La synthèse vocale

OBJECTIF: GÉNÉRER NUMÉRIQUEMENT UNE VOIX HUMAINE À PARTIR DE D'UN TEXTE

Sommaire

I. Etude de la voix humaine

1)- Origine et caractérisation de la voix humaine

- Voix

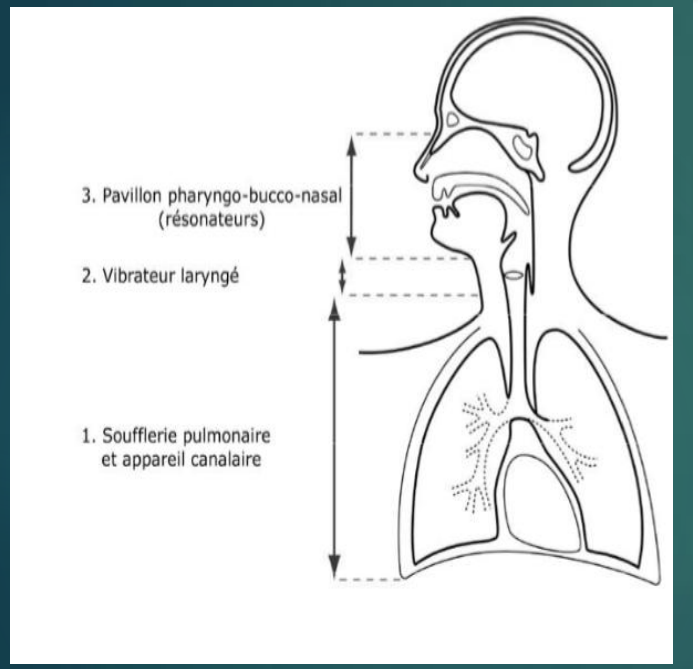
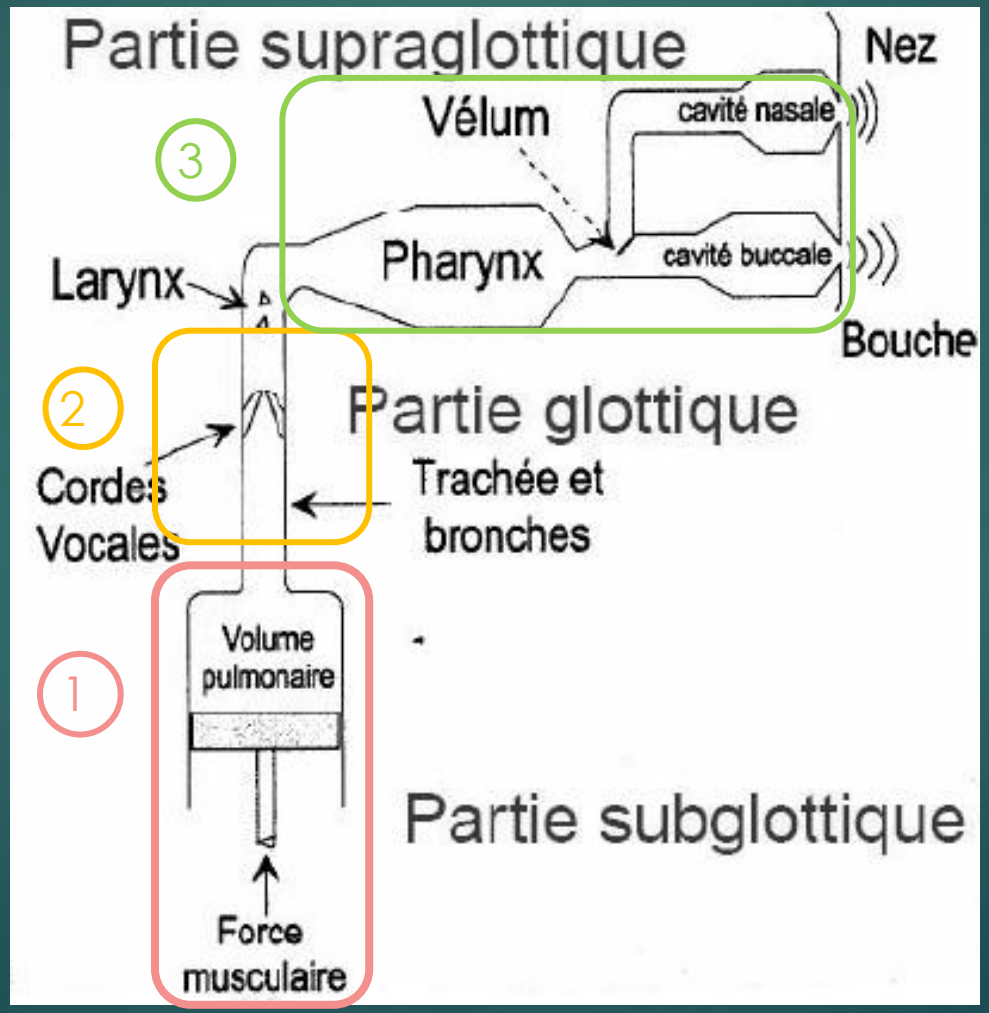


Schéma de l'appareil phonatoire

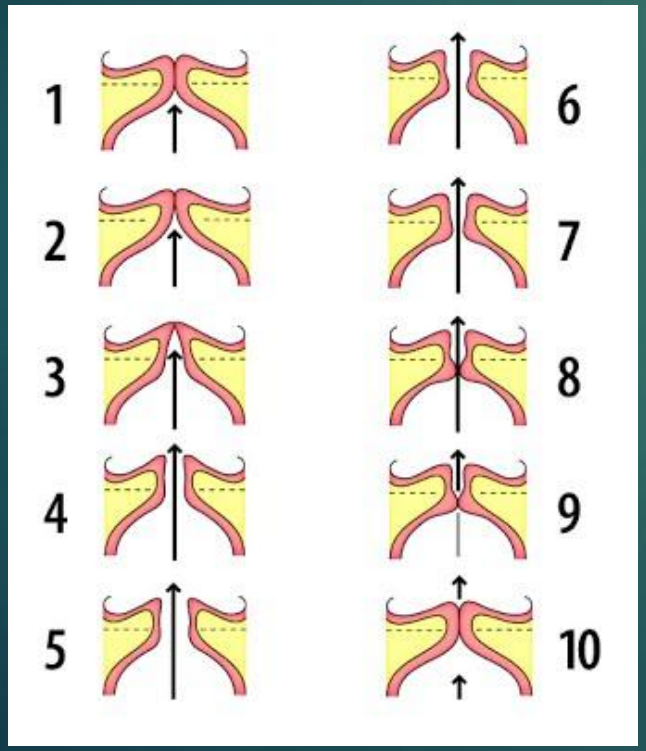


Mécanisme de production de la parole :

- Contraction du diaphragme (1)
- Libération d'air dans la trachée
- Passage de l'air au niveau:
  - des cordes vocales (2)
  - des résonateurs (3)

Les cordes vocales

- 3 positions fondamentales
- Rôle de valve régulant le débit d'air
- Mouvement cyclique

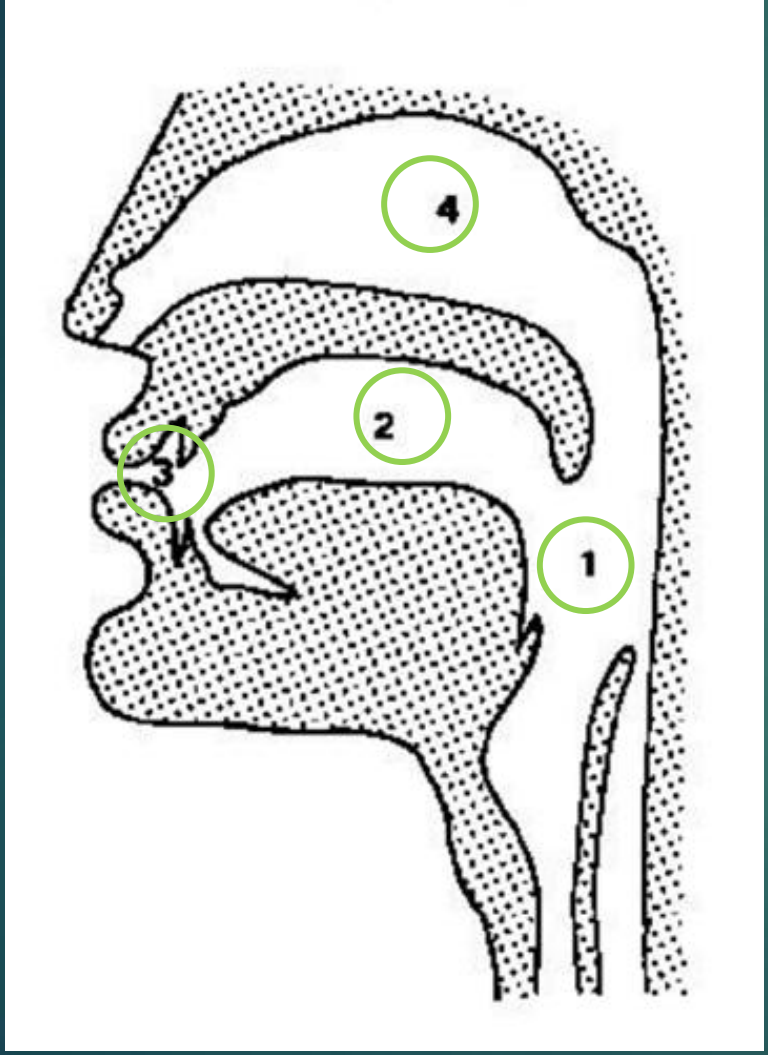


(2) écoulement turbulent

(4) et (5) écoulement parfait stationnaire incompressible et homogène

(6) Effet Venturi

Les résonateurs



Les principaux :

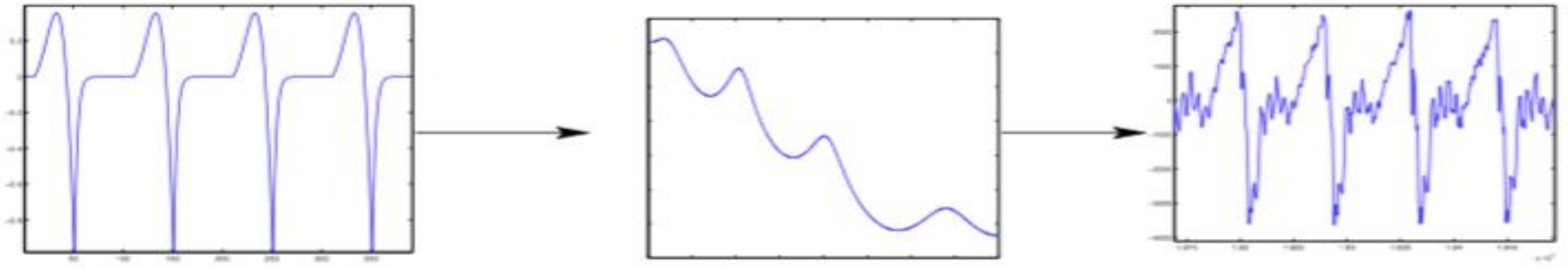
- Le pharynx (1)
- La cavité buccale (2)
- La cavité labiale (3)
- Les fosses nasales (4)

- Fréquence de résonance propre

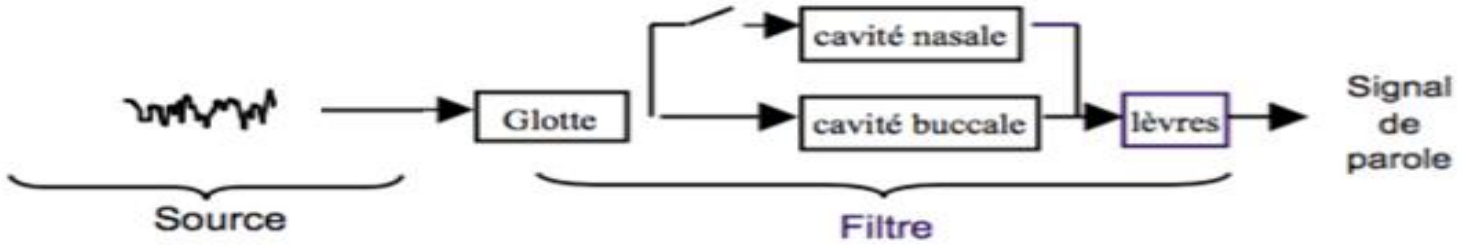
- Condition de résonance: $f=f_{résonateur}$

  f = formant

Modélisation source-filtre de l'appareil phonatoire



(a) source (b) Filtre (c) signal de parole



Production de la parole : le modèle source-filtre.

2)-Paramètres pertinents à étudier

La prosodie :

- Durée

- Intensité (niveau sonore)

$I_{dB} = 10 \log\left(\frac{I}{I_0}\right)$

I_{dB} en dB

I intensité sonore en $W.m^{-2}$

$I_0 = 10^{-12}\ W.\ m^{-2}$

- Hauteur

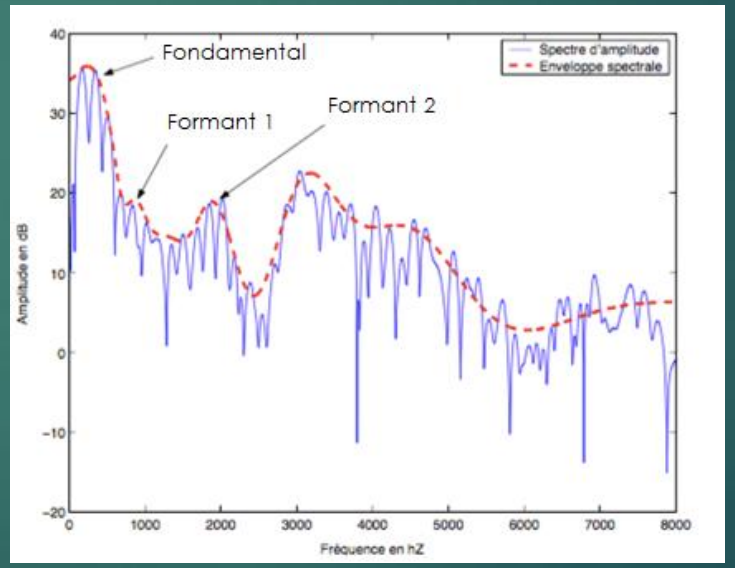
→ Pitch (fondamental)

→ Formants

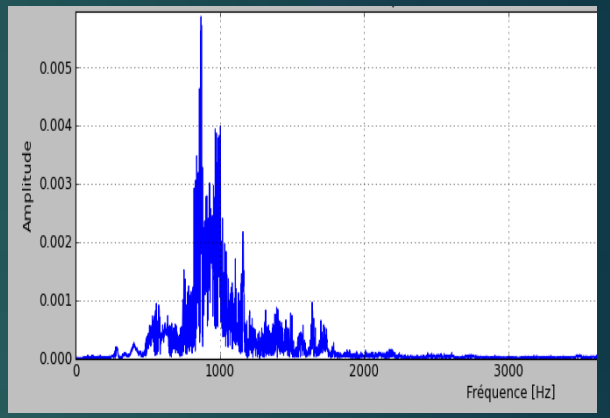
F1 : ouverture de la bouche

F2 : position de la langue

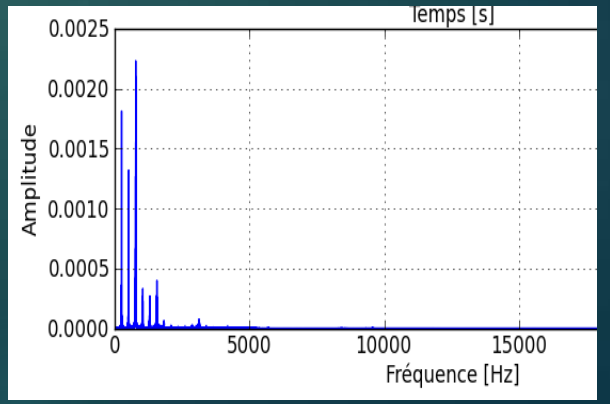
F3 : configuration des lèvres



Google



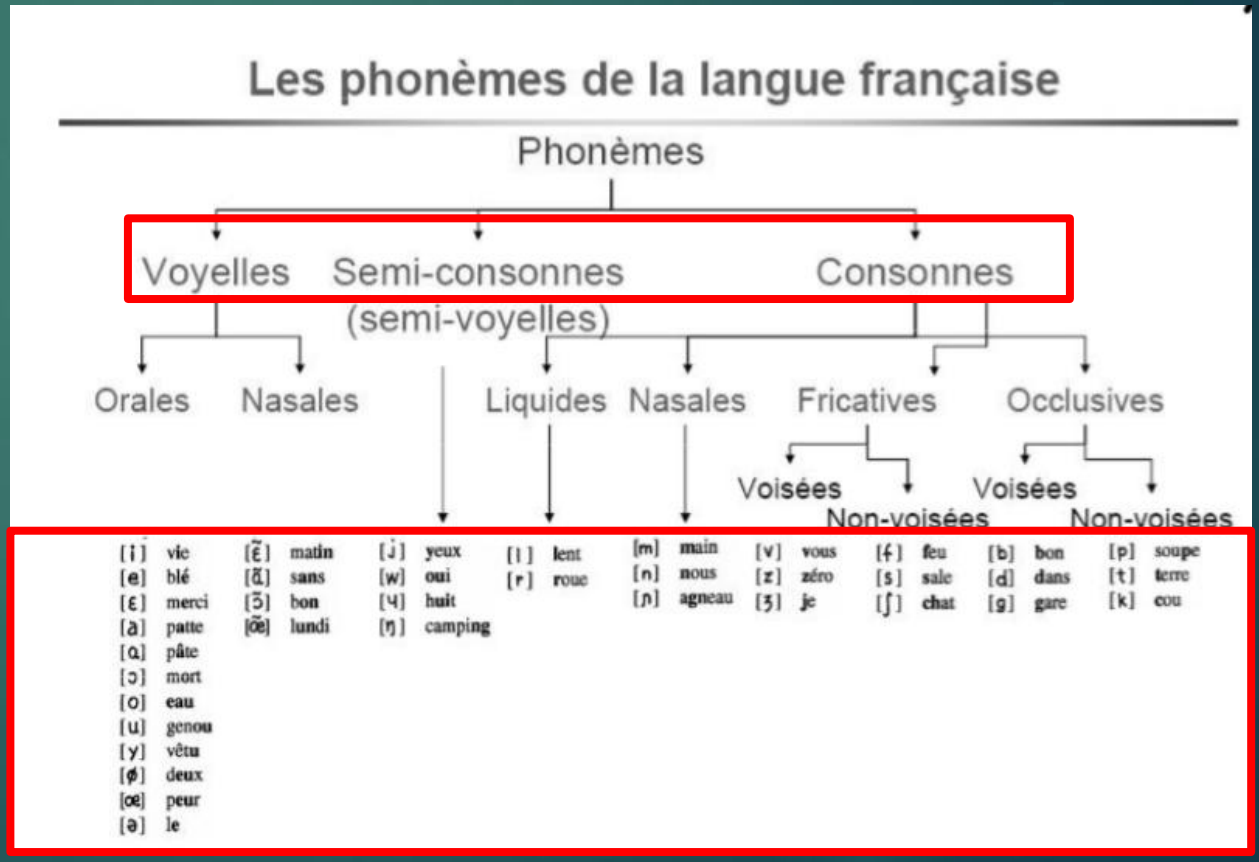
Humain



3)-Étude phonétique: décomposition du signal

- Phonème

	AVANTAGES	INCONVENIENTS
MOT	coarticulation incluse indépendant de la langue reconnaissance simple	grand vocabulaire adaptation au locuteur
SYLLABE	facile à localiser centrée sur la voyelle coarticulation incluse	nombre total élevé frontières difficiles à localiser
DIPHONE	contient une partie de coarticulation	nombre total de l'ordre de 1 000 problèmes de segmentation
PHONEME	nombre total peu élevé codage aisé des mots dans le lexique	très dépendant du contexte difficile à localiser algorithmes complexes pour la segmentation et la reconnaissance
PHONE	peut correspondre à un segment acoustique facile à localiser et à reconnaître	nombre de phones élevé peut dépendre du contexte

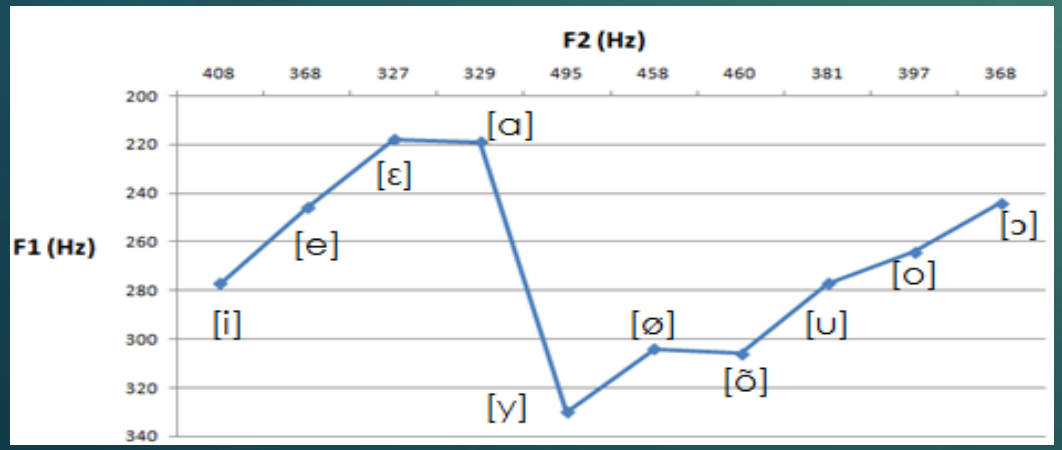


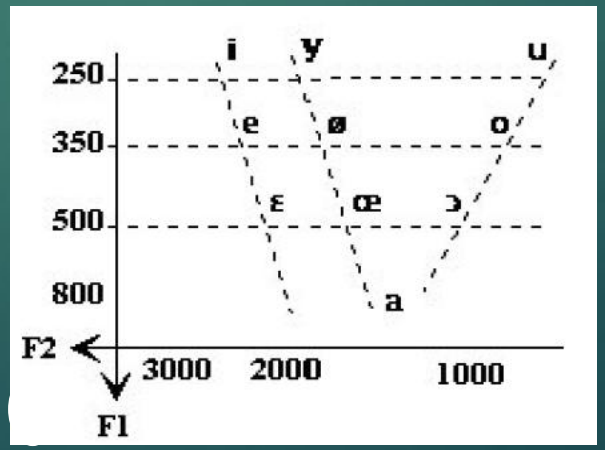
Phonème	Orateur	Fondamental	Amplitude 0	Formant 1	Amplitude 1	Formant 2	Amplitude 2	Formant 3	Amplitude 3
[a]	Lisa	262	0.00221	524	0.0015	784	0.0027	1046	0.00038
[a]	Sarah	241	0.00473	476	0.0003	722	0.0016	983	0.00022
[a]	Nicolas	109	0.004	219	0.007	329	0.002	437	0.003
[a]	Stéphane	179	0.0046	266	0.0041	449	0.0026	539	0.0049
[ɲ]	Lisa	244	0.017	493	0.013	742	0.0018	986	0.001
[ɲ]	Sarah	210	0.009	425	0.0015	673	0.0004	861	0.00015
[ɲ]	Nicolas	167	0.01	337	0.013	525	0.0045	658	0.0029
[ɲ]	Stéphane	97	0.0034	189	0.01	280	0.0035	409	0.0015
[ã]	Lisa	254	0.0022	508	0.0015	763	0.00050	1016	0.00056
[ã]	Sarah	249	0.0062	499	0.00038	748	0.001	998	0.00017
[ã]	Nicolas	122	0.032	245	0.0208	369	0.0097	491.9	0.019
[ã]	Stéphane	101	0.0045	202	0.01	303	0.0029	409	0.003
[e]	Lisa	231	0.00225	463	0.0025	697	0.0002	927	0.000039
[e]	Sarah	245	0.0047	499	0.0045	742	0.0034	1001	0.00011
[e]	Nicolas	122	0.021	246	0.0198	368	0.021	489	0.0026
[e]	Stéphane	113	0.004	228	0.009	336	0.0076	449	0.0034
[ɔ]	Lisa	279	0.00338	558	0.0022	837	0.0036	1115	0.0005
[ɔ]	Sarah	257	0.0036	510	0.0024	691	0.00169	1468	0.0006
[ɔ]	Nicolas	121.5	0.020	244	0.0226	368	0.0012	491	0.052
[ɔ]	Stéphane	87	0.00239	177	0.003	267	0.00239	539	0.00088

Extrait du tableau contenant les valeurs mesurées sur les spectres

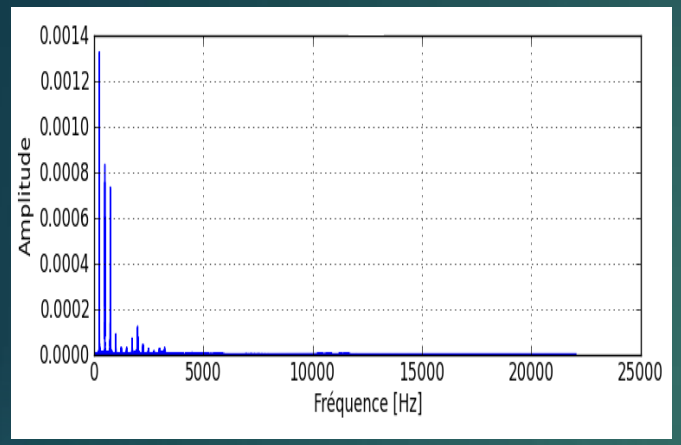
Individu	Âge
Lisa	19 ans
Sarah	20 ans
Nicolas	20 ans
Stéphane	50 ans

132 échantillons

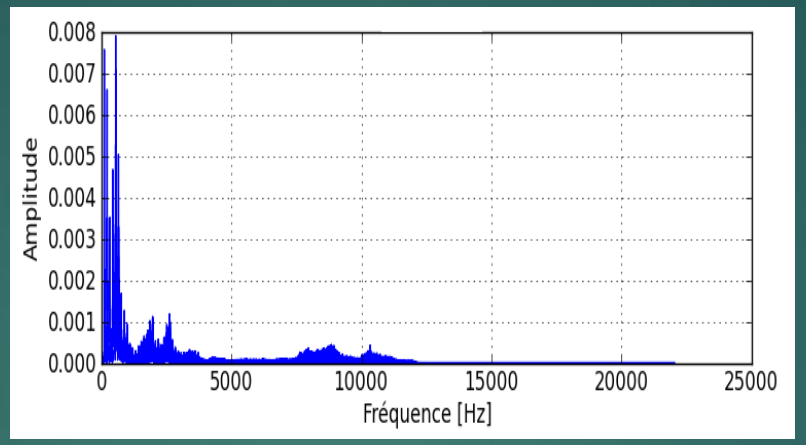




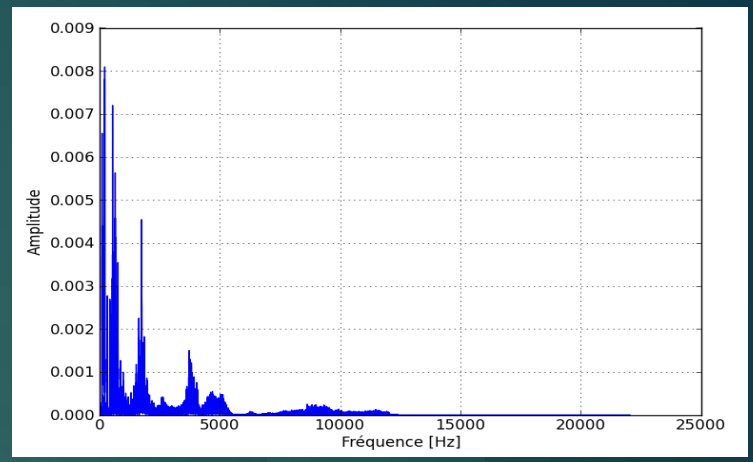
- Plus d'harmoniques chez l'homme Phonème [ɛ] (merci)



Lisa



Nicolas



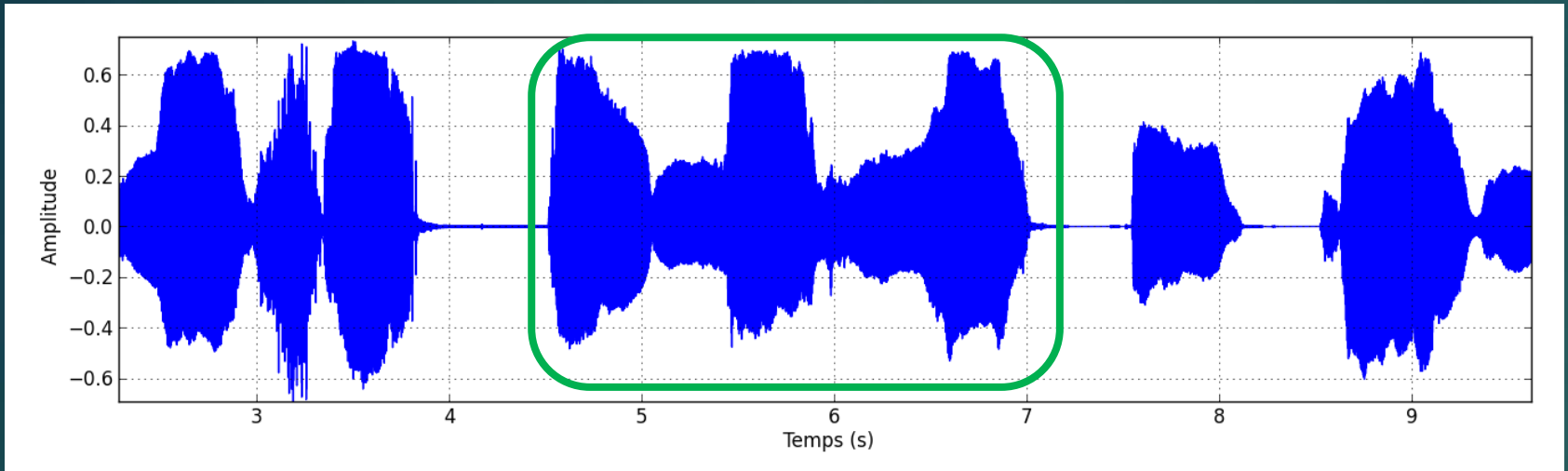
Stéphane

- Fondamental et formants féminins plus haut

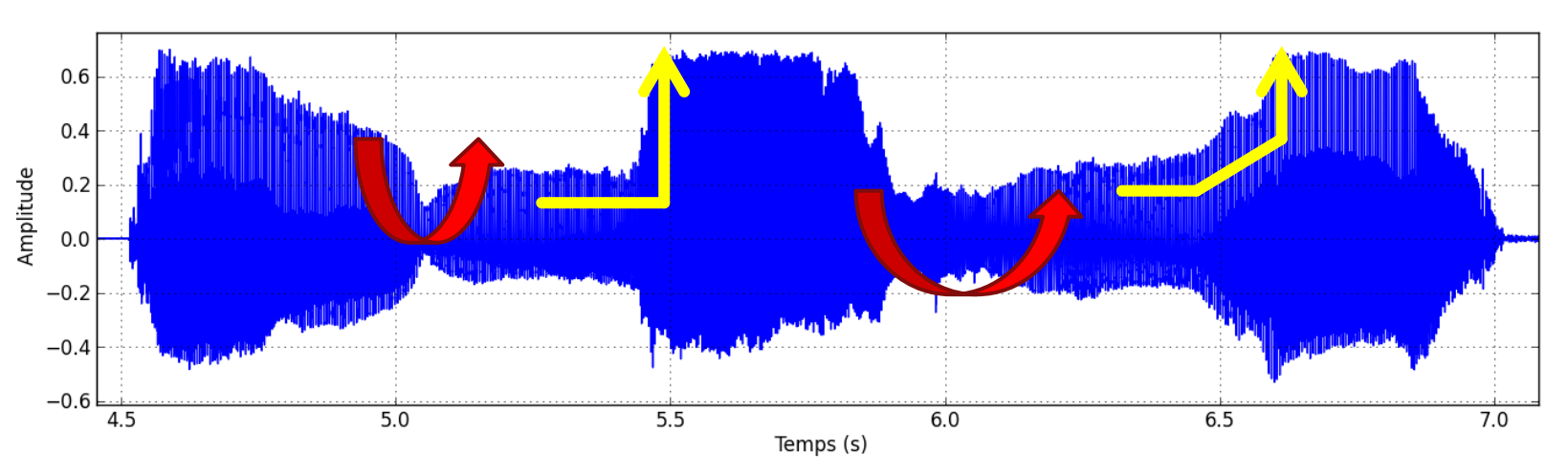
  Accroissement de l'écart

[ʁ]	Lisa	244	0.017	493	0.013	742	0.0018	986	0.001
[ʁ]	Sarah	210	0.009	425	0.0015	673	0.0004	861	0.00015
[ʁ]	Nicolas	167	0.01	337	0.013	525	0.0045	658	0.0029
[ʁ]	Stéphane	97	0.0034	189	0.01	280	0.0035	409	0.0015

« Le chat et l'agneau ont tous les deux quatre pattes »



« et l' a gn eau »



- Transitions

Transition voyelle-consonne

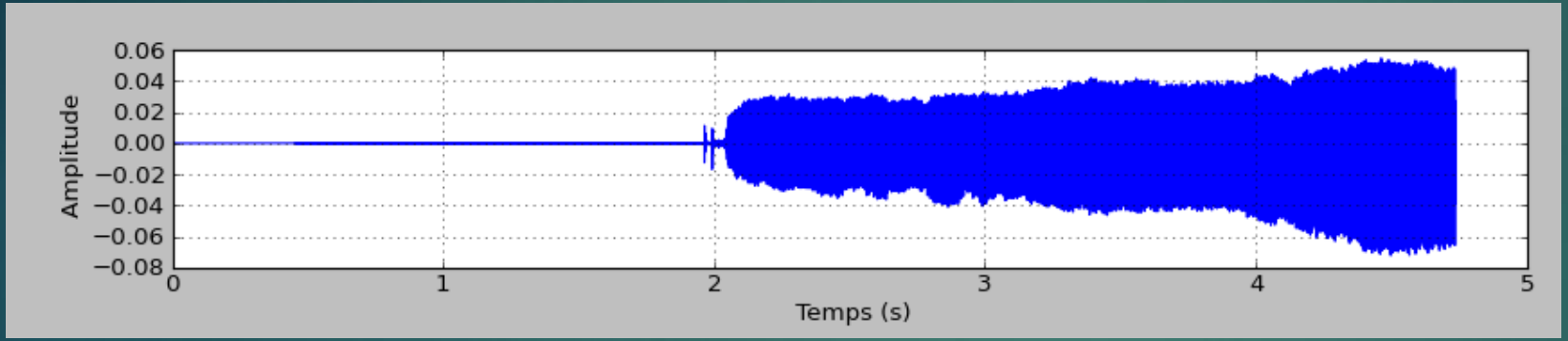
Transition consonne-voyelle

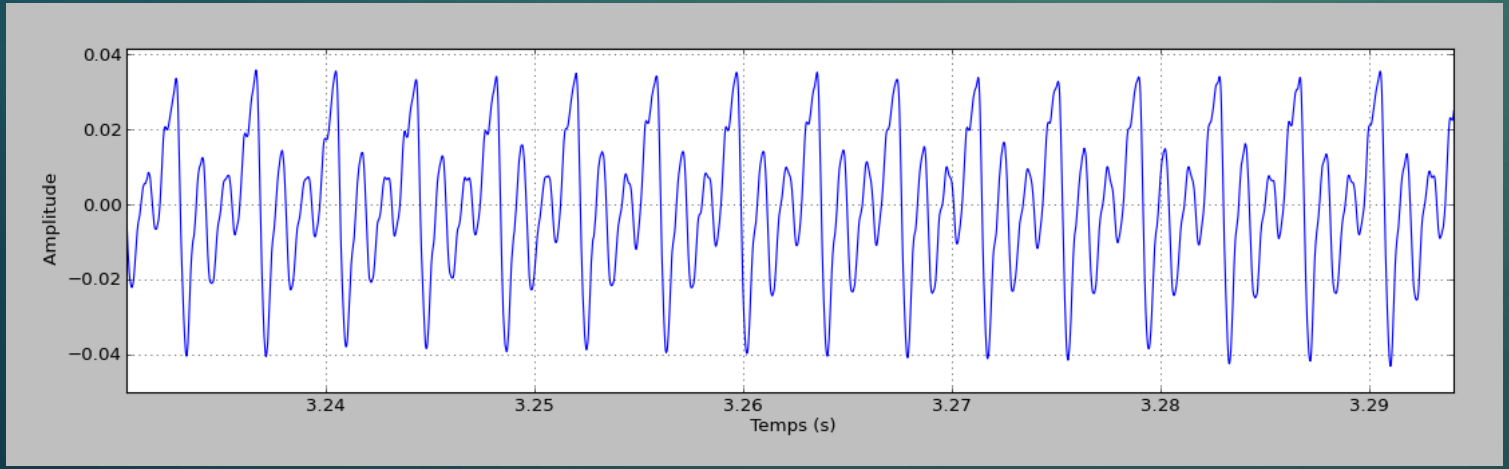
II. Synthèse

1)-Synthèse d'une voyelle

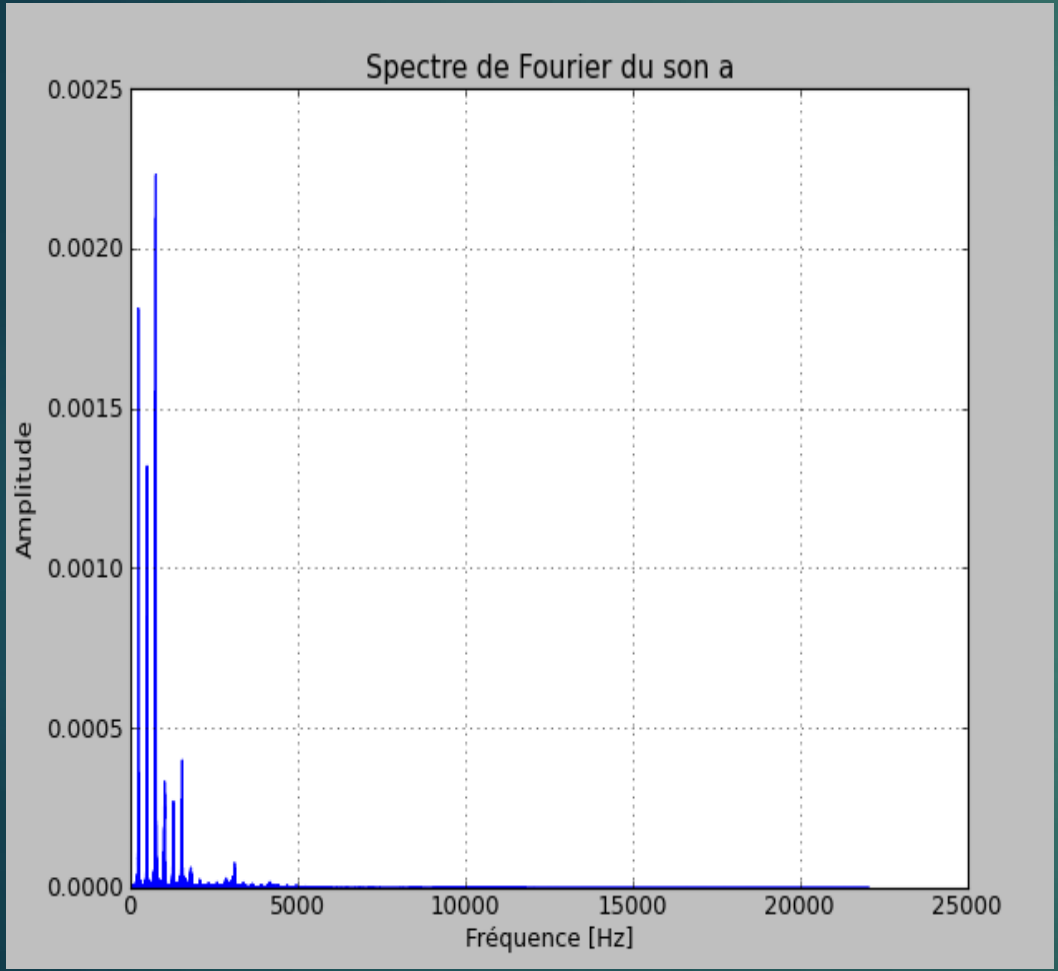
- Principe de Fourier :

Tout signal périodique peut se décomposer en somme de signaux sinusoïdaux



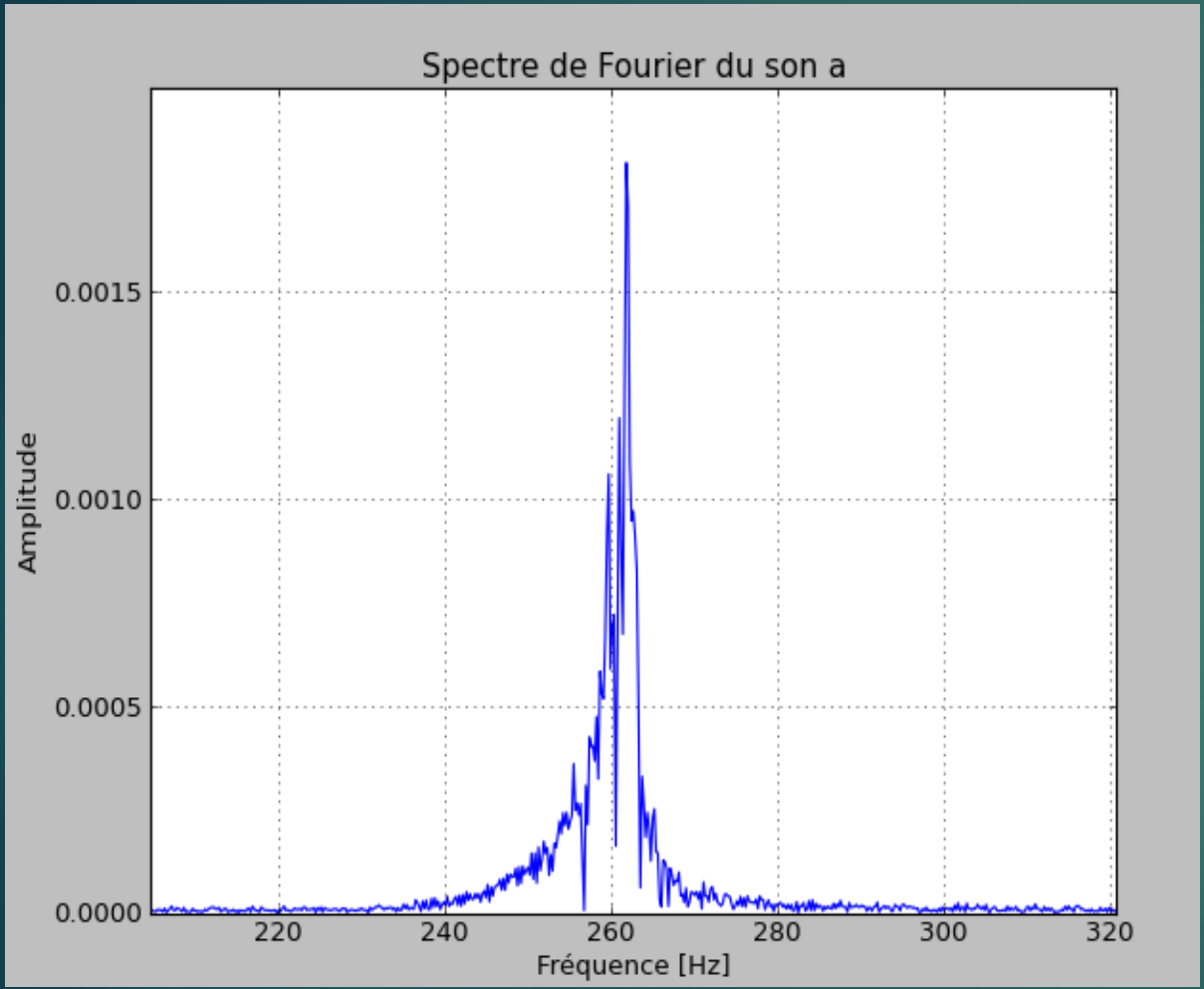


Synthèse additive



- Amplitudes égales
  → Absence d'humanité
- Amplitudes avec rapport de proportionnalité
  → Différence notable
- Amplitudes précises
  → Diffère du cas précédent
- Influence de la phase
  → Inaudible

Après un zoom…



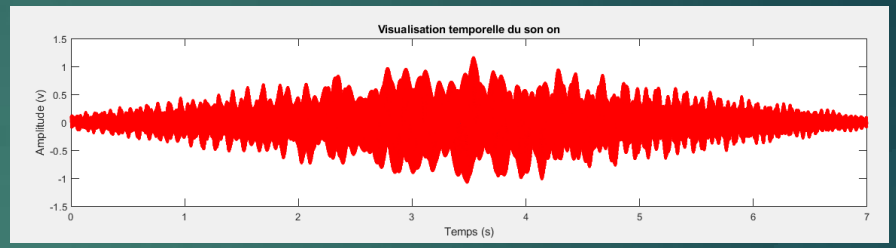
- Pas une mais plusieurs fréquences autour d'un pic
- Fréquence principale + quelques fréquences alentours
  → phénomène de battement
- Plages de fréquences
  → amélioration la plus importante
- Plages d'amplitudes (ascendantes puis descendantes)
  → Résultat proche du son réel

Difficultés rencontrées:

- Lecture simultanée des listes
  - → montée vers l'aigu
  - → variation du niveau sonore
  - ➔ Plages plus étroites

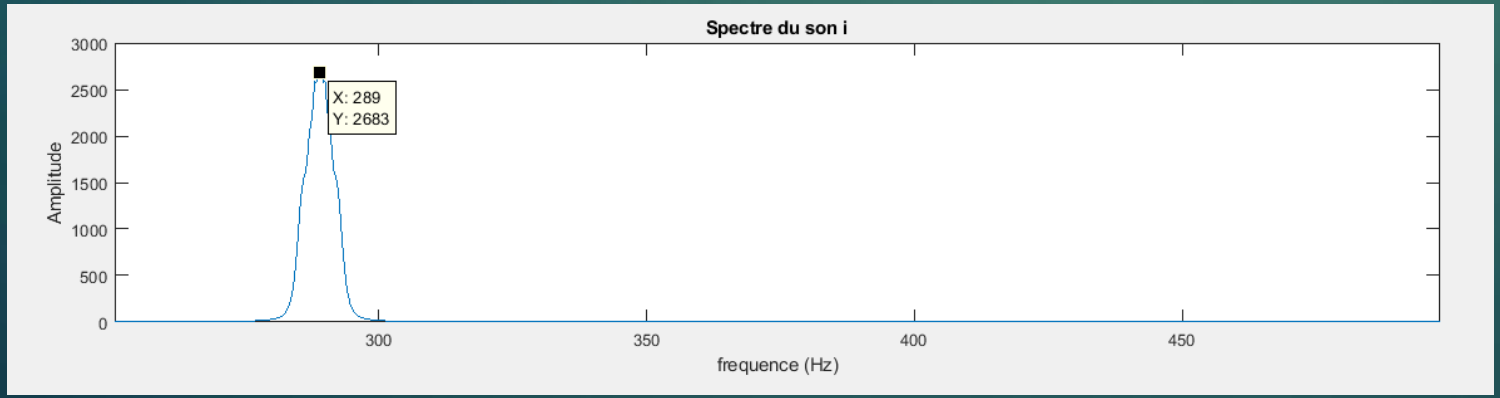
listes de même longueur

```
y0   = A0.*sin(2.*pi.*F00.*t);
```



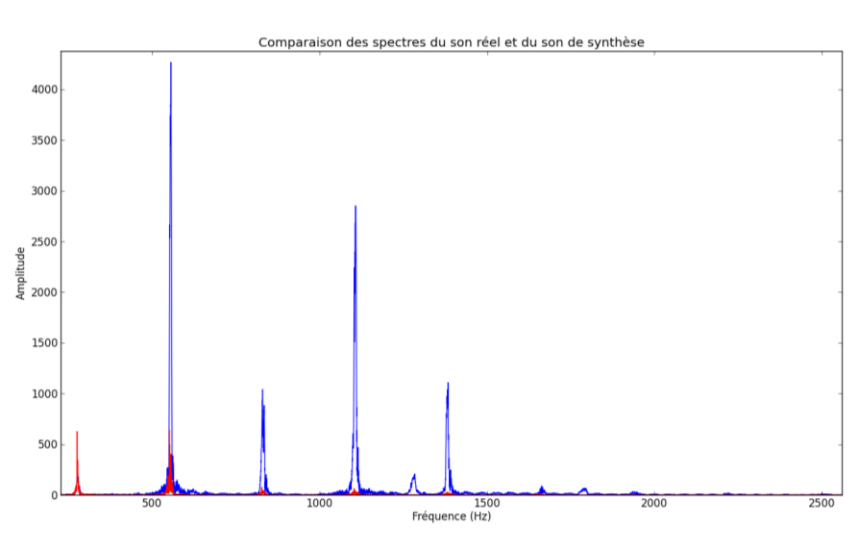
- Plages de fréquences élargies par Matlab

```
F00=284:(5/132300):289;
```

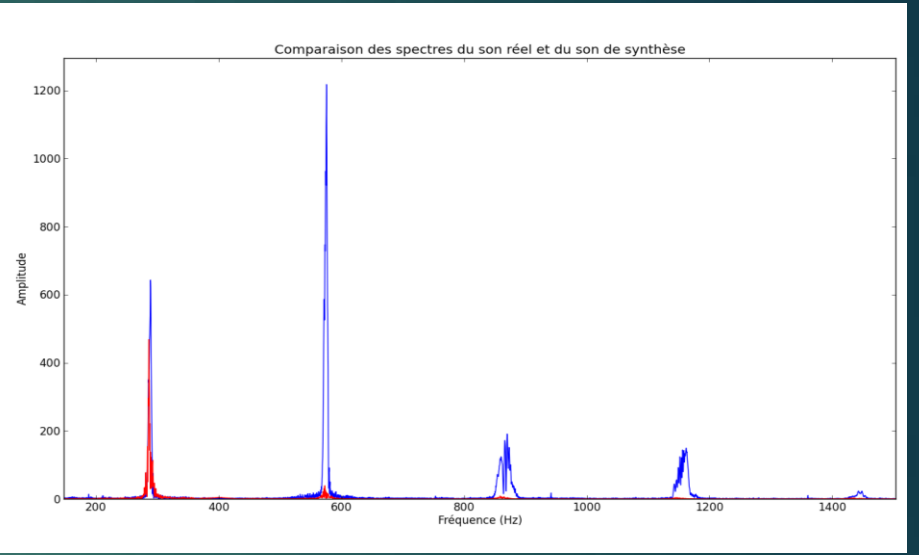


Résultats obtenus

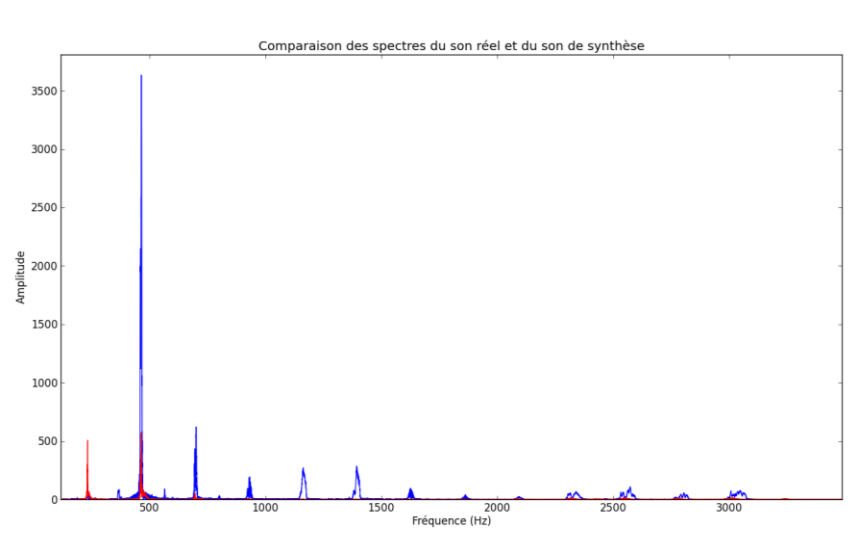
[o]



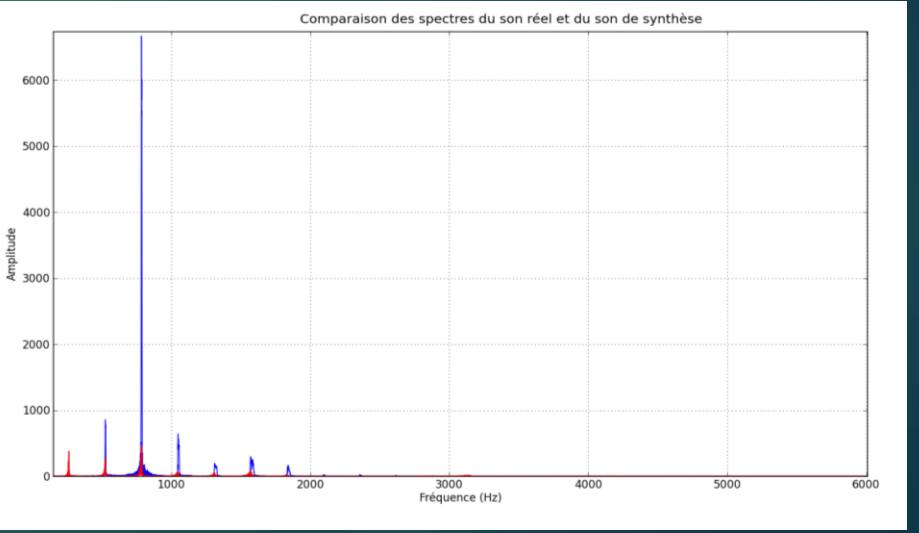
[i]

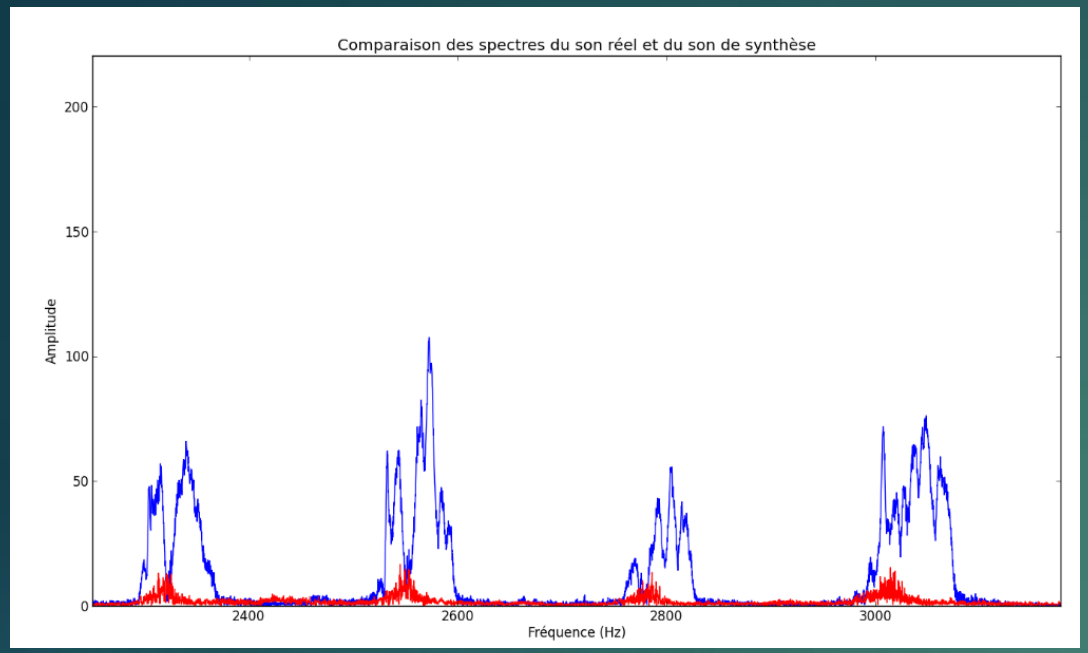


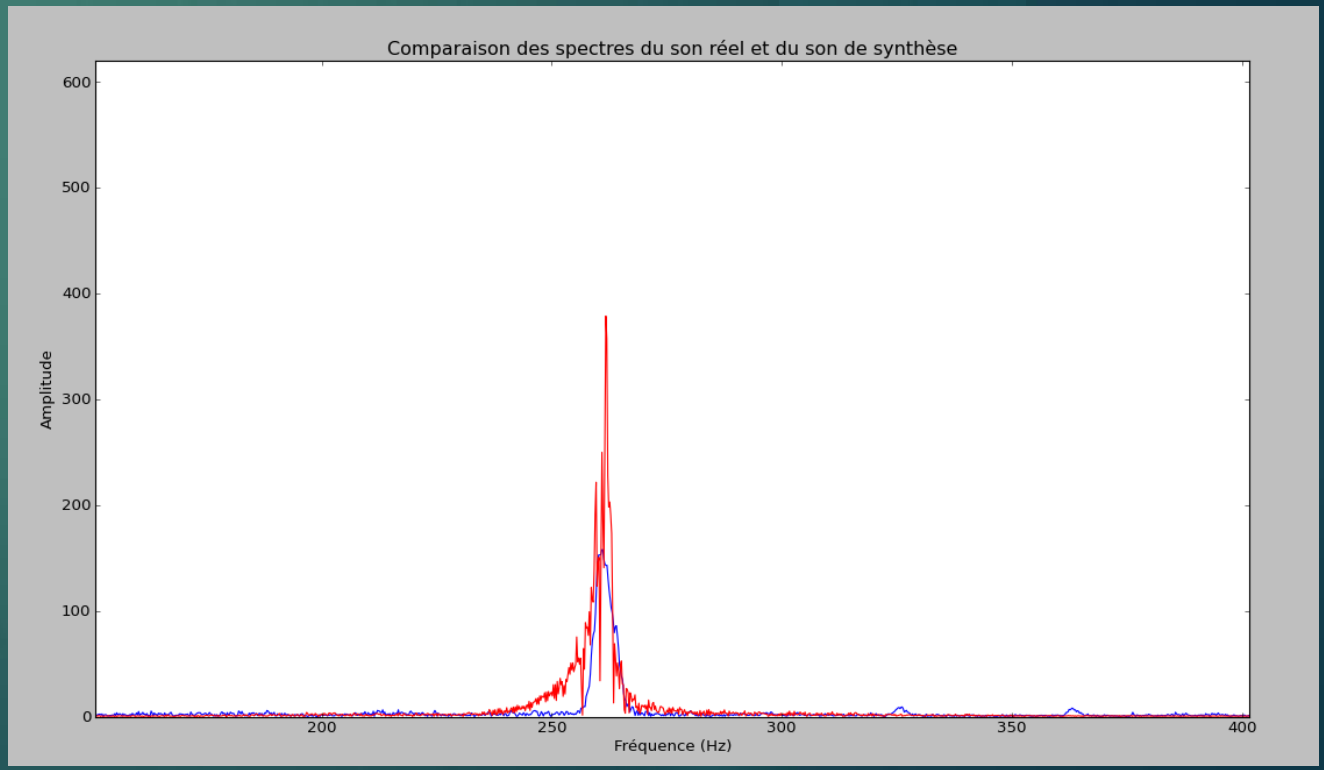
[e]



[a]

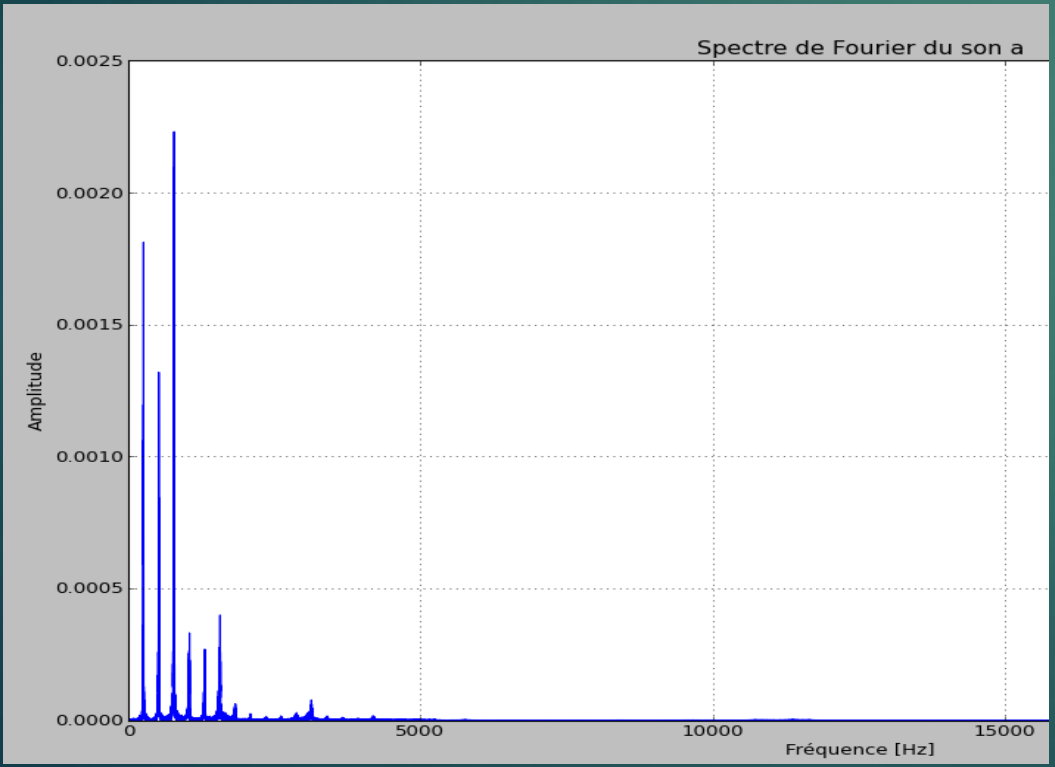


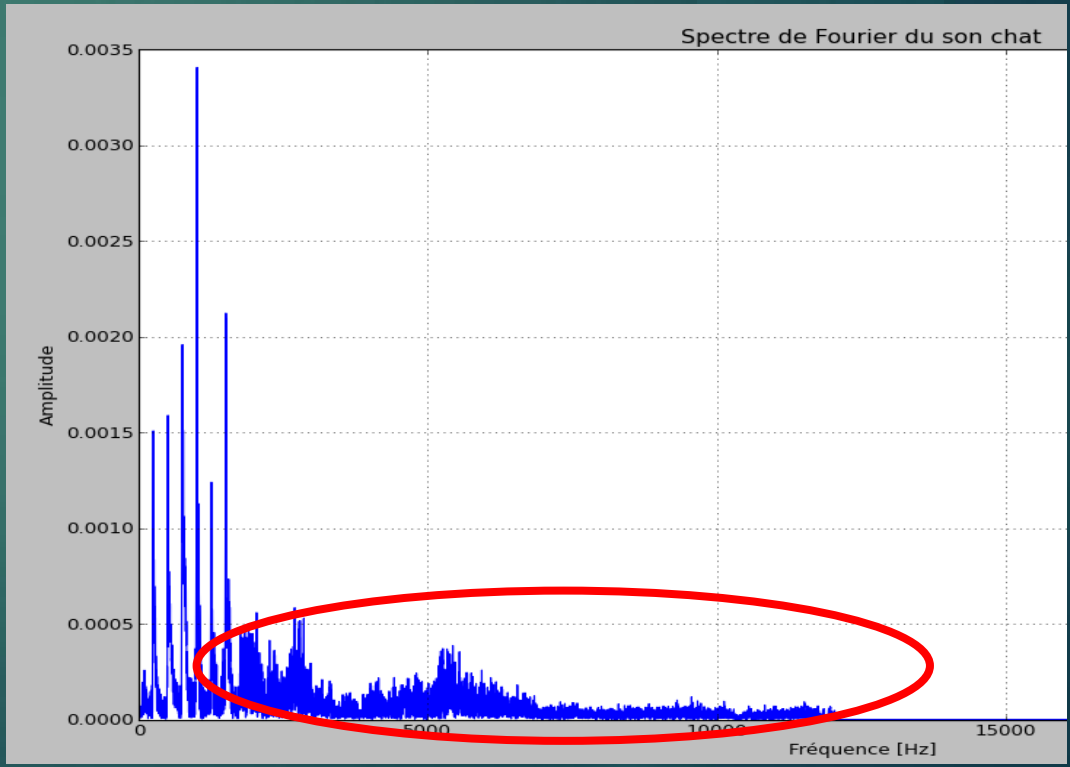

Comparaison des spectres du son réel et du son de synthèse
Amplitude
Fréquence (Hz)
0
50
100
150
200
2400
2600
2800
3000


Comparaison des spectres du son réel et du son de synthèse
Amplitude
Fréquence (Hz)
0
100
200
300
400
500
600
200
250
300
350
400

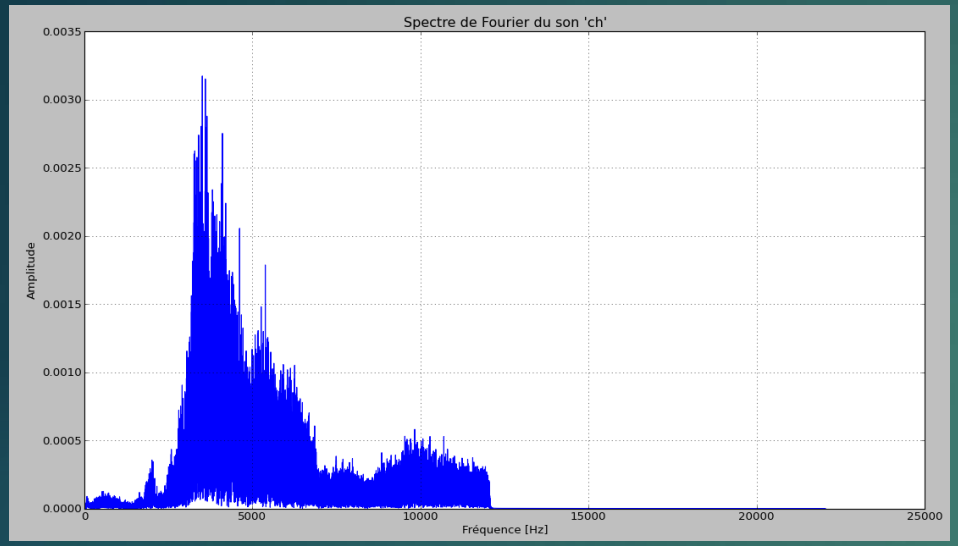
2)- Synthèse d'une consonne

Spectre	Fondamental (Hz)	Formant 1 (Hz)	Formant 2 (Hz)	Formant 3 (Hz)	Formant 4 (Hz)	Formant 5 (Hz)
a	262	524	784	1046	1308	1570
chat	255	505	756	1009	1259	1511

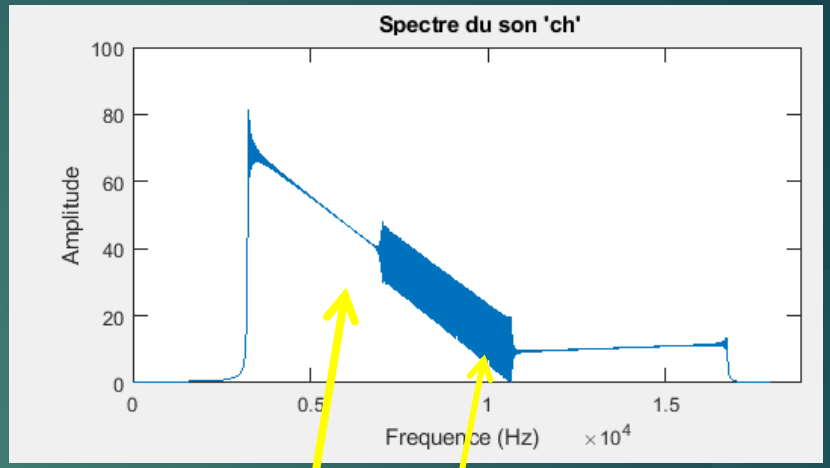


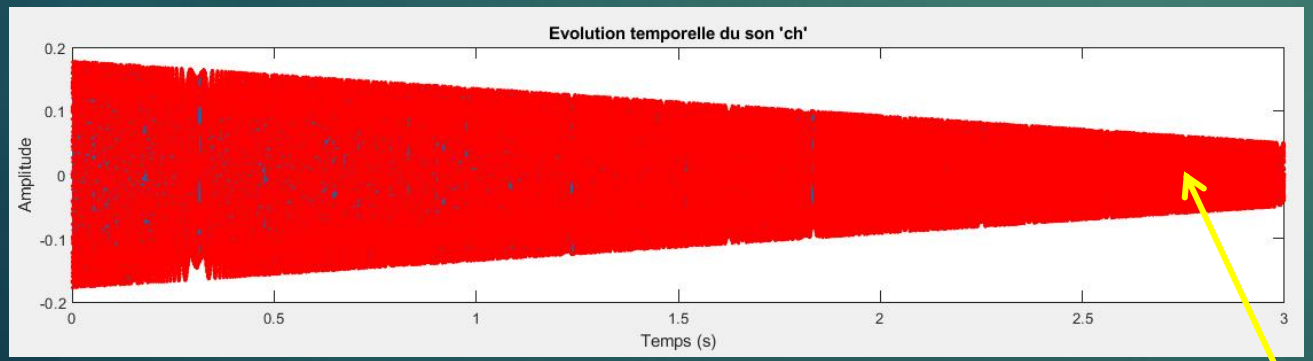


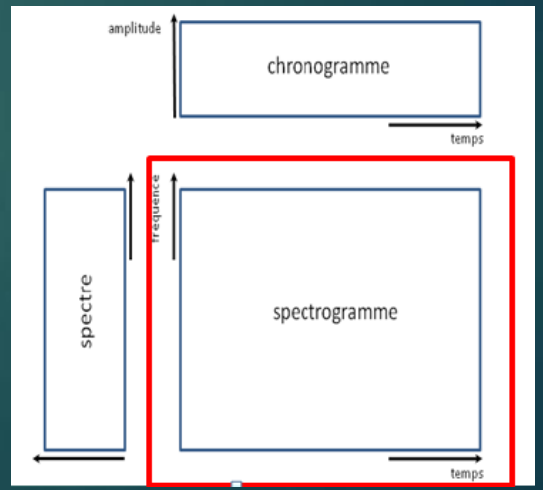
Spectre désiré
(réellement prononcé)



Spectre obtenu
(son synthétisé)



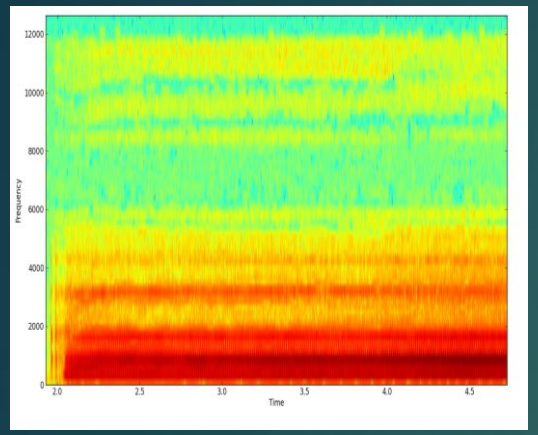




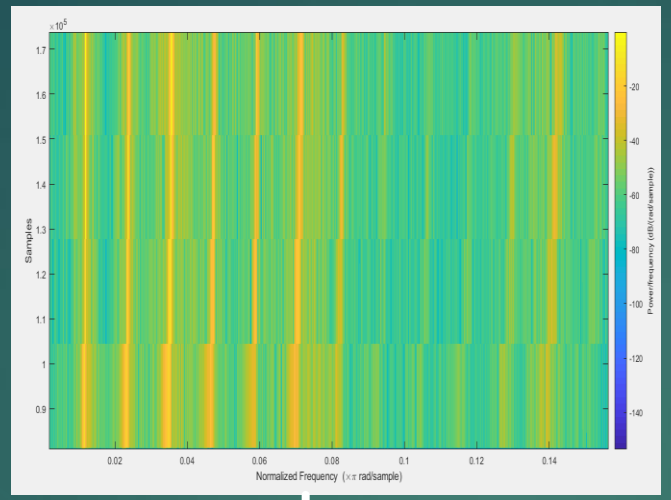
Difficultés

Alternative

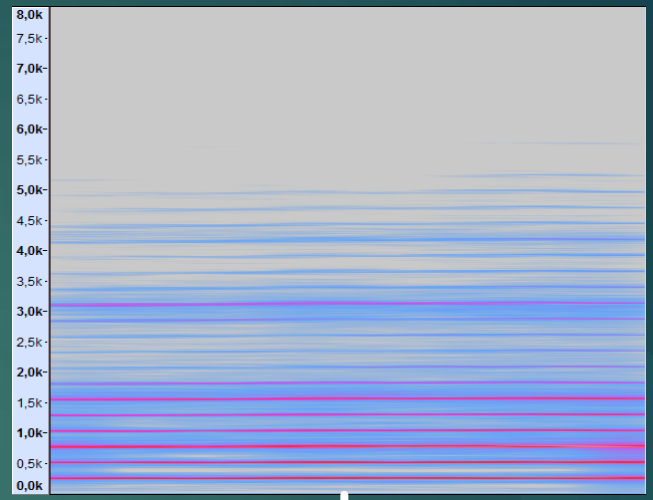
Spectrogramme Python



Spectrogramme Matlab

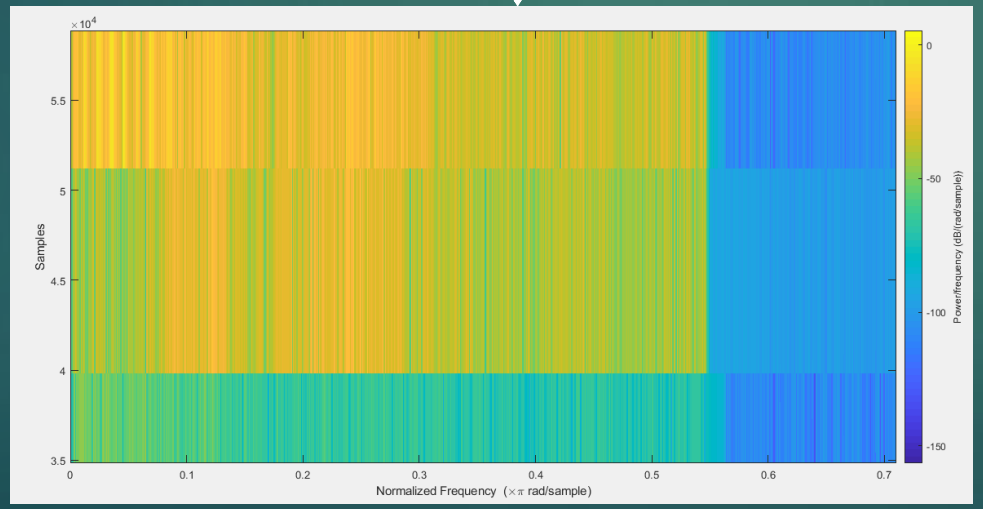


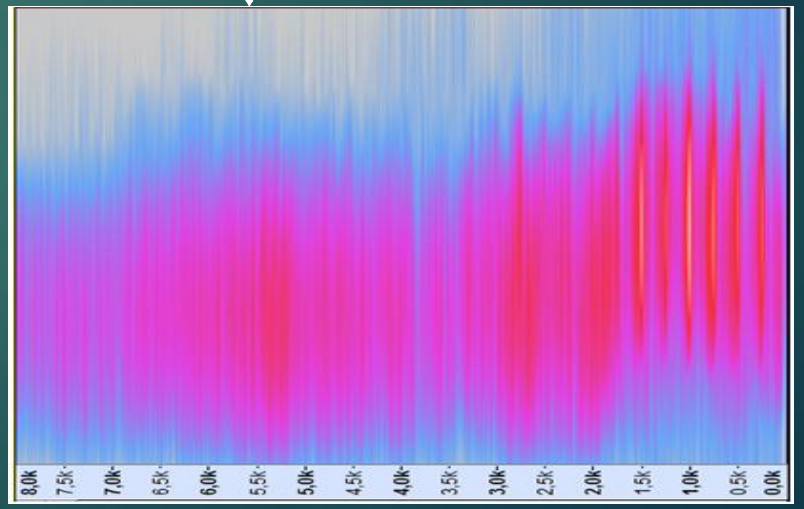
Spectrogramme Audacity



Phonème [a]

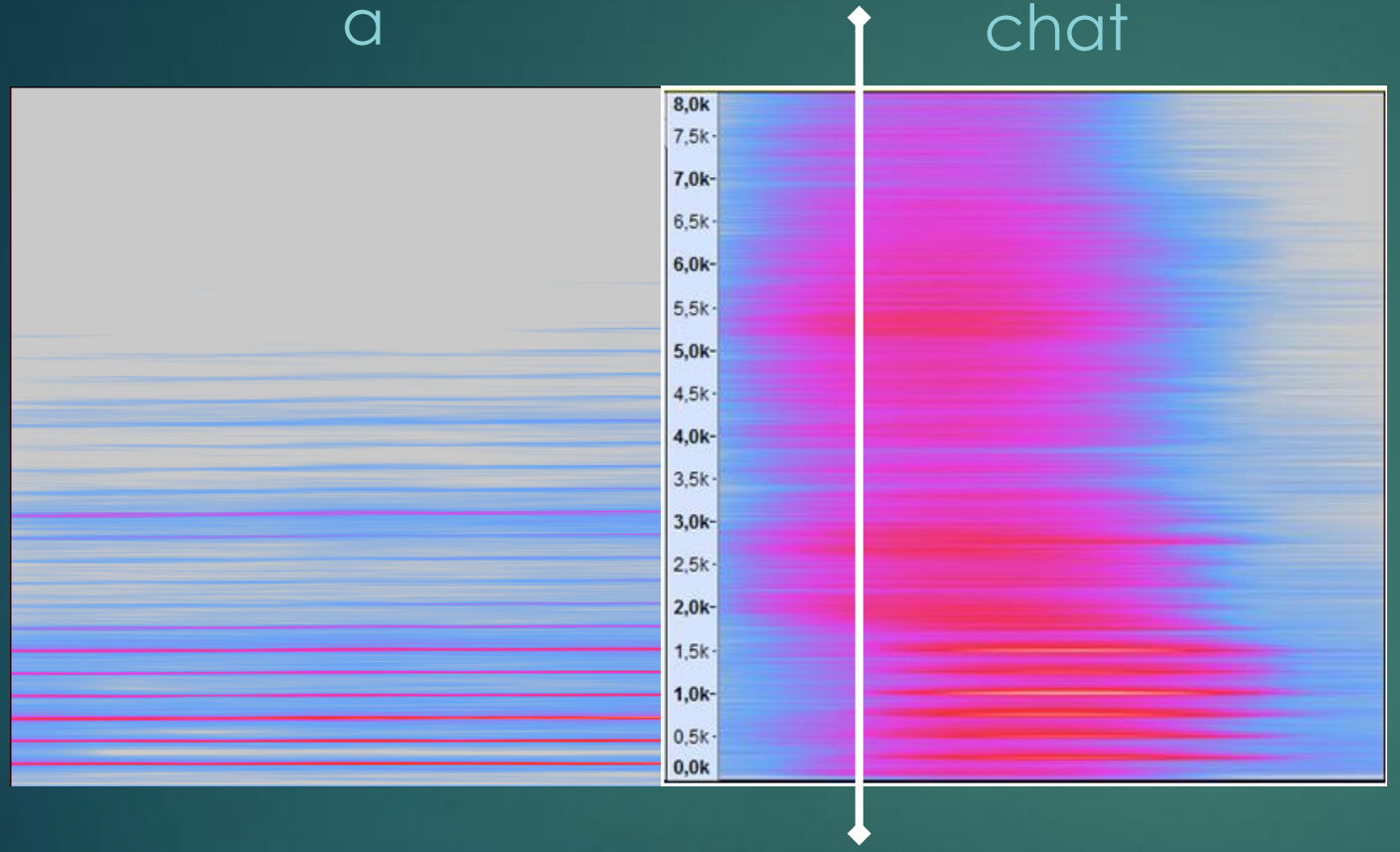
Phonème [ʃ]
Chat





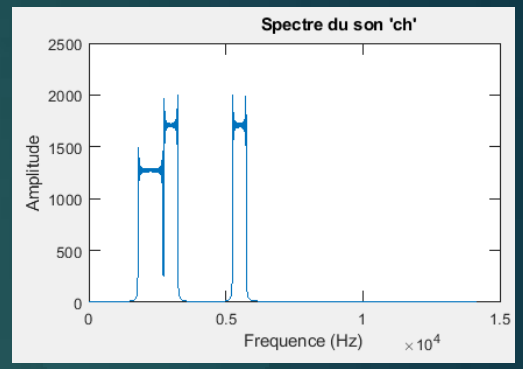
Méthode d'identification des fréquences associées aux consonnes

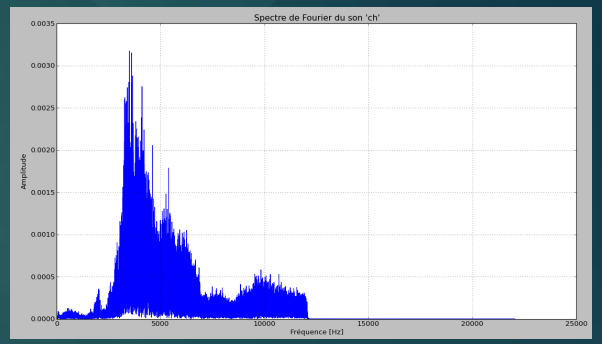
a chat



[∫]

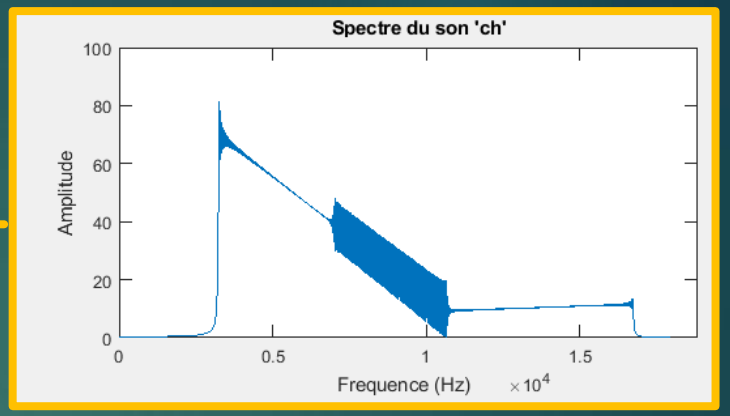
[a]

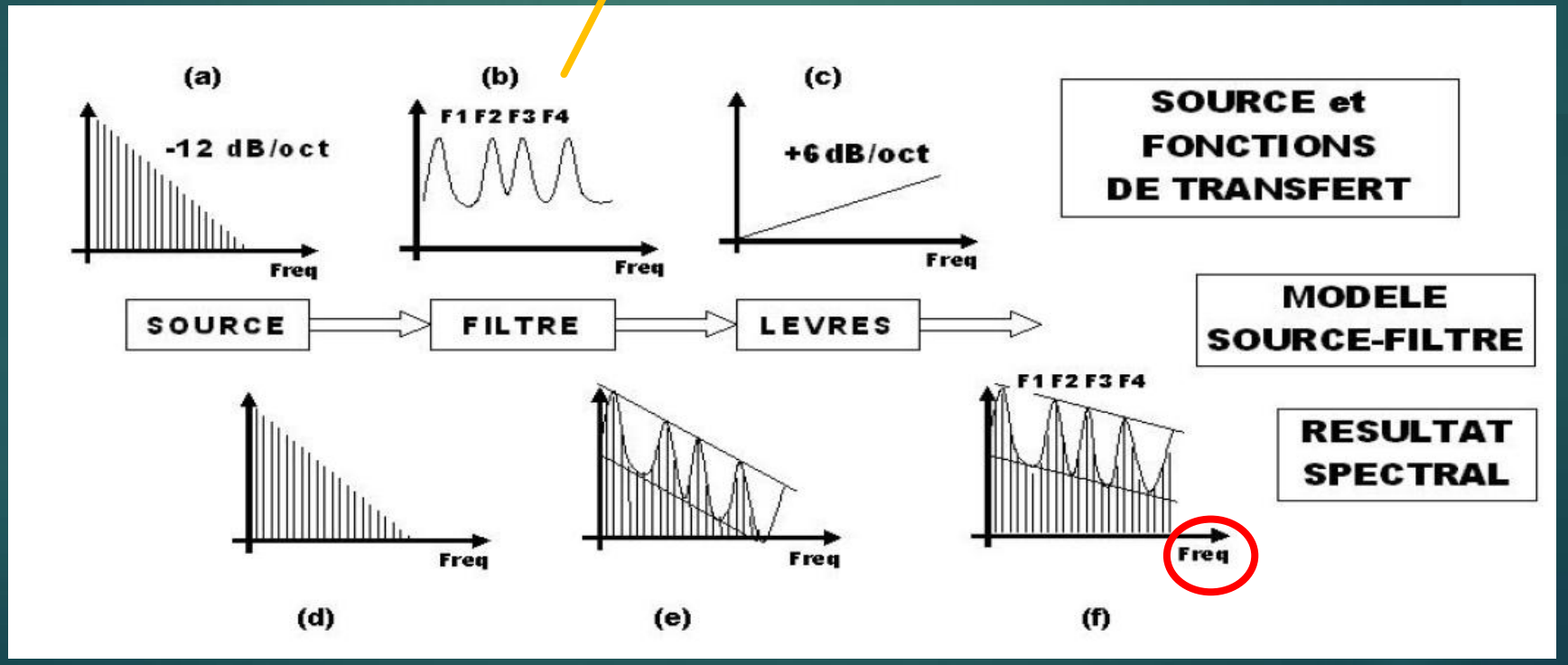




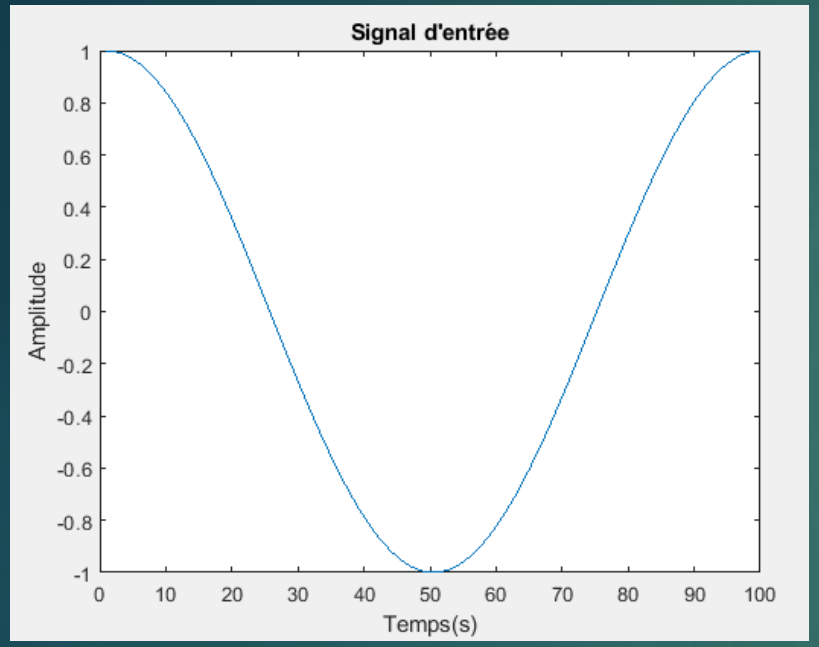
→ Synthèse additive incompatible avec les consonnes

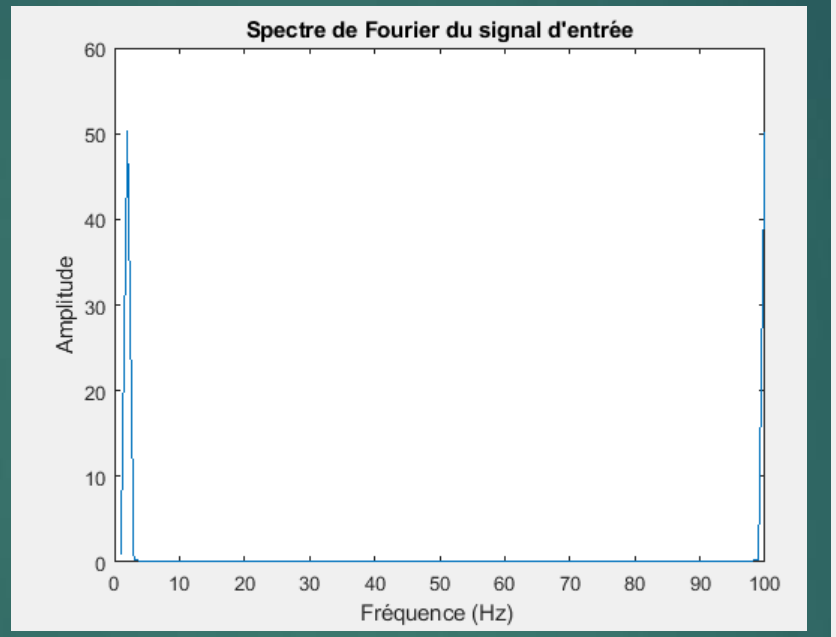
La méthode soustractive

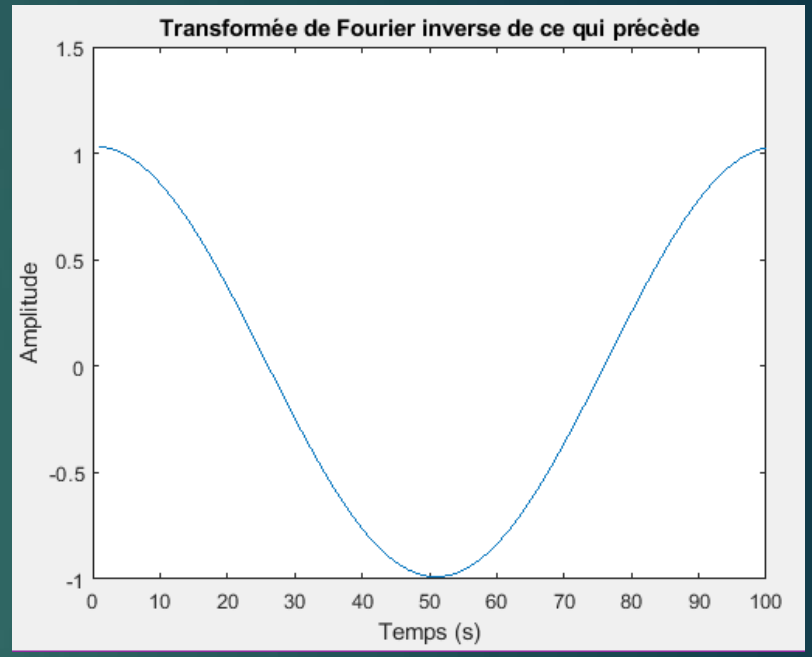




➢ Passage fréquentiel→temporel : ifft



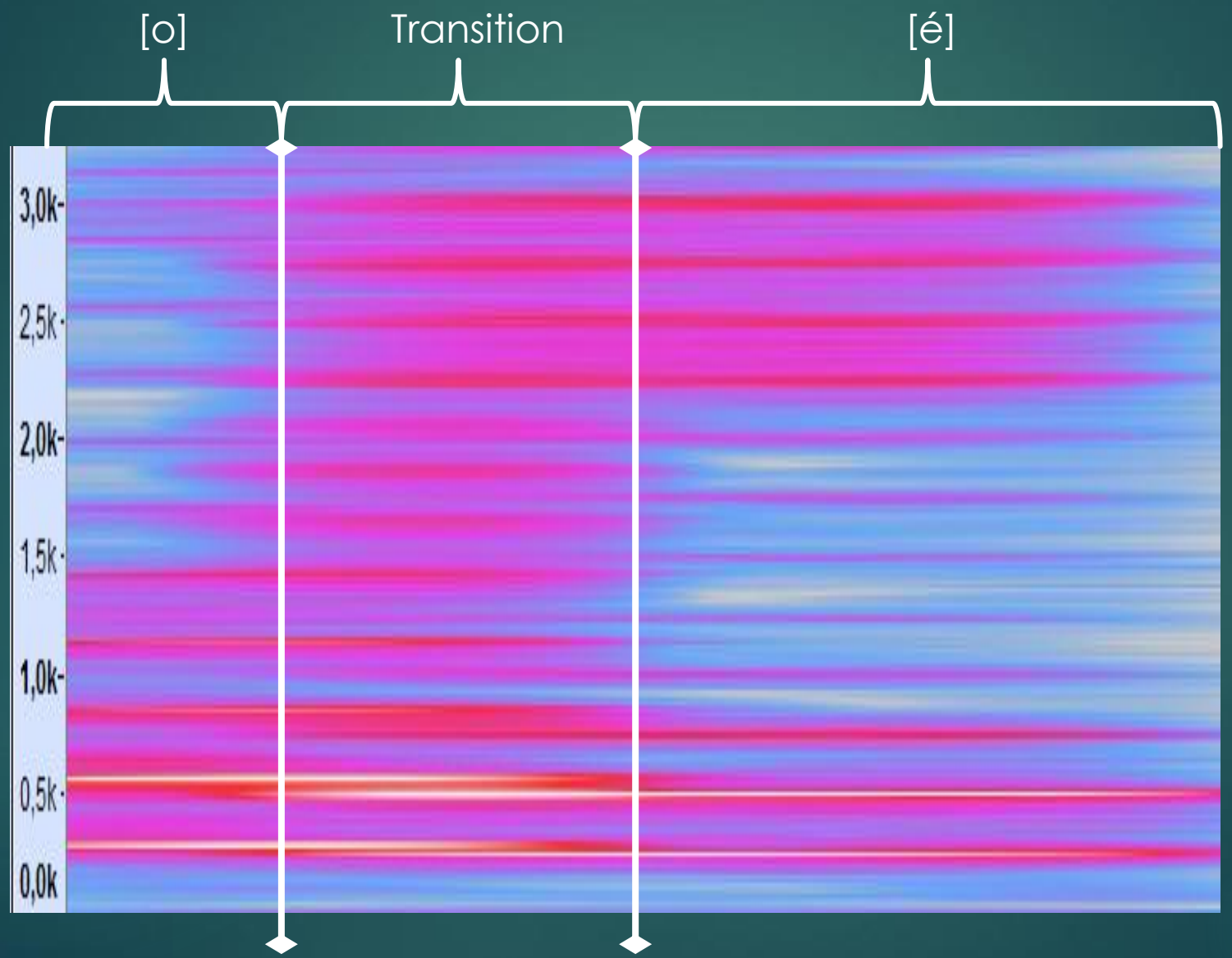




Domaine temporel → FFT → Domaine fréquentiel → IFFT → Domaine temporel

III. Concaténation

1)-Concaténation de phonèmes



Concaténation de deux voyelles

- Fréquences communes
- Fréquences disparues
- Fréquences apparues

Résultats obtenus

Sans transition

Avec transition

2)-Algorithme d'identification des phonèmes à concaténer

Objectif: comprendre la méthode du décodage accoustico-phonétique

Le décodage acoustico-phonétique

1- Segmentation

2-Nettoyage du texte

3- Etiquetage

Exemple: Je suis ton père.
[Sujet][Verbe][Adjectif Possessif][Nom Masculin]

4- Phonétisation

Exemple: Je suis ton père -> /ʒə sɥi tɔ̃ pɛʁ/

Segmentation	Nettoyage	Etiquetage	Phonétisation
Dans	dans	dans [Préposition]	da~ [Préposition]
la	la	la [Déterminant Féminin Singulier]	la [Déterminant Féminin Singulier]
coll.	collection	collection [Nom Féminin Singulier]	k0lEksjo~ [Nom Féminin Singulier]
Dupond	Dupond	dupond [Nom Propre]	dypo~ [Nom Propre]
l'	l'	l [Déterminant Masculin Singulier]	l [Déterminant Masculin Singulier]
ouvrage	ouvrage	ouvrage [Nom Masculin Singulier]	uvRaZ [Nom Masculin Singulier]
Pourquoi	pourquoi	pourquoi [Adverbe]	puRkwa [Adverbe]
Moi	moi	moi [Pronom Personnel]	mwa [Pronom Personnel]
?	?	? [Point d'interrogation]	_ [Silence]
à	à	a [Préposition]	a [Préposition]
pour	pour	pour [Préposition]	puR [Préposition]
code	code	code [Nom Masculin Singulier]	kOd [Nom Masculin Singulier]
IV	quatre	quatre [Chiffre]	katR@ [Chiffre]
12	douze	douze [Chiffre]	duz [Chiffre]
14	quatorze	quatorze [Chiffre]	katORz@ [Chiffre]

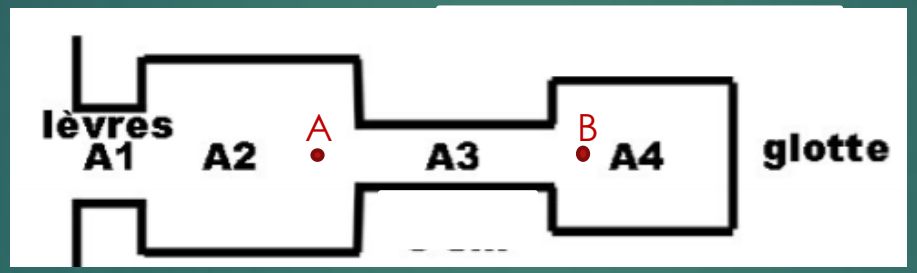
Exemple d'application du décodage acoustico-phonétique

IV. Conclusion

Explications physiques: Effet VENTURI

Hypothèse: l'écoulement de l'air dans le conduit vocal est considéré comme parfait, stationnaire, homogène, et incompressible.

Modélisation du conduit vocal:



Avec:- **la section des lèvres (A1)**

- **la cavité frontale (A2)**
- **l'articulation linguale (A3)**
- **la cavité arrière (A4)**

En considérant par exemple la partie de A4 vers A2 on a:

En appliquant la relation de Bernoulli sur une ligne de courant allant de A vers B:

$$P_A + \mu g z_A + \frac{1}{2}\mu {v_A}^2 = P_B + \mu g z_B + \frac{1}{2}\mu {v_B}^2$$

Or A et B étant à la même altitude on obtient:

$$P_B = P_A + \frac{1}{2}\mu({v_A}^2 - {v_B}^2)$$

Comme $S_B < S_A$ alors $v_B > v_A$, par conséquent $P_B < P_A$. Il y a diminution de la pression.

Interprétation physique:

L'air provenant de la trachée tente de sortir et accélère. Sa vitesse augmente et la pression au niveau des cordes vocales diminue, ce qui permet leur ouverture. Puis elles se referment et la situation se répète.

Etapes majeures du code

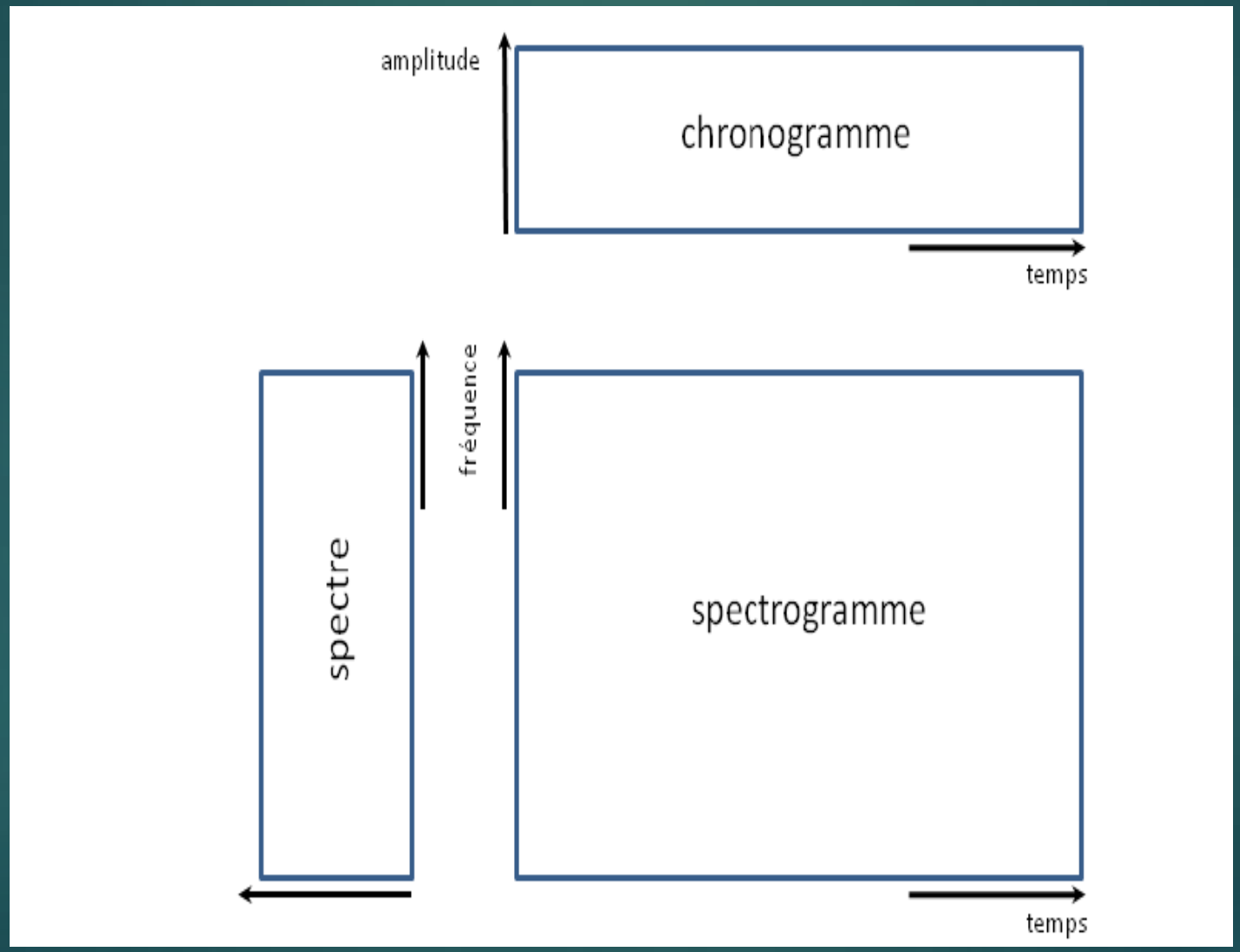

```
Fs= 44100;
tmax=7;
Nbdt=floor((tmax*132301)/3);
T= 0:1/Fs:tmax;
t=[0 T];
```

```
F00=254:(2/(Nbdt-1)):256;
```

```
A0_m=0.000337:((0.00378-0.000337)/mt):0.00378;
A0_d=0.00378:((0.000333-0.00378)/mt):0.000333;
A0=[A0_m,A0_d];
```

```
y0   = A0.*sin(2.*pi.*F00.*t);
```

Les différentes représentations possibles des paramètres d'un signal:



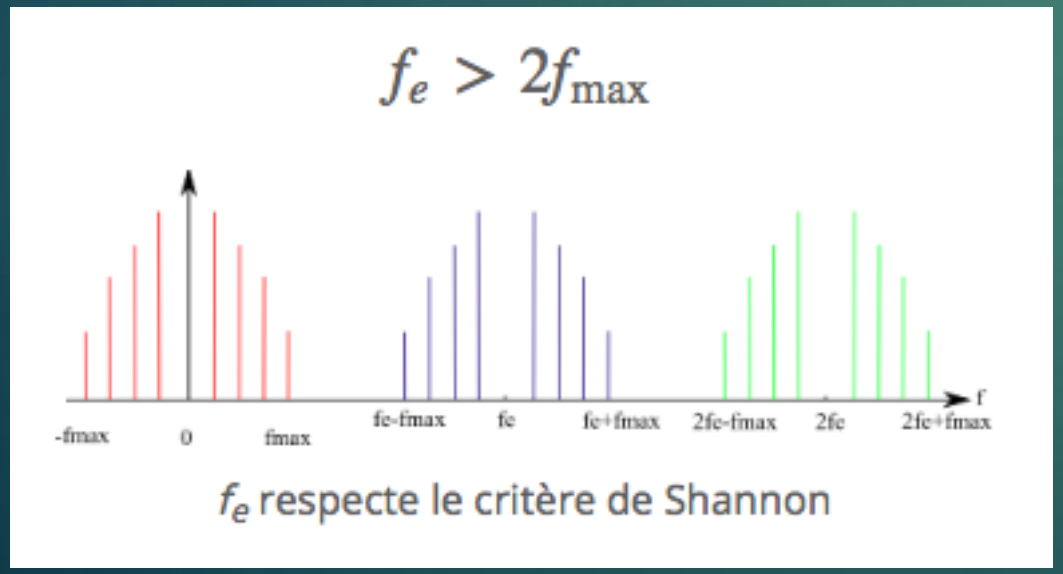
Le théorème de Shannon :

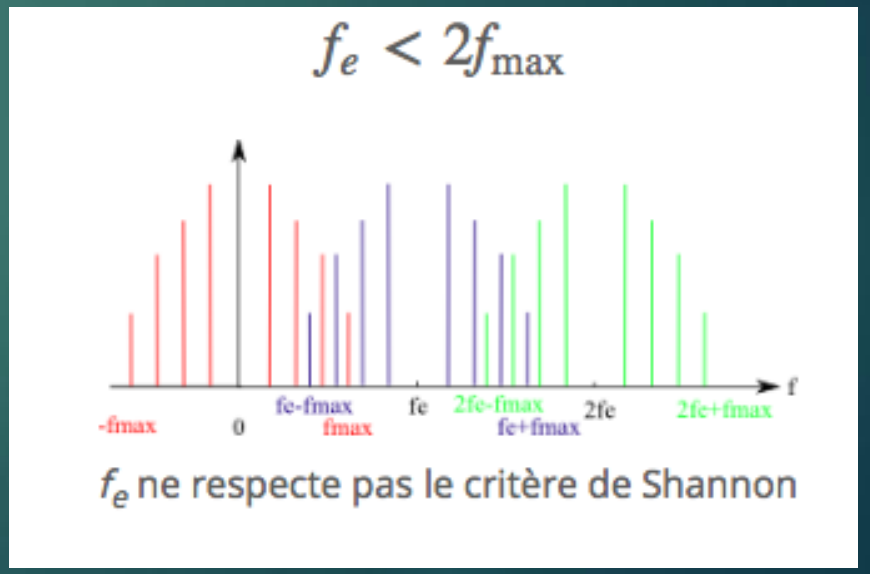
Permet de connaître la fréquence d'échantillonnage à choisir pour un signal donné :

Pour reconstruire un signal de sortie de manière fidèle au signal d'entrée, il faut choisir une fréquence d'échantillonnage au moins deux fois supérieure à la fréquence maximale contenue dans le signal d'entrée.

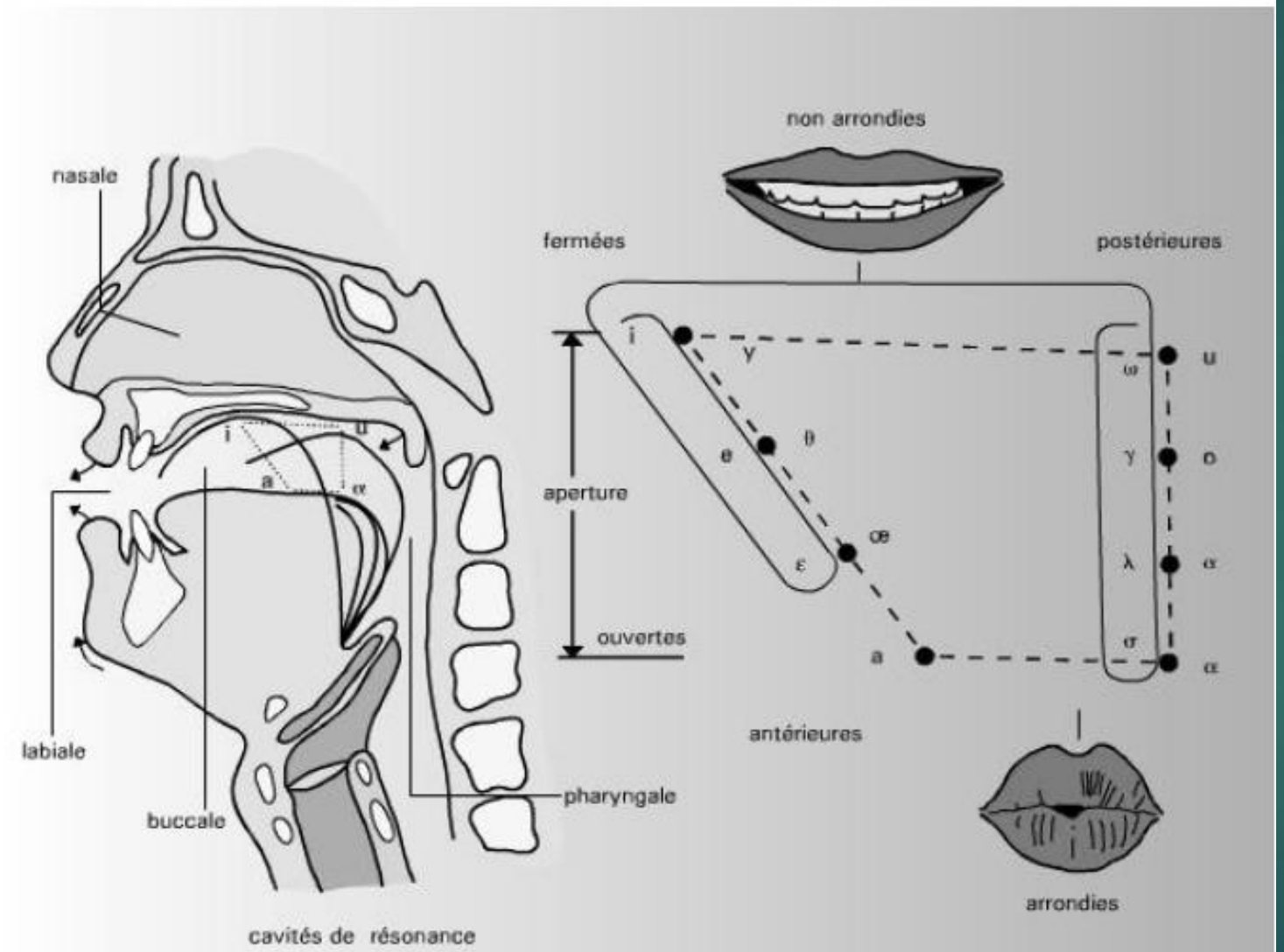
$$f_e > 2f_{max}$$

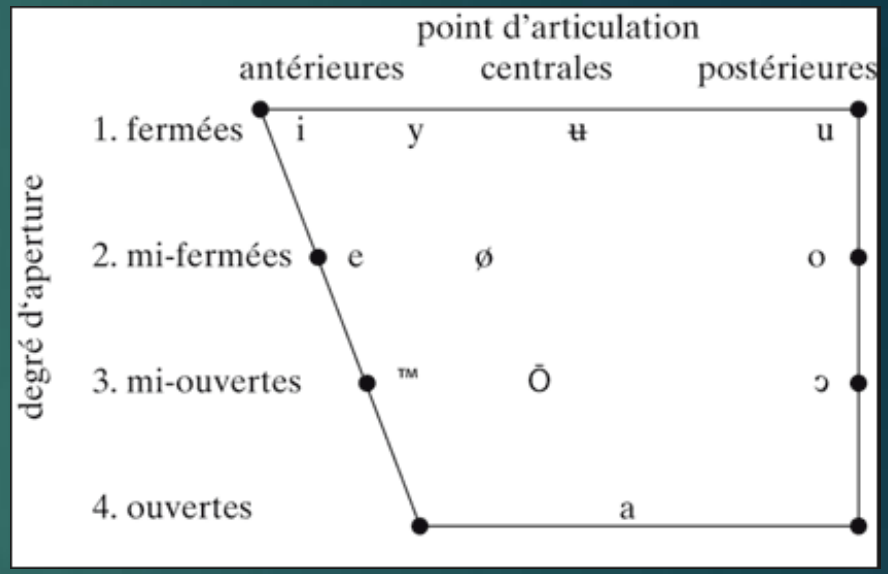
Si cette règle n'est pas respectée, des fréquences parasites qui n'appartiennent pas au signal de départ apparaissent.
→ repliement du spectre





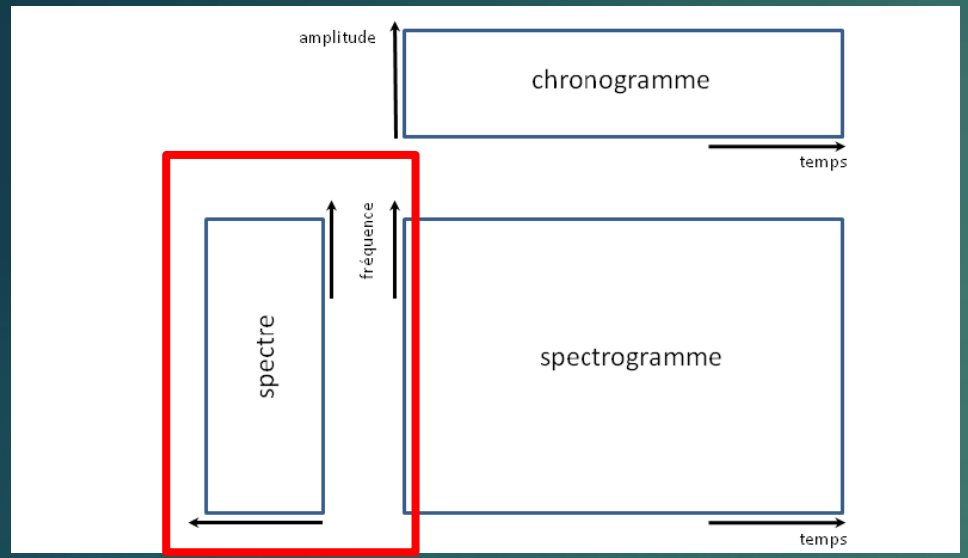
Trapèze vocalique des voyelles



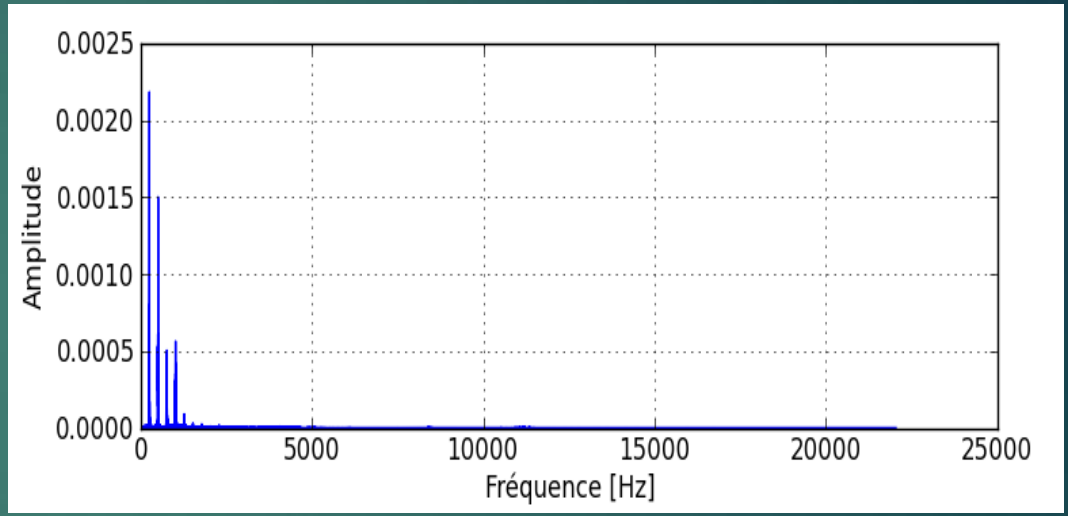


Spectres de Fourier

- Représentation de l'amplitude d'un son en fonction des fréquences qu'il contient



- Spectre de Fourier du son an (réalisé par un programme python)



- Principe de Fourier : Tout signal périodique se décompose en somme de signaux sinusoïdaux
- FFT: $S(jf)=\int_{-\infty}^{+\infty} s(t).e^{-j.2.\pi.f.t}\ dt$

- Principe de Fourier:

Tout signal périodique peut se décomposer en somme de signaux sinusoïdaux.

- Utilisation de la transformé de Fourier discrète:

La méthode consiste à échantillonner un signal u(t) à valeurs réelles

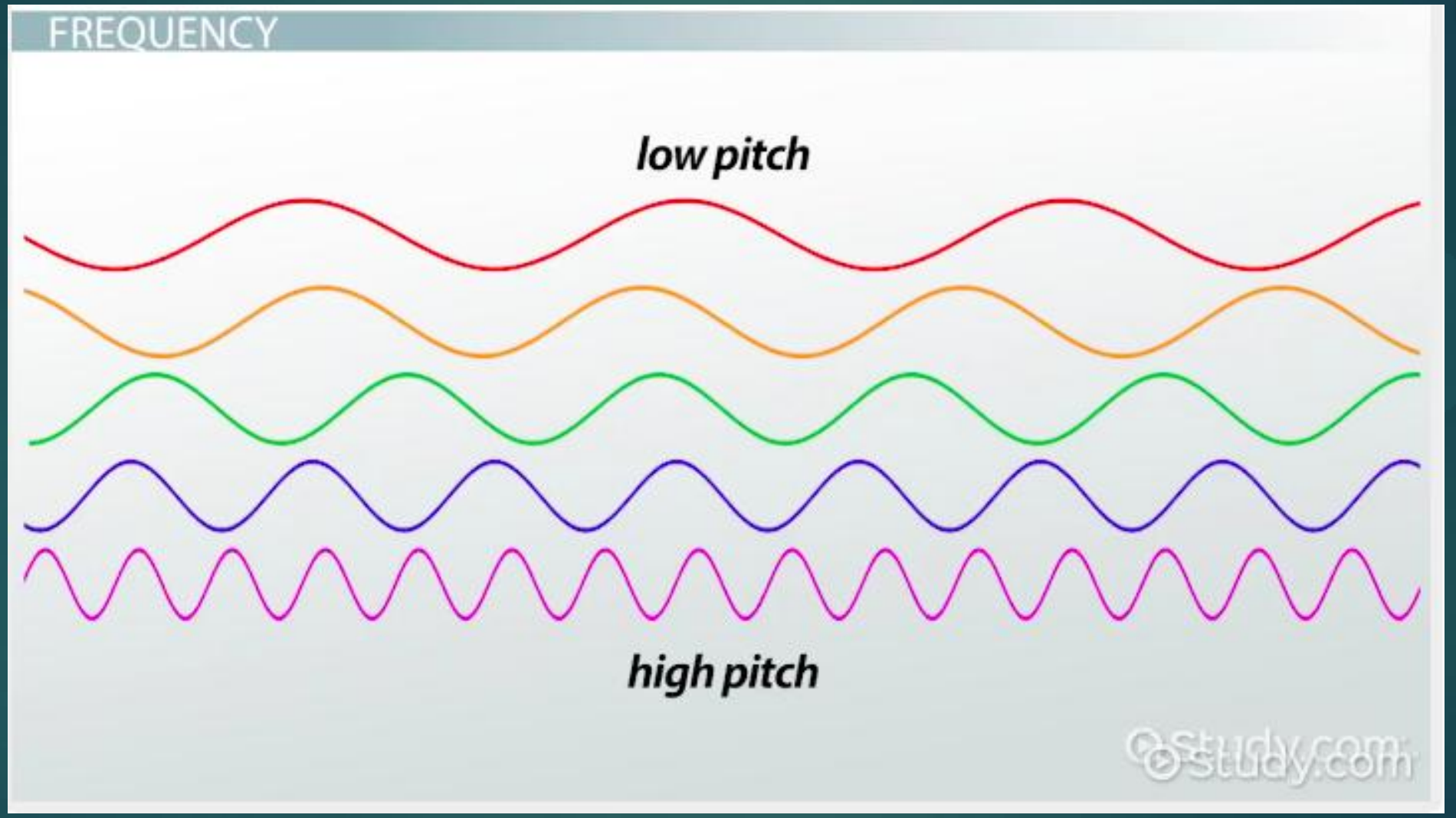
en N échantillons u_n, avec : $\begin{cases} f_e = 2f_{max} \\ N = \frac{f_e}{\Delta f} \end{cases}$ avec Δf: résolution fréquentielle en Hz

La transformée de Fourier discrète est alors donnée par:

$$TU(k) = \frac{1}{N}\sum_{n=0}^{N-1} u_n e^{\frac{-2i\pi kn}{N}}$$

Dans la pratique, on implémente la transformée de Fourier rapide (FFT, pour Fast Fourier Transform).

Pertinence de la transformée Fourier: permet de réaliser une analyse temps-fréquence avec une résolution suffisante pour le signal vocal.


FREQUENCY
low pitch
high pitch
Study.com