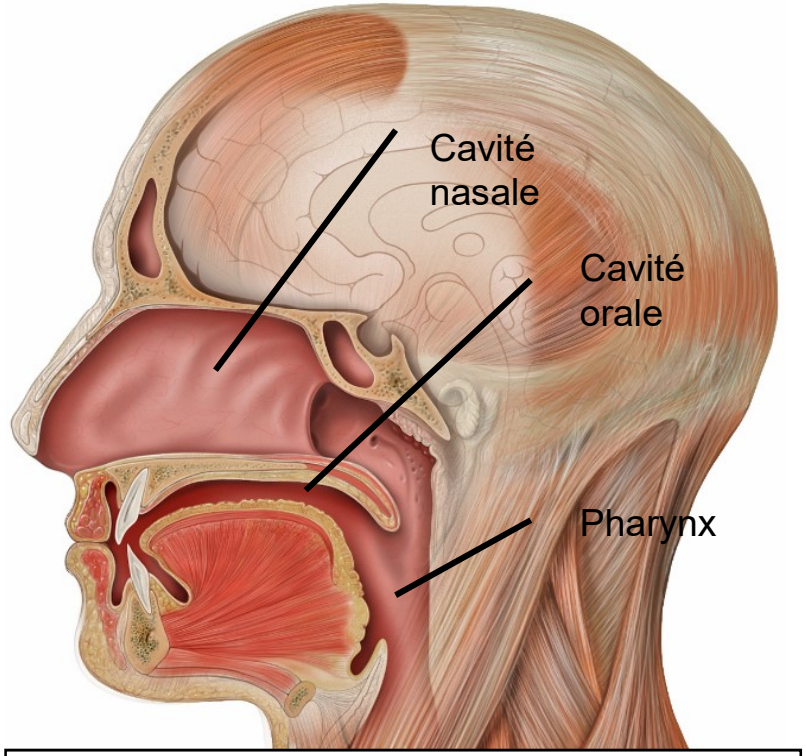
Reconnaissance Automatique de la Parole

1) Modélisation de l’appareil vocal
2) Formant Filter
3) Algorithme d’étude de spectrogrammes pour la reconnaissance vocale

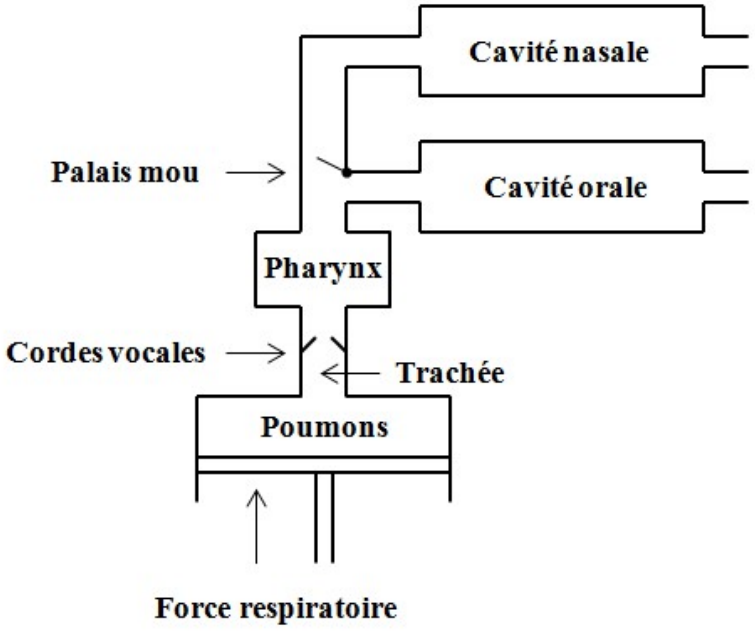
1) Modélisation de l’appareil vocal

a) Les cavités de l'appareil vocal

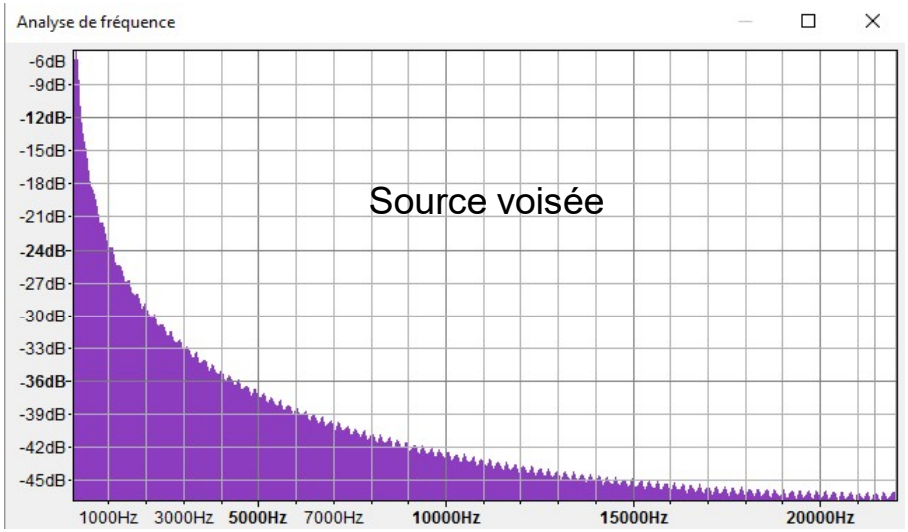


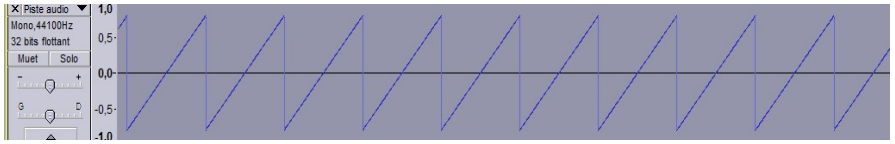
Head lateral view with sagittal mouth anatomy,

23 December 2006, Patrick J. Lynch, medical illustrator

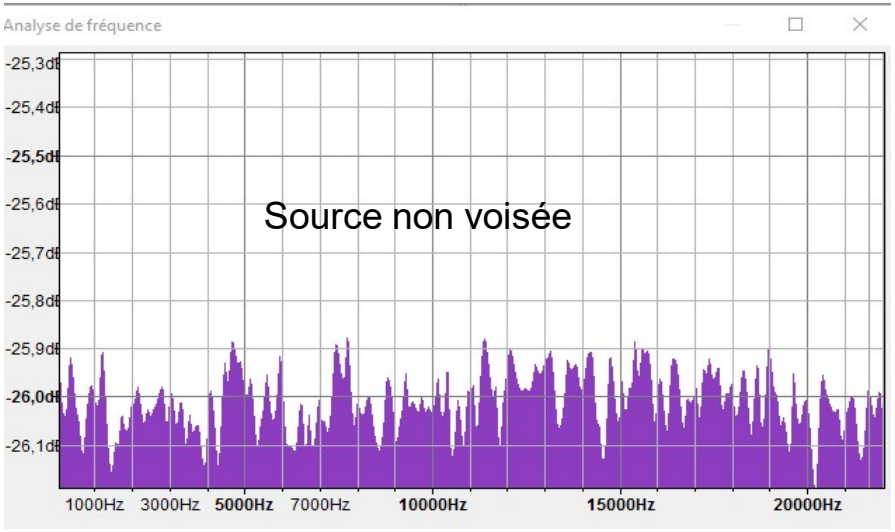


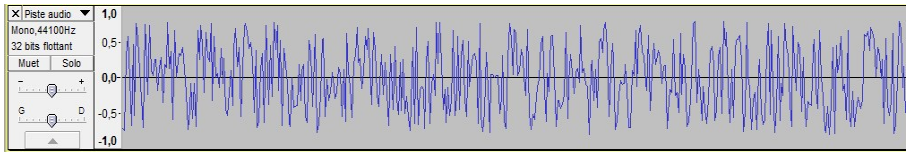
b) Différentes sources de son





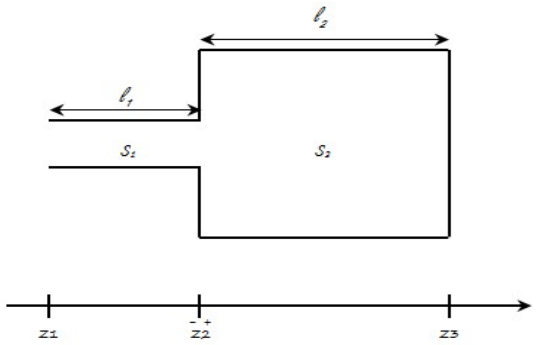
Signal en dents de scie à140 Hz





Bruit Blanc

c) Oscillateur de Helmoltz (pour les voyelles)



Les matrices de transferts acoustiques nous donnent:

$$\begin{pmatrix} P(z_1) \\ V(z_1) \end{pmatrix} = \begin{pmatrix} 1 & j\rho_0\omega l_1 \\ j\frac{\omega l_1}{\rho_0 c^2} & 1 \end{pmatrix} \begin{pmatrix} P(z_2^-) \\ V(z_2^-) \end{pmatrix} \quad (1)$$

$$\begin{pmatrix} P(z_2^+) \\ V(z_2^+) \end{pmatrix} = \begin{pmatrix} 1 & j\rho_0\omega l_2 \\ j\frac{\omega l_2}{\rho_0 c^2} & 1 \end{pmatrix} \begin{pmatrix} P(z_3) \\ V(z_3) \end{pmatrix} \quad (2)$$

Les conditions de pression :

$$P(z_2^-) = P(z_2^+) = P(z_2)$$

$\forall x$ tel que $z_2 < x < z_3$,

on a $P(x) = P(z_2) = P(z_3) = P_c$

Le débit acoustique est constant:

$$S_1 V(z_2^-) = S_2 V(z_2^+) = U$$

Condition limite:

$$V(z_3) = 0$$

Avec (1) on obtient:

$$P(z_1) = P_c + j\omega L_a U \text{ avec } L_a = \frac{l_1 \rho_0}{S_1}$$

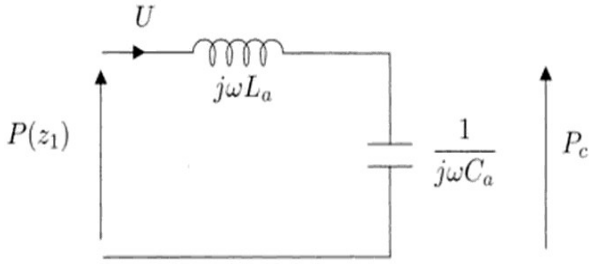
Avec (2) on obtient:

$$U = P_c j \frac{\omega l_2 S_2}{\rho c^2} = P_c j \frac{\omega V}{\rho c^2} \qquad C_a = \frac{V}{\rho c^2}$$

$$U = P_c j\omega Ca$$

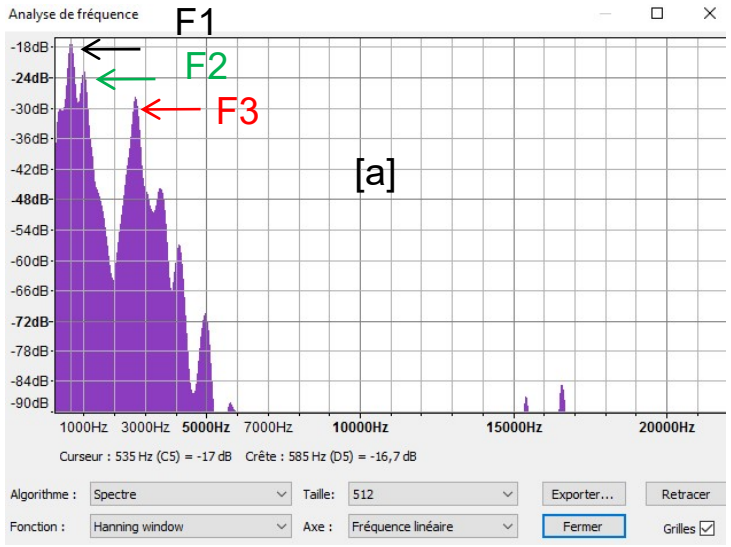
En couplant les 2 résultats: $P(z_1) = P_c(1 + (j\omega)^2 LaCa)$

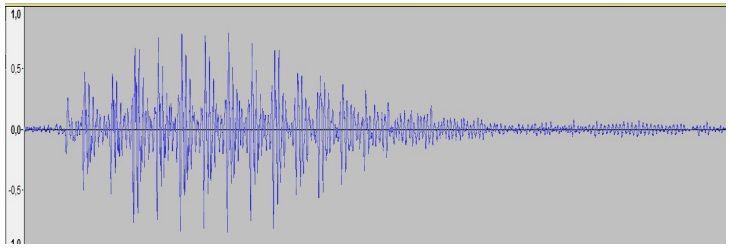
Nous pouvons poser le schéma électrique équivalent au résonateur de Helmoltz

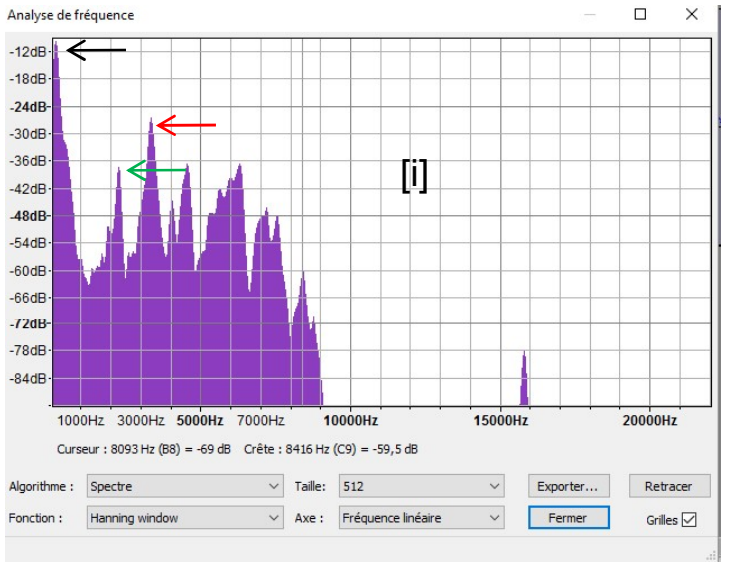


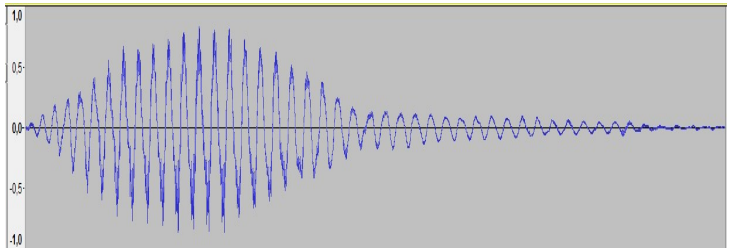
$$\frac{P_c}{P(z_1)} = \frac{1}{1 - \frac{\omega^2}{\omega_r^2}} \text{ avec } \omega_r = 2\pi f_r = c\sqrt{\frac{S_1}{l_1 V}}$$

b) Les Formants caractéristiques des voyelles









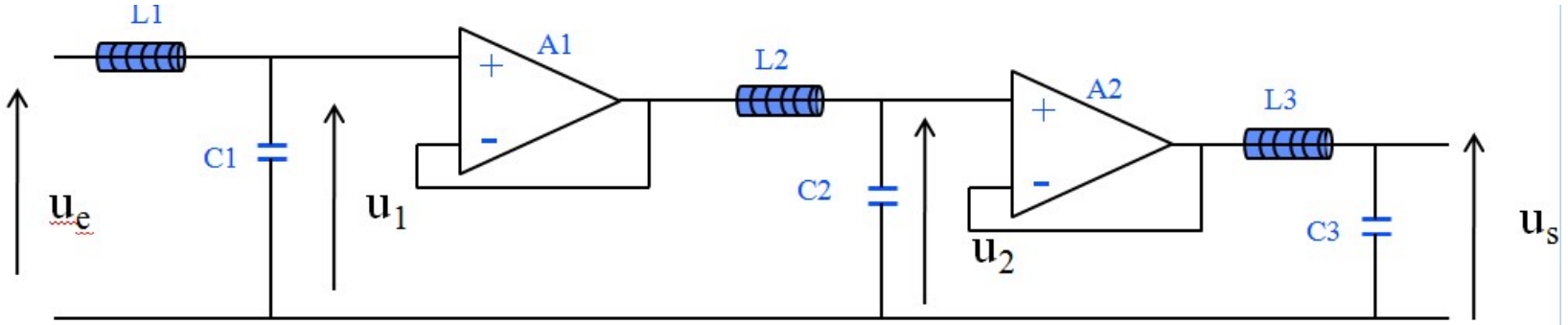
		F1	F2	F3
voy. fermées	i	308	2064	2976
	y	300	1750	2120
	u	315	764	2027
voy. mi-fermées	e	365	1961	2644
	ø	381	1417	2235
	o	383	793	2283
voy. mi-ouvertes	ɛ	530	1718	2558
	œ	517	1391	2379
	ɔ	531	998	2399
voy.ouv.	a	684	1256	2503

Figure1) Valeurs formantiques moyennes des voyelles orales du français (d'après Tubach, 1989)

Source: Christine Meunier. Phonétique acoustique : Phonétique acoustique. Auzou P. Les dysarthries, Solal, pp.164-173, 2007.

2) Formant Filter

a) Circuit électronique

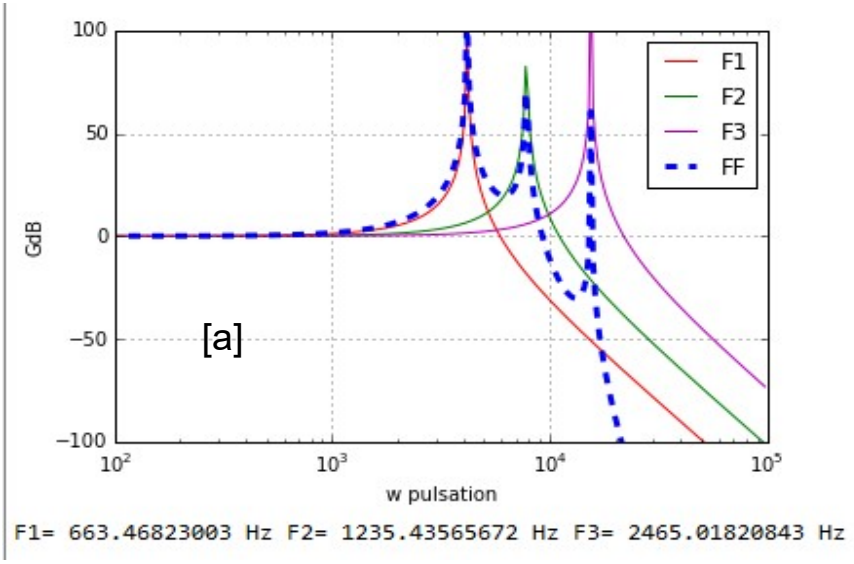


$$\frac{us}{ue} = \frac{us}{u2} \times \frac{u2}{u1} \times \frac{u1}{ue}$$

$$H(j\omega) = \frac{us}{ue} = \frac{1}{(1 - L1C1\omega^2)} \frac{1}{(1 - L2C2\omega^2)} \frac{1}{(1 - L3C3\omega^2)}$$

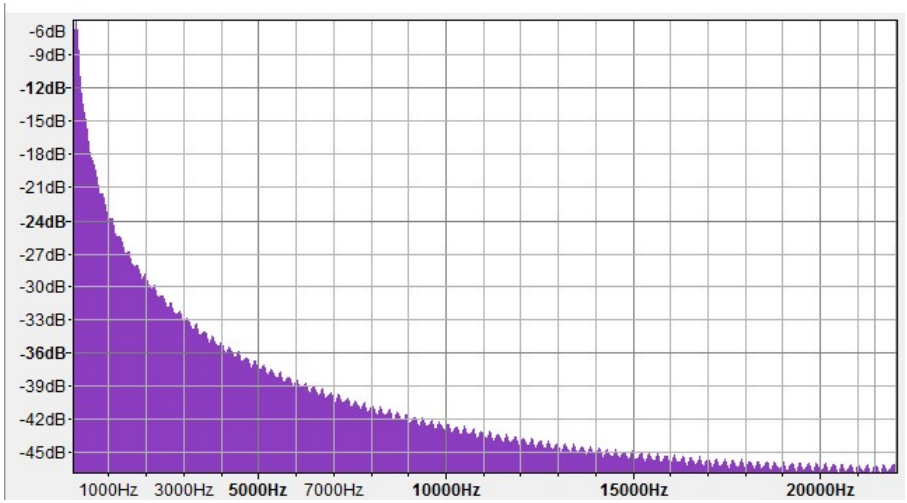
b) Diagramme de Bode

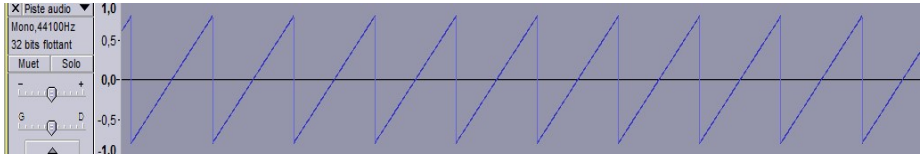
$$GdB = 20\log(\|H(j\omega)\|) = 20\log(\|H1(j\omega)\|) + 20\log(\|H2(j\omega)\|) + 20\log(\|H3(j\omega)\|)$$



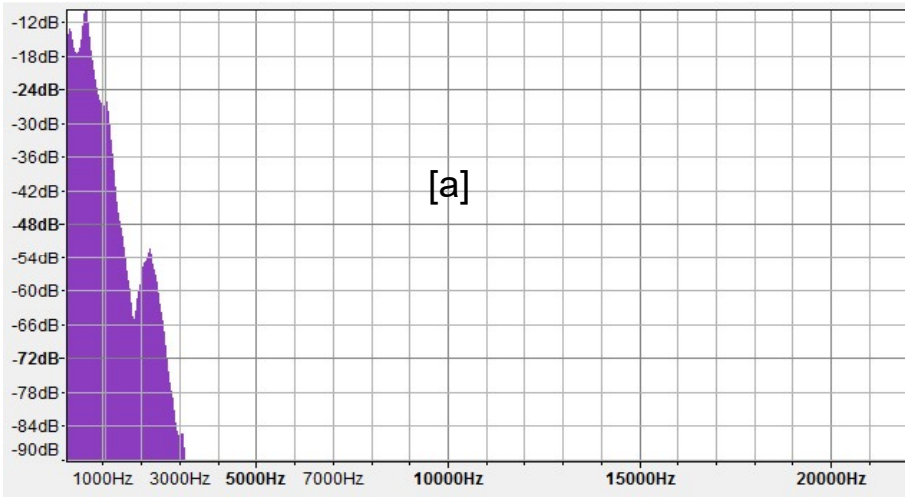
F1= 663.46823003 Hz F2= 1235.43565672 Hz F3= 2465.01820843 Hz

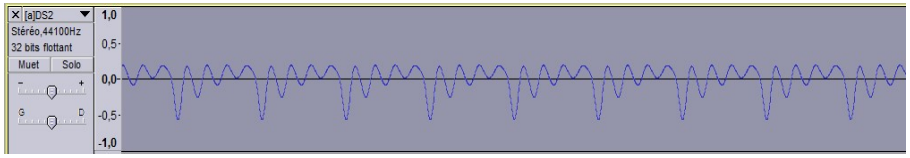
c) Spectre en sortit du filtre





Signal en dents de scie à140 Hz





Signal en sortit du filtre

d)Limite du modèle

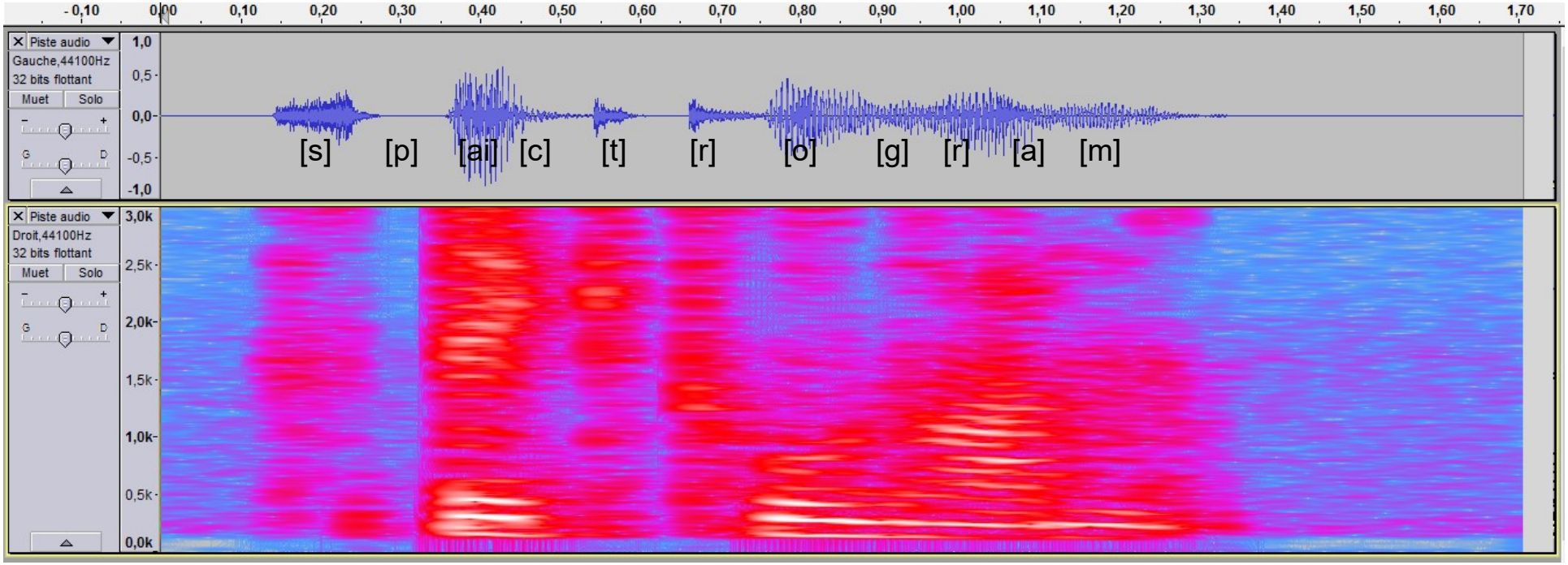
	F1 (Hz)	V1 (cm^3)	F2 (Hz)	V2 (cm^3)	F3 (Hz)	V3 (cm^3)
[i]	308	33.0	2064	0.70	2976	0.33
[a]	684	7.0	1256	2.0	2503	0.51
[ɛ̃]= [in]	250	50.0	600	9.0	1750	1.0

Figure 2 .a) Valeurs du volume de la cavité pour chaque formant en prenant $l1 = 10^{-1}m \; et \; S1 = 10^{-4}m^2$

Vmoyen = 10 cm^3

3) Algorithme d’étude de spectrogramme pour la reconnaissance vocale

a) Spectrogramme

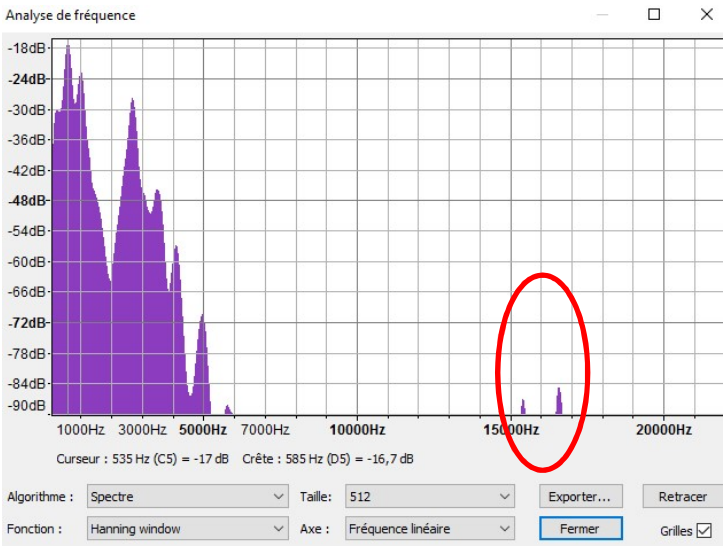


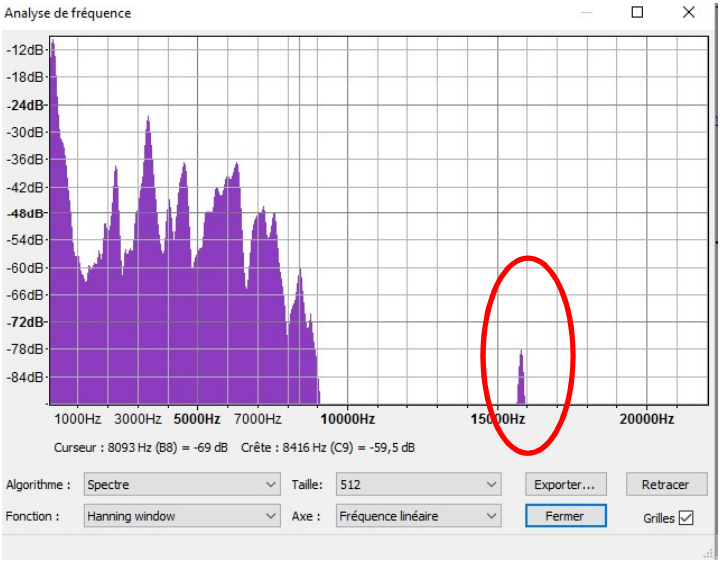
b) Algorithme de reconnaissance d'empreintes vocales

PHONEMES																	
VOYELLES				SEMI-CONSONNES		CONSONNES											
ORALES		NASALES				LIQUIDES		NASALES		FRICATIVES				OCCLUSIVES			
										voisées		non voisées		voisées		non voisées	
[i]	lit	[ɛ̃]	lin	[j]	yéyé	[l]	lait	[m]	mais	[v]	vais	[f]	fait	[b]	baie	[p]	paie
[e]	les	[ɑ̃]	lent	[w]	wais	[R]	raie	[n]	nez	[z]	zéros	[s]	sait	[d]	dais	[t]	taie
[ɛ]	lait	[ɔ̃]	long	[y]	huer			[h]	gagner	[z]	jeux	[ʃ]	chez	[g]	gai	[k]	quai
[a]	là	[œ̃]	un														
[u]	loup																
[o]	lot																
[y]	lu																
[∅]	leu																
[œ]	leur																
[ə]	le																

Source:
BOITE R. et KUNT M., *Traitement de la parole*, Lausanne, PRESSES POLYTECHNIQUES ROMANDES, 1987.

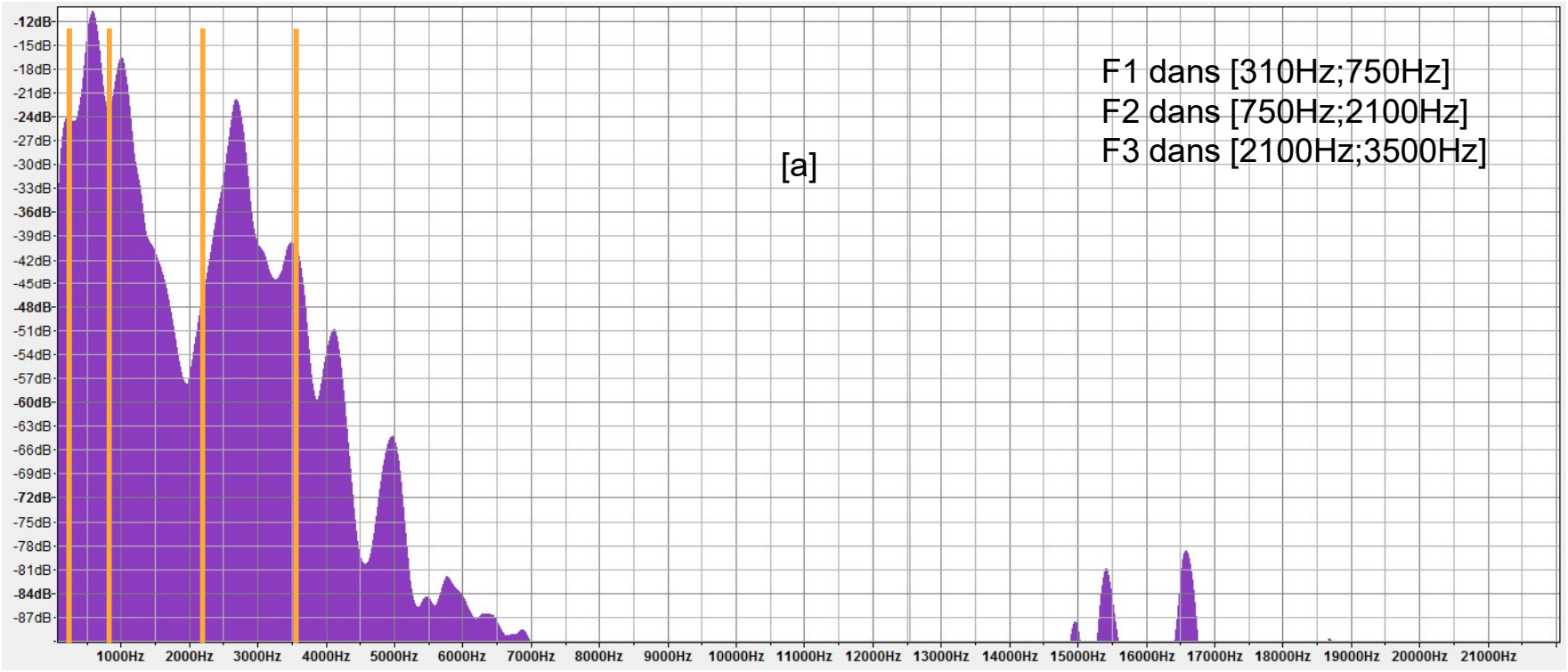
c) Reconnaissance de sons Voisés ou non Voisés





Fref=16000Hz

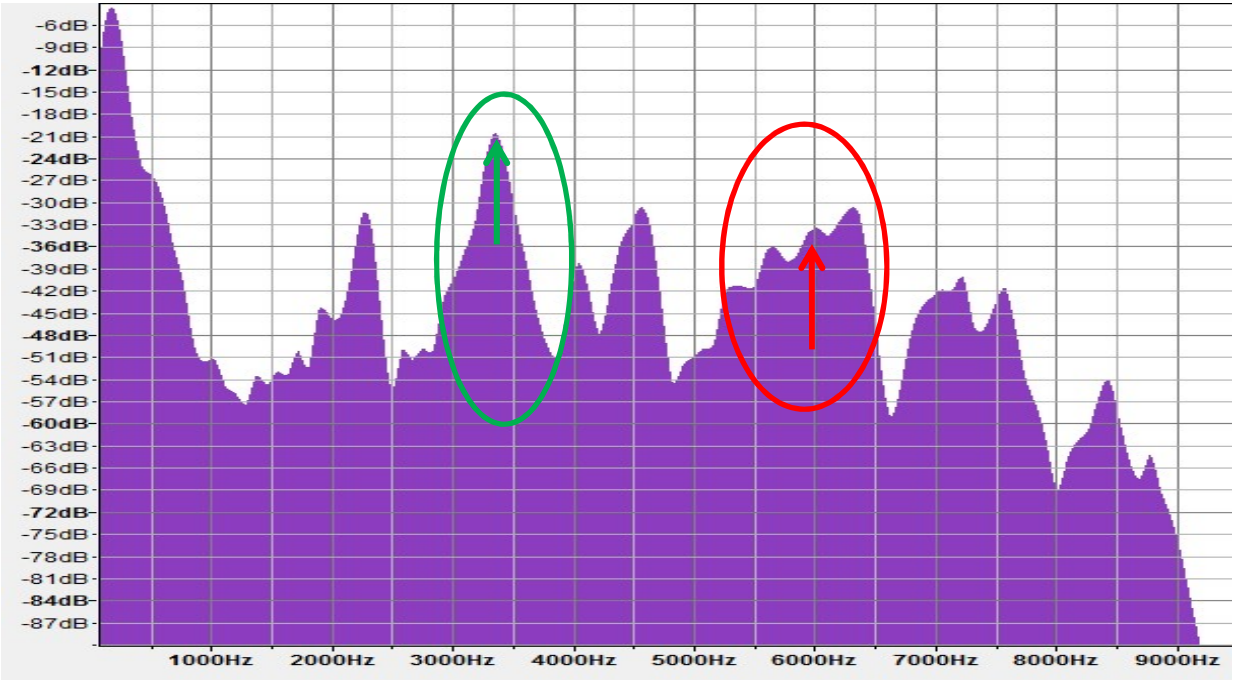
b) Reconnaissance des Voyelles



Etape1: découpage de la recherche

Etape2: recherche des formants dans chaque intervalle

a) Recherche les maximums globaux de chaque intervalle
b) On détermine si ces maximums sont sur des sommets ou des plateaux



Voyelles testées (5 échantillons)	E-F [a] En(Hz)	E-F [eu] En (Hz)	E-F [i] En (Hz)	E-F [o] En (Hz)	E-F [u] En (Hz)	Nb de Reconnaissances
[a]	2002	3418	9231	2987	6063	5/5
[eu]	4182	1356	7654	4232	2157	4/5
[i]	8638	7539	3376	9403	6746	3/5
[o]	4351	3922	11091	1198	4488	5/5
[u]	5996	3044	5199	6298	1702	5/5

Figure3) Test de l'algorithme de reconnaissance de voyelles sur 5 échantillons de chaque son

*E-F : la somme des écarts formantiques entre les formants théoriques et les formants observés

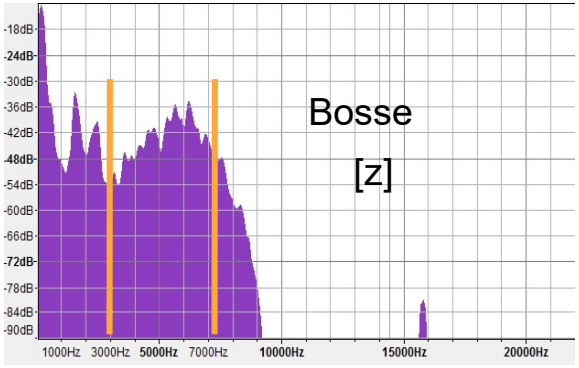
*Nb de Reconnaissances : le nombre de fois où l'algorithme a reconnu le bon phonème

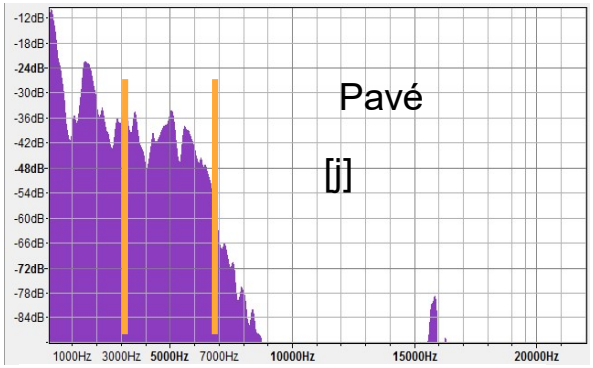
Voyelles testées	E-F [a] En(Hz)	E-F [eu] En (Hz)	E-F [i] En (Hz)	E-F [o] En (Hz)	E-F [u] En (Hz)	Reconnaissance
[a]ff	547	479	1940	539	1002	[eu]
[eu]ff	817	469	2210	529	768	[eu]
[i]ff	979	845	678	1419	708	[i]
[o]ff	749	749	2210	249	1048	[o]
[u]ff	809	189	1650	809	488	[eu]

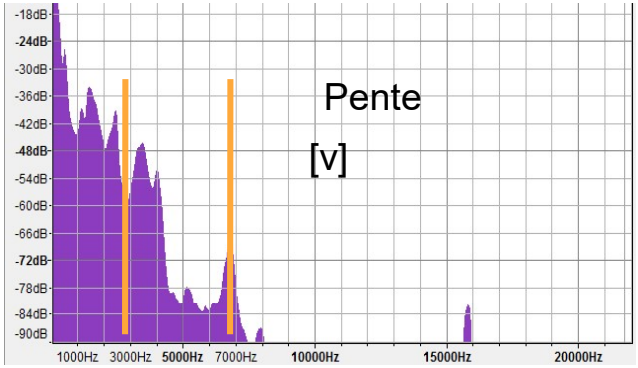
Figure4) Test de l'algorithme de reconnaissance de voyelles sur des voyelles fabriquées avec un « formant filter »

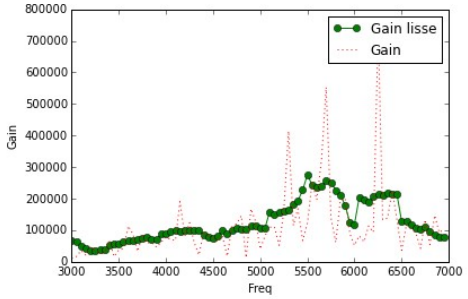
*Reconnaissance : Phonème reconnu par l'algorithme

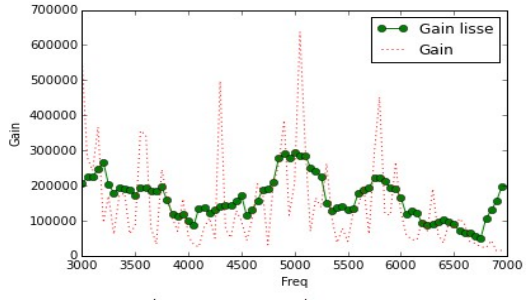
e) Reconnaissance des Fricatives voisées

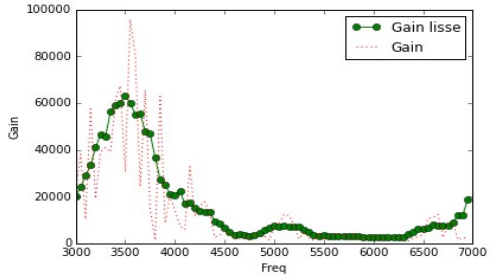








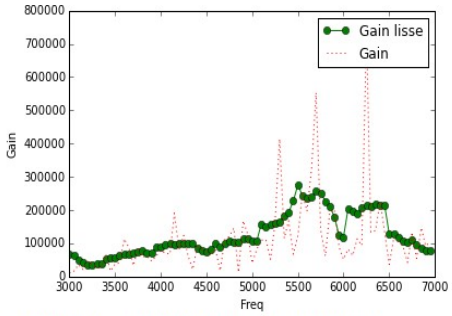




Etape1: découpage de la recherche

[3000Hz;7000Hz]

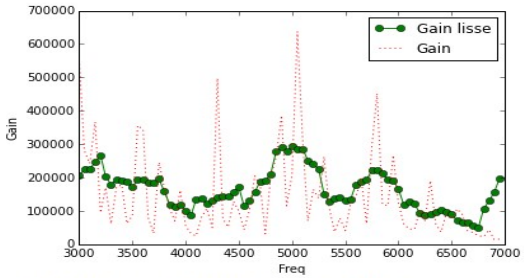
Bosse a1>0




```
a=3.492937e+01, b=-4.901151e+04, rho=6.381543e-01
bosse

In [600]: Fric
Out[600]: '[z],alveolaire'
```

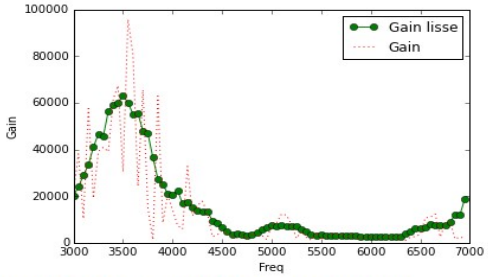
Pavé a1<0 et s<0.5




```
a=-2.044280e+01, b=2.661430e+05, rho=-3.953710e-01
0.363121982737
pave

In [7]: Fric
Out[7]: '[j],platale'
```

Pente a1 <0 et s>=0.5




```
a=-1.084752e+01, b=6.985470e+04, rho=-7.133191e-01
1.10604672019
pente

In [21]: Fric
Out[21]: '[v],labiodentale'
```

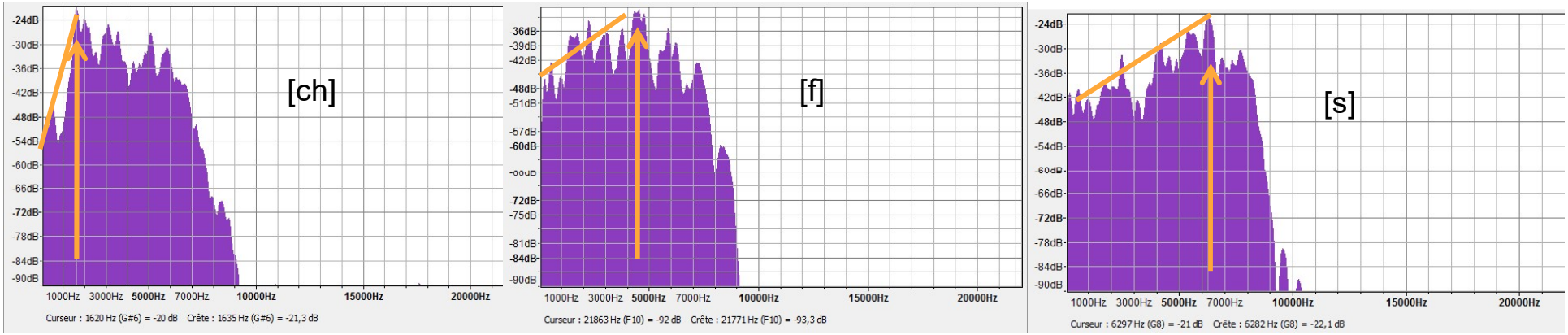
Etape2: Calcul du coefficient directeur de la courbe lissée

Si a1 >=0 c'est une bosse [z]

Etape3: Calcul du coefficient directeur de la courbe lissée

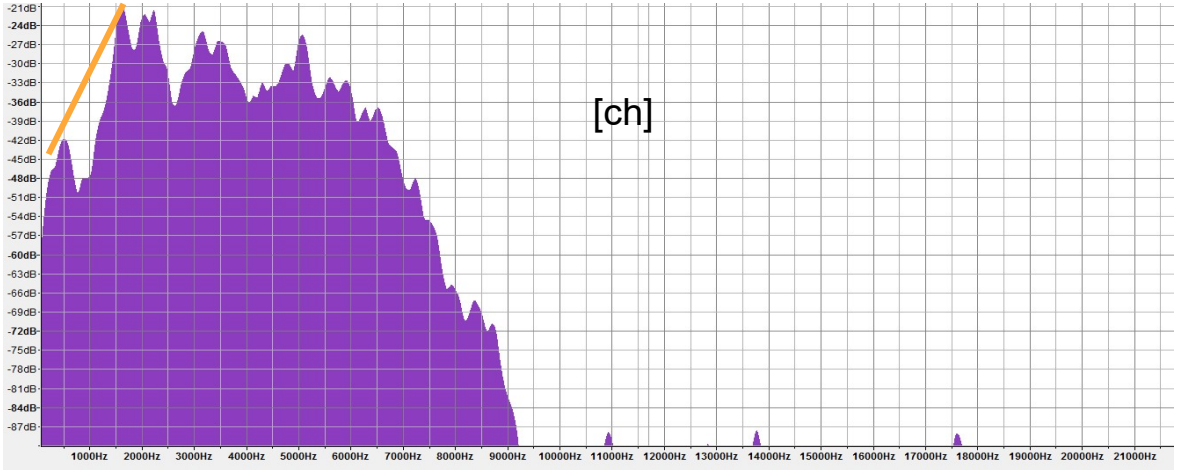
Si a1<=0
Si s < 0.5 c'est un pavé [j]
s>=0.5 c'est une pente [v]

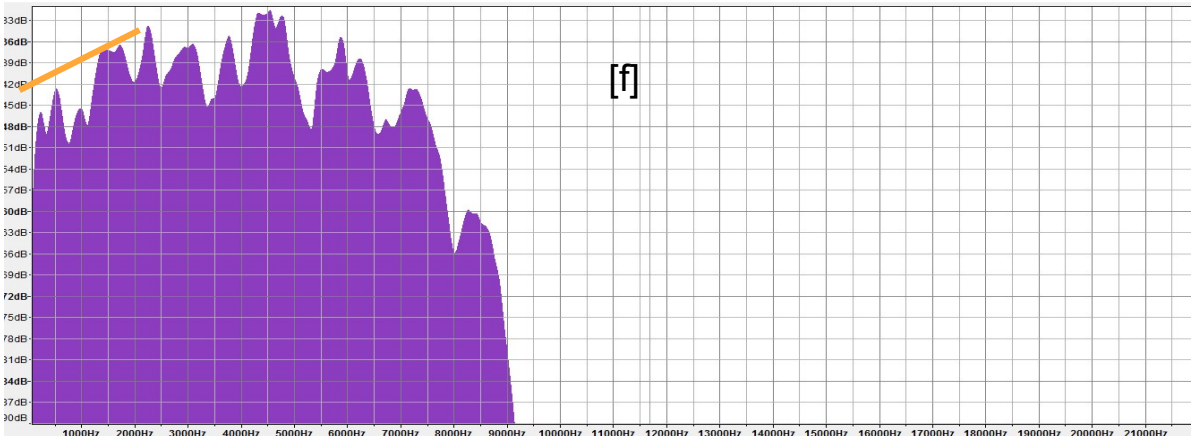
d) Reconnaissance des fricatives non voisées



Étape 1 : recherche de fmax

Si fmax >= 5000Hz alors la consonne est [s]





Etape 2: étude de pente

On s'intéresse au coefficient directeur de la courbe (a2)
Sur [0Hz;2000Hz]

Si celui-ci est supérieur à 100:
    On reconnait [ch]
Sinon :
    c'est un [f]

consonne	Echantillon1	Echantillon2	Echantillon3	Echantillon4	Echantillon5	Nb
[v]	a1=-1.6 s=0.9 Voisée Pente[v]	a1=-2.6 s=0.7 Voisée Pente[v]	a1=-1.8 s=0.7 Voisée Pente[v]	a1=-2.1 s=1.6 Voisée Pente[v]	a1=-1.2 s=1.2 Voisée Pente[v]	5/5
[f]	a2=28 fmax=4420 Non voisée Labiodentale[f]	fmax=39525 Non voisée alvéolaire[s]	a2=29 fmax=2289 Non voisée Labiodentale[f]	a2=52 fmax=4789 Non voisée Labiodentale[f]	a2=28 fmax=4418 Non voisée Labiodentale[f]	4/5
[z]	a1=1.1 s=0.6 Voisée Bosse [z]	a1=4 s=0.44 Voisée Bosse[z]	a1=7 s=0.6 Voisée Bosse[z]	a1=4 s=0.4 Voisée Bosse[z]	a1=6 s=0.54 Voisée Bosse[z]	5/5
[s]	fmax=39391 Non voisée alvéolaire[s]	fmax=39453 Non voisée alvéolaire[s]	a2=17 fmax=4741 Non voisée Labiodentale[f]	a2=15 fmax=4481 Non voisée Labiodentale[f]	fmax=39799 Non voisée alvéolaire[s]	3/5
[j]	a1=-9 s=0.4 Voisée Pavé[j]	a1=-1.5 s=0.5 Voisée Pente[v]	a1=-1.2 s=0.6 Voisée Pente[v]	a1=-2 s=0.4 Voisée Pavé[j]	a1=-1.3 s=0.5 Voisée Pente[v]	2/5
[ch]	a2=337 fmax=2223 Non voisée platale[ch]	a2=172 fmax=3323 Non voisée platale[ch]	a2=165 fmax=3100 Non voisée platale[ch]	a2=814 fmax=3653 Non voisée platale[ch]	fmax=48127 Non voisée alvéolaire[s]	4/5

Figure5) Test de l'algorithme de reconnaissance des consonnes

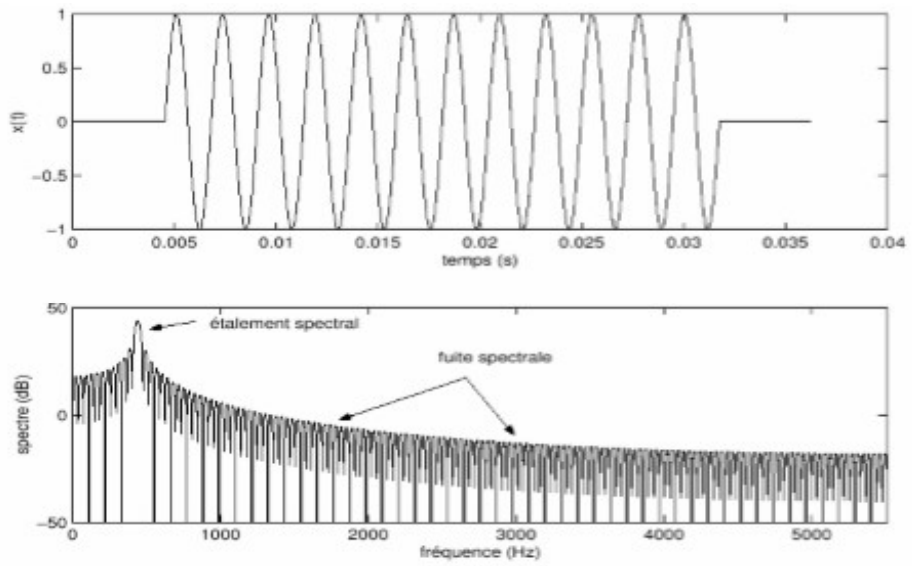
*pour les fricatives voisées on s'intéresse au coefficient directeur du spectre sur [3000Hz,7000Hz] (a1), et à l'écart type (s)

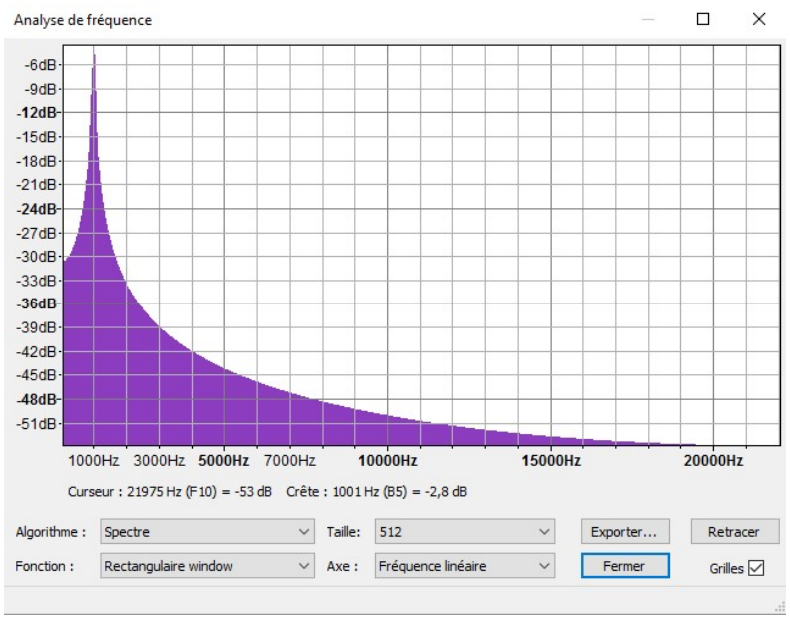
*pour les non voisées on s'intéresse à la fréquence avec le gain maximal puis au coefficient directeur sur [0Hz,2000Hz] (a2)

*Nb: Nombre de fois où l'algorithme a reconnu le bon phonème

FIN DU DIAPORAMA

*) Fuite et étalement spectral





Analyse et représentation de signaux acoustiques modulés

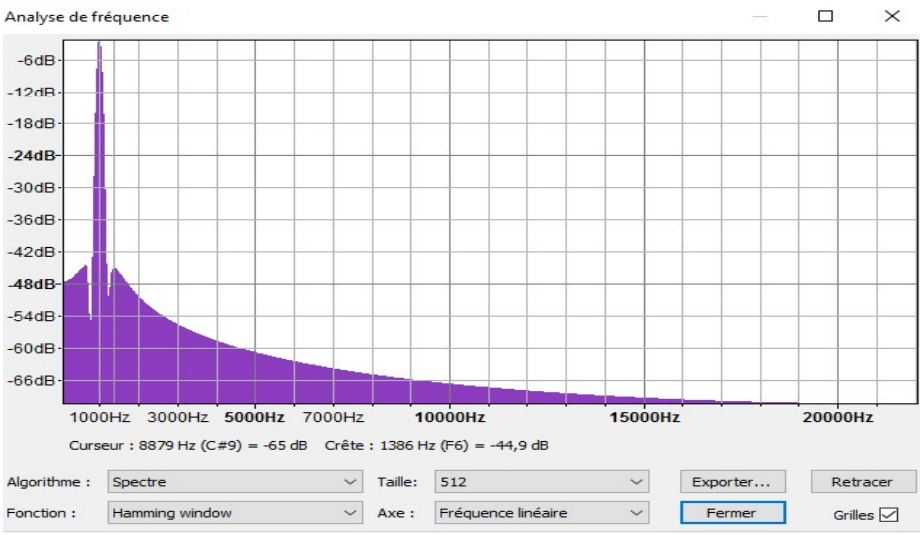
Bertrand DAVID

Département du Traitement de Signal et des Images Télécom Paris (ENST)

Observation a travers une fenêtre rectangulaire f=1000Hz

**)Fênetrage de Hamming

$$h(t) = \begin{cases} 0,54 - 0,46\cos\left(2\pi\frac{t}{T}\right) & \text{si } t \in [0,T] \\ 0 & \text{sinon.} \end{cases}$$



Spectre d'une sinusoïde f=1000Hz

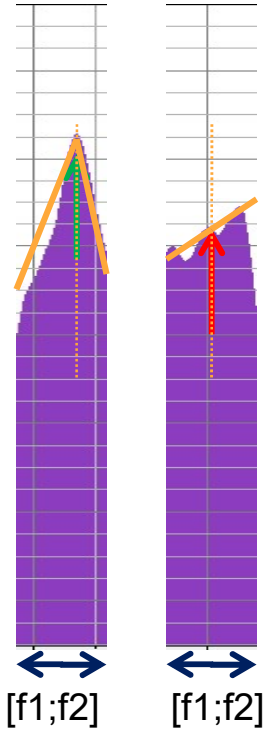
On étudie le signale tronquer

Dans le domaine temporel:
sh(t)=s(t)*h(t)

Dans le domaine fréquentiel:
Sh(f)=(S*H)(f)

***) Fonction sommet

b) On détermine si les fréquences maximales (fmax) sont sur des sommets ou des plateaux



Étape 1 : Etude des deux coefficients directeurs

Si ag>0 et ad<0 on passe à l'étape suivante, sinon c'est un plateau

Étape 2 : Etude des différences de hauteur

Et Si diff = (abs(Gfmax-Gf1)+abs(Gfmax-Gf2))<= epsilon de précision

Alors c'est un sommet