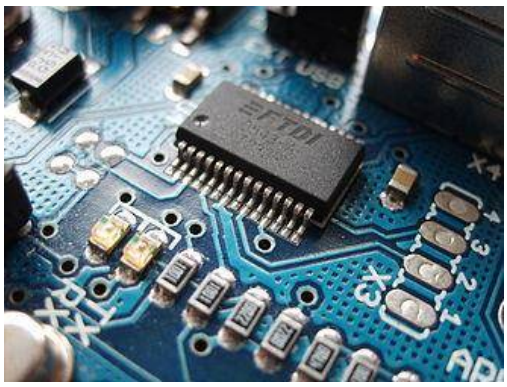
**18ième siècle**
Tentatives de création de synthétiseurs de parole (Wolfgang von kempelen)

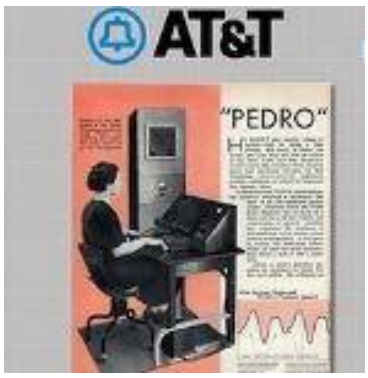
**Invention du téléphone 1876**
(Début des recherches sérieuses sur le traitement de la parole)



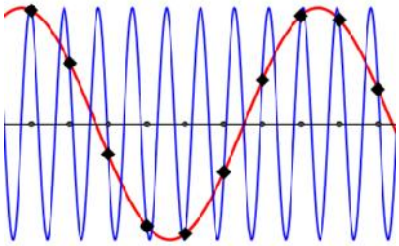
**L'électronique 1920**
Augmentation des possibilités de traitement du signal par des composants (Bande passante Hz, volume db, multiplexage …)

**1936**
Nouvelles tentatives de réalisation d'une machine de synthèse de la parole (projet AT&T)






**1948-1949 Shannon**
Théorie de l'échantillonnage
Théorie de l'information
(Numérisation du signal, étude spectrale, études phonétique)



**1970-1971**
Recherches en reconnaissance de la parole par plusieurs compagnies

DARPA IBM

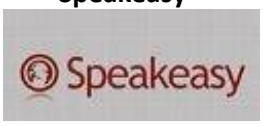
L'avancée technologique dans le domaine du TAP a favorisé l'apparition de solutions commerciales dont les plus connues :

**1997**
**Solution DRAGON**


**1998**
**Learnout & Hauspie**

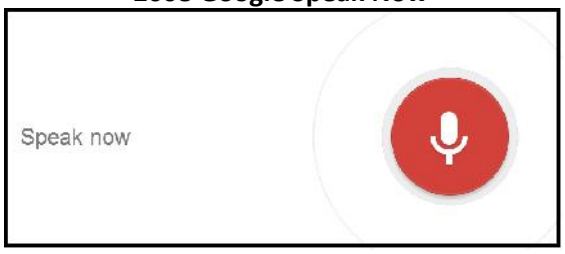

**2000**
**NetByTel**


**2007**
**Speakeasy**




**2007 Siri de Apple**


**2008 Google Speak Now**



**2018 Smart speakers:**

- Amazon Echo Spot
- Google Home
- …

## I.4 Les sous domaines du TAP

Le TAP comporte plusieurs sous domaines dont :

- **Codage, compression et transmission** de la parole qui est surtout utilisé en télécommunication et les logiciels de voip, ça permet de réduire la taille des données générée par la parole afin de le transmettre à très bas débit.
- **Débruitage et amélioration** de la parole qui trouve son utilité dans le domaine de l'audiovisuel, la musique et la téléphonie …
- **Reconnaissance de la parole** dans des applications d'intelligence artificielle, elle peut comporter l'une des applications suivantes :
  - Commandes vocales : pour la commande de logiciels et de robots.
  - Parole continue : dans des applications de dictée vocale.
  - Reconnaissance labiale : lecture sur les lèvres.
  - Reconnaissance parole/non parole : détection de la parole dans un signal.
- **Reconnaissance vocale** en biométrie pour identifier ou ouvrir l'accès à une personne grâce à sa voix.
- **Manipulation de la voix** surtout utilisée en musique pour éliminer les défauts de la voix.
- **Reconnaissance des émotions** afin de de reconnaitre l'état émotionnel d'une personne à partir de ce qu'il dit.
- **Synthèse de la parole** afin de permettre aux logiciels/machines de générer de la parole le plus naturellement possible et de communiquer avec l'homme avec la parole, lire du contenu textuel ...

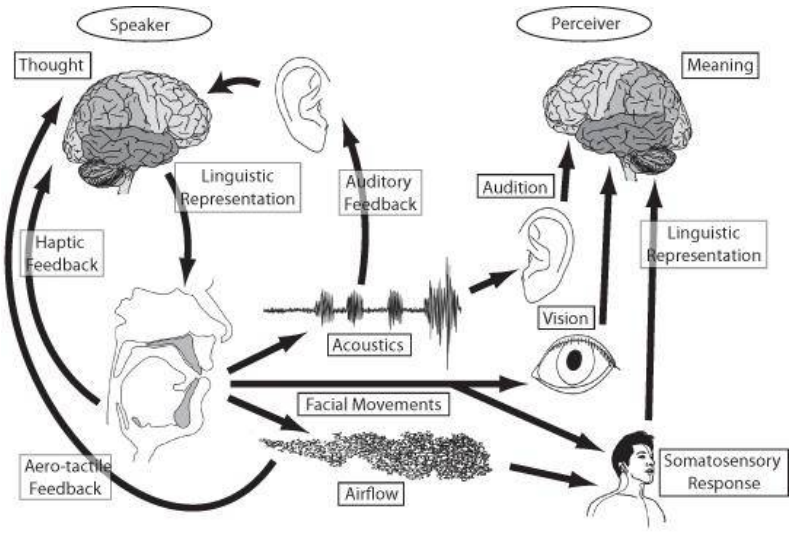
## I.5 Les outils du TAP

La recherche en TAP s'appuie sur plusieurs outils théoriques, informatiques et mathématiques.

### I.5.1 Outils théoriques – La phonétique

Branche de la linguistique qui étudie les caractéristiques physiques des sons de la parole (phones), elle se divise en trois branches principales :

- **La phonétique articulatoire** qui s'intéresse au mécanisme de production de la parole.
- **La phonétique acoustique** qui s'intéresse au signal de parole
- **La phonétique auditive** qui s'intéresse aux caractéristiques de l'oreille




### I.5.2 Outils théoriques – La phonologie

Branche de la linguistique qui étudie l'organisation des sons de la parole dans une langue sans prendre en considération leurs caractéristiques physiques.

**Les phonèmes de la langue arabe en symboles API**

| L'alphabet arabe | | | | | |
|---|---|---|---|---|---|
| Dad | [dˤ] | ض | alif | [a] | ا |
| Ta | [tˤ] | ط | ba | [b] | ب |
| Dha | [ðˤ] | ظ | ta | [t] | ت |
| Ayn | [ʕ] | ع | tha | [θ] | ث |
| ghayn | [ɣ] | غ | jim | [ʒ] | ج |
| fa | [f] | ف | Ha | [ħ] | ح |
| Qaf | [q] | ق | kha | [x] | خ |
| kaf | [k] | ك | dal | [d] | د |
| lam | [l] | ل | dhal | [ð] | ذ |
| mim | [m] | م | ra | [r] | ر |
| nun | [n] | ن | za | [z] | ز |
| ha | [h] | ه | sin | [s] | س |
| waw | [w] & [u] | و | shin | [ʃ] | ش |
| ya | [j] & [i] | ي | Sad | [sˤ] | ص |

Exemples de transcription phonétique (API) :

- **Les voyelles en arabe (6):**

  **a/â** : كتَبَ kataba / بـاب bâb
  **i/î** : مِنْ min / في f î
  **u/û** : جـُلوس ʒulûs / جـُلوس ʒulûs

- **Chedda:**

  ad-dâr (la maison) الـدّار

- **Tafkhim:**

  a**s-s**ayf (l'épée) السّيْف
  a**sˁ-sˁ**ayf (l'été) الصّيْف

### I.5.3 Outils théoriques – prosodie

La prosodie représente le changement en : ton/tonalité/intonation/modulation, accent, rythme (vitesse d'élocution) de l'expression orale, afin de transmettre nos émotions et intentions à nos interlocuteurs.

**Exemples :**

- Ton déclaratif : صَبَاحُ الْخَيْرْ
- Ton interrogatif : مَا اسْمُكَ ؟
- Ton exclamatif: مَا أَجْمَلَ...

### I.5.4 Outils théoriques – Traitement Automatique du Langage Naturel

Le langage naturel est traité selon différents niveaux :

- **Lexique :** l'ensemble de ses lemmes/mots d'une langue (sorte de vocabulaire).
- **Syntaxe :** branche de la linguistique qui étudie la combinaison des mots dans une langue pour former des phrases ou des énoncés.
- **Sémantique :** branche de la linguistique qui étudie le sens du message qu'on veut transmettre dans une langue.

Le problème qui se pose est que l'être humain ne respecte pas toujours les règles de la langue qu'il parle.

### I.5.5 Outils informatiques et mathématiques - Traitement du signal

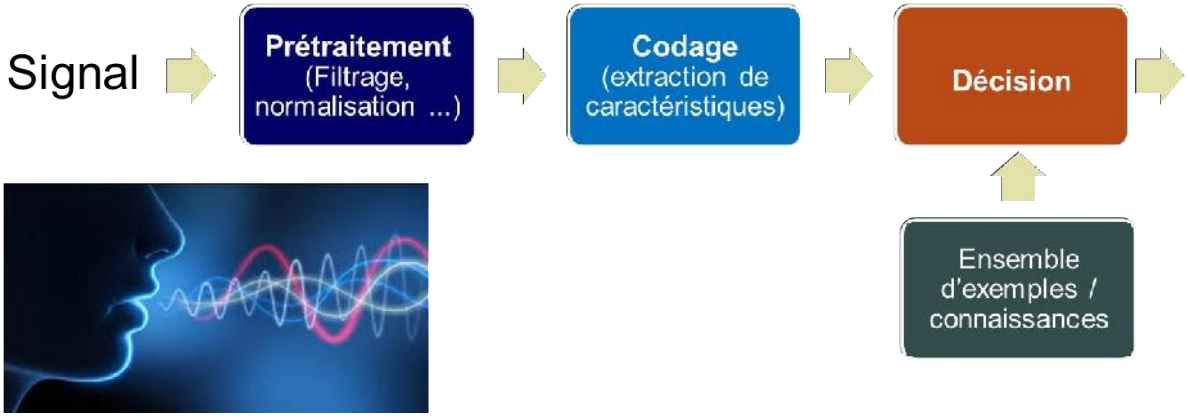
Un signal est une variation d'une grandeur de nature quelconque, transportant de l'information. Le traitement du signal est une discipline qui permet de manipuler, d'analyser et d'interpréter des signaux généralement en s'appuyant sur des moyens mathématiques, informatiques et électroniques.

**Exemples :**

- **La transformée de Fourier :** elle permet le passage d'une représentation temporelle vers une représentation fréquentielle du signal.
- **Autocorrélation :** elle permet l'étude de la périodicité du signal.

### I.5.6 Outils informatiques et mathématiques - Intelligence Artificielle

Le TAP fait principalement appel aux outils de la reconnaissance des formes tels que : la distance DTW, le HMM ou les réseaux de neurones afin d'identifier et de labéliser des composantes du signal de parole. Le processus se divise principalement en 3 phases : le prétraitement su signal, l'extraction de ses caractéristiques et la prise de décision qui repose sur un ensemble d'exemples ou des connaissances.




## I.6 Difficultés en TAP

En TAP on est confronté à plusieurs difficultés, principalement :

- Le bruit : dû à l'environnement d'acquisition ou le matériel utilisé.
- La variabilité inter et intra locuteurs : variation du signal de parole par rapport à lui-même et aux autres locuteurs.
- L'accent (variabilité phonétique) : pour les personnes étrangères la prononciation des phonèmes est souvent approximative.
- Variabilité linguistique : phrases avec la même signification mais avec une prononciation différente.
- Parole continue : les mots sont connectés sans frontières détectables afin de les identifier.
- La coarticulation : due à la vitesse d'élocution, chaque phonème altère les caractéristiques des phonèmes voisins.

## I.7 Conclusion

La parole est le moyen de communication le plus naturel et le plus simple entre les êtres humains. Permettre à la machine d'interagir avec la parole ouvre beaucoup de possibilités d'applications futuristes afin d'améliorer notre quotidien. Cependant, beaucoup de connaissances sont requises afin de réaliser cette interaction. Bien que le domaine soit très évolué actuellement, il subsiste toujours certaines difficultés à surmonter.

# II La phonétique articulatoire

## II.1 Introduction

La phonétique articulatoire est une branche de la phonétique qui classe les sons de la parole selon leur mode de production par le système articulatoire humain. Elle permet de créer des modèles articulatoires artificiels afin de modéliser et de générer de la parole.

## II.2 Le rôle de la phonétique articulatoire en TAP

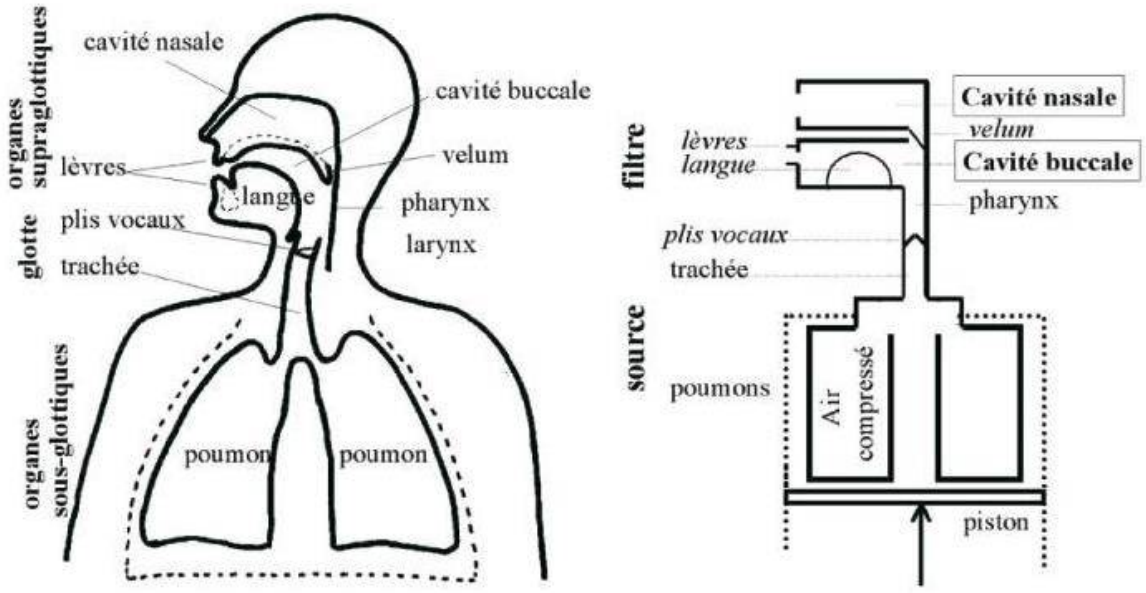
Etablir des relations entre les caractéristiques acoustiques des sons émis et leur mode de production par le système articulatoire permet à la fois :

- D'extraire les caractéristiques articulatoires à partir d'un signal de parole et de faire de la reconnaissance ou de la compression.
- De produire un signal de parole synthétisé à partir des caractéristiques articulatoires.

## II.3 La production de la parole

La production de la parole est un mécanisme complexe qui résulte de l'interaction entre le système neurologique et les muscles qui entrent dans l'articulation des sons des langues (phones) :

1. Les poumons fournissent de l'air.
2. Le passage de l'air à travers la glotte produit un son voisé ou bruité (selon les cordes vocales).
3. Le son final est dessiné par les organes supra-glottiques (pharynx, langue, lèvres …)



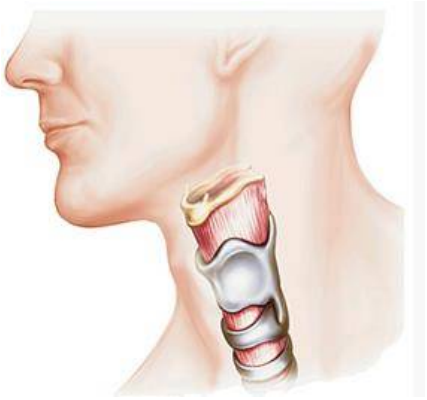

## II.4 Les caractéristiques articulatoires

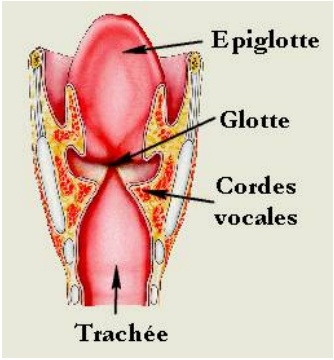
Les caractéristiques articulatoires permettent de faire une distinction entre les phonèmes selon leur mode d'élocution.

### II.4.1 La distinction voyelle/consonne

Du point de vue articulatoire on peut diviser les phonèmes en deux grandes classes :

- **Les voyelles :** l'air passe librement à travers la glotte (les cordes vocales sont ouvertes et vibrent), les articulations sont quasi fixes. Exemples : a, o, i, ou, an, am …
- **Les consonnes :** l'air est obstrué en passant par la glotte (fermeture totale ou partielle de la glotte). Exemples : s, ch, t, d, k …






**Larynx**

### II.4.2 La distinction consonne occlusive/fricative

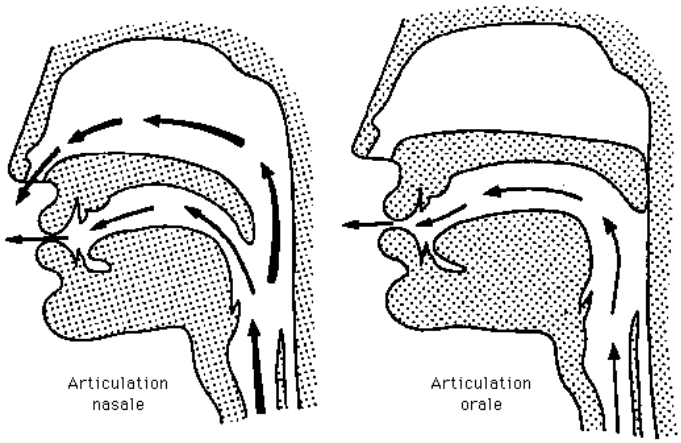
- **Fricative :** La glotte est fermée partiellement avec une petite ouverture qui génère un bruit de friction, il est possible de rallonger sa prononciation dans le temps.
  Exemples : ف, ث, ذ, س, ز, ش, خ, غ, ح, ع,ه …
- **Occlusive :** La glotte est fermée complètement puis s'ouvre brusquement pour produire un bruit de courte durée, il est impossible de rallonger sa prononciation.
  Exemples : ب, ت, د, ك, ق …

### II.4.3 La distinction voisé (sonore)/non voisé (sourde)

- **Voisé :** avec vibration des cordes vocales.
  Exemples : د ,ز,ب , ع, ي, و …
- **Non voisé :** sans vibration des cordes vocales.
  Exemples : ت, ط, ف, س, ح …

### II.4.4 Autres distinctions

- **Nasale/orale :** passage de l'aire par la cavité nasale ou non.
  Exemples : م, ن , an, am, em…

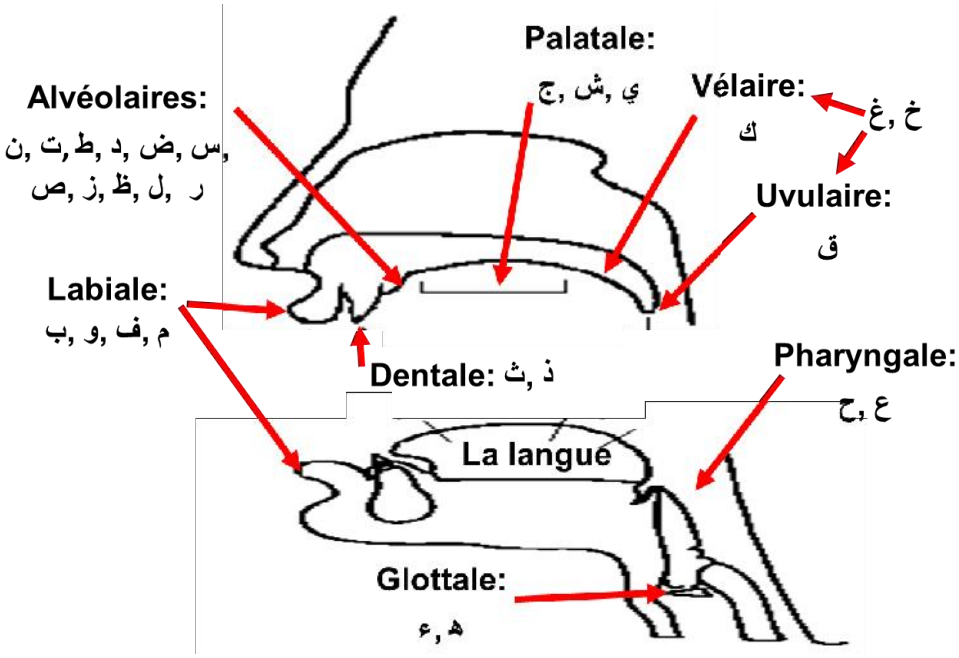



- **Vibrante :** avec vibration de la langue.
  Exemples : ر…

- **Semi-voyelles :** passage d'une voyelle à une autre.
  Exemples : و, ي…
- **Emphatique/non :** dotée de l'articulation d'emphase.
  Exemples : ذ/ظ , د/ض

### II.4.5 Points d'articulation

Le point d'articulation désigne généralement la localisation de la pointe de la langue lors de la prononciation du phonème, pour certains phonèmes qui ne nécessitent pas la participation de la langue c'est l'endroit où s'effectue l'obstruction au passage de l'air éjecté lors de leur prononciation, on distingue les points d'articulation suivants : labiale, alvéolaire, dentale, palatale, vélaire, uvulaire, pharyngale, glottale.




## II.5 Classification articulatoire des phonèmes de la langue arabe

Ce tableau résume les caractéristiques articulatoires de chaque phonème de la langue arabe avec son symbole API.

| | | Labiale/ Labio-dentale | Dental | Alvéolaire | | Palat. | Vél. | Uvul. | Pharyng. | Glot. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Ordinaire | emphatique | | | | | |
| Nasale | | m م | | n ن | | | | | | |
| Occlusive | soude | | | t ت | tˤ ط | | k ك | q ق | | ʔ ء |
| | Sonore | b ب | | d د | dˤ ض | ʒ ج | | | | |
| Fricative | sourde | f ف | θ ث | s س | sˤ ص | ʃ ش | x خ | | ħ ح | h ه |
| | Sonore | | ð ذ | z ز | ðˤ ظ | | ɣ غ | | ʕ ع | |
| semi-voyelle | | w و | | | | j ي | | | | |
| Latérale | | | | l ل | lˤ ل | | | | | |
| Vibrante | | | | r ر | | | | | | |

## II.6 Conclusion

La phonétique articulatoire nous permet de diviser l'ensemble des phonèmes en des catégories selon leur mode d'élocution, cette distinction offre une première caractérisation théorique qui pourra nous aider par la suite afin d'expliquer les caractéristiques acoustiques des phonèmes. La phonétique articulatoire ouvre la voie vers la création de systèmes artificiels de synthèse de la parole, ainsi que des systèmes de compression à très bas débit.

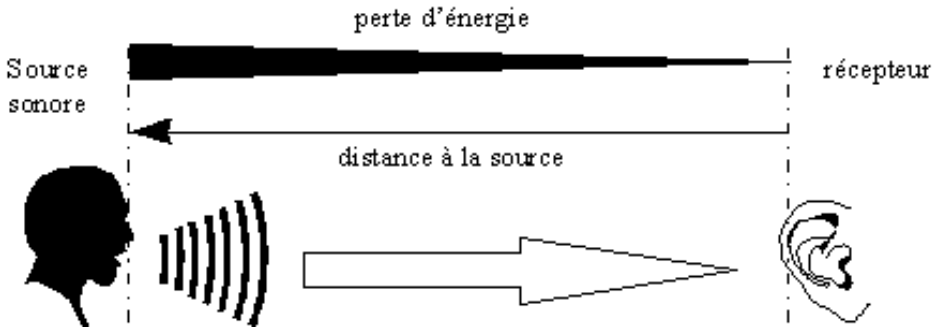
# III La phonétique acoustique

## III.1 Introduction

La phonétique acoustique est une branche de la phonétique qui étudie l'onde sonore produite par les organes phonatoires. Elle propose une classification des sons de la parole basée sur des caractéristiques physiques de l'onde qu'on visualise généralement grâce à des outils informatiques.

## III.2 Définition du son

Physiquement, un son est causé par des vibrations dans l'air. Ces vibrations peuvent se propager et sont sujette à la résistance de l'air et des obstacles traversés qui font atténuer leur énergie.




## III.3 La numérisation du son

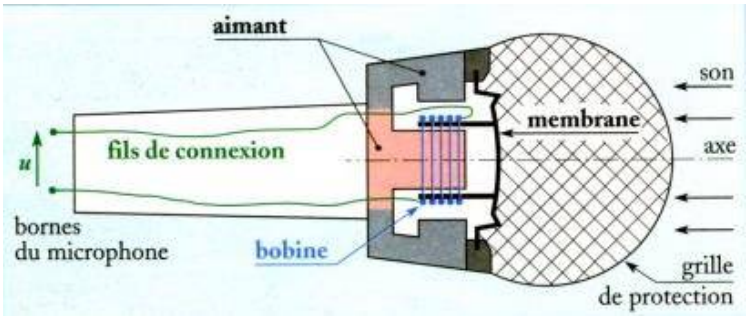
La numérisation du son consiste à capturer et à enregistrer le son dans un support numérique (Disque dure, CDROM, carte mémoire, RAM …). Une quantification des grandeurs physiques de l'onde sonore vers une représentation binaire est nécessaire.

### III.3.1 Capture et rendu du son

La capture de l'onde sonore sur ordinateur se fait généralement via un microphone et son rendu via les haut-parleurs.

### III.3.2 Principe du microphone

- Les vibrations de l'onde sonore traversent la grille de protection et font bouger la membrane (morceau de plastique très fin).
- Les mouvements de la membrane font bouger la bobine à l'intérieur d'un aimant permanent.
- Ces mouvements génèrent un faible courant électrique. Sur les bornes du microphone.




### III.3.3 La carte son

Le courant électrique généré du microphone est traité dans la carte son comme suite :

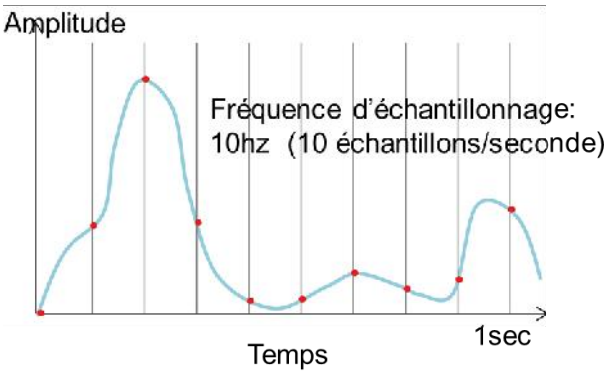
- Étape1 : échantillonnage du signal
- Étape2 : quantification du signal
- Étape3 : codage du signal

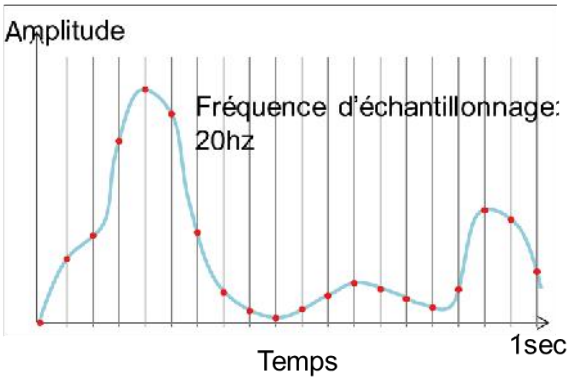
Ce sont les étapes de la numérisation du son.

### III.3.4 L'échantillonnage

L'échantillonnage consiste à prendre des mesures de l'amplitude du signal capté à intervalles de temps réguliers :

- Plus les intervalles sont courts, plus le signal est finement représenté.
- Le nombre d'échantillons pris dans une seconde représente la fréquence d'échantillonnage.



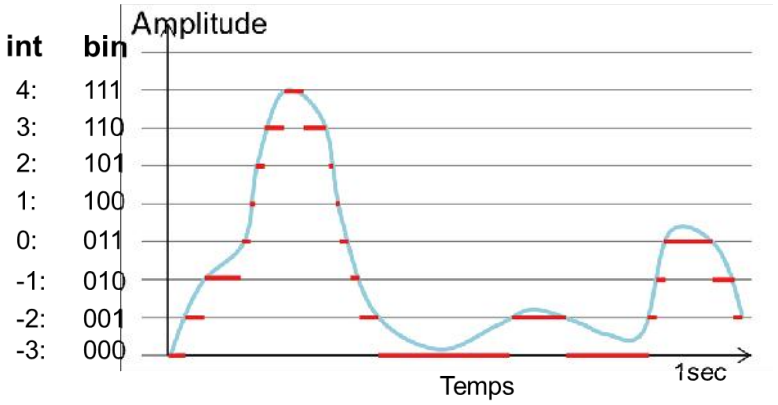
### III.3.5 La quantification

La quantification consiste à représenter les amplitudes des échantillons avec un certain nombre de bits :

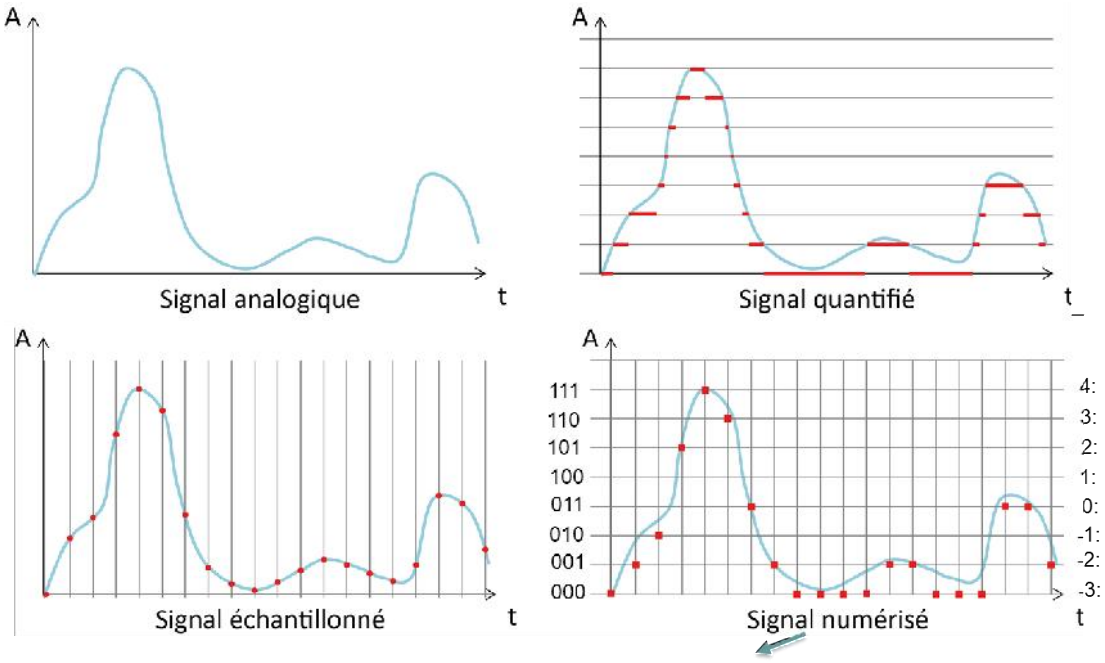
- Augmenter le nombre de bits permet d'améliorer la qualité.
- 2 bits → 4 valeurs, 4bits → 16 valeurs, 8bits → 256 valeurs
- 16bits → 65536 valeurs

Signal quantifié sur 3bits → 8 valeurs

(En rouge → signal quantifié)




**En combinant l'échantillonnage + la quantification :**




[000, 001, 010, 101, 111,110, 011, 001, 000, 000, 000, 000, 001, 001, 000,000, 000, 011, 011, 001 ] (binaire)

= [-3, -3, -1, 2, 4, 3, 0, -2, -3, -3, -3, -3, -2, -2, -3, -3, -3, 0, 0, -2] (en entier)

### III.3.6 Codage du signal

Le signal représenté par un vecteur de valeurs après les étapes d'échantillonnage et de quantification passe par l'étape de codage qui désigne le format de stockage des valeurs (type de compression appliqué) :

- Le codage PCM (sans compression) qu'on trouve dans les formats wav ou aiff, les valeurs sont stockées telles quelles.
- Un codage MPEG3 les valeurs sont compressées (format mp3).

**Exemple :**

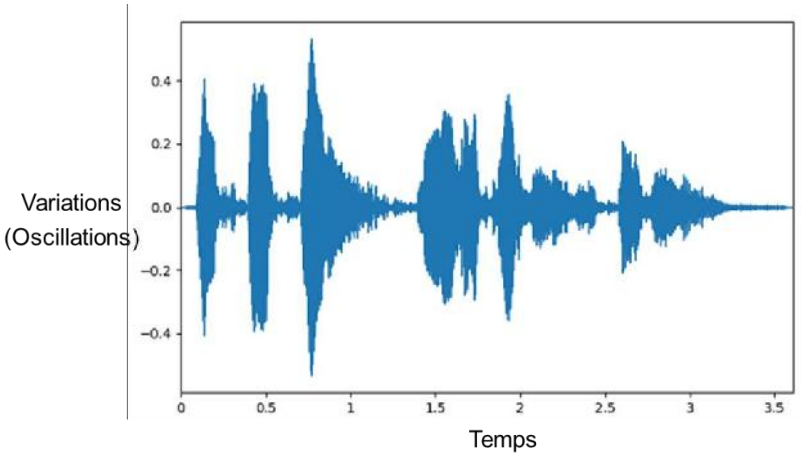
- Un CD audio :
  - Codage PCM (sans compression)
  - Quantifié sur 16bits : (chaque valeur nécessite 2 octets)
  - Fréquence d'échantillonnage : 44khz (44000 échantillons enregistrés à la seconde)
  - Stéréo (2 canaux : canal droit et gauche)

1 seconde = 2 canaux x 1seconde x 44000 échantillons x 2 octets = 176000 octets

1 heure = 2 x 3600 x 44000x 2 = 633600000 octets ≈ 604Mo

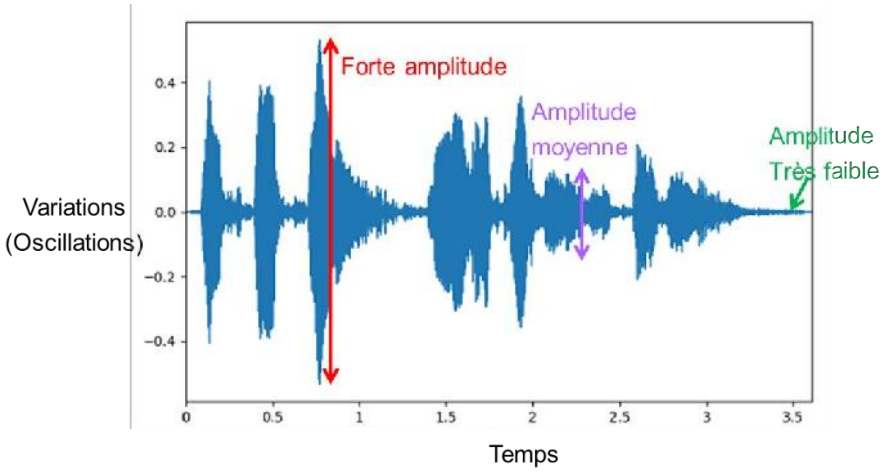
## III.4 Signal temporel / oscilloscope

L'oscilloscope est un simple dessin des valeurs numérisées (non compressées) du signal sonore sous forme d'une courbe. Cependant, il est difficile de visualiser des informations sur le signal avec cet outil.



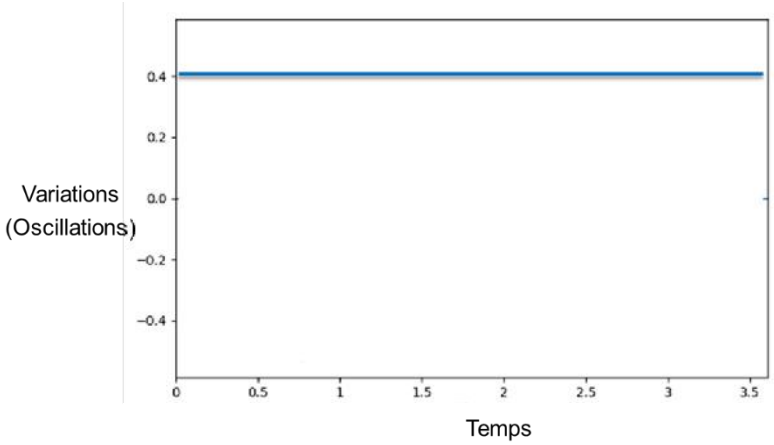

### III.4.1 L'amplitude

L'amplitude représente la force d'oscillation (vibration) du signal, elle est représentatif du volume/énergie du son.



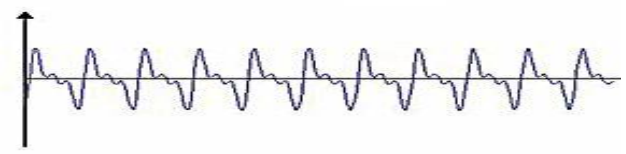
### III.4.2 Signal constant

Un signal constant représente un silence (0 variations → amplitude = 0).

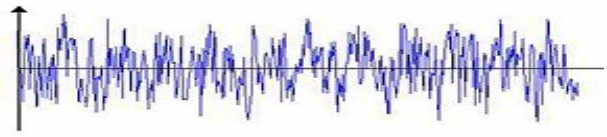



### III.4.3 Signal périodique/apériodique

Un signal de parole périodique est indicateur de vibration des cordes vocales, on parle de signal voisé.

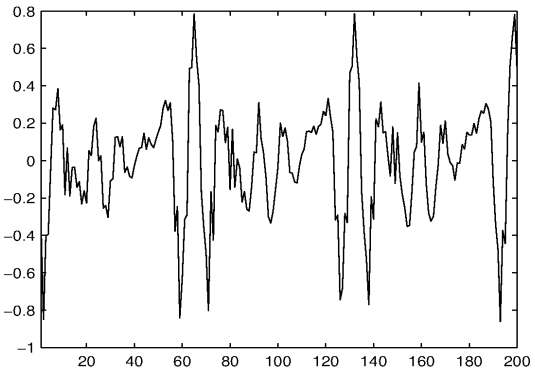


Un signal de parole apériodique (aléatoire) est caractéristique des fricatifs et occlusifs non voisés.



### III.4.4 Signal pseudopériodique

Pour un signal réel on parle de signal pseudopériodique (avec une périodicité non parfaite).




### III.4.5 La fréquence fondamentale (F0)

Pour un signal de parole pseudopériodique, la fréquence de ce dernier indique la fréquence de vibration des cordes vocales, qui permet de savoir s'il s'agit d'une voix grave (homme) ou aigue (femme/enfant).

**Exemple de calcul :**

- 3 répétitions durant 200 échantillons (voir figure précédente).
- On suppose qu'on a échantillonné à 6000 échantillons/seconde
- Combien de répétitions/seconde ?

**Réponse :**

- F0= 3 x 6000 / 200 = 90Hz.

**Comparaison :**

- F0 d'un homme : [80hz, 180hz]
- F0 d'une femme : [130hz, 300hz]
- F0 d'un enfant : [250hz, 400hz]

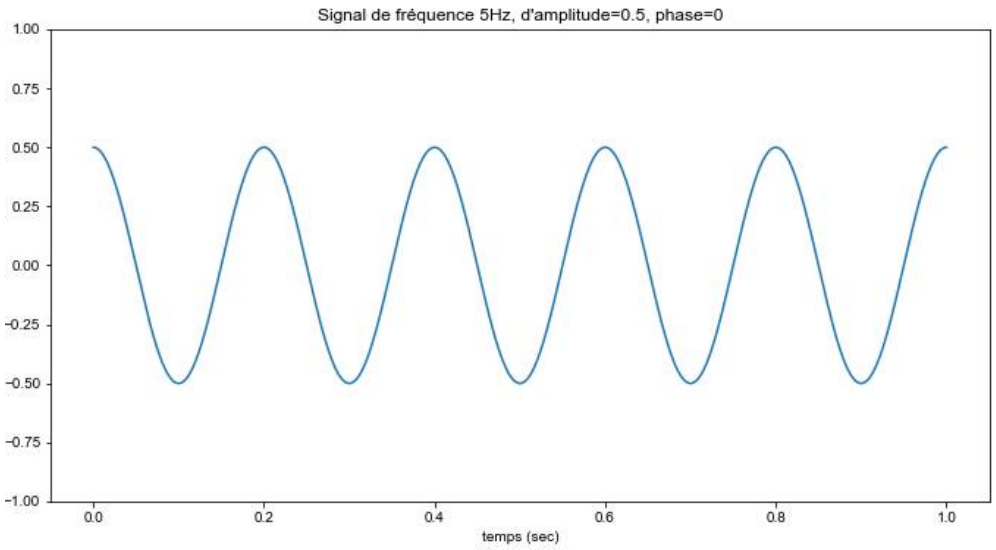
## III.5 L'analyse fréquentielle

L'analyse fréquentielle d'un signal consiste à le décomposer en plusieurs ondes de propriétés différentes (fréquences, amplitudes …). Il s'agit le plus souvent de le décomposer en sommation d'ondes de fréquences variées. Cette analyse permet une meilleure visualisation des caractéristiques d'un signal qui sont affichées sous forme de spectres.

**Exemple :** On génère une onde d'amplitude $a$, de fréquence $f$, de phase $d$ et $t$ représente le temps en secondes.

$$y = a\cos(2\pi ft + d)$$

Pour $a$ =0.5 et $f$=5 et $d$ =0 on obtiens le signal suivant pour $t\epsilon$[0,1]

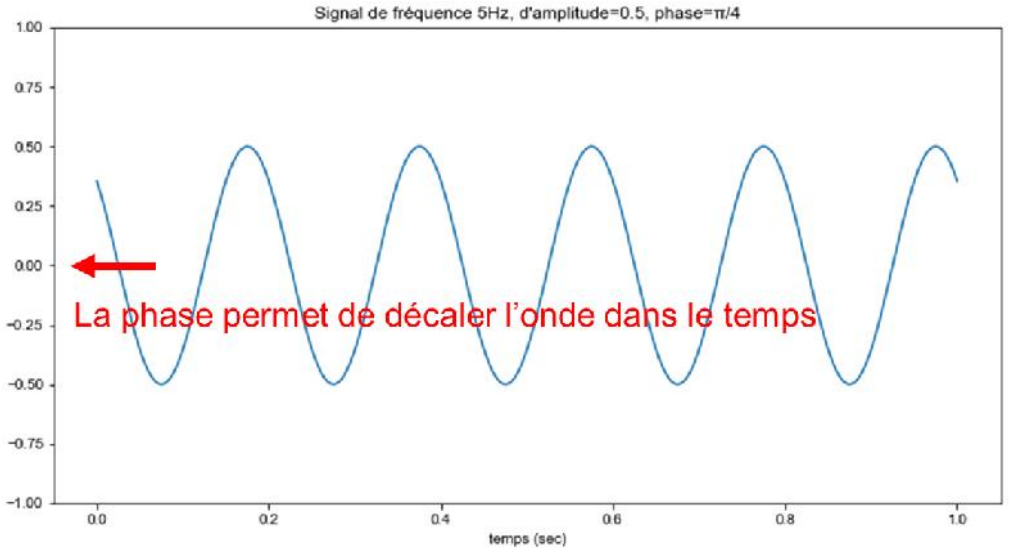



### III.5.1 La phase

La phase permet d'estimer le retard de l'onde périodique en radian. Par exemple pour l'onde :

$$y = a\cos(2\pi ft + d)$$

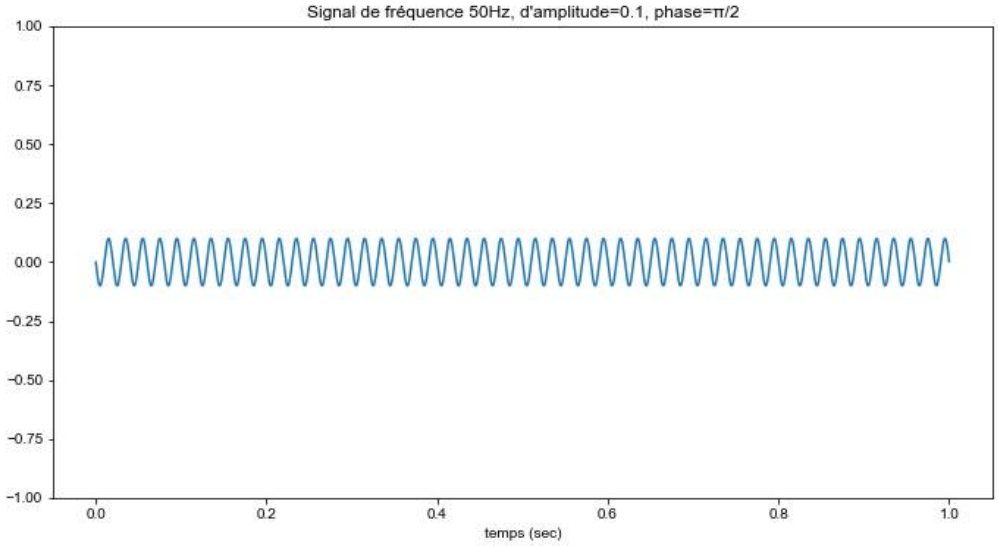
Pour $a$ =0.5 et $f$=5 et $d$ =$\pi/4$ on obtiens le signal suivant pour $t$ϵ[0,1]




### III.5.2 L'amplitude et la fréquence

L'amplitude correspond à la hauteur de l'onde (grande hauteur → onde énergétique→ son plus fort), qu'on a la fréquence, elle correspond au nombre d'oscillations par seconde, elle se mesure en hertz (Hz).

**Exemple** : pour $a$ =0.1 et $f$=50 et $d$ =$\pi/2$ on obtiens le signal suivant pour $t$ϵ[0,1]



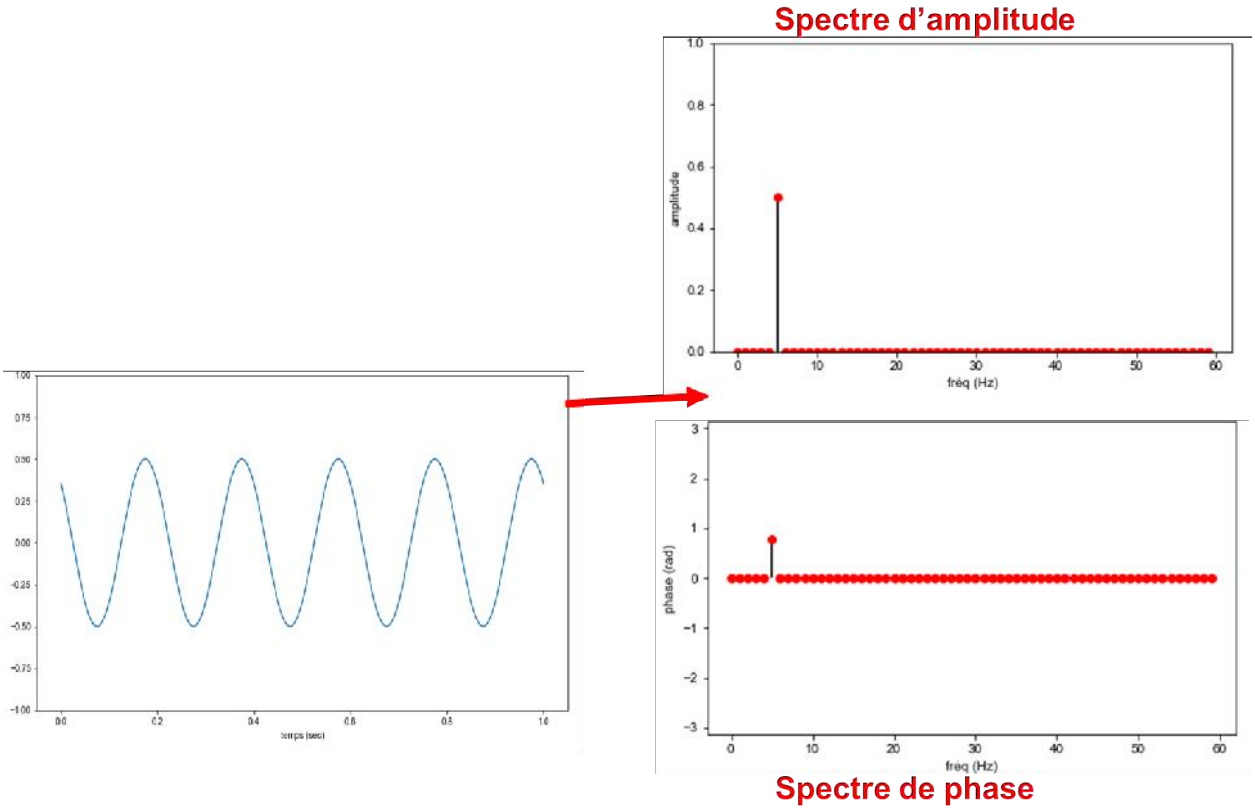

### III.5.3 Passage du domaine temporel au fréquentiel par transformée de Fourier

Le passage du domaine temporel au domaine fréquentiel permet de visualiser le signal sous forme de sommation d'ondes de différentes fréquences en l'affichant sous forme de spectres amplitude/phase.

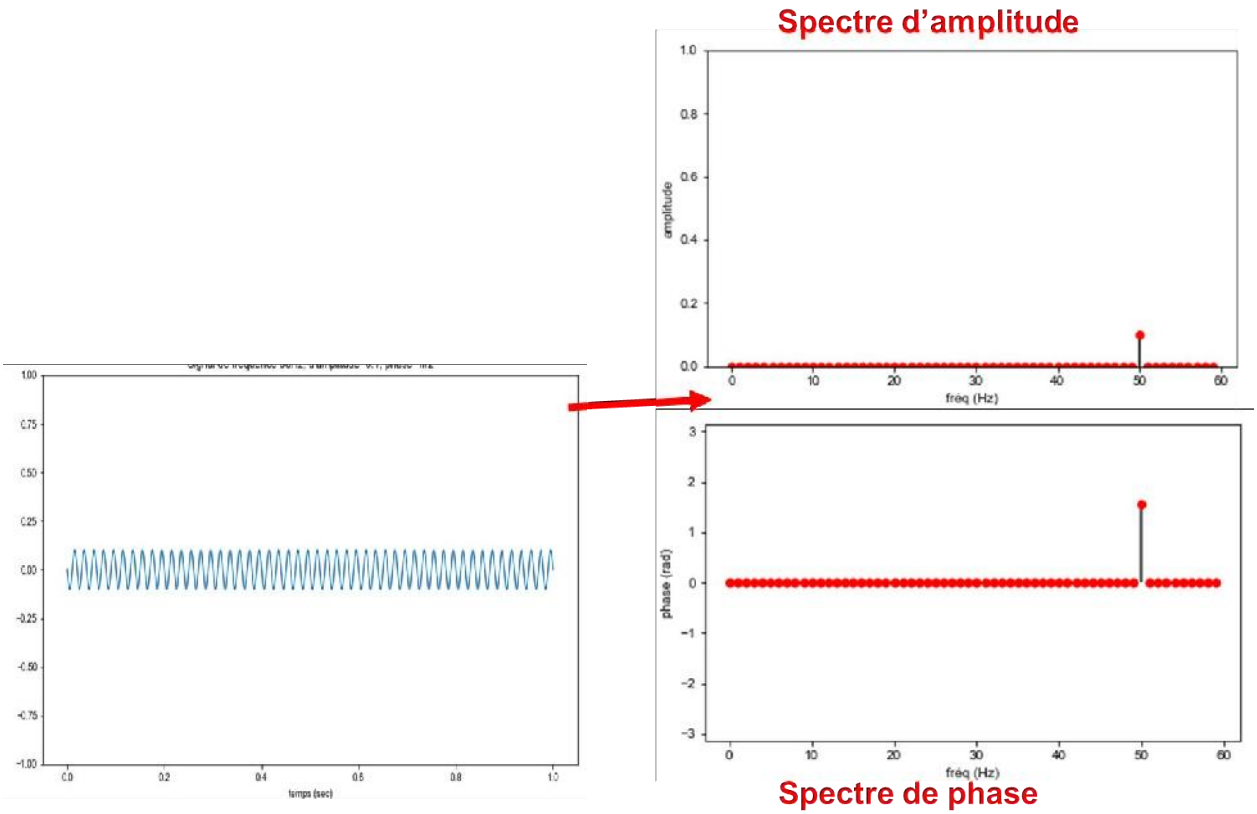
**Exemple :** pour un signal composé d'une seule onde générée avec la formule suivante

$$y = a\cos(2\pi ft + d)$$

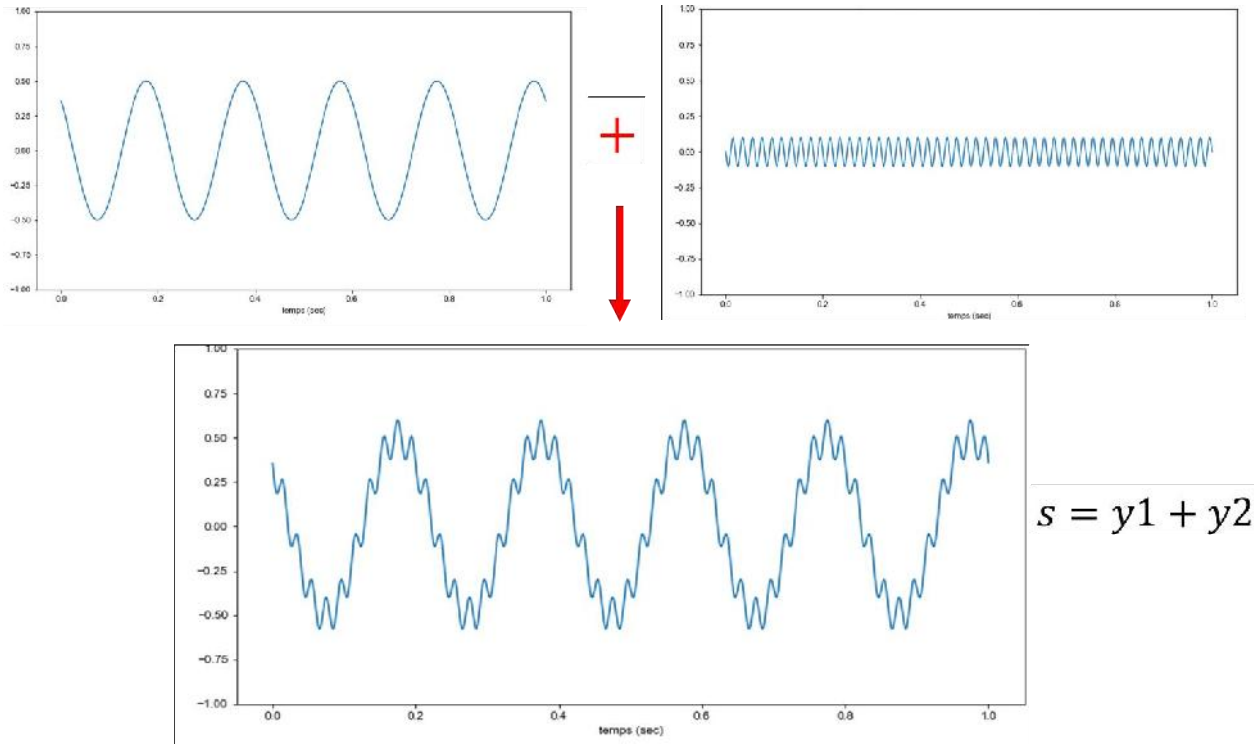
Les paramètres suivants : $a = 0.5,\ f = 5Hz,\ d = \frac{\pi}{4}$



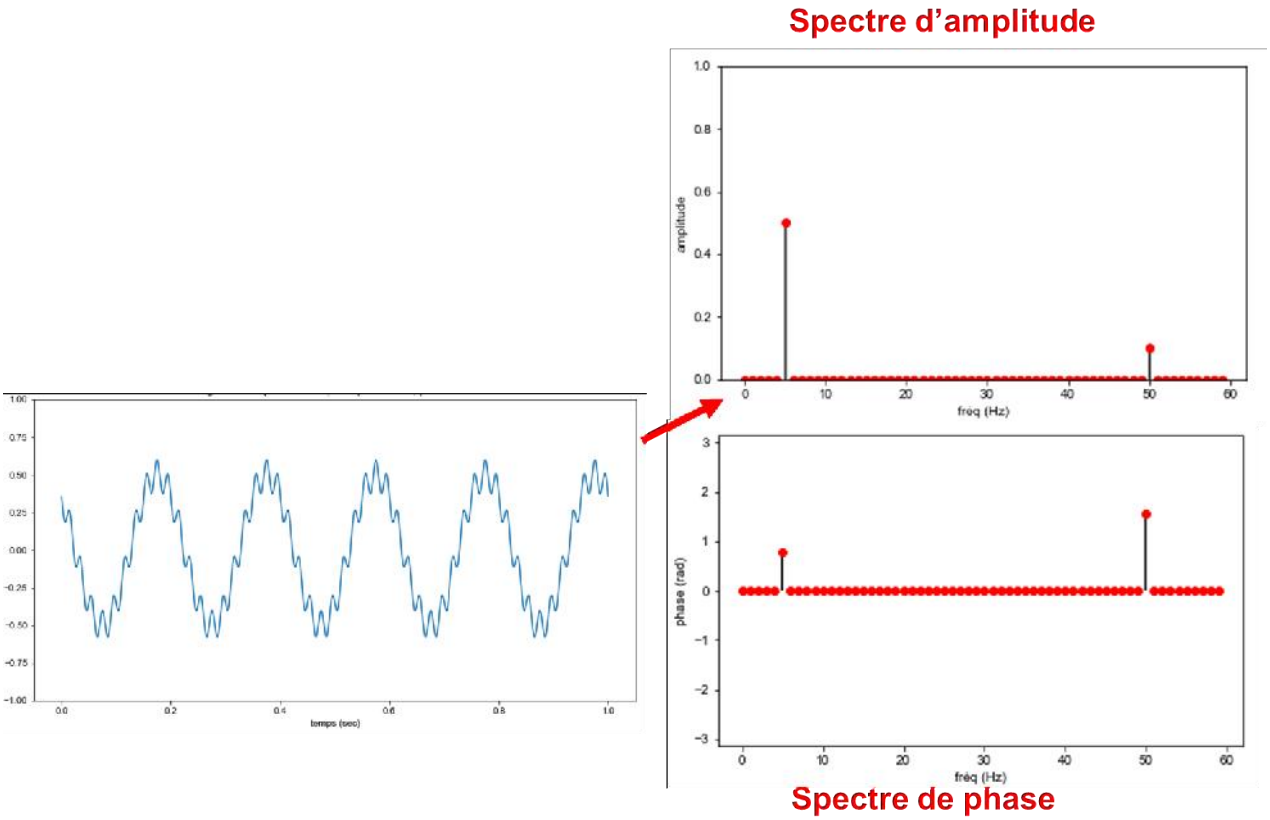

Les paramètres suivants : $a = 0.1,\ f = 50Hz,\ d = \frac{\pi}{2}$



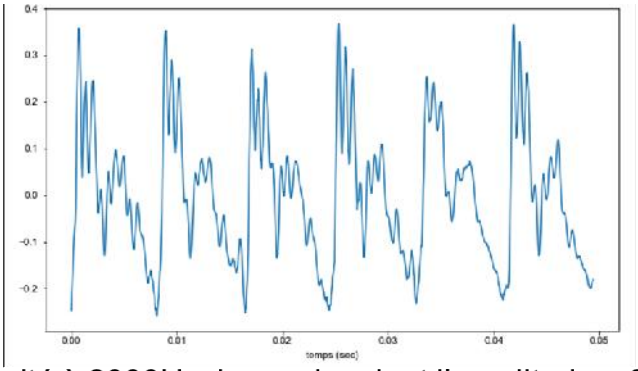
Les paramètres suivants → $y1$: $a = 0.5$, $f = 5Hz$, $d = \frac{\pi}{4}$ et $y2$: $a = 0.1$, $f = 50Hz$, $d = \frac{\pi}{2}$



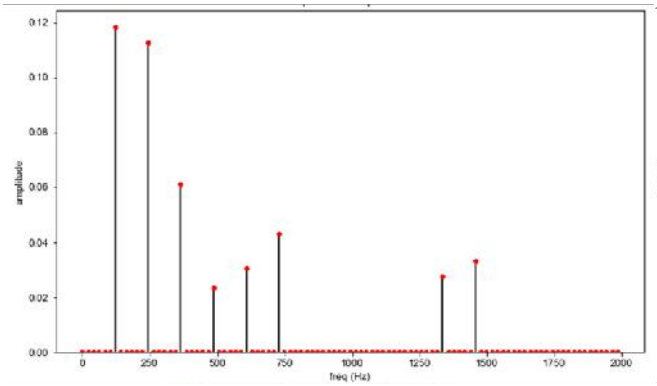

L'analyse fréquentielle de *s*=y1+y2 donne :



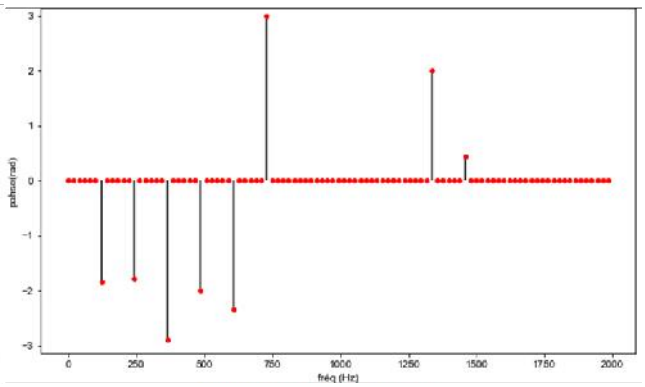
L'analyse fréquentielle d'un signal réel de parole (phonème « un »prononcé par un homme) par une transformée de Fourier :




Affichage limité à 2000Hz, les ondes dont l'amplitude < 0.02 ont été annulées

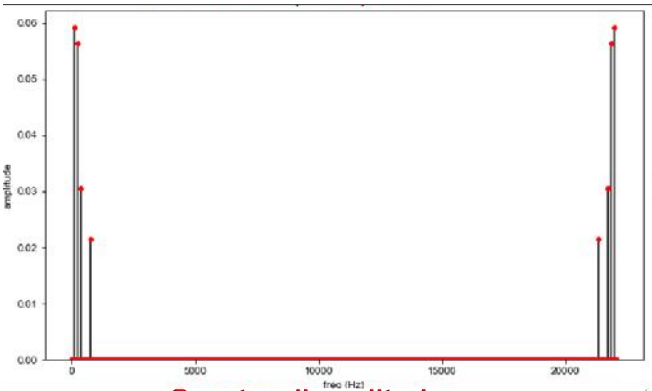



**Spectre d'amplitude**

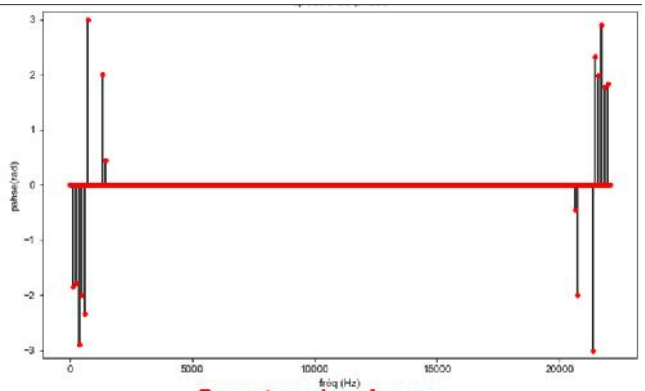



**Spectre de phase**

**Affichage complet**, les ondes dont l'amplitude < 0.02 ont été annulées




**Spectre d'amplitude**




**Spectre de phase**

**Remarques :**

- On remarque un effet miroir dans les 2 spectres, cet effet miroir est due à l'échantillonnage du signal et aux propriétés mathématiques de la transformée de Fourier.
- La représentation spectrale avec la transformée de Fourier ne peut représenter que les ondes dont la fréquence est inférieur à la moitié de la fréquence d'échantillonnage.

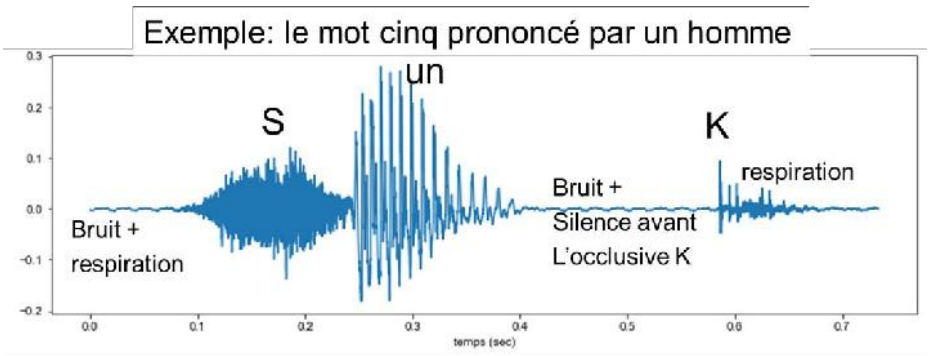
Fréquence Max = ½ Fréquence d'échantillonnage -1 →

La fréquence d'échantillonnage doit être supérieure à 2 fois la fréquence maximale des ondes qu'on veut étudier

**Théorème de Shannon :** Fréquence d'échantillonnage $>$ 2 $\times$ Fréquence max

## III.6 Le spectrogramme

Les caractéristiques fréquentielles d'un signal peuvent changer dans le temps, on dit que le signal n'est pas stationnaire.

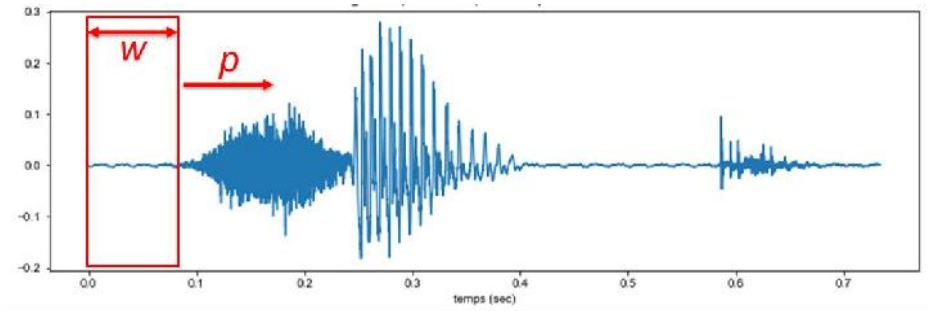



### III.6.1 Le fenêtrage

L'étude fréquentielle d'un signal non stationnaire se fait sur de petites périodes où **on suppose** qu'il est pseudo-stationnaire.

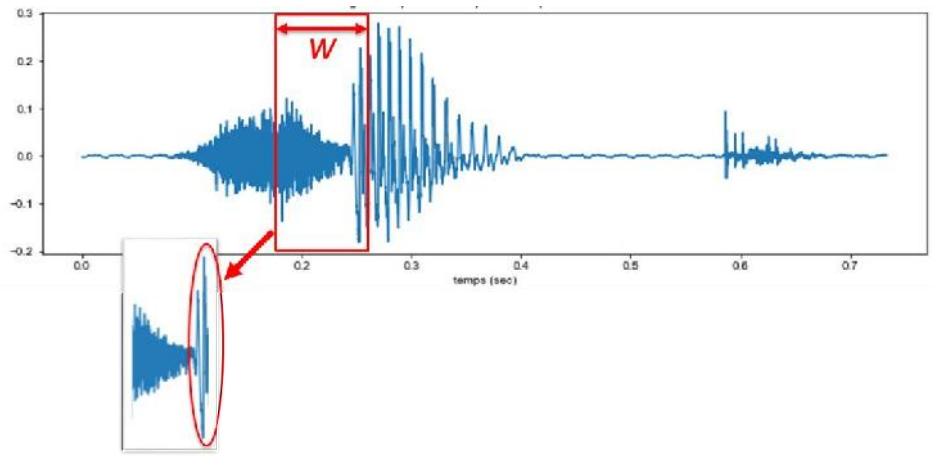
Cette solution est appelée fenêtrage, où on utilise une fenêtre glissante sur le signal de taille fixe $w$, qui avance avec un pas $p$.

Pour $p < w$ on aura des fenêtres qui se chevauchent.

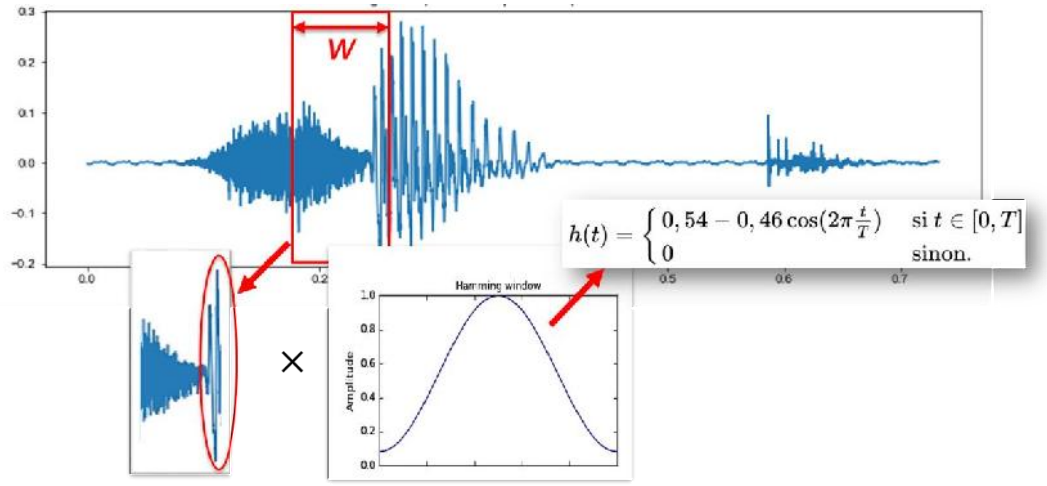



### III.6.2 L'effet de bord

La supposition de la pseudo-stationnarité n'étant pas toujours vrai (effet de bord), les bords de la fenêtre peuvent prendre des parties non stationnaires.



L'effet de bord induit du bruit sur le spectrogramme, afin de réduire ce bruit, on multiplie la fenêtre par des fonctions telle que Hamming, Hanning, Blackman ce qui permettent de lisser le spectrogramme...




### III.6.3 La préaccentuation du signal

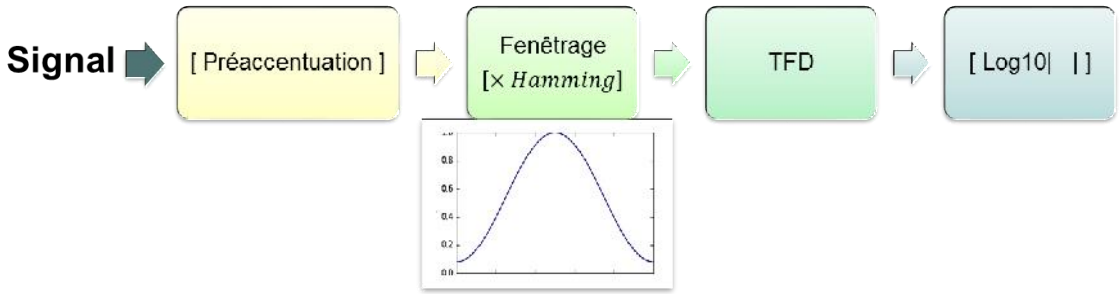
Les hautes fréquences ont naturellement une amplitude faible par rapport aux basses fréquences, un filtre passe haut (rehausseur des hautes fréquences) permet d'améliorer la visibilité des hautes fréquences sur le spectrogramme.

**Exemple :** Filtre de préaccentuation → $x(t) = x(t) - a \times x(t-1)$

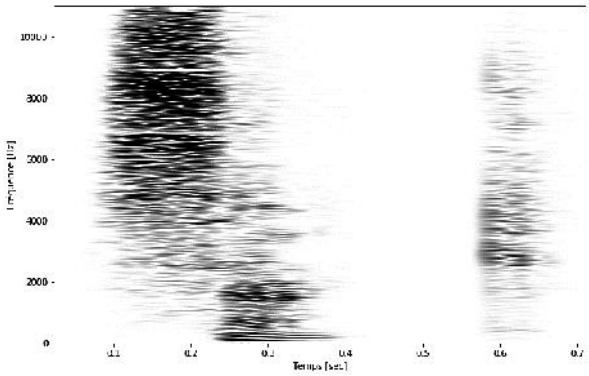
avec $a = 0.98$ $(valeur\ usuelle)$

### III.6.4 Les étapes du calcul du spectrogramme

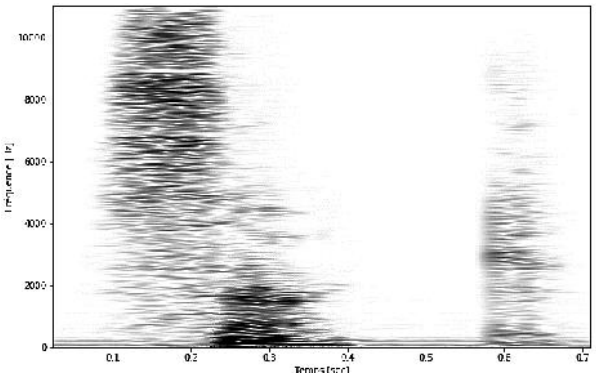
Le calcul du spectrogramme comprend des étapes optionnelles (afin d'améliorer la qualité d'affichage telles que la préaccentuation, la multiplication par une fenêtre lissante et le logarithme), ainsi que des étapes nécessaires telles que le fenêtrage, la transformée de Fourier discrète (DFT).



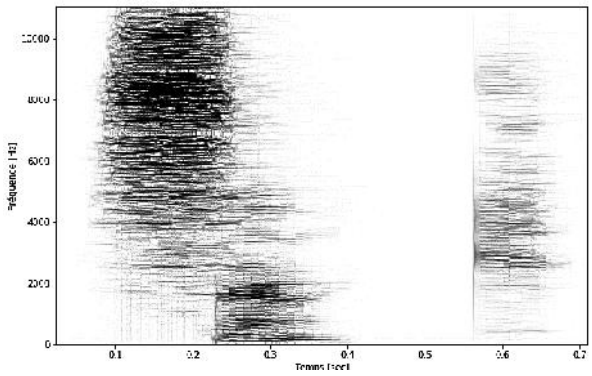
Exemple du spectrogramme du mot « cinq » prononcé par un locuteur masculin :



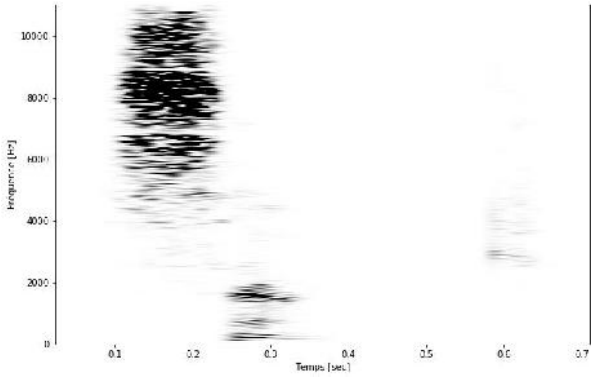
Avec préaccentuation, Hamming et log



Sans préaccentuation



Sans Hamming (bruit → effet de bord)



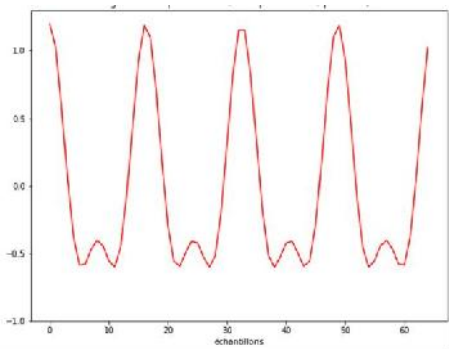
Sans log| |

### III.6.5 Les limites de la transformée de Fourier

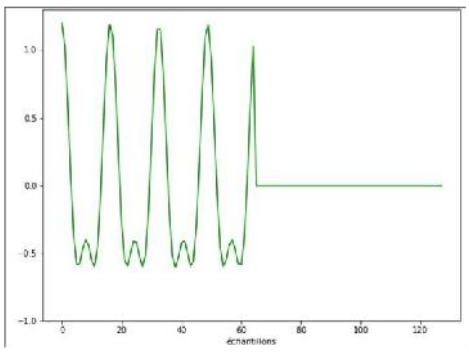
- Une grande fenêtre temporelle implique plus de précision en fréquences et moins de précision dans le temps (les caractéristiques des sons de courte durée sont mélangées avec d'autres sons →risque de non stationnarité de la fenêtre).
- Une petite fenêtre temporelle implique moins de précision en fréquences et plus de précision dans le temps (résolution fréquentielle faible).
- La transformée de Fourier rapide (FFT: Fast Fourier Transform) est une version modifiée et beaucoup plus rapide en calcul que la transformée de Fourier discrète, cependant elle exige une taille de fenêtre de puissance 2.

### III.6.6 Le zero padding (remplissage par des zéros)

L'idée est de remplir la fin du signal par des zéros afin d'obtenir une fenêtre puissance de 2 et d'augmenter la résolution fréquentielle (meilleure précision en fréquences sans perdre la précision dans le temps).

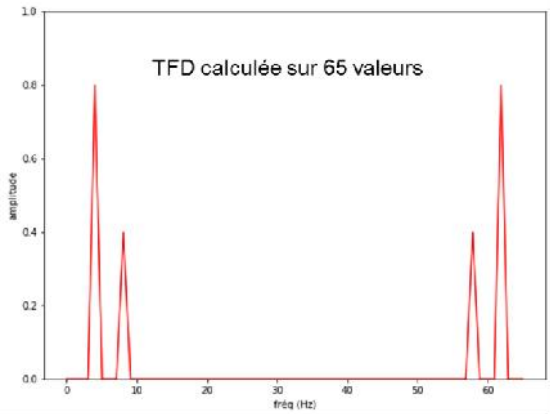


Exemple d'une fenêtre sur 65 valeurs

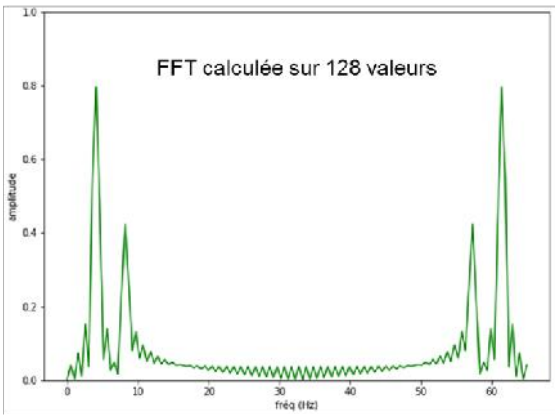


Fenêtre de 65 valeurs avec Zero padding jusqu'à 128

L'idée est de remplir la fin du signal par des zéros afin d'obtenir une fenêtre puissance de 2 et d'augmenter la résolution fréquentielle (meilleure précision en fréquences sans perdre la précision dans le temps).




Spectre d'amplitude sans zero padding

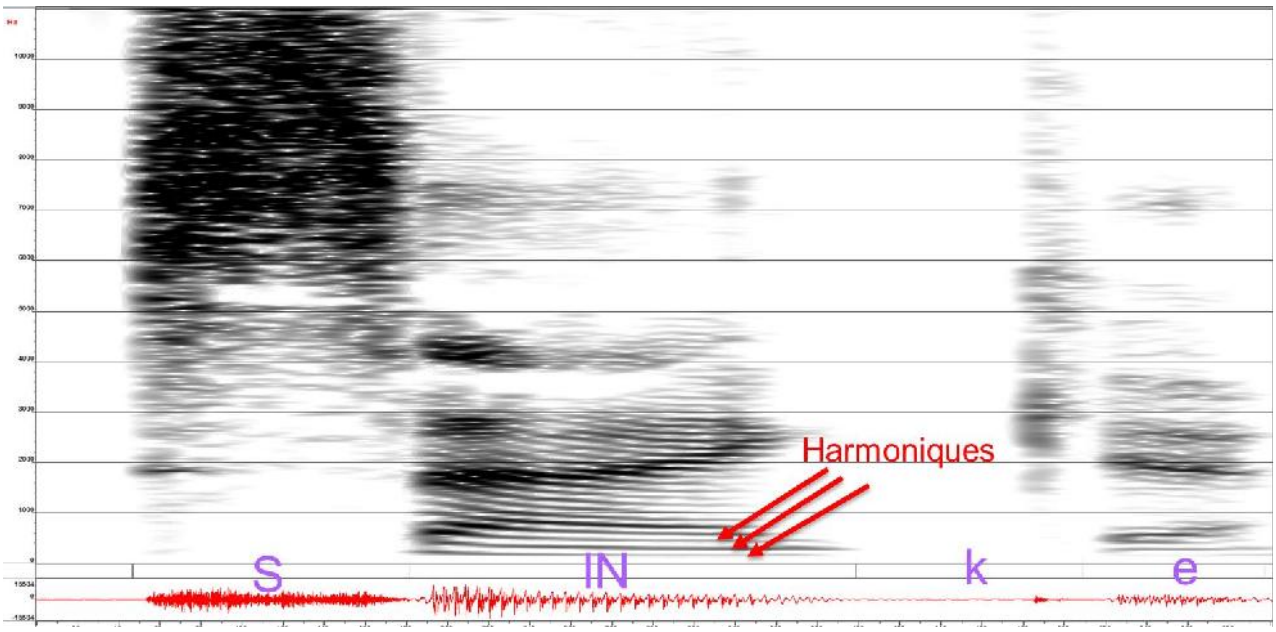



Spectre d'amplitude avec zero padding

On remarque d'après les figures des spectres avec et sans zero padding, l'apparition d'un léger bruit sur le spectre avec zero padding.

### III.6.7 Spectrogramme à bande étroite

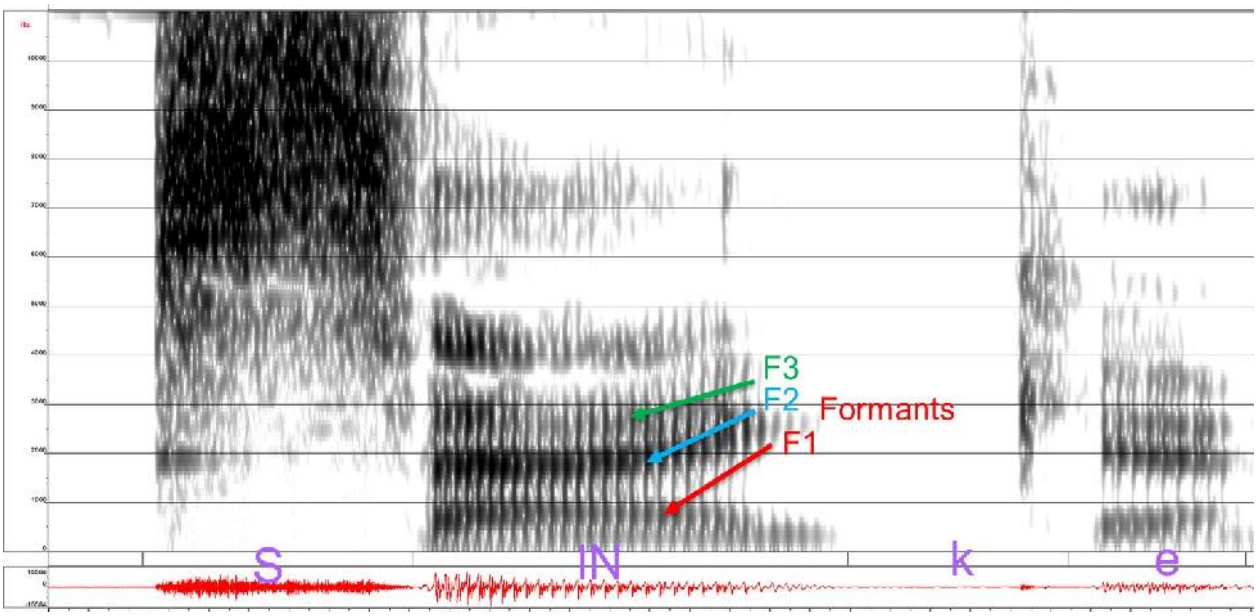
Ce type de spectrogramme permet de visualiser les harmoniques (traits parallèles équidistants représentant des piques d'énergies dans des fréquences) qui indiquent un signal voisé → vibration des cordes vocales, il est calculé avec une fenêtre de grande taille qui permet une meilleure résolution fréquentielle.




Exemple d'un spectrogramme à bande étroite du mot « cinq » FS : 22Khz prononcé par une femme
Fenêtre : 705 échantillons (32ms), pas : 23 échantillons, zero padding (taille FFT)→ 1024 échantillons

### III.6.8 Spectrogramme à large bande

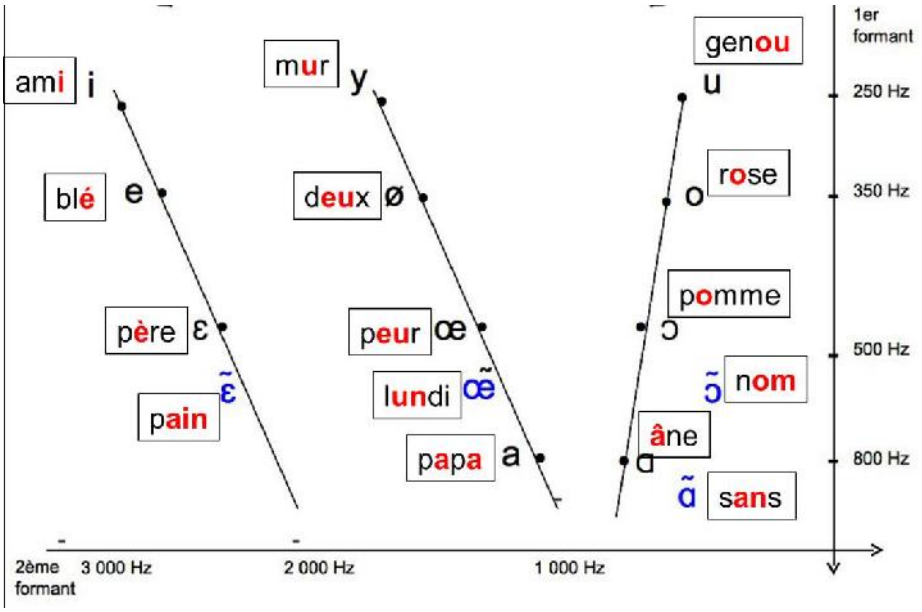
Il est caractérisé par une fenêtre temporelle de petite taille offrant une meilleure résolution temporelle et une résolution fréquentielle faible. Ce type de spectrogramme permet de visualiser les formants (nuages d'énergies localisés dans certaines fréquences), utilisés pour la reconnaissance de la parole (surtout des voyelles).




Exemple d'un spectrogramme à large bande du mot « cinq » FS : 22Khz prononcé par une femme
Fenêtre : 88 (4ms), pas : 7, zero padding (taille FFT)→ 256

### III.6.9 Triangle vocalique des voyelles F1/F2

Le triangle vocalique est une estimation des positions des formants (F1/F2) pour chaque voyelle d'une langue faite sur un échantillon d'individus. Il permet de positionner n'importe qu'elle voyelle sur un plan 2D afin de pouvoir la rapprocher à une des voyelles estimées.



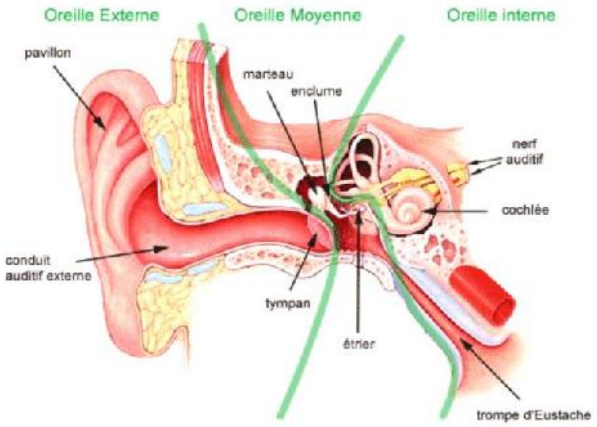

## III.7 Conclusion

La phonétique acoustique est un outil indispensable à tout module de TAP, elle offre des outils très pratiques afin de visualiser et étudier les caractéristiques temporelles et fréquentielles de l'onde sonore qui résulte des phonèmes prononcés par le locuteur.

# IV La phonétique auditive

## IV.1 Introduction

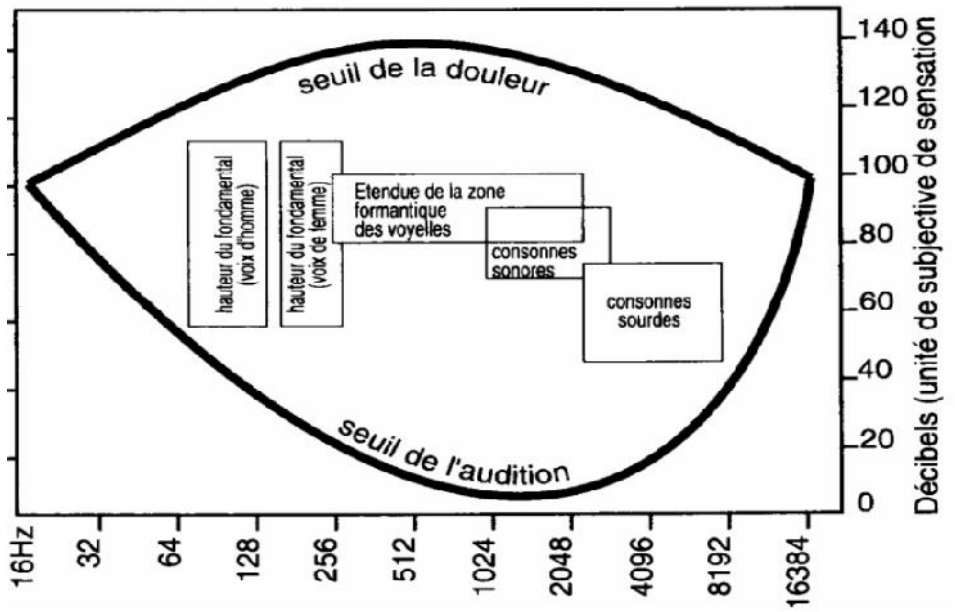
La phonétique auditive est une branche de la phonétique qui s'intéresse à la façon dont l'oreille humaine perçoit le son en étudiant l'appareil auditif humain.




## IV.2 La perception de l'oreille humaine

La perception de l'oreille humaine est caractérisée par certaines caractéristiques :

- La perception de l'oreille humaine est limitée en fréquences et en intensités.
- Le bruit peut réduire la perception de l'oreille en produisant des effets masque.
- L'âge peut influer sur la perception des sons, ex: les personnes âgées entendent souvent mal les fréquences aigües (hautes) des fricatives /s/, /f/.
- L'oreille perçoit l'onde sonore avec une intensité minimale qui diffère selon sa fréquence.
- Un son avec une intensité trop forte peut endommager l'oreille (seuil de douleur).



### IV.2.1 La perception de l'intensité

L'intensité d'un son correspond à sa force (volume). L'unité de mesure de l'intensité la plus utilisée est le décibel (db), une unité logarithmique proposée par Graham Bell. L'intensité d'une voix se situe entre :

- 10 db pour un son murmuré.
- 60-80 db voix d'un chanteur qui chante fort.
- Les voyelles ont besoin de moins d'intensité pour être entendues.

### IV.2.2 La perception de la durée

Un phonème n'est perçu par l'oreille qu'à condition de dépasser une certaine durée (longueur physique) :

- 30ms est nécessaire pour identifier un phone.
- 50ms est nécessaire pour identifier une variation de fréquences.
- 80 à 150ms est suffisante pour identifier une syllabe.

### IV.2.3 La perception des fréquences

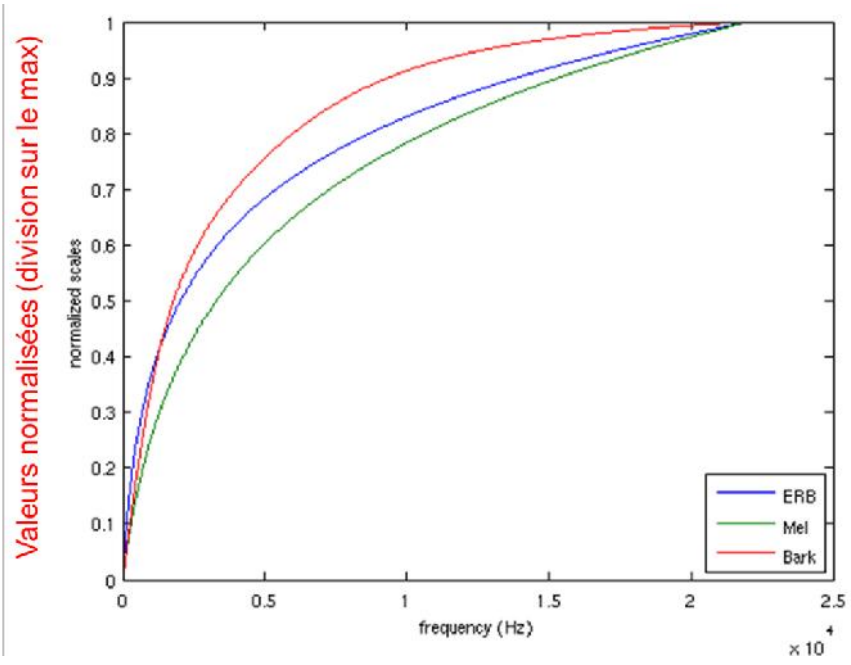
L'oreille humaine est plus sensible aux basses fréquences qu'aux hautes fréquences. Des estimations sur cette perception sont fournies sous forme d'échelles logarithmes, ex : échelle de Bark, des Mels, ERB. Les formules qui permettent le changement d'échelle d'une fréquence $f\ (Hz)$ sont les suivantes :

$$Bark = 13\,\mathrm{atan}\big(0.00076\,f\big) + 3.5\,\mathrm{atan}\Big(\Big(\frac{f}{7000}\Big)^2\Big)$$

$$Mel = 2595\,\log10\Big(1 + \frac{f}{700}\Big)$$

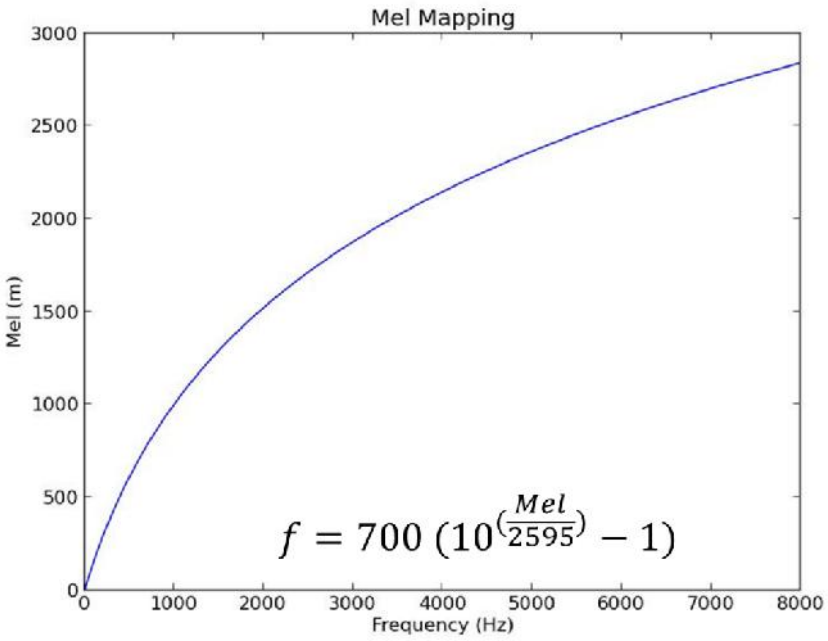
$$ERB = 214\,log10(0.00437\,f + 1)$$

Courbes psycho-acoustiques normalisées des différentes échelles auditives :



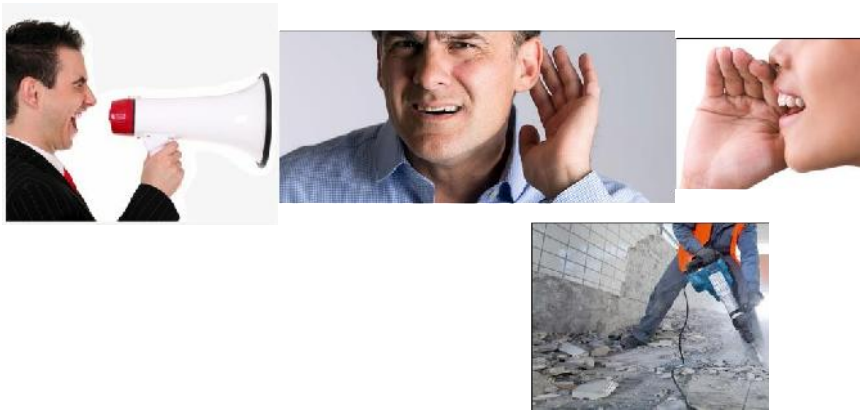
Courbes psycho-acoustiques de l'échelle des Mels :

$$Mel = 2595 \log 10(1 + \frac{f}{700})$$

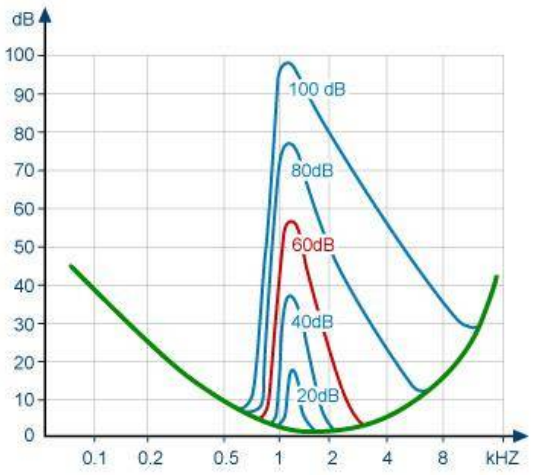



## IV.3 Principe de masquage des sons

Un son peut être masqué par un autre son de plus forte intensité, le premier son n'est pas perçu par l'oreille.

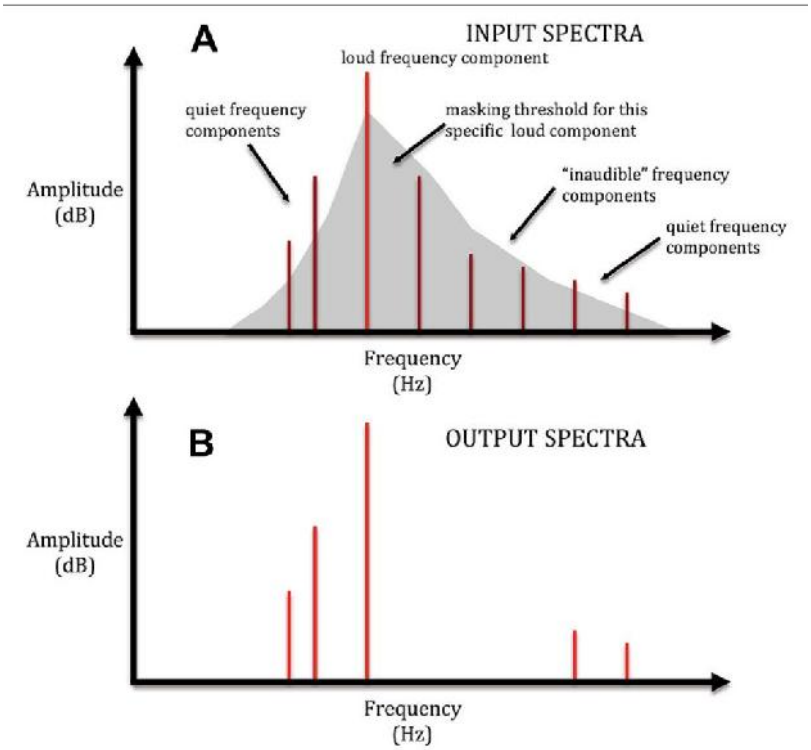


Exemple de masquage obtenu avec un bruit d'une intensité variable (20, 40, 60, 80, 100db) de fréquence [1100-1300] Hz sur les fréquences voisines. En vert le seuil d'audition sans bruit.



La figure précédente montre en rouge la modification des seuils auditifs obtenus pour un niveau de bruit de 60 dB. Par exemple, un son pur de 1000 Hz ne sera perçu qu'à une intensité de 45dB au lieu de 3dB en l'absence de bruit.

La figure ci-dessous montre dans un spectre, les fréquences audibles par l'oreille (partie B) et les fréquences masquées par une fréquence trop intense (représenté en rouge dans la partie A).




## IV.4 Conclusion

L'oreille humaine ne perçoit pas tous les sons, elle est sujette à des contraintes d'intensité, de fréquence et de masquage, l'utilisation des connaissances en phonétique auditive en TAP nous permet de :

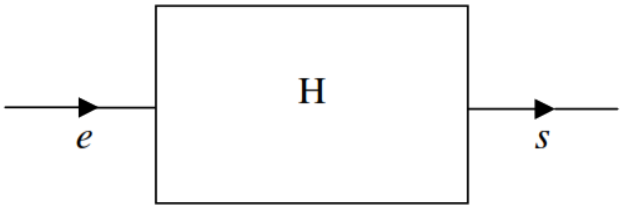
- **Principe 1 :** réduire la taille des données en éliminant du traitement tout son inaudible par l'oreille humaine.
- **Principe 2 :** traiter la parole sur une échelle fréquentielle proche de l'oreille humaine (logarithmique).

# V Outils mathématiques pour le traitement du signal de la parole

## V.1 Introduction

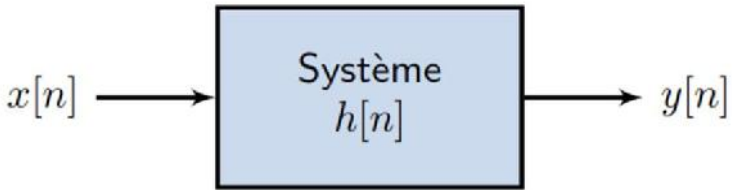
De nombreux signaux physiques ($s = H(e)$) peuvent être décrits sous forme d'une transformée($H$) appliquée sur un signal d'entrée ($e$).

En théorie, le signal de sortie $s$ est souvent obtenu par un produit de convolution entre le signal d'entrée e et le filtre $H$ : $s = e * H$



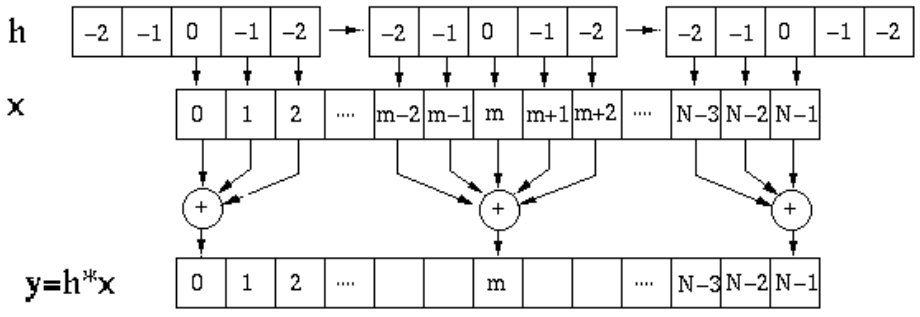

## V.2 Le produit de convolution des signaux discrets

Le produit de convolution entre deux signaux discrets ($x[n] * h[n]$) permet de représenter un filtrage linéaire d'un signal en entrée $x[n]$ par le signal filtre $h[n]$ pour obtenir un signal de sortie $y[n]$.



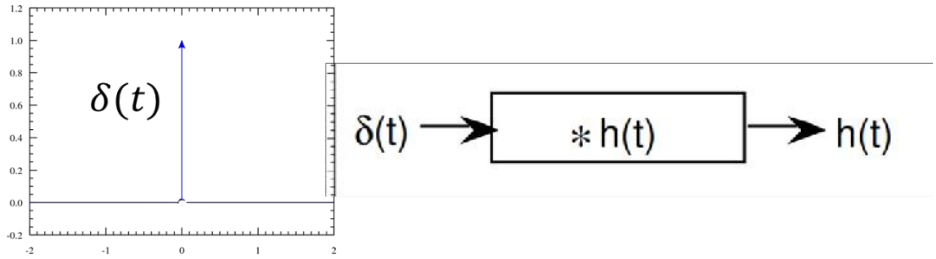

L'opération consiste à calculer une moyenne glissante de $x[n]$ pondérée par $h[n]$ :

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n] * h[n]$$



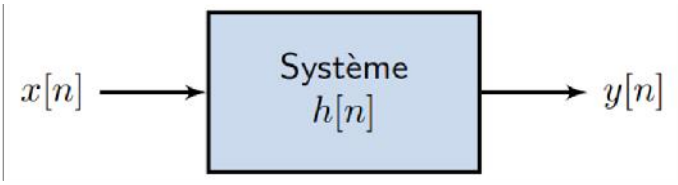

### V.2.1 La réponse impulsionnelle

La réponse impulsionnelle d'un système est le signal de sortie obtenu pour une impulsion de Dirac en entrée.




### V.2.2 La fonction de transfert

La fonction de transfert est la transformée de Fourier de la réponse impulsionnelle.




Transformée de Fourier Discrète

DFT

$$h[n] \quad \Leftrightarrow \quad H[\omega]$$

Domaine temporel (réponse impulsionnelle) ⇔ Domaine fréquentiel (Fonction de transfert)

**Propriété intéressante :** Dans le domaine fréquentiel le produit de convolution devient une multiplication simple entre la fonction de transfert du filtre et la transformée de Fourier du signal d'entrée.

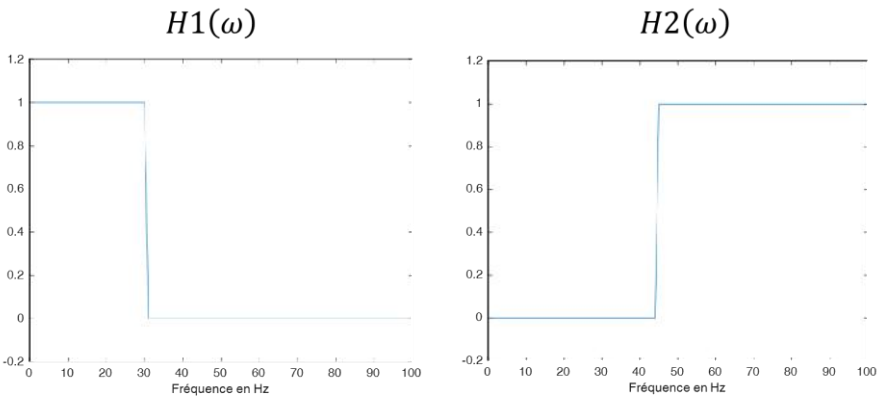
$$y[n] = x[n] * h[n] \quad \Leftrightarrow \quad Y[\omega] = X[\omega] \times H[\omega]$$

Domaine temporel ⇔ Domaine fréquentiel

### V.2.3 Le filtrage des signaux discrets

Le filtrage des signaux permet d'en éliminer/atténuer certaines composantes et d'en rehausser d'autres pour ne garder que la partie utile au traitement.

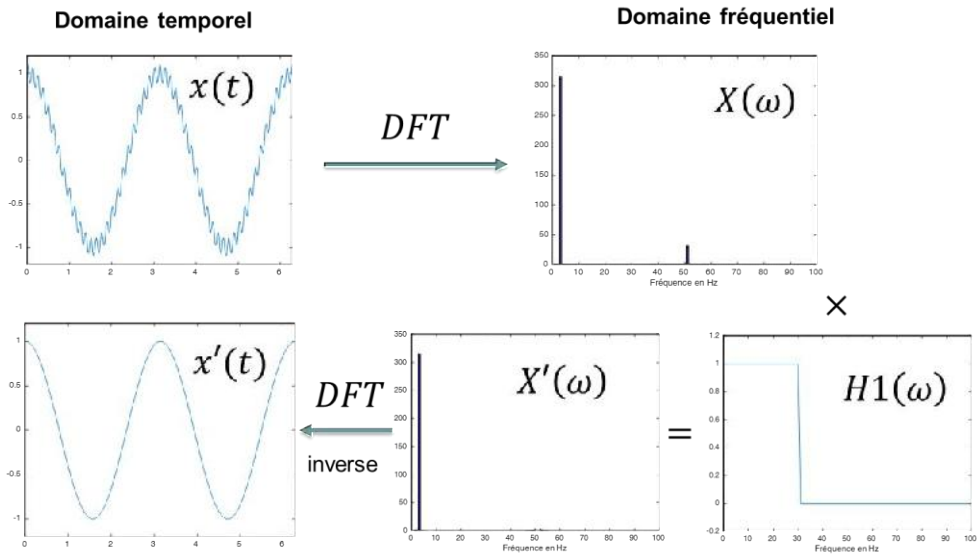
- **Filtre passe bas :** filtre qui atténue les hautes fréquences par rapport au basses, il est utilisé surtout pour réduire les parasites (bruit en hautes fréquences).
- **Filtre passe haut :** filtre qui atténue les basses fréquences par rapport au hautes, il est utilisé surtout pour isoler les détails et les variations rapides du signal.
- **Filtre passe bandes :** filtre qui atténue certaines bandes fréquentielles qui peuvent être localisées en hautes et en basses fréquences.




Exemple d'un filtre passe bas qui annule les fréquences > 30Hz

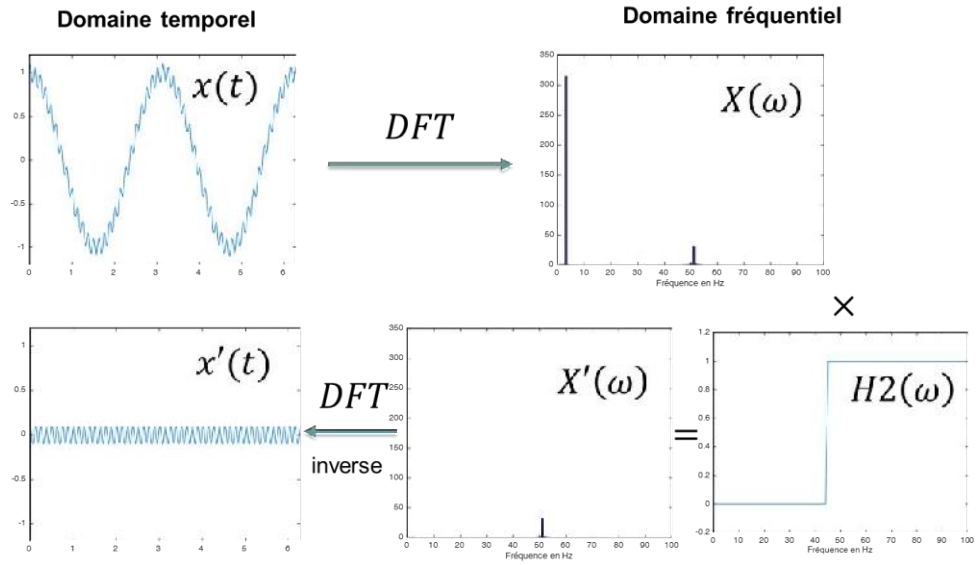
Exemple d'un filtre passe haut qui annule les fréquences < 45Hz

Exemple d'un filtrage passe bas :




$$x'(t) = x(t) * h1(t) = DFT^{-1}(X(\omega) \times H1(\omega))$$

Exemple d'un filtrage passe haut :




$$x'(t) = x(t) * h1(t) = DFT^{-1}(X(\omega) \times H1(\omega))$$

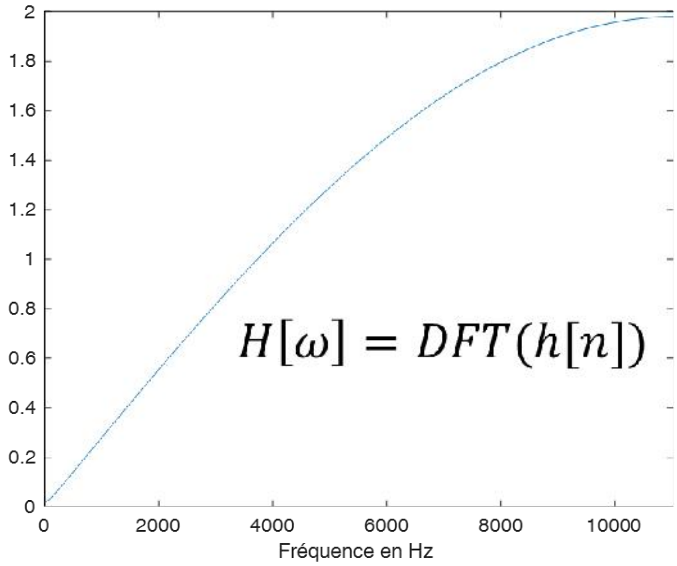
**Question :** le filtre de préaccentuation est-il un filtre passe bas ou un filtre passe haut ?

$$y[n] = x[n] - 0.98 \times x[n-1]$$

**La réponse impulsionnelle du filtre :** $x = \delta$

| $h[n]$: | 1 | -0.98 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |
|---|---|---|---|---|---|---|---|---|---|---|

Estimation de la fonction de transfert $H[\omega]$ pour un signal x[n] échantillonné à 22KHz



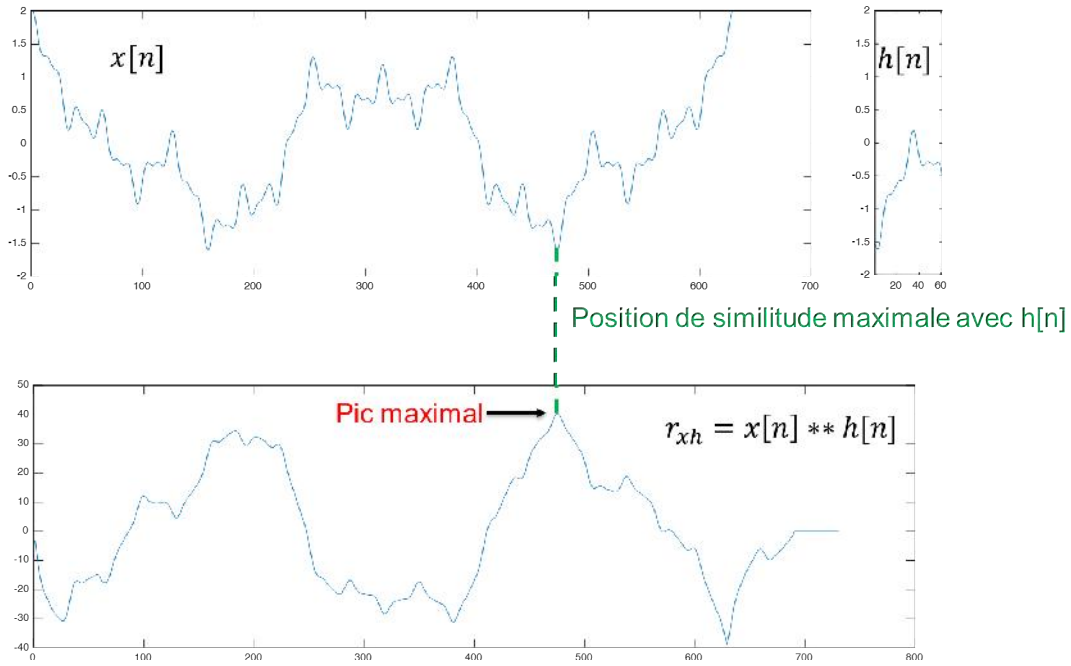

La fonction de transfert du filtre montre qu'il rehausse les hautes fréquences et atténue les basses fréquences, donc c'est un filtre passe haut.

## V.3 La corrélation

La corrélation permet de comparer la similitude entre 2 signaux, ou de chercher un signal à l'intérieur d'un autre signal plus grand.

$$r_{xh}[n] = x[n] ** h[n] = \sum_{k=-\infty}^{\infty} x[n]h\,[k-n] = \sum_{k=-\infty}^{\infty} x[n+k]h\,[k]$$

- La corrélation peut être obtenue par convolution entre $x[n]$ et $h[-n]$.
- On peut donc calculer la corrélation par une multiplication dans le domaine fréquentiel.




### V.3.1 L'autocorrélation

L'autocorrélation est un calcul de corrélation entre un signal avec lui-même, on l'utilise souvent pour déterminer si un signal est périodique, ainsi que la période.

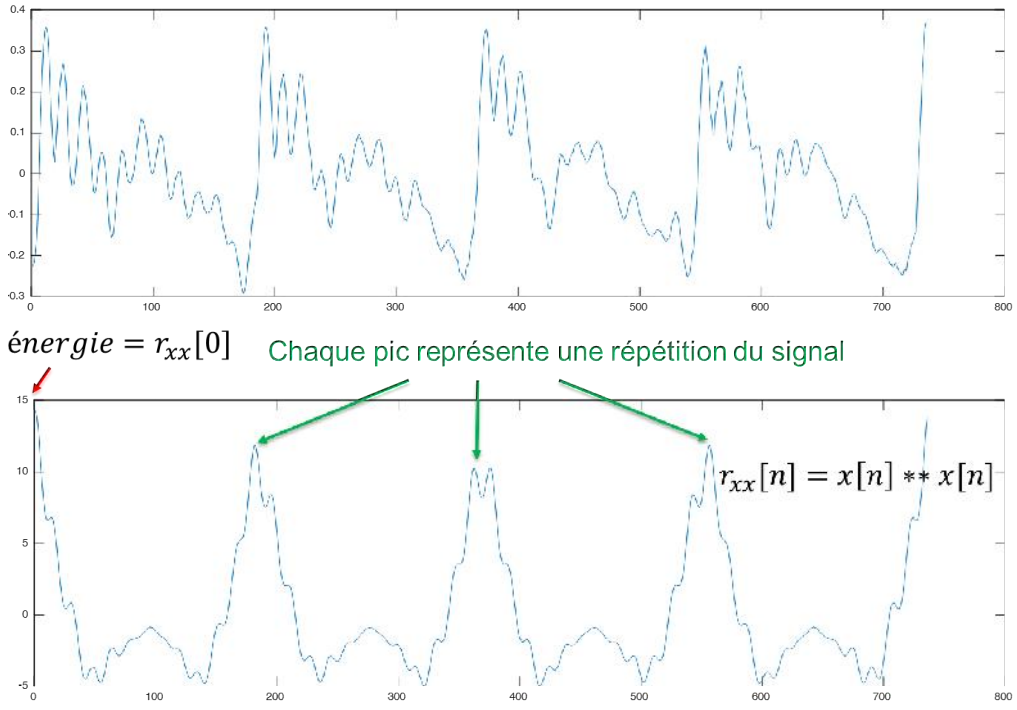
$$r_{xx}[n] = x[n] ** x[n]$$

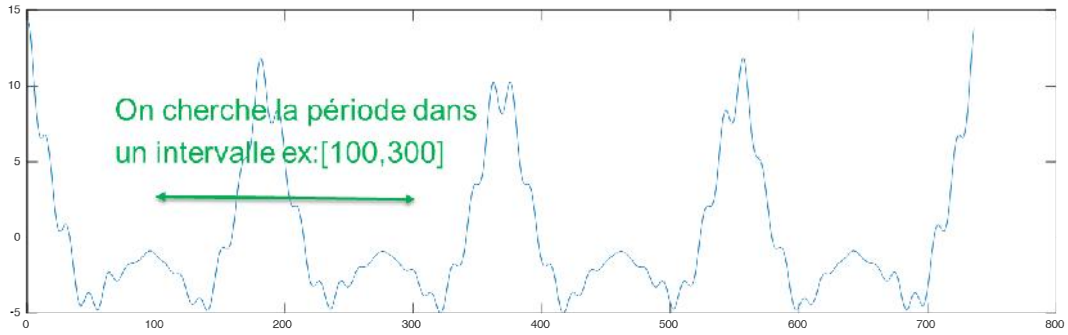
Dans le domaine fréquentiel : ($X^*$ est le conjugué complexe)

$$R_{xx}[\omega] = X[\omega] \times X^*[\omega] = |X[\omega]|^2$$

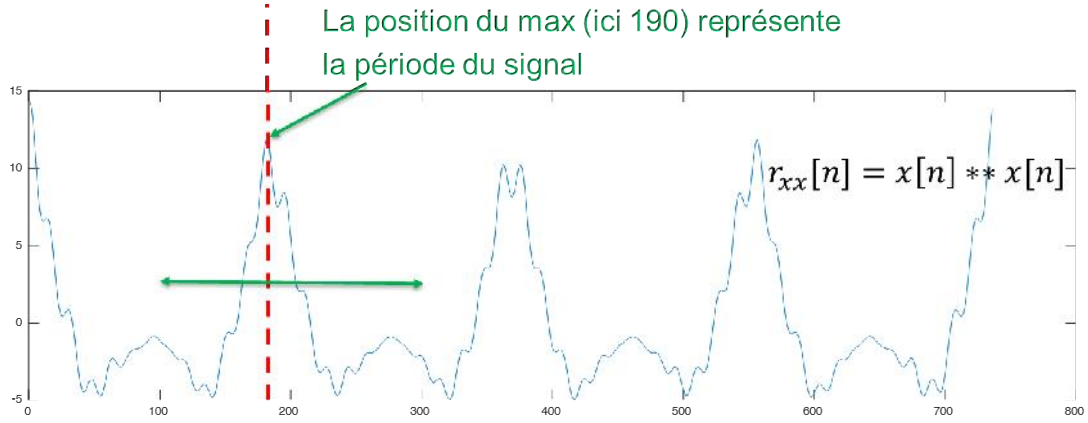
$$\Leftrightarrow\ r_{xx}[n] = DFT^{-1}(|X[\omega]|^2)$$

Exemple de l'autocorrélation d'un signal pseudopériodique $x[n]$ de *N* échantillons :










Dans un signal de parole la périodicité indique une vibration des cordes vocales, la période du signal permet de déduire la fréquence fondamentale $F0$ du locuteur :

$$période\ \ T = 190\ échantillons\ et\ à\ supposé\ que\ x[n]\ est\ échantillonné\ à\ 22050Hz$$

$$22050\ échantillons\ \rightarrow 1\ seconde$$

$$190\ échantillons\ \rightarrow T = ?$$

$$T = \frac{190}{22050} = 0.0086\ secondes \quad \Leftrightarrow \quad F0 = \frac{1}{T} \approx \mathbf{116Hz}$$

## V.4 L'énergie du signal

Le calcul de l'énergie du signal de parole permet de déterminer les zones silence (à faible volume ou faiblement audible) des zones audible (à fort volume).

- Dans le cas d'un système de commandes vocales avec une prononciation de mots isolés par un silence, le silence permet de déterminer le début et la fin de chaque mot.
- On utilise un seuil pour déterminer les positions du silence, ce seuil peut être fixe ou variable selon le bruit de fond.

$x[n]: un\ signal\ discret\ de\ N\ échantillons$

$$E_x = \sum_{k=0}^{N-1} x[k]^2 \qquad (energie\ du\ signal: formule\ 1)$$

$X[\omega]: transformée\ de\ Fourier\ de\ x[n],\ on\ a\ aussi:$

$$E_x = \frac{1}{N} \sum_{\omega=0}^{N-1} |X[\omega]|^2 \qquad (energie\ du\ signal:\ formule\ 2)$$

$sachant\ que\ dans\ le\ cas\ d'un\ signal\ réel:$

$$r_{xh}[n] = \sum_{k=-\infty}^{\infty} x[n+k]h\,[k] \quad \rightarrow \quad r_{xx}[0] = \sum_{k=0}^{N-1} x[k]x[k] = E_x \quad (selon\ formule\ 1)$$

## V.5 Conclusion

Le domaine du traitement du signal nous offre des outils mathématiques très pratiques afin d'analyser et de manipuler des signaux de parole, que ça soit à travers le filtrage par produit de convolution, la corrélation pour la comparaison des signaux, l'autocorrélation pour la détection du voisement et l'estimation de la fréquence fondamentale ainsi que le calcul de l'énergie pour la détection des parties silence/parole du signal.

# VI Filtrage du bruit d'un signal de parole

## VI.1 Introduction

On définit par bruit, toute donnée inutile qui vient perturber ou dégrader l'information utile et son interprétation. En TAP, on trouve plusieurs sortes de bruits, par exemple :

- Un bruit ambiant : bruit d'environnement, bruit de fond …
- Un bruit issu du matériel d'acquisition : microphone, carte son …
- Une ou plusieurs autres sources sonores : d'autres personnes qui parlent, instruments musicaux, véhicules …
- Une transformation du signal de parole : compression, robotisation …

## VI.2 Les différents types de bruits en TAP

On peut classer les bruits selon différents critères :

- Bruit stationnaire/non : un bruit stationnaire est un bruit qui garde les mêmes caractéristiques à travers le temps et reste inchangé tout au long du signal, par contre, un bruit non stationnaire est bruit qui change dans le temps. Le premier type de bruit est plus facile à filtrer car il est plus facile à modéliser.

- Bruit continu/non : un bruit continu est un bruit qui touche tout le signal de parole, par contre un bruit non continu et un bruit qui touche seulement certaines parties du signal. Le premier type de bruit est plus facile à filtrer car on n'a pas besoin de le détecter avant de faire le filtrage.

- Bruit additif/convolutif … : selon l'opération qui entraine le mélange entre le signal du bruit et le signal sans bruit. L'opération peut être une addition des 2 signaux (bruit additif) : c'est le cas pour le bruit de fond ou le mélange de plusieurs sources de sons. L'opération peut aussi être une convolution entre les 2 signaux (bruit convolutif) : c'est le cas pour la robotisation de la parole et le bruit de compression par exemple. Le premier type de bruit est plus facile à filtrer car une simple soustraction avec le profil du bruit permet de l'atténuer, pour les bruits convolutifs l'opération est plus complexe.

- Localisation fréquentielle : le bruit peut être localisé exclusivement sur certaines fréquences, un filtre passe bandes permet de le filtrer. Le bruit peut toucher plusieurs ou toutes les fréquences et est plus difficile à filtrer.

- Bruit corrélé/aléatoire : le signal de bruit peut présenter certaines corrélations, ce qui facilite sa modélisation pour son filtrage, par contre, un bruit aléatoire est un bruit difficile à modéliser et est donc plus difficile à filtrer.

## VI.3 Le filtrage par soustraction spectrale

Selon la typologie des bruits évoquée, le type de bruit le plus simple à filtrer est le bruit stationnaire additif, on utilise à cette fin le filtre par soustraction spectrale. Cette méthode repose sur l'estimation du profil fréquentiel du bruit (concentration énergétique du bruit dans chaque fréquence) sur certaines zones du signal, puis de faire une soustraction avec son spectre d'amplitude, elle fonctionne selon deux modes :

- Le mode manuel : une personne sélectionne manuellement des zones du signal de parole qui contiennent du bruit, il est préférable de sélectionner des zones avec le bruit seul.
- Le mode automatique : l'algorithme de filtrage détecte les zones de silence du signal en calculant son énergie et en comparant avec un seuil (prédéfini ou calculé), ces zones à faible énergie contiennent le bruit seul, elles sont utilisées afin d'estimer le profil fréquentiel du bruit.

Dans le cas d'un bruit additif, un signal bruité $y(t)$ est l'addition du signal sans bruit $x(t)$ avec le signal du bruit $b(t)$:

$$y(t) = x(t) + b(t)$$

Dans le domaine fréquentiel : $|Y(f)| = |X(f)| + |B(f)|$ où $|B(f)|$ est le profil du bruit.

Sachant que le bruit est stationnaire, le profil est le même (ou quasi pour pseudo-stationnaire) pour chaque trame du signal bruité.

En supposant que $\left|\hat{B}(f)\right|$ est le profil du bruit estimé par une des méthodes manuelle ou automatique :

$$\left|\hat{X}(f)\right| = |Y(f)| - \alpha \times \left|\hat{B}(f)\right|$$

Où $\alpha$ est un nombre réel positif qui représente le degré du filtrage

Le signal filtré est reconstitué à l'aide du spectre d'amplitude filtré $\left|\hat{X}(f)\right|$ et le spectre de phase du signal bruité $\varphi_y$ :

$$\hat{x}(t) = FFT^{-1}(|\hat{X}(f)| \times e^{i\varphi_y})$$

**Exemple :**

On dispose d'un signal bruité avec un bruit pseudo-stationnaire additif dont le spectre d'amplitude est le suivant :

$$|Y(f)| = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 4 & 4 & 0 & 0 & 2 & 2 & 2 & 0 \\ 3 & 5 & 4 & 1 & 1 & 6 & 6 & 5 & 3 \end{bmatrix}$$

On suppose que le seuil de silence = 4, on veut estimer le profil du bruit et filtrer ce spectre d'amplitude.

- Etape 1 : estimation de l'énergie du signal $E = \frac{1}{N}\sum_f |X(f)|^2$

$$E = \frac{1}{3}\begin{bmatrix} 9 & 41 & 33 & 1 & 1 & 44 & 44 & 33 & 9 \end{bmatrix}$$

- Etape 2 : comparaison avec le seuil ($E$<=4) pour localiser les zones silence

$$s = [vrai \quad faux \quad faux \quad vrai \quad vrai \quad faux \quad faux \quad faux \quad vrai]$$

- Etape 3 : découper la partie silence du spectre d'amplitude (cette partie contient le bruit seul)

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 3 & 1 & 1 & 3 \end{bmatrix}$$

- Etape 4 : estimer le profil du bruit en calculant la moyenne de chaque fréquence (ligne) de la matrice découpée

$$\left|\hat{B}(f)\right| = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

- Etape 5 : filtrer le spectre d'amplitude bruité en utilisant le profil du bruit estimé, dans cet exemple on suppose que $\alpha = 1$

$$\left|\hat{X}(f)\right| = |Y(f)| - \alpha \times \left|\hat{B}(f)\right|$$

$$\left|\hat{X}(f)\right| = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 4 & 4 & 0 & 0 & 2 & 2 & 2 & 0 \\ 3 & 5 & 4 & 1 & 1 & 6 & 6 & 5 & 3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 4 & 4 & 0 & 0 & 2 & 2 & 2 & 0 \\ 1 & 3 & 2 & -1 & -1 & 4 & 4 & 3 & 1 \end{bmatrix}$$

- Etape 6 : remplacer les valeurs négatives par des zéros pour obtenir le spectre d'amplitude filtré

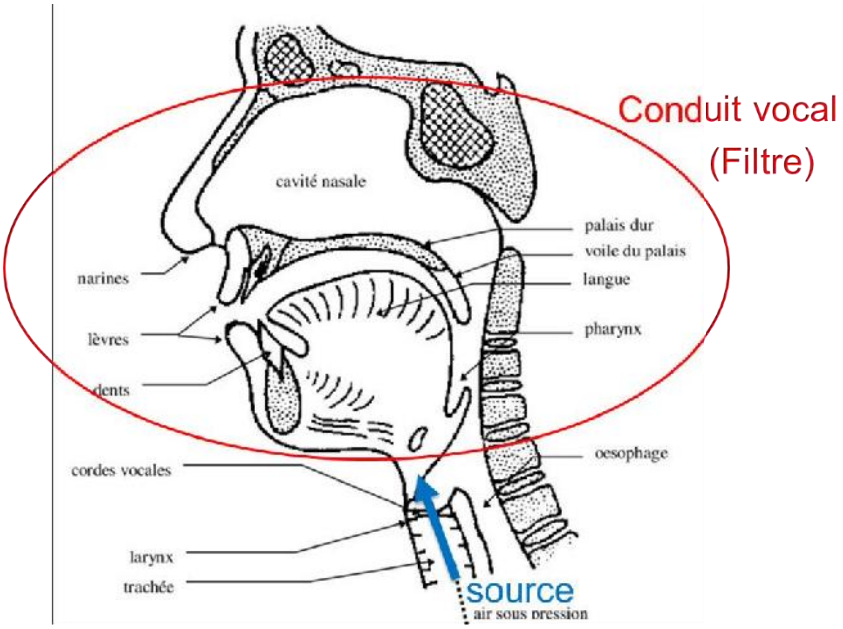
$$\left|\hat{X}(f)\right| = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 4 & 4 & 0 & 0 & 2 & 2 & 2 & 0 \\ 1 & 3 & 2 & 0 & 0 & 4 & 4 & 3 & 1 \end{bmatrix}$$

On peut atténuer le bruit d'avantage en augmentant la valeur de $\alpha$.

# VII Modélisation du signal de parole

## VII.1 Introduction

Les études en phonétique articulatoire montrent que la parole résulte d'un filtrage (de fréquences) d'un signal source (issu des cordes vocales). Le filtrage est fait par le conduit vocal qui regroupe tous les organes du système articulatoire en dessus des cordes vocales (pharynx, langue, lèvres, cavité nasale, dents ...)

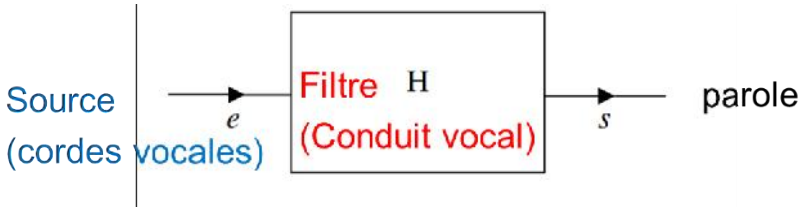



La modélisation du signal de parole nous permet à la fois de :

- Simplifier le processus de production de la parole pour pouvoir le reproduire → synthèse de parole.
- Décrire la parole par un ensemble restreint de paramètres (les paramètres du filtre et du source) → compression de la parole.
- Estimer les paramètres de la parole à partir du signal permet d'inverser le processus → reconnaissance de la parole.

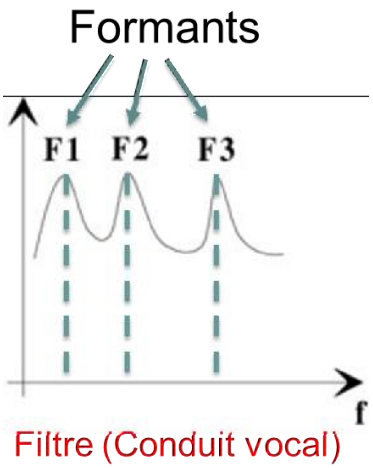
## VII.2 Séparation source/filtre

En TAP, le signal source est représentatif des caractéristiques vocales (voix) du locuteur et résulte du passage de l'air à travers les cordes vocales. Par contre, le signal filtre est une transformation appliquée sur le signal source afin de générer de la parole et est le résultat des mouvements des résonateurs du conduit vocale. Usuellement, on suppose en TAP que le filtrage fait par le conduit vocal sur le signal source est un produit de convolution.

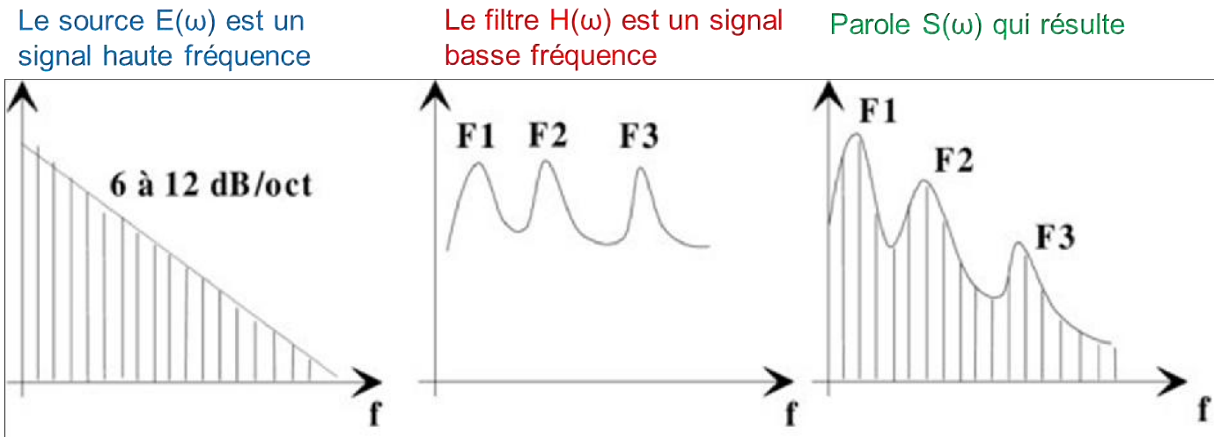


Domaine temporel : $s(t) = e(t) * h(t)$

Le signal filtre $h(t)$ est généralement modélisé en estimant les fréquences de ses piques (F1/F2/F3 …) dans le domaine fréquentiel $|H(\omega)|$ qu'on désigne par le terme « formants ».




Domaine fréquentiel (spectre d'amplitude) : $|S(\omega)| = |E(\omega)| \times |H(\omega)|$




### VII.2.1 Le lissage Cepstral (Cepstre)

Afin de faciliter l'estimation des formants (paramètres du signal filtre) une séparation du signal source à partir du signal de parole est nécessaire suivant les remarques suivantes :

- Un filtre passe bas sur le spectre d'amplitude $|S(\omega)|$pourrait faire la séparation.
- Note1 : le filtre passe bas permet de séparer les signaux à basses fréquences des signaux à hautes fréquences en cas d'addition (et non une multiplication).
- Note2 : le filtrage nécessite un deuxième passage en domaine fréquentiel du spectre d'amplitude (noté domaine quéfrentiel) puis un retour au domaine fréquentiel.

Sachant que :

$$|S(\omega)| = |E(\omega)| \times |H(\omega)|$$

Transformer la multiplication en addition via un logarithme :

$$log|S(\omega)| = log|E(\omega)| + log|H(\omega)|$$

Passage au domaine quéfrentiel pour le filtrage (ici noté liftage) :

$$DFT^{-1}(log|S(\omega)|) = DFT^{-1}(log|E(\omega)| + log|H(\omega)|) = DFT^{-1}(log|E(\omega)|) + DFT^{-1}(log|H(\omega)|)$$

Attention (sachant que): $S(\omega) = DFT\big(s(t)\big)$

$$DFT^{-1}(S(\omega)) = s(t) \text{ mais } DFT^{-1}(|S(\omega)|) \neq s(t)$$

A partir de :

$$DFT^{-1}(log|S(\omega)|) = DFT^{-1}(log|E(\omega)|) + DFT^{-1}(log|H(\omega)|)$$

On simplifie les formules avec un remplacement de variables (domaine quefrentiel) :

$$\hat{s}(t) = \hat{e}(t) + \hat{h}(t)$$

Liftage :

$$l(t) = 1 \quad si\ t < t0 \quad \text{sinon } 0$$

Choisir t0<f0 (fréquence fondamentale)

On obtient :

$$\hat{s}(t) \times l(t) = \hat{e}(t) \times l(t) + \hat{h}(t) \times l(t)$$

$E(\omega)$ est un signal à hautes fréquences, alors : $\hat{e}(t) \times l(t) = 0$

$H(\omega)$ est un signal à basse fréquence, alors : $\hat{h}(t) \times l(t) = \hat{h}(t)$

On conclue que :

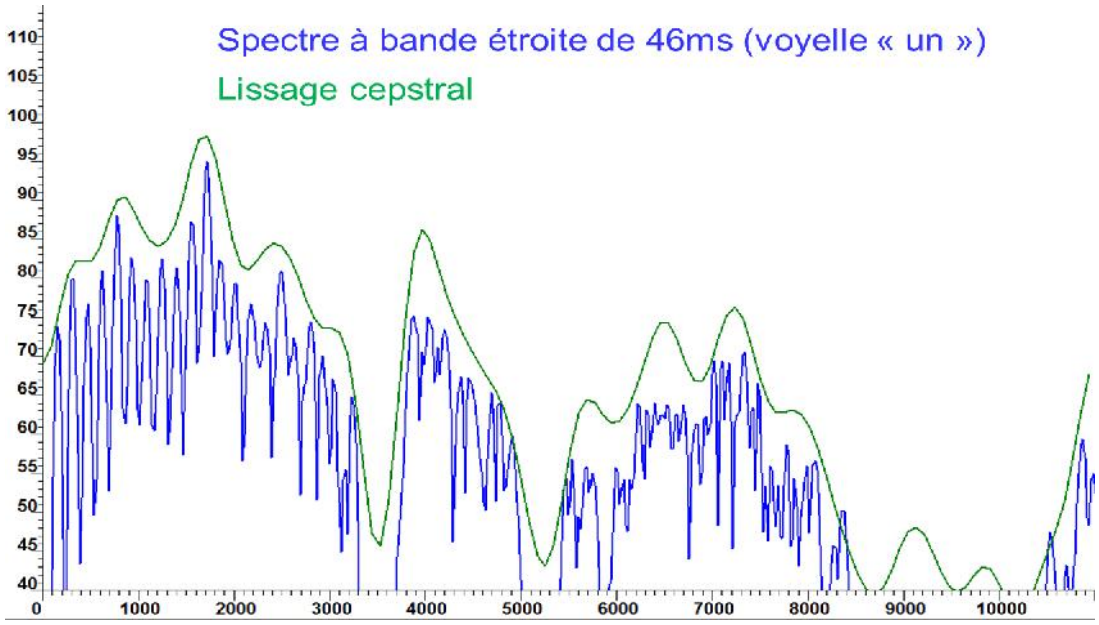
$$\hat{s}(t) \times l(t) = \hat{h}(t)$$

Alors :

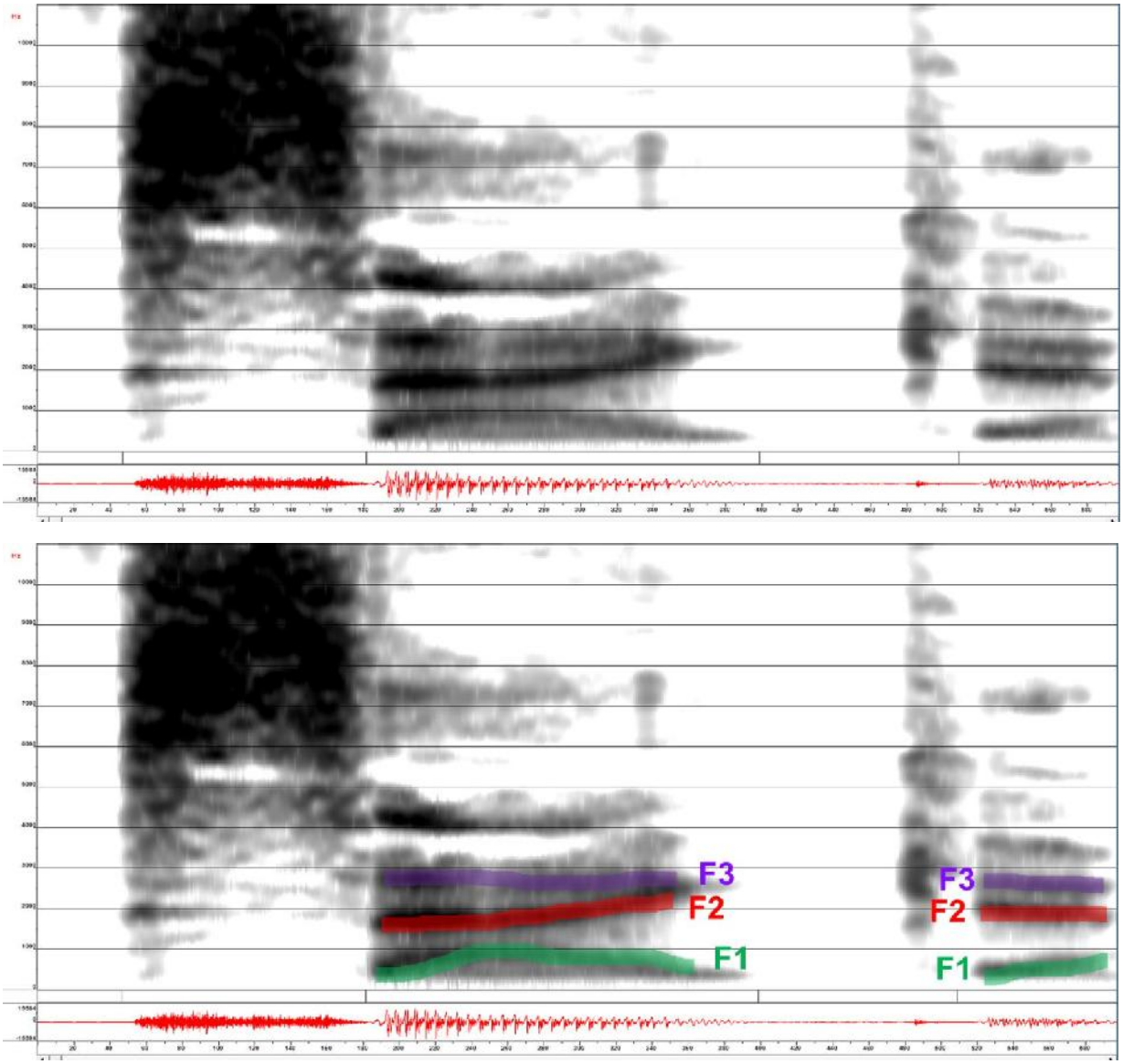
$$DFT(\hat{s}(t) \times l(t)) = \log(|H(\omega)|) \rightarrow |H(\omega)| = e^{DFT\,(\hat{s}(t) \times l(t))}$$

Le Cepstre est affiché généralement avec le log :

$$\log(|H(\omega)|) = DFT(DFT^{-1}(log|S(\omega)|) \times l(t))$$



La figure suivante montre le Cepstre du mot « cinq » prononcé par un locuteur féminin, la seconde figure montre les formants F1/F2/F3 qui modélisent le conduit vocale lors de la prononciation des voyelles « ɛ̃ » et « ə » du mot cinq transcrit phonétiquement «sɛ̃kə ».

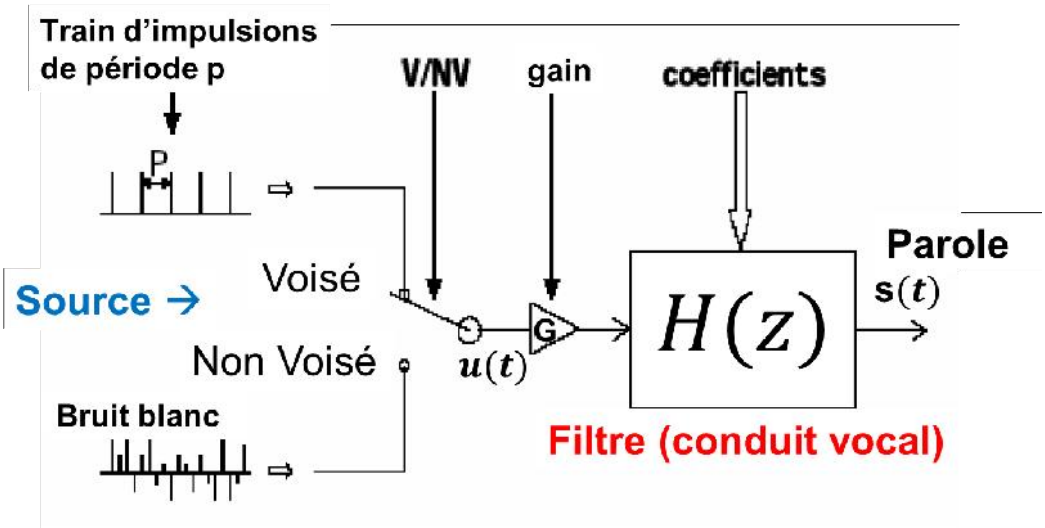



### VII.2.2 Le codage linéaire prédictif (LPC)

Le codage prédictif linéaire et ses variantes est une méthode largement utilisée en codage et compression de la parole à très bas débit dans les télécommunications. Elle est très utilisée en synthèse et reconnaissance de la parole. Elle repose sur l'hypothèse que la parole peut être prédite à partir des échantillons passés via un filtre linéaire.

Certaines remarques se posent :

- La parole n'étant pas linéairement prédictible, une erreur de prédiction (résidu) est à minimiser lors de l'estimation des paramètres du filtre.
- Plus le taux de compression est élevé, plus l'erreur est importante, plus on aperçoit une perte de qualité du son.
- Le codage LPC fait la distinction entre les fenêtres voisées (périodiques) et non voisées.




Le signal de parole est le résultat d'un filtre appliqué sur les échantillons précédents et d'un signal excitateur *u(t)*.

$$s(t) = \sum_{k=1}^{p} a_k \times s(t-k) + G \times u(t)$$

- $a_k$: les coefficients (paramètres) du filtre (à estimer).
- *p*: ordre du filtre, on utilise de coefficients pour chaque formant à estimer, valeur usuelle=10
- *G*: gain à minimiser
- L'estimation des $a_k$, passe généralement par une résolution d'un système d'équations avec la méthode « Levinson-Durbin » sur les coefficients d'autocorrélation du signal (cependant il existe d'autres méthodes d'estimation).

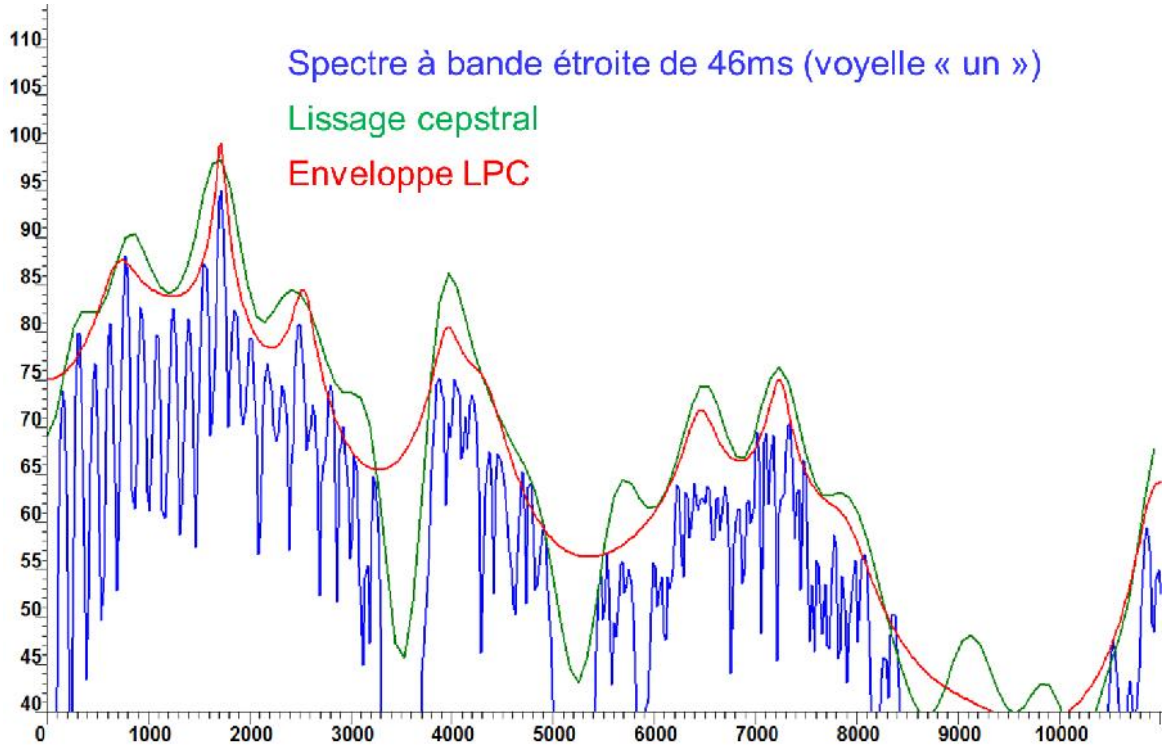
La réponse impulsionnelle du filtre est sous la forme :

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k \, z^{-1}}$$

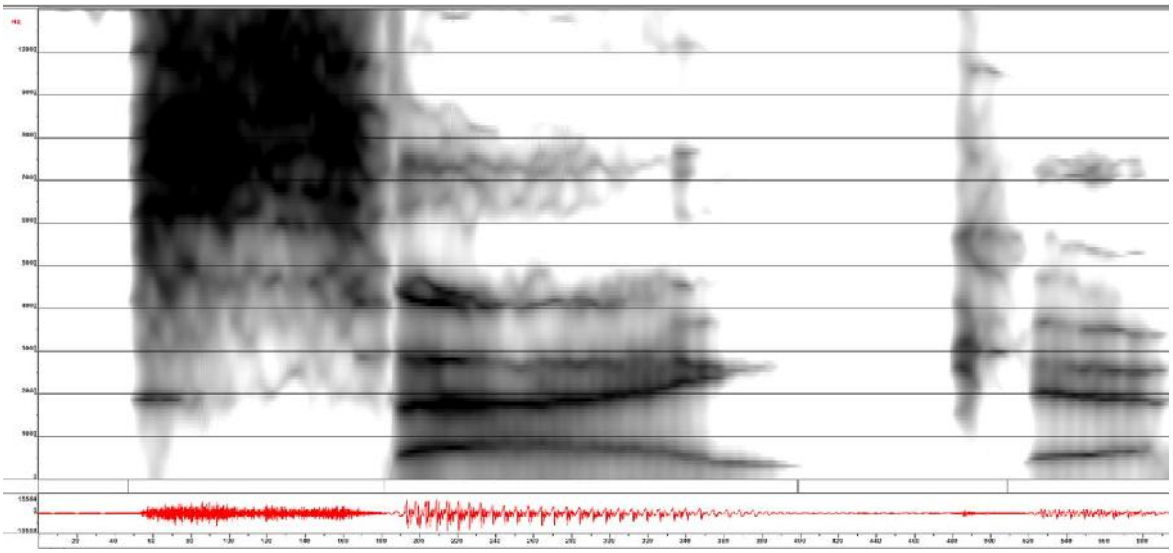
On met : $H(z) = \frac{1}{A(z)} \quad avec \;\; A(z) = 1 + \sum_{k=1}^{p} a_k \, z^{-1}$

Le spectre LPC (enveloppe LPC) peut être calculé facilement avec une DFT sur les coefficients LPC estimés :

$$|H(\omega)| = |DFT\left(\frac{1}{a_k}\right)| \quad avec \;\; \omega: onde\ de\ fourier$$




La figure suivante représente l'enveloppe LPC calculé sur l'ensemble des fenêtres du mot « cinq » prononcé par un locuteur féminin.



On remarque que la visibilité des formants F1/F2/F3 est bien plus nette que celle du Cepstre.

Le codage LPC inclut plusieurs variantes : CELP, PLP, RASTA PLP, RELP et NPC , afin d'améliorer la qualité du son avec un meilleur taux de compression elles utilisent des techniques basées sur:

- La quantification vectorielle
- Les échelles auditives non linéaires (ex : échelle de Bark)
- Les réseaux de neurones

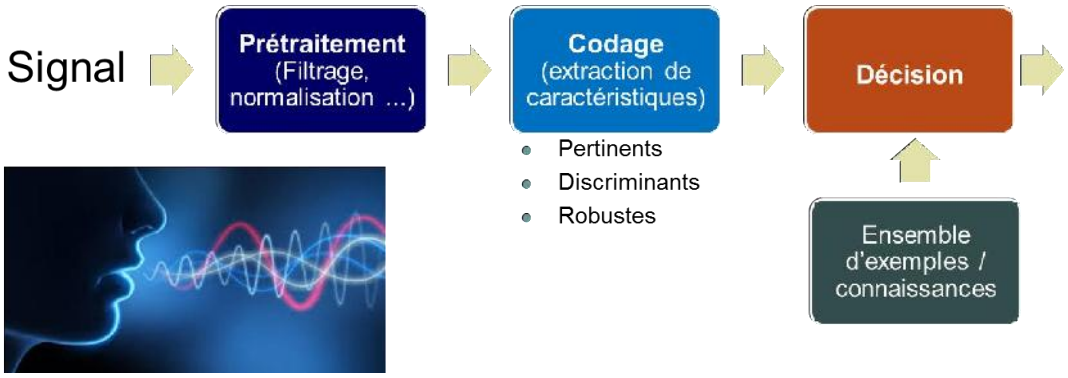
## VII.3 Conclusion

La modélisation de la parole permet de décrire le signal de parole à l'aide d'un nombre réduit de paramètres ce qui se trouve très utile en compression et codage de la parole. L'estimation des paramètres repose sur des modèles théoriques issues des études en phonétique articulatoire, acoustique et auditive. Le lissage Cepstral et le codage LPC sont des exemples de modèles basés sur la théorie source/filtre, ils permettent notamment de visualiser les formants qui modélisent la transformation faite par les résonateurs du conduit vocal sur le source et sont utilisés en reconnaissance des voyelles.

# VIII La reconnaissance de la parole

## VIII.1 Introduction

La reconnaissance de la parole est un sous domaine du TAP qui consiste à donner des labels à des segments du signal de parole permettant d'identifier ce qui a été dit, en bref, transformer la parole en texte ou commandes. Cette faculté bien qu'elle parait très simple pour l'être humain, est une tâche très fastidieuse pour la machine qui nécessite l'enchainement de plusieurs étapes : amélioration et normalisation du signal de parole, extraction des caractéristiques, apprentissage et classification à partir d'une base d'exemple ...




## VIII.2 Le prétraitement de la parole

Cette étape a pour but de mettre les signaux de parole dans un format amélioré et normalisé afin de faciliter les étapes suivantes, elle peut inclure :

- La décompression de la parole
- Le filtrage du bruit
- La segmentation de la parole (délimitation et extraction des mots/phonèmes/phrases)
- Normalisation du volume (amplitude)
- Normalisation temporelle

### VIII.2.1 La segmentation de la parole

La parole est un flux de données (signal) continu (qui peut être infini), la reconnaissance de la parole consiste en la localisation des unités de base (phonèmes/ syllabes/ mots/ phrases …) puis l'étiquetage de ces dernières (reconnaissance).

Ces 2 opérations peuvent soit être faites ensemble, ou dans certaines pratiques séparément. La segmentation de la parole consiste à trouver le début et la fin de chaque unité à partir du signal de parole.

Dans le cas simple de parole isolée, c'est à dire de présence de silence entre les mots/unités il est possible d'utiliser l'énergie du signal ainsi qu'un seuil afin de localiser le silence. Cependant plusieurs problèmes se posent :

- Comment déterminer le seuil ?
- Le seuil est-il fixe ou variable ?
- La présence de silence à l'intérieur des mots

**Estimation de l'énergie du bruit de fond :**

Dans un signal de parole, le bruit rond sa segmentation difficile vu qu'on a plus un silence absolu (d'énergie 0) mais plus un bruit de fond au lieu du silence (d'énergie variable $\beta_t$).

La détermination de l'énergie du bruit de fond initiale peut se faire facilement si le début du signal est constitué de silence en calculant l'énergie moyenne de cette partie (de *t=0* jusqu'à *t0*).

$$\beta_0 = \frac{1}{t0}\sum_{t=0}^{t<t0} E(t)$$

$$E(t): l'énergie\ du\ signal \text{ à } l'instant\ t$$

$$\beta_0: l'énergie\ du\ bruit\ de\ fond\ initiale$$

**Mise à jour de l'énergie du bruit :**

$$\begin{cases} \beta_t = \alpha\beta_{t-1} + (1-\alpha)E(t) & si\ E(t) < seuil \\ \beta_t = \beta_{t-1} & sinon \end{cases}$$

$\alpha$: Typiquement dans [0.95, 1] représente la vitesse de variation de l'énergie du bruit.

$\beta_t: l'énergie\ du\ bruit\ de\ fond \text{ à } l'instant\ t$

$E(t): l'énergie\ du\ signal \text{ à } l'instant\ t$

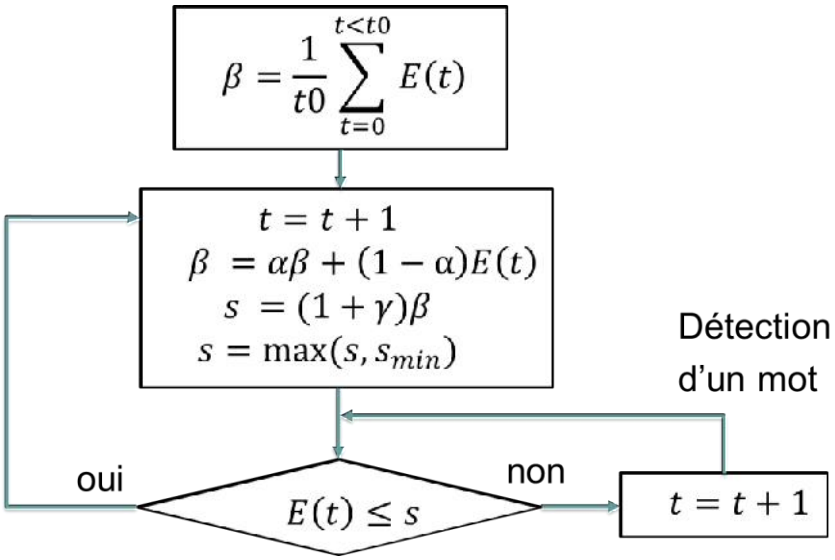
**Détermination du seuil à l'instant t :**

$$s_t = (1+\gamma)\beta_t$$

$$s_t = \max(s_t, s_{min})$$

$\gamma$: Typiquement 1 afin que le seuil soit le double de l'énergie moyenne du bruit de fond.

$s_{min}$ : est le seuil minimal



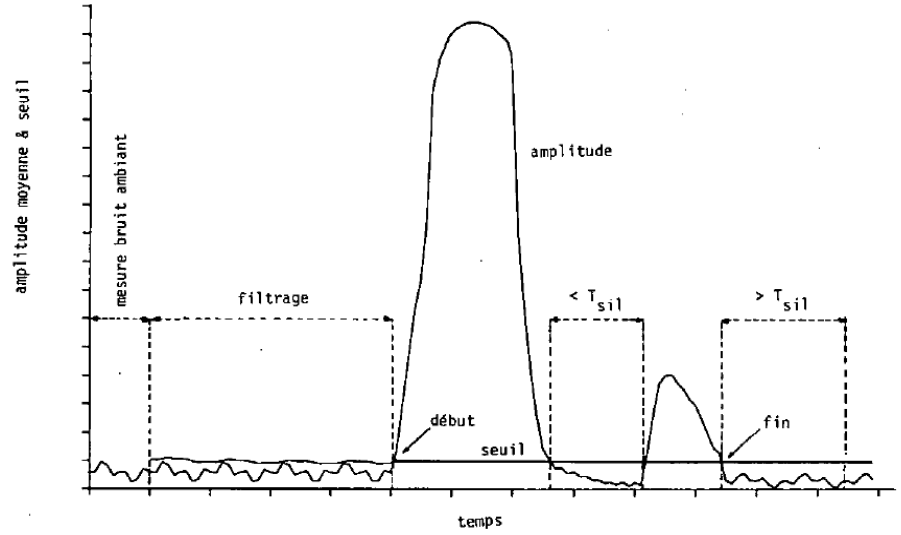
**Détection du début et la fin d'un mot :**

Dans un signal de parole il est possible de trouver du silence à l'intérieur des mots, dû au mode d'élocution des occlusifs. Cependant, ce silence est de très courte durée. Dans un système de reconnaissance de mots isolés, on impose un silence de longue durée entre les mots (1 seconde minimum) et on définit 3 constantes :

$t_{sil}$ : Temps de silence minimal

$t_{min}$ : Temps de mot minimal

$t_{max}$ : Temps de mot maximal




$s = initialiser\ le\ seuil$
**Répéter**
  $t = t + 1$  *#si t=fin du signal alors sortir*
  $s = mise\ à\ jour\ du\ seuil$
  **si** $E(t) > s$
    $debut\_mot = t$
    **Répéter**
      **si** $E(t) > s$ **alors** $compt_{sil} = 0$ **sinon** $compt_{sil} += 1$
      $t = t + 1$  *#si t=fin du signal alors sortir*
    **Jusqu'à ce que** $compt_{sil} \geq t_{sil}$
    $fin\_mot = t - t_{sil}$
    $duree\_mot = fin\_mot - debut\_mot + 1$
    **si** $t_{min} \leq duree\_mot \leq t_{max}$ **alors** $mot\_detecté(debut\_mot, fin\_mot)$
  **fin si**
**jusqu'à la fin du signal**

## VIII.2.2 Normalisation de l'amplitude de la parole

Dans un signal de parole les mots ne sont pas prononcés avec la même amplitude (volume), il est souvent nécessaire de les normaliser avant de passer à l'étape de reconnaissance. Une méthode simple est de diviser le signal du mot sur son énergie moyenne (le signal est fenêtré avec une fenêtre de taille *w*) :

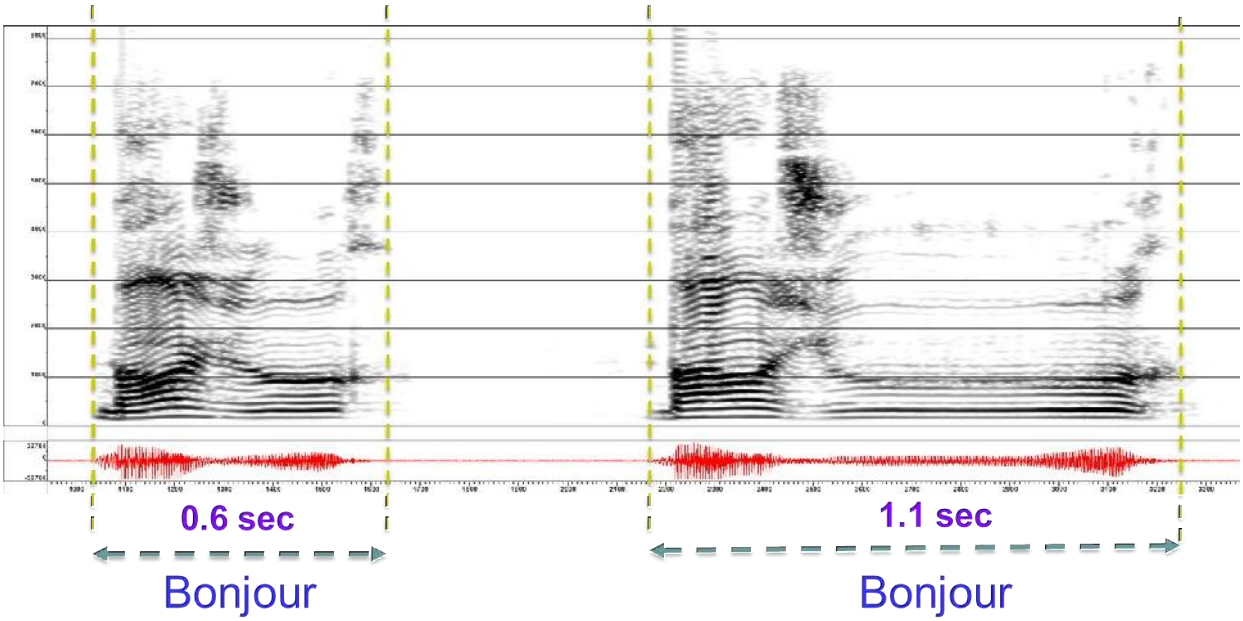
$$signal(debut_{mot}: fin_{mot}) = \frac{signal(debut_{mot}: fin_{mot})}{E_{moyenne}}$$

$$E_{moyenne} = moyenne(energie(debut_{mot}: fin_{mot}))$$

### VIII.2.3 Normalisation temporelle de la parole

Dans un signal de parole les mots ne sont pas prononcés avec la même durée. Beaucoup de méthodes en reconnaissance des formes sont statiques (nécessitent que les signaux à comparer soit de même durée). La normalisation temporelle consiste à convertir tous les mots sur la même durée.

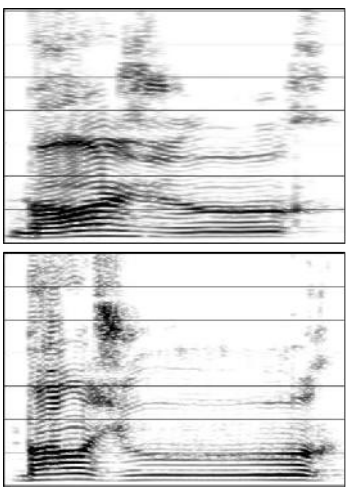
Exemple du mot bonjour prononcé par le même locuteur sur des durées différentes :




#### *VIII.2.3.1 Normalisation temporelle linéaire*

La normalisation linéaire consiste à faire une simple interpolation (étirer ou comprimer linéairement) sur les signaux des mots afin de le réarranger dans une longueur fixe.

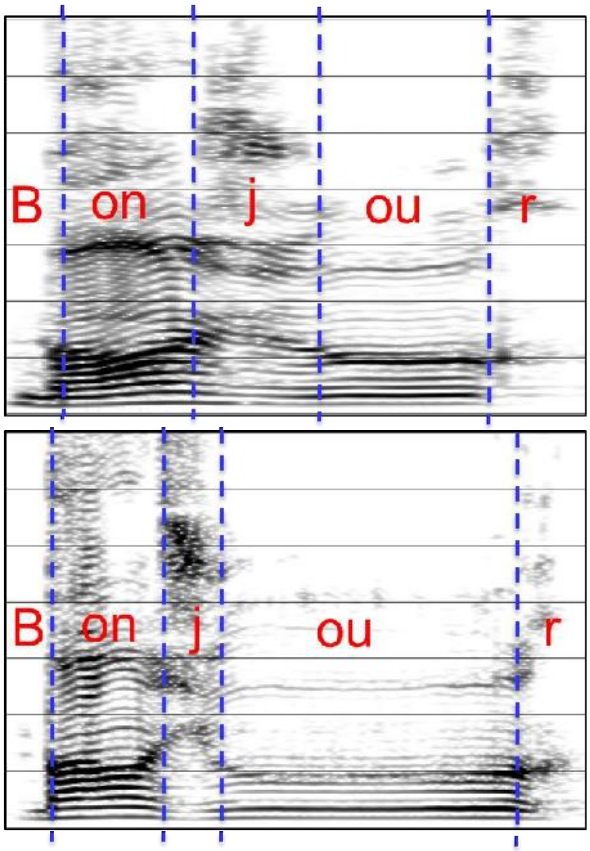
**Problème :** même si les mots sont de même durée, les phonèmes risquent d'être décalés les uns par rapport aux autres.



Normaliser les 2 mots à 1 sec:

- Étirer « bonjour 0.6sec».
- Comprimer « bonjour 1.1sec »

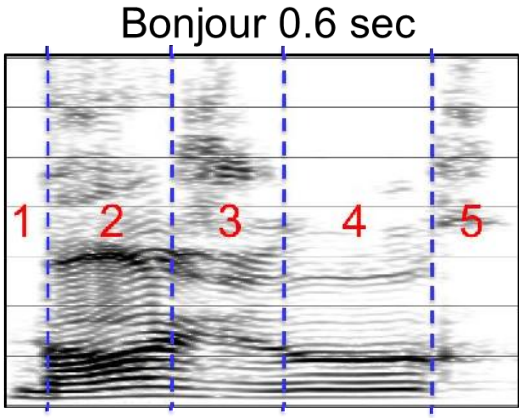
**Problème :** Même après la normalisation linéaire, les phonèmes sont décalés par rapport aux deux prononciations ce qui risque de fausser la comparaison. La deuxième prononciation, le locuteur avait rallongé la durée du « ou » dans « bonjour ».




### *VIII.2.3.2 Normalisation temporelle curviligne*

La normalisation curviligne consiste à faire une interpolation non linéaire du signal de parole sur une durée fixe. Cette technique consiste à harmoniser la durée des parties stationnaires du signal.

L'exemple ci-dessous montre les zones pseudo-stationnaires (5) pour une prononciation du mot « bonjour ».



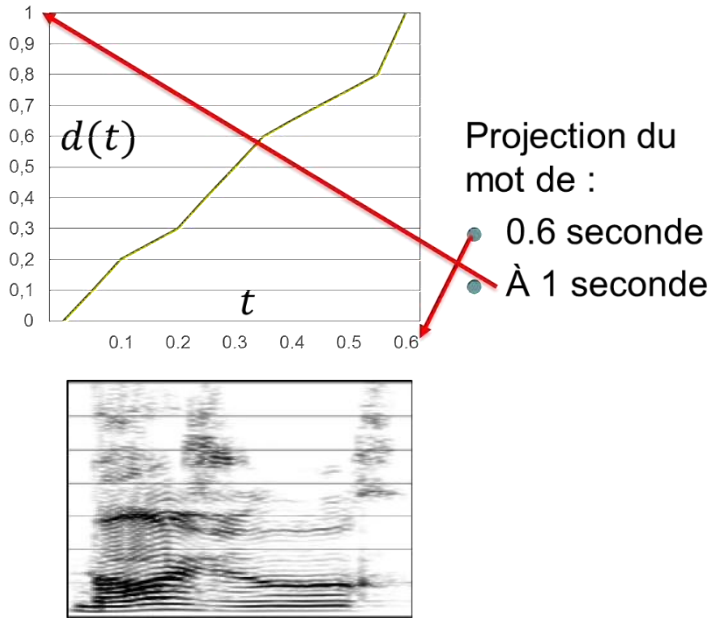

- On dispose de 5 zones pseudo-stationnaires
- On veut normaliser à 1 seconde
- Chaque zone sera normalisée linéairement à une durée de $\frac{1}{5}$=0.2 secondes

$$f(t) = distance\left(S_\omega(t),\, S_\omega(t-1)\right) \;\; avec \;\; f(0) = 0$$

$S_\omega(t)$: Vecteur de parole à l'instant $t$

$$C(t) = \sum_{i=0}^{t} f(i) \quad \text{(Somme cumulée des différences)}$$

$$d(t) = \frac{N}{C(T)} C(t) \quad \text{(Vecteur de projection d'un mot de durée } T \text{ vers un mot de durée } N\text{)}$$



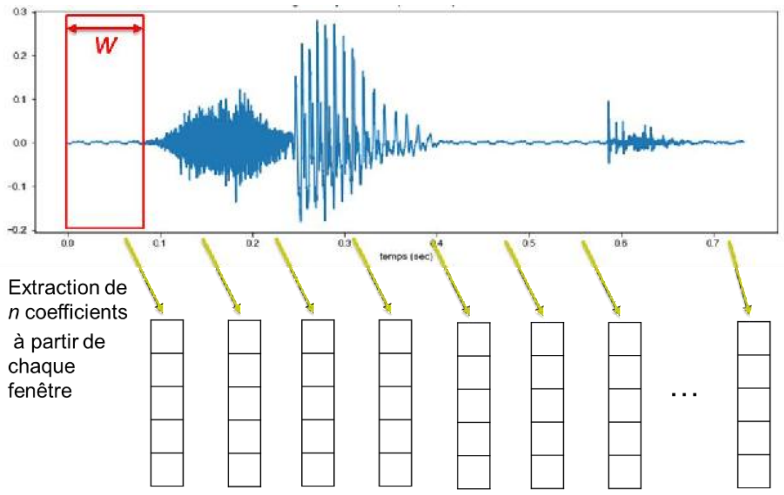

## VIII.3 Extraction des caractéristiques

Le signal de parole est un flux continu non stationnaire, il est nécessaire d'extraire des coefficients caractéristiques en nombre limité sur de courtes durées du signal.

Les coefficients extraits du signal se doivent d'être :

- **Pertinents :** caractéristiques et distinctifs des catégories avec un minimum de redondance.
- **Discriminants :** en nombre suffisant pour permettre une séparation entre les catégories.
- **Robustes :** résistance au bruit.

Le codage permet de réduire les données d'une fenêtre de taille *w* à *n* coefficients pertinents seulement, par exemple de : w= 1024 échantillons à 13 coefficients MFCCs.



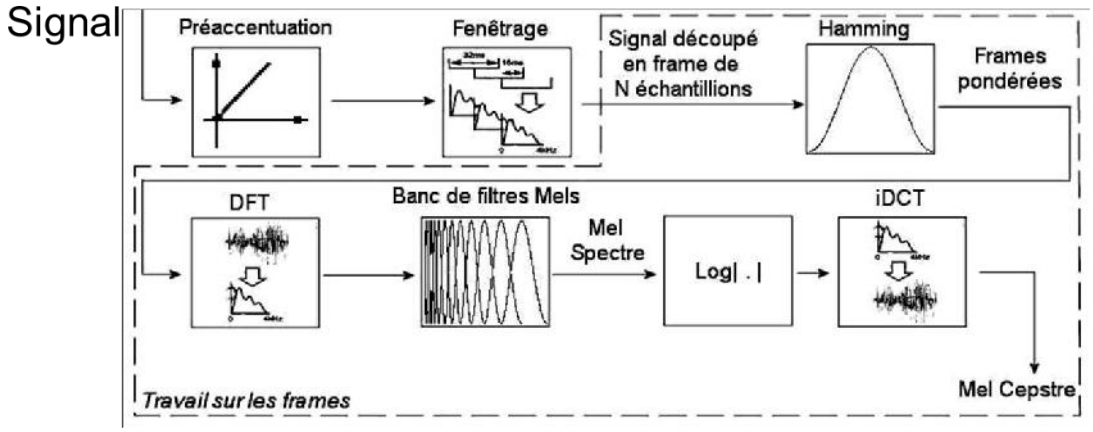
Les techniques de codage de la parole reposent sur des études théoriques (phonétique articulatoire, acoustique & auditive) les plus utilisés en reconnaissance de la parole sont :

- **Le codage LPC :** basé théorie de l'articulation (séparation source/filtre), vue dans le cours modélisation de la parole.
- **Le codage MFCC :** basé à la fois sur la théorie de l'articulation + la théorie auditive (échelle Mel, voir cours phonétique auditive).

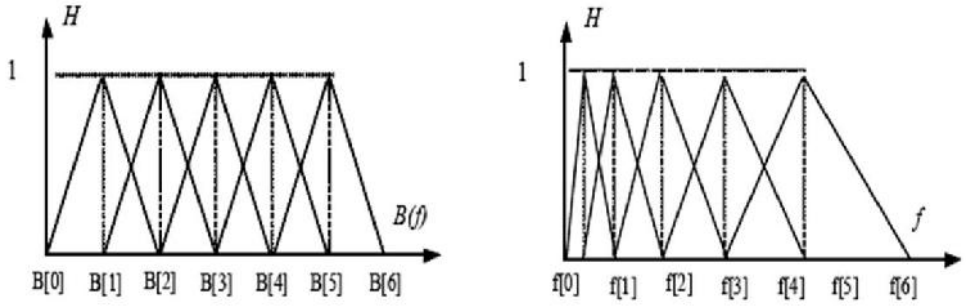
### *VIII.3.1.1 Le codage MFCC*

Le codage MFCC est l'un des codages des plus utilisés et des plus robustes en reconnaissance de la parole. Le codage MFCC effectue une séparation source/filtre en utilisant le calcul du Cepstre, cependant 2 modifications ont été apportées afin de réduire le nombre de coefficients par rapport au lissage cepstral classique :

- Le spectrogramme du signal passe par un banc de filtres passe-bandes en Mel simulant le filtrage fait par l'oreille humaine, usuellement (26 filtres).
- La transformée de Fourier inverse (après le $log$) est remplacée par une transformée en cosinus inverse ($IDCT$) afin de simplifier les calculs (éviter les nombres complexes).




Les filtres triangulaires passe-bandes en Mel-Fréquence (B(f)) sont disposés de façons linéaire et en fréquence Hz (f) de façon non linéaire



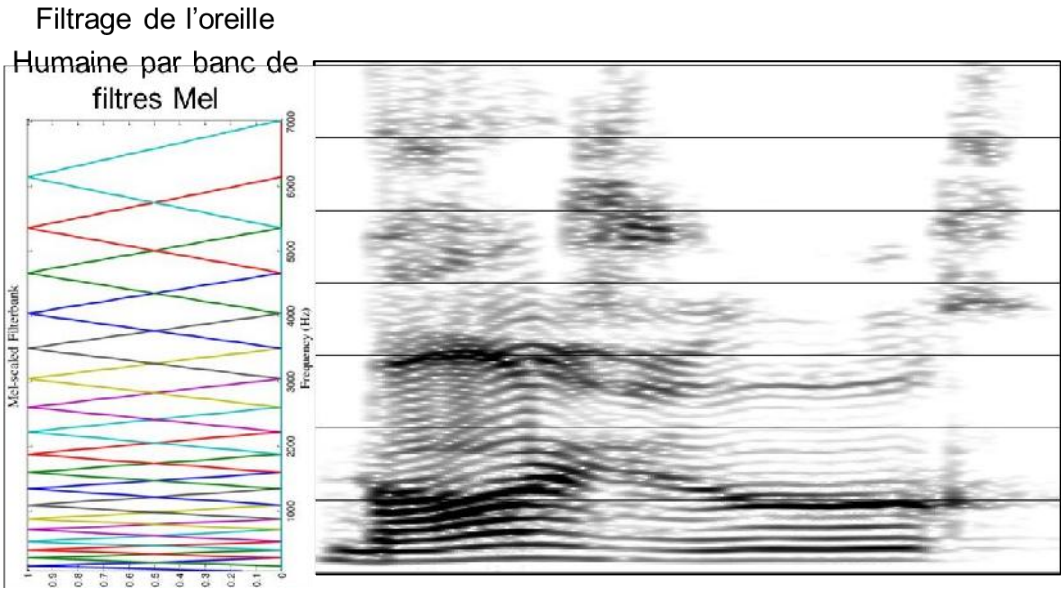

Les formules de conversion entre fréquences Hz et fréquences Mel sont les suivantes :

Convertir une fréquence $f$ en Hz en Mel :

$$B(f) = 2595 \times \log_{10}(1 + \frac{f}{700})$$

Convertir d'une fréquence Mel $b$ vers une fréquence Hz :

$$B^{-1}(b) = 700 \times (10^{\frac{b}{2595}} - 1)$$




Chaque fenêtre du spectre d'amplitude passe par *n* filtres à l'échelle Mel, chaque filtre écoute une bande de fréquence limitée, un produit vectoriel entre le filtre et la fenêtre nous permet d'obtenir une sortie par filtre. Donc, par exemple pour une fenêtre de 1024 échantillons et 26 filtres, chaque fenêtre sera réduite à 26 valeurs au lieu de 1024 (réduction de la taille des données).

**Transformée de cosinus inverse(IDCT) :**

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left[\frac{\pi n\left(m + \frac{1}{2}\right)}{M}\right] \qquad 0 \leq n \leq M$$

La transformé de cosinus inverse se substitue à la transformé de Fourier pour le passage au domaine quefrentiel, on ne prend usuellement que les 13 premières valeurs au lieu de *n* (nombre de filtres Mel) afin de simuler le liftage (Cepstre), ceci permet de réduire encore plus la taille des données et de faire en même temps la séparation source/filtre pour ne garder que les paramètres du conduit vocal.

## VIII.4 Prise de décision (classification)

La prise de décision ou la classification en reconnaissance de la parole est le faite d'associer à une séquence de vecteurs caractéristiques un label choisi parmi plusieurs (représentant une partie de ce qui a été dit). La décision repose sur une base d'exemples établie au préalable, soit par comparaison de l'entrée actuelle avec une mesure de similarité soit par la synthèse de modèles/classifieurs par

apprentissage. La difficulté de la parole réside dans la longueur des séquences de vecteurs caractéristiques qui est dynamique, à défaut d'utiliser une technique de normalisation temporelle, l'utilisation d'une technique dynamique est nécessaire, dans la littérature de la reconnaissance de la parole on trouve :

- La distance DTW
- Le HMM
- Les réseaux de neurones récurrents/temporels

### VIII.4.1 La distance DTW

La distance DTW est une mesure de dissimilarité qui permet de comparer des séquences de vecteurs de longueurs différentes et est donc idéale pour la reconnaissance de la parole, cependant elle exige une comparaison avec toute la base d'exemples, le temps de prise de décision risque d'accroitre avec l'accroissement de la taille de cette dernière, sachant que plus il y a d'exemples, plus la classification est précise.

**Algorithme DTW :** On souhaite calculer la distance entre la séquence $X$ et la séquence référence $R$

$n$: nombre de vecteurs de $R$, $m$: nombre de vecteurs de $X$

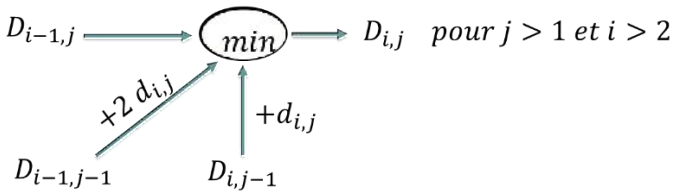
On note $d_{i,j} = d(R_i, X_j)$ la distance (exemple Hamming) entre la vecteur $i$ de $R$ et le vecteur $j$ de $X$

$$d(A,B) = \sum_i |A_i - B_i| \quad (distance\ de\ Hamming)$$

1. On initialise :

$$D_{1,1} = 2 \times d_{1,1}$$
$$D_{i,1} = \infty \quad pour\ i > 1$$

2. On calcul :




3. Pour finir :

$$DTW(R,X) = \frac{D_{n,m}}{n+m}$$

**Exemple :** Soit la matrice de coefficients suivante obtenue après codage du signal de parole.

$$\begin{bmatrix} 0&0&0&1&0&0&0&0&1 \\ 0&1&0&1&0&0&1&0&1 \\ 0&0&1&0&1&0&0&1&0 \end{bmatrix}$$

Question : Faites la segmentation en mots de ce signal de parole avec un seuil=0.1, puis reconnaissez les mots trouvés avec la distance DTW.

Mots de références :

$$\begin{bmatrix}0&0&0&1&0\\1&0&0&1&0\\0&1&1&0&1\end{bmatrix}$$ **café**

$$\begin{bmatrix}0\\0\\1\end{bmatrix}$$ **au**

$$\begin{bmatrix}0&0\\1&0\\0&1\end{bmatrix}$$ **lait**

$$\begin{bmatrix}0&0&1&1\\1&0&1&1\\0&1&0&0\end{bmatrix}$$ **noir**

Segmentation :

$$\begin{bmatrix}0&0&0&1&0&0&0&0&1\\0&1&0&1&0&0&1&0&1\\0&0&1&0&1&0&0&1&0\end{bmatrix}$$

$$\text{Energie}=\frac{1}{3}\begin{bmatrix}0&1&1&2&1&0&1&1&2\end{bmatrix}$$

$$\text{Energie<seuil}=\begin{bmatrix}1&0&0&0&0&1&0&0&0\end{bmatrix}$$ 1 : True (silence), 0 : False (mot)

$$\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix}$$ mot1 $$\begin{bmatrix}0&0&1\\1&0&1\\0&1&0\end{bmatrix}$$ mot2

Comparaison :

$$\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix}$$ mot1

$$\begin{bmatrix}0&0&0&1&0\\1&0&0&1&0\\0&1&1&0&1\end{bmatrix}$$ café $$\begin{bmatrix}0\\0\\1\end{bmatrix}$$ au $$\begin{bmatrix}0&0\\1&0\\0&1\end{bmatrix}$$ lait $$\begin{bmatrix}0&0&1&1\\1&0&1&1\\0&1&0&0\end{bmatrix}$$ noir

Références

Comparaison « mot1 » et « café » :

café $\begin{bmatrix}0&0&0&1&0\\1&0&0&1&0\\0&1&1&0&1\end{bmatrix}$

| | | | | |
|---|---|---|---|---|
| ∞ | 0 3 | 3 3 | 0 0 | |
| ∞ | 3 3 | 0 0 | 3 3 | |
| ∞ | 0 0 | 3 3 | 0 3 | |
| ∞ | 0 0 | 3 3 | 0 3 | |
| 0 0 | 2 2 | 1 3 | 2 5 | |

$$\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix}$$ mot1

$$dtw(mot1, café) = \frac{0}{4+5} = 0$$

Comparaison « mot1 » et « au » :

au $\begin{bmatrix}0\\0\\1\end{bmatrix}$

| | | | |
|---|---|---|---|
| 2 4 | 0 4 | 3 7 | 0 7 |

$$\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix}$$ mot1

$$dtw(mot1, au) = \frac{7}{4+1} = 1.4$$

Après comparaison avec toutes les références :

$$dtw(mot1, café) = \frac{0}{4+5} = 0$$

$$dtw(mot1, au) = \frac{7}{4+1} = 1.4$$

$$dtw(mot1, lait) = \frac{3}{4+2} = 0.5$$

$$dtw(mot1, noir) = \frac{3}{4+4} = 0.375$$

Minimum→ mot1 est reconnu en tant que « café »

$$dtw(mot2, café) = \frac{3}{3+5} = 0.375$$

$$dtw(mot2, au) = \frac{7}{3+1} = 1.75$$

$$dtw(mot2, lait) = \frac{3}{3+2} = 0.6$$

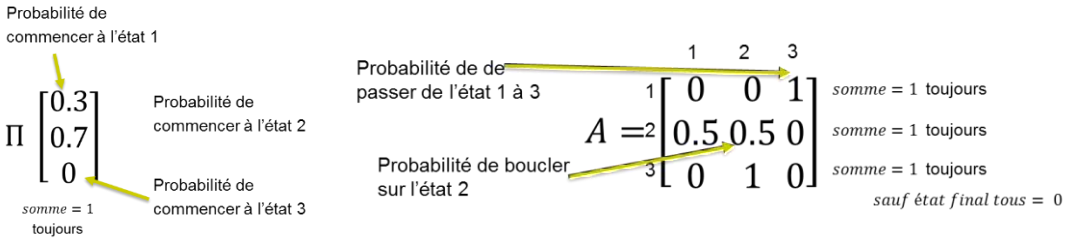
$$dtw(mot2, noir) = \frac{0}{3+4} = 0$$

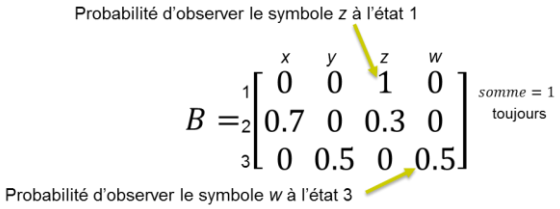
Minimum→ mot2 est reconnu en tant que « noir »

### VIII.4.2 Le HMM

Le HMM est un modèle probabiliste qui comprend une suite d'états et des symboles observables, les probabilités de transitions entre états ne dépendent que de l'état précédent. Dans chaque état on peut observer chaque symbole avec une certaine probabilité. Un HMM comprend la matrice des probabilités initiales notée $\Pi$, la matrice des probabilités de transitions $A$ et la matrice des probabilités d'observations $B$.

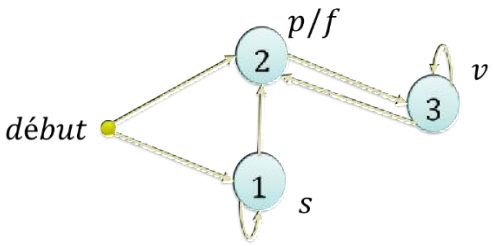
**Exemple :** $HMM(\Pi, A, B)$



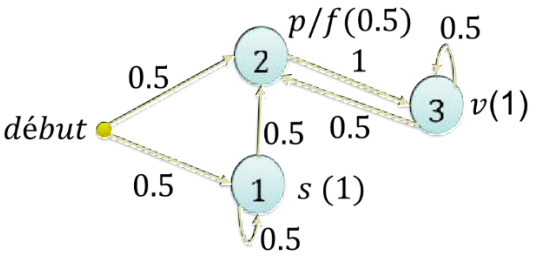



### VIII.4.2.1 Exercice 1

Complétez ce HMM sachant que toutes les probabilités sont équiprobables (les chemins/observations possibles partagent les probabilités à parts égales).




**Réponse :** $HMM(\Pi, A, B)$




$$\Pi = \begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}0.5\\0.5\\0\end{bmatrix} \qquad A = \begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}0.5 & 0.5 & 0\\0 & 0 & 1\\0 & 0.5 & 0.5\end{bmatrix} \qquad B = \begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}1 & 0 & 0 & 0\\0 & 0.5 & 0.5 & 0\\0 & 0 & 0 & 1\end{bmatrix}$$

### VIII.4.2.2 Algorithme de Viterbi

Soit un $HMM(\Pi, A, B)$ et une suite d'observations : $o = o_0 o_1 o_2 \dots o_n$ de longueur *n+1.*

On veut trouver la meilleure suite d'états et sa probabilité du HMM défini qui a généré la suite d'observations o.

$$a_0 = \Pi \times B(o_0)$$

$$a_i = max_{ligne}(A^T \times a_{i-1}^T) \times B(o_i)$$

$$I_i = argmax_{ligne}(A^T \times a_{i-1}^T)$$

Probabilité du meilleur chemin : $P_o = max\ (a_n)$

Le meilleur chemin est : $c_0 c_1 c_2 \dots c_n$ Tel que : $c_n = argmax\ (a_n)$ et $c_{i-1} = I_i(c_i)$

**Remarque :** ici $A^T \times a_{i-1}^T$ consiste à dupliquer les lignes de $a_{i-1}^T$ autant de fois qu'il y a de lignes dans la matrice $A^T$ et de faire une multiplication élément par élément.

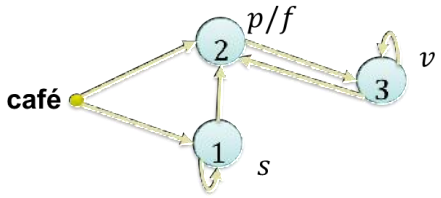
### VIII.4.2.3 Exercice 2

Soit les mots trouvés à partir du signal après segmentation :

$$\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix} \qquad \begin{bmatrix}0&0&1\\1&0&1\\0&1&0\end{bmatrix}$$

mot1 mot2

Soit le HMM du mot café obtenu après apprentissage sur des exemples (on suppose que toutes les probabilités sont équiprobables) :




Soit les symboles observés trouvé avec une quantification vectorielle sur les exemples d'apprentissage :

$$\underset{s}{\begin{bmatrix}0\\0\\0\end{bmatrix}} \quad \underset{p}{\begin{bmatrix}0\\1\\0\end{bmatrix}} \quad \underset{f}{\begin{bmatrix}1\\1\\0\end{bmatrix}} \quad \underset{v}{\begin{bmatrix}0\\0\\1\end{bmatrix}}$$

**Question :** quelle est la probabilité que le mot1 soit le mot café ? Faites pareil pour mot2.

**Réponse pour mot1:**

Transformer mot1 en suite d'observations en calculant la distance entre chaque vecteur du mot1 et les symboles du HMM, puis à remplacer par le plus proche, puis Calculer la probabilité de O avec Viterbi.

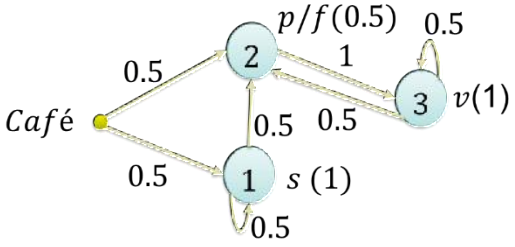
$$\underset{\text{mot1}}{\begin{bmatrix}0&0&1&0\\1&0&1&0\\0&1&0&1\end{bmatrix}} \qquad \underset{s}{\begin{bmatrix}0\\0\\0\end{bmatrix}} \quad \underset{p}{\begin{bmatrix}0\\1\\0\end{bmatrix}} \quad \underset{f}{\begin{bmatrix}1\\1\\0\end{bmatrix}} \quad \underset{v}{\begin{bmatrix}0\\0\\1\end{bmatrix}}$$

$$O = p\ v\ f\ v$$

$$O_{mot1} = pvfv$$

Compléter le HMM avec les probabilités (dans ce cas supposées équiprobables) :

$$\Pi = \begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}0.5\\0.5\\0\end{bmatrix} \qquad A = \begin{matrix}1\\2\\3\end{matrix}\overset{\begin{matrix}1&2&3\end{matrix}}{\begin{bmatrix}0.5&0.5&0\\0&0&1\\0&0.5&0.5\end{bmatrix}} \qquad B = \begin{matrix}1\\2\\3\end{matrix}\overset{\begin{matrix}s&p&f&v\end{matrix}}{\begin{bmatrix}1&0&0&0\\0&0.5&0.5&0\\0&0&0&1\end{bmatrix}}$$




**Application de l'algorithme de Viterbi pour:** $O_{mot1} = pvfv$

$$a_0 = \Pi \times B(p) = \begin{bmatrix}0.5\\0.5\\0\end{bmatrix} \times \begin{bmatrix}0\\0.5\\0\end{bmatrix} = \begin{bmatrix}0\\0.25\\0\end{bmatrix}$$

| | |
|---|---|
| $a_1 = max_{ligne}(A^T \times a_0^T) \times B(v)$<br>$= max_{ligne}(\begin{bmatrix}0.5 & 0 & 0\\ 0.5 & 0 & 0.5\\ 0 & 1 & 0.5\end{bmatrix} \times [0 \;\; 0.25 \;\; 0]) \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$= max_{ligne}(\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0.25 & 0\end{bmatrix}) \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$= \begin{bmatrix}0\\0\\0.25\end{bmatrix} \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$a_1 = \begin{bmatrix}0\\0\\0.25\end{bmatrix}$ | $I_1 = argmax_{ligne}(A^T \times a_0^T)$<br>$I_1 = \begin{bmatrix}1\\1\\2\end{bmatrix}$<br>Indice du max de chaque ligne, si égalité on prend le premier. |
| $a_2 = max_{ligne}(A^T \times a_1^T) \times B(f)$<br>$= max_{ligne}(\begin{bmatrix}0.5 & 0 & 0\\ 0.5 & 0 & 0.5\\ 0 & 1 & 0.5\end{bmatrix} \times [0 \;\; 0 \;\; 0.25]) \times \begin{bmatrix}0\\0.5\\0\end{bmatrix}$<br>$= max_{ligne}(\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0.125\\ 0 & 0 & 0.125\end{bmatrix}) \times \begin{bmatrix}0\\0.5\\0\end{bmatrix}$<br>$= \begin{bmatrix}0\\0.125\\0.125\end{bmatrix} \times \begin{bmatrix}0\\0.5\\0\end{bmatrix}$<br>$a_2 = \begin{bmatrix}0\\0.0625\\0\end{bmatrix}$ | $I_2 = argmax_{ligne}(A^T \times a_1^T)$<br>$I_2 = \begin{bmatrix}1\\3\\3\end{bmatrix}$ |
| $a_3 = max_{ligne}(A^T \times a_2^T) \times B(v)$<br>$= max_{ligne}(\begin{bmatrix}0.5 & 0 & 0\\ 0.5 & 0 & 0.5\\ 0 & 1 & 0.5\end{bmatrix} \times [0 \;\; 0.0625 \;\; 0]) \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$= max_{ligne}(\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0.0625 & 0\end{bmatrix}) \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$= \begin{bmatrix}0\\0\\0.0625\end{bmatrix} \times \begin{bmatrix}0\\0\\1\end{bmatrix}$<br>$a_3 = \begin{bmatrix}0\\0\\0.0625\end{bmatrix}$ | $I_3 = argmax_{ligne}(A^T \times a_2^T)$<br>$I_3 = \begin{bmatrix}1\\1\\2\end{bmatrix}$ |

La probabilité que le mot1 soit le mot café : $P_o = max\,(a_3) = \mathbf{0.0625}$

$$c_3 = 3,\;\; c_2 = I_3(c_3) = 2,\; c_1 = I_2(c_2) = 3, c_0 = I_1(c_1) = 2$$

La suite d'états la plus probable qui a généré cette séquence d'observations est : 2 - 3 - 2 - 3

### *VIII.4.2.4 HMM hiérarchique*

Dans le cas de la parole continue (sans silence entre les mots), on établit un HMM hiérarchique en combinant plusieurs HMMs (de chaque mot/phonème par exemple) via un modèle de langage et en utilisant l'algorithme du jeton afin de trouver le chemin le plus probable et du coup les mots prononcés.

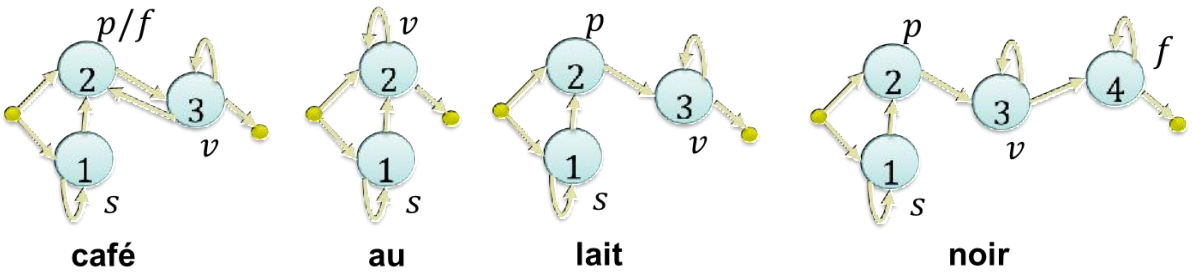
**Exercice 1 :** Trouvez la phrase qui a été dites à partir des coefficients suivants

$$\begin{bmatrix} 0&0&0&1&0&0&0&0&1 \\ 0&1&0&1&0&0&1&0&1 \\ 0&0&1&0&1&0&0&1&0 \end{bmatrix}$$

On dispose des symboles observés suivants (estimés par k-means) :

$$s \begin{bmatrix}0\\0\\0\end{bmatrix} \quad p \begin{bmatrix}0\\1\\0\end{bmatrix} \quad f \begin{bmatrix}1\\1\\0\end{bmatrix} \quad v \begin{bmatrix}0\\0\\1\end{bmatrix}$$

Et des HMMs suivants (dont les probabilités sont équiprobables et disposant d'un état initial et d'un état final) :




On suppose qu'on peut reconnaitre une des deux phrases suivantes :

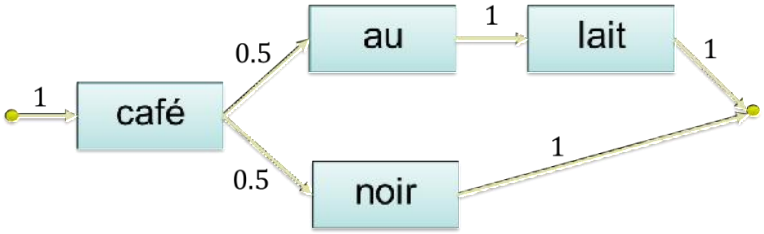
- Café au lait
- Café noir

**Réponse :**

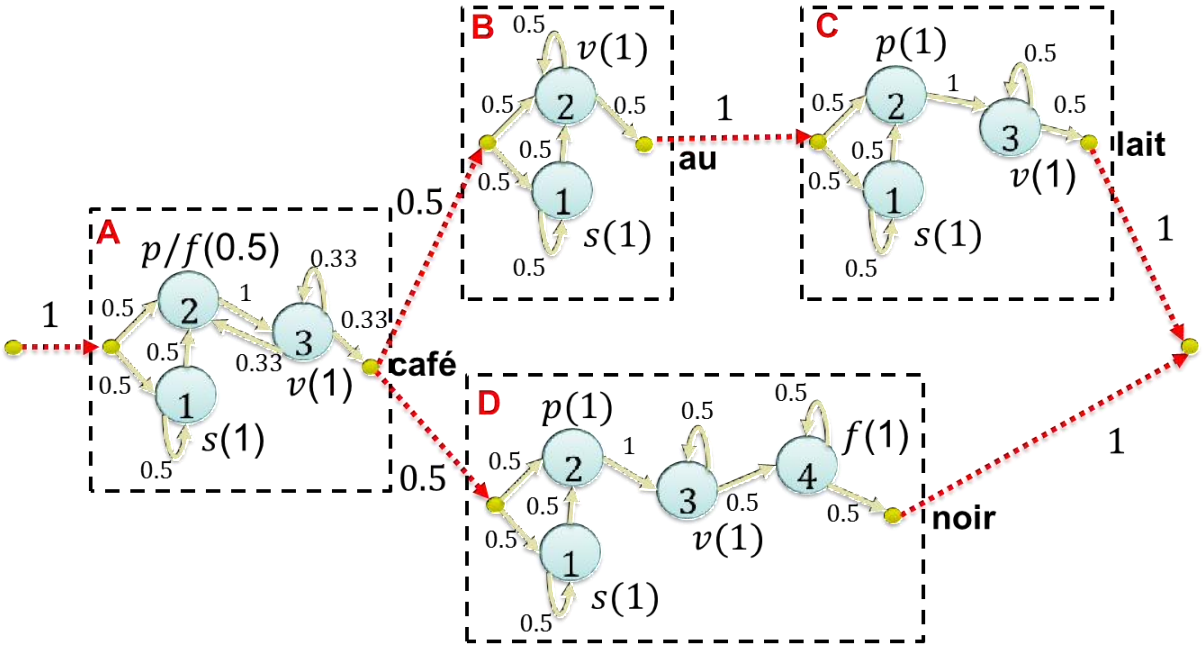
1. Quantifier les coefficients en symboles :

$$\begin{bmatrix} 0&0&0&1&0&0&0&0&1 \\ 0&1&0&1&0&0&1&0&1 \\ 0&0&1&0&1&0&0&1&0 \end{bmatrix} \longrightarrow spvfvspvf$$

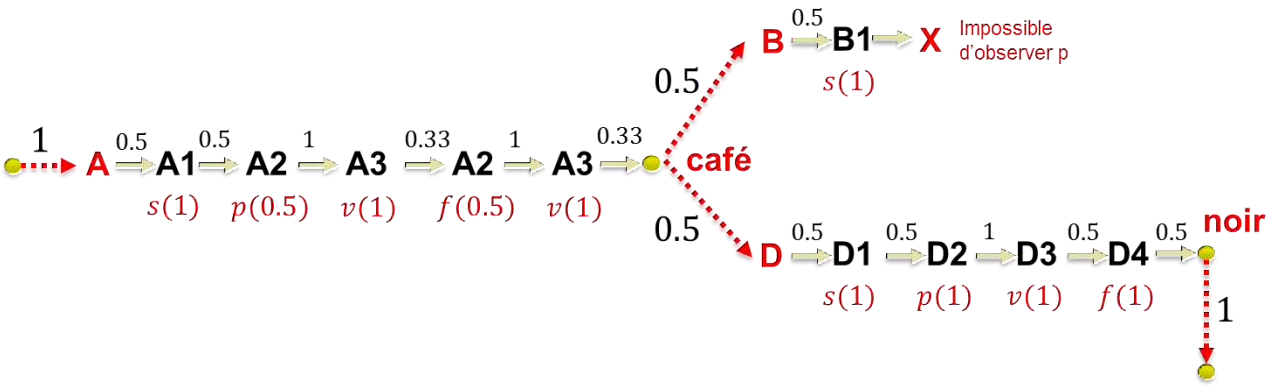
2. Établir un modèle de langage à partir des phrases à reconnaitre :



3. Établir le HMM hiérarchique à partir du modèle de langage et compléter les probabilités :




4. Trouver le chemin avec la probabilité maximal pour la suite : $spvfvspvf$

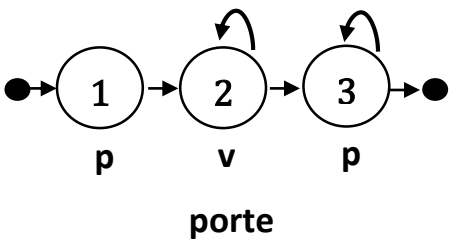



Un seul chemin trouvé et donc « $spvfvspvf$ » correspond à la phrase « café noir » avec la probabilité :

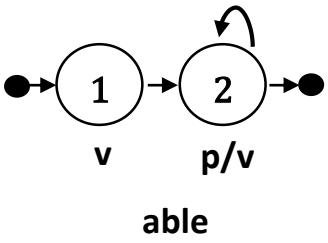
$$P = 0.5^7 \times 0.33^2 = 0.00085078125$$

**Exercice 2 :**

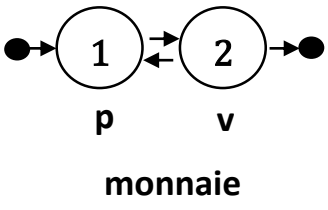
On veut réaliser un système de reconnaissance de la parole qui reconnait les termes suivants : (**porte**, **portable**, **porte-monnaie**), on dispose des HMMs suivants (avec des probabilités équiprobables) obtenus après apprentissage sur des exemples :




porte

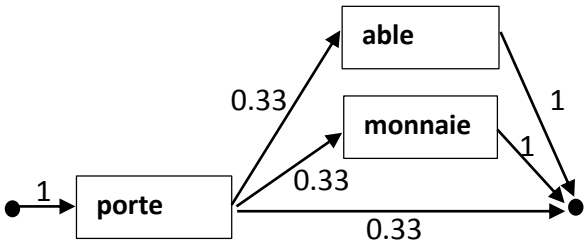



able

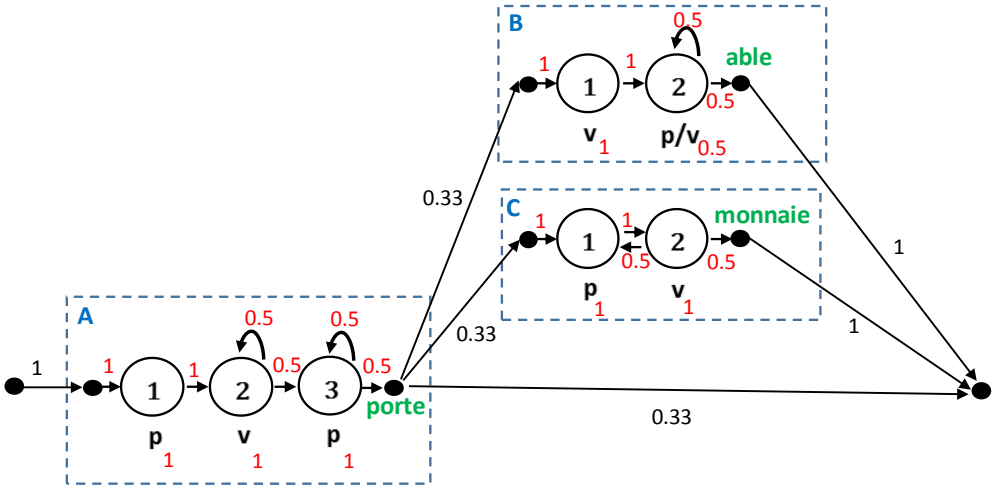



monnaie

a: Réalisez un modèle de langage pour reconnaitre les termes (**porte**, **portable**, **porte-monnaie**) avec les labels des HMMs si dessus.




b: Déduire le HMM hiérarchique pour notre système de reconnaissance en complétant les probabilités sur les HMMS.



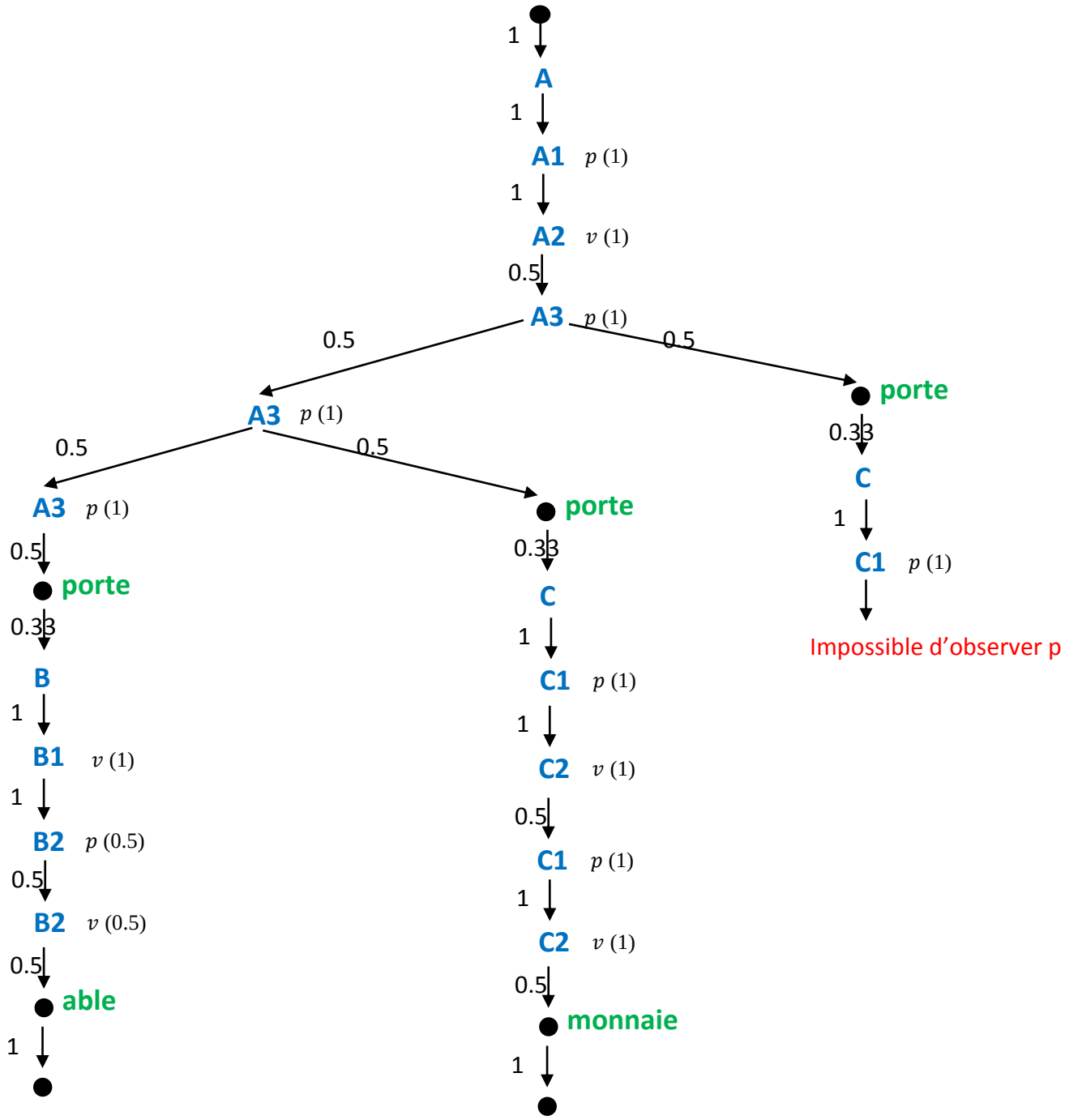

c: Trouvez le terme prononcé à partir de ces coefficients :

$$\begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.1 & 1.0 & 1.0 & 1.1 & 0.0 & 1.0 & 0.0 \\ 0.1 & 1.0 & 0.1 & 0.1 & 0.1 & 1.1 & 0.1 & 1.0 \end{pmatrix}$$

Sachant que : $p = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ $v = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$p \quad v \quad p \quad p \quad p \quad v \quad p \quad v$

**Conversion en symboles**




**Portable** reconnu avec une probabilité : $p1 = 0.5^8 \times 0.33 = 0{,}0012890625$
**Porte-monnaie** reconnu avec une probabilité : $p2 = 0.5^5 \times 0.33 = 0{,}0103125$

p2 > p1, donc le terme reconnue est porte-monnaie.

# IX Liste des abréviations

API : Alphabet Phonétique International

DFT : Discrete Fourier Transform (Transformée de Fourier discrète)

DTW : Dynamic Time Warping

FFT : Fast Fourier Transform (Transformée de Fourier rapide)

FS : Frequency Sampling (fréquence d'échantillonnage)

HMM : Hidden Markov Model (Modèle de Markov caché)

IDCT : Inverse Discrete Cosine Transform

LPC : Linear Predictive Coding

MFCC : Mel-scale Frequency Cepstral Coefficients

TAP : Traitement Automatique de la Parole

# X Bibliographie


1. L. Buniet. Traitement automatique de la parole en milieu bruité, étude de modèles connexionnistes statiques et dynamiques. Thèse présentée et soutenue publiquement le 10 février 1997 pour l'obtention du Doctorat de l'Université Henri Poincaré – Nancy 1 spécialité informatique, France.

2. N. Quoc Cuong. Reconnaissance de la parole en langue Vietnamienne. Thèse présentée et soutenue publiquement le 19 juin 2002 pour l'obtention du grade de Docteur de l'INPG de l'Institut National Polytechnique de Grenoble, France.

3. L. Besacier. Parole et technologies vocales. Cours de M2R informatique présenté à l'université J. Fourier de Grenoble, France.

4. M. Kowalski. Transmission multimedia : codage et traitement de la parole. Cours présenté à l'université de Paris-sud, France.

5. M. Tahon. Acoustique de la voix. Cours niveau M2 de Traitement Automatique de la parole, Master ATAL, Université le Mans / Nantes, France.

6. W. Chou & B. H. Juang, Eds., Pattern Recognition in Speech and Language Processing. Boca Raton, FL, USA: CRC Press, Inc., 2002.