Résultats d'analyse
Version 1.0

Evaluation automatique de la prononciation

ECH-CHOUQI Rim

AALAOU Samia

BIOY Marianne

GUERY Samuel

YU Shaofeng

3IN

2015-2016

Sommaire

1 Analyse avec les coefficients de Fourier

Les résultats sur les tableaux sont les pourcentages de chaque catégorie dans chaque classe (par exemple le pourcentage des voisés dans la classe 1)

Une première étape a été de travailler avec kmeans en version non supervisée sur 3 classes :

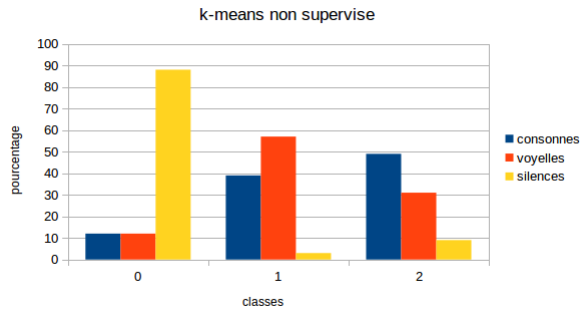

FIGURE 1.1 – K-means non supervisé sur les paramètres de fourier

On constate qu'il n'y a rien de très distinguable. Ensuite, on a essayé de faire une initialisation des centres en choisissant comme centres une lettre voisée, une non voisée et un silence quelconques.

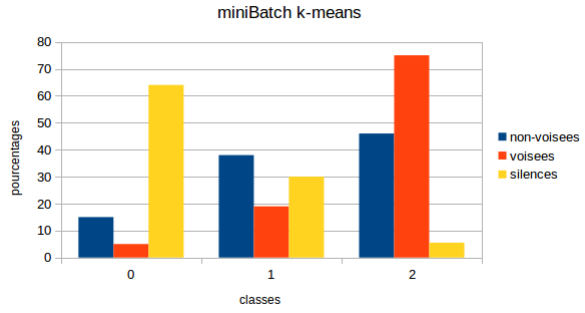

FIGURE 1.2 – K-means supervisé avec les catégories voisés, non-voisés silences

On remarque qu'on arrive à mieux distinguer les voisés dans la 3ème classe.

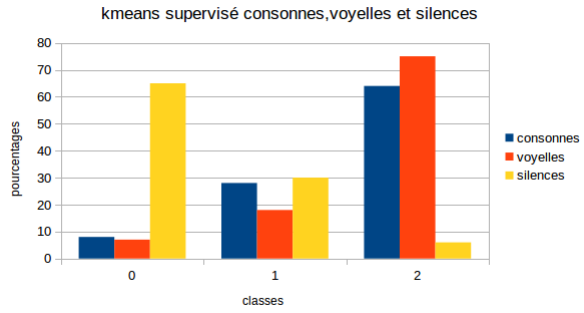

FIGURE 1.3 – K-means supervisé avec les catégories consonnes, voyelles et silences

On remarque qu'on arrive à bien distinguer le silence dans la première classe. On s'apperçoie directement que le résultat est un peu mieux qu'un kmeans non supervisé.

Après, on a essayé de voir ce que ça donnait si on procédait à une catégorisation à 6 classes : Occlusives, fricatives, nasales, semi-consonnes, voyelles et silences :

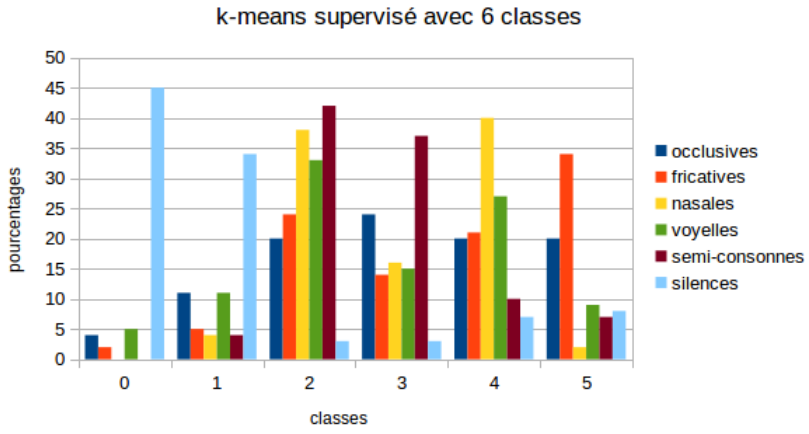

FIGURE 1.4 – K-means supervisé avec les catégories occulusives, fricatives, voyelles...

On a essayé aussi les algorithmes affinity propagation qui a donné un nombre trop grand de classes (environs 80) et spectral clustering qui prend trop de temps pour l'éxécution.

Et comme autre méthode de visualisation des résultats, on a fait des histogrammes qui affichent dans chaque classe le nombre d'occurences de chaque phonème :

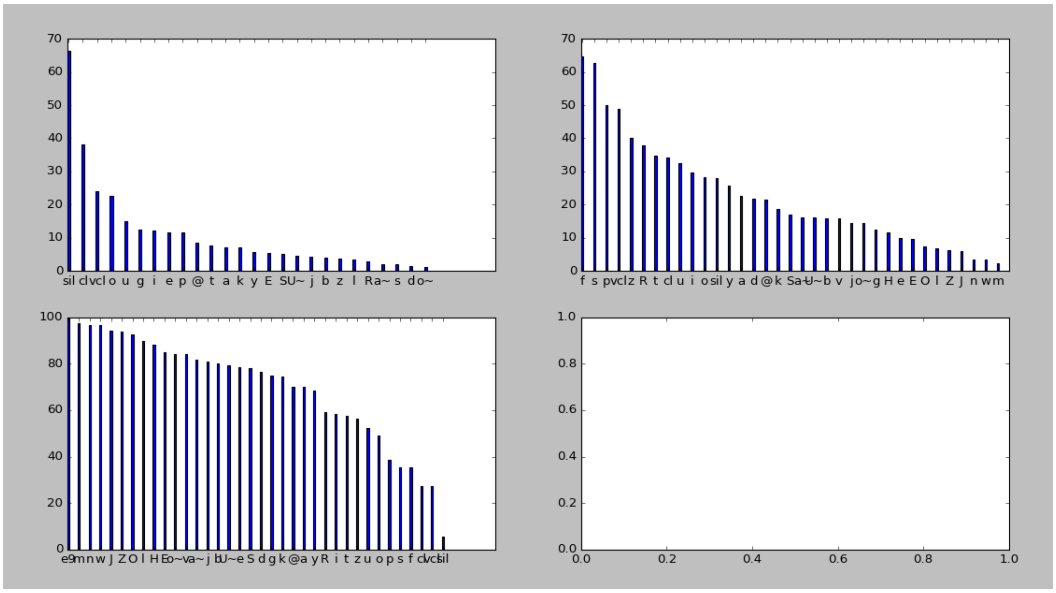

FIGURE 1.5 – Histogramme issue de kmeans avec 3 classes

2 Analyse avec les coefficients cepstraux de Mel (MFCC)

Plusieures méthodes de clustering ont été essayées : kmeans avec les centres initialisés ou non, agglomerative clustering, et le meanShift. Les deux premiers clusterings ont été établis pour 3 et 6 clusters, et le meanShift a lui, produit 2 classes. On vérifie l'efficacité de ces clusterisations en faisant apparaître le pourcentage de consonnes/voyelles/silences, de phonèmes voisés/non-voisés/silences et de catégories (occlusives, fricatives...) de phonèmes classés dans chaque cluster. Toute cette analyse a étée effectuée pour 13 et 40 coefficients de Mel. Les résultats montrés ici le sont ceux obtenus pour 13 coefficients.

En première analyse, on peut dire que les résultats entre le k-means initialisé et le k-means non initialisé sont sensiblement les mêmes, que ce soit pour 3 ou pour 6 clusters, à part, éventuellement pour les catégories de phonèmes avec 6 clusters, où les semi-consonnes sont un peu mieux reconnues avec l'algorithme supervisé.

Voici ce qu'on obtient en moyenne pour ces deux clusterings (on montre les résultats du supervisé)

KMEANS initialise 3 clusters						
classes	**consonnes**	**voyelles**	**silences**			
0	17	11	82			
1	43	35	11			
2	40	54	7			
classes	**non-voisees**	**voisees**	**silences**			
0	25	11	82			
1	43	39	11			
2	32	50	7			
classes	**occlusives**	**fricatives**	**nasales**	**voyelles**	**mi-consonn**	**silences**
0	47	7	4	13	4	62
1	30	52	37	35	64	16
2	23	41	59	53	31	21

FIGURE 2.1 – kmeans pour 3 clusters

KMEANS initialise 6 clusters						
classes	consonnes	voyelles	silences			
0	9	6	58			
1	13	7	27			
2	19	31	3			
3	20	16	8			
4	23	26	4			
5	16	14	0			
classes	non-voisees	voisees	silences			
0	17	5	58			
1	12	10	27			
2	17	26	3			
3	11	21	8			
4	17	26	4			
5	26	12	0			
classes	occlusives	fricatives	nasales	voyelles	mi-consonn	silences
0	26	3	1	7	0	43
1	27	8	5	9	5	23
2	8	21	29	31	5	12
3	13	18	27	16	23	11
4	17	23	32	24	36	10
5	10	27	5	13	30	2

FIGURE 2.2 – kmeans pour 6 clusters

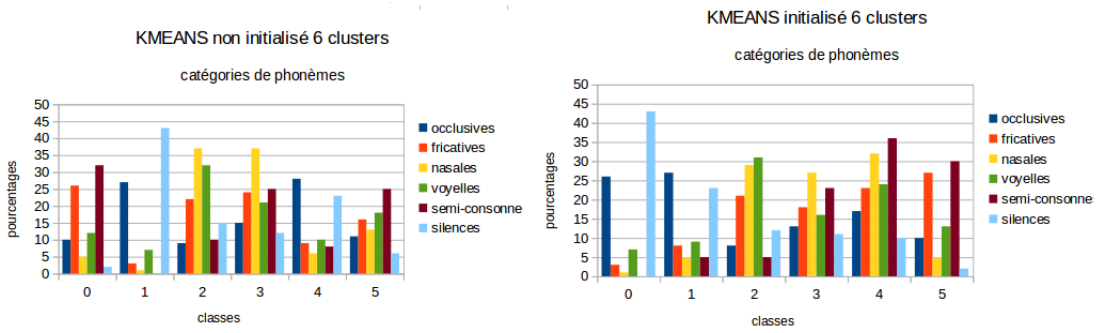

FIGURE 2.3 – catégories de phonèmes reconnues par les kmeans initialisé et non-initialisé

D'une façon générale, les phonèmes voisés sont très bien regroupés ensemble, et les voyelles sont très bien reconnues par le k-means supervisés (plus de 60% dans une seule classe, soit un meilleur résultat que les silences). De plus, les occlusives tendent à être rangées avec les silences, et les semi-consonnes à être bien reconnues. Les fricatives, les consonnes nasales et les voyelles sont rangées ensemble.

Les résultats de l'agglomerative clustering sont eux plus mitigés : Pour 3 classes, les consonnes nasales sont regroupées avec les autres phonèmes, mais sont tout de même bien reconnues (elles sont regroupées à plus de 80% dans la même classe). On voit toujours les occlusives reconnues comme proches des silences :

Agglomerative clustering 3 clusters						
classes	consonnes	voyelles	silences			
0	58	68	15			
1	28	24	27			
2	14	8	58			
classes	non-voisees	voisees	silences			
0	41	69	15			
1	37	23	27			
2	22	8	58			
classes	occlusives	fricatives	nasales	voyelles	semi-consonn	silences
0	37	55	87	66	57	32
1	20	41	11	24	40	24
2	43	4	2	10	3	44

FIGURE 2.4 – résultats pour 3 clusters de l'agglomerative clustering

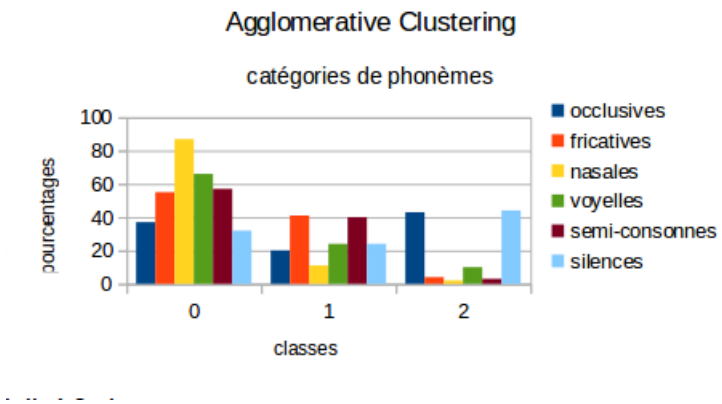

FIGURE 2.5 – catégories de phonèmes reconnues par l'agglomerative clustering

Enfin, on constate que le MeanShift ne crée que deux classes où il oppose les silences aux autres. On remarque cependant toujours l'association d'une partie non négligeable des occlusives au silences, et les consonnes nasales mieux discriminées que les autres :

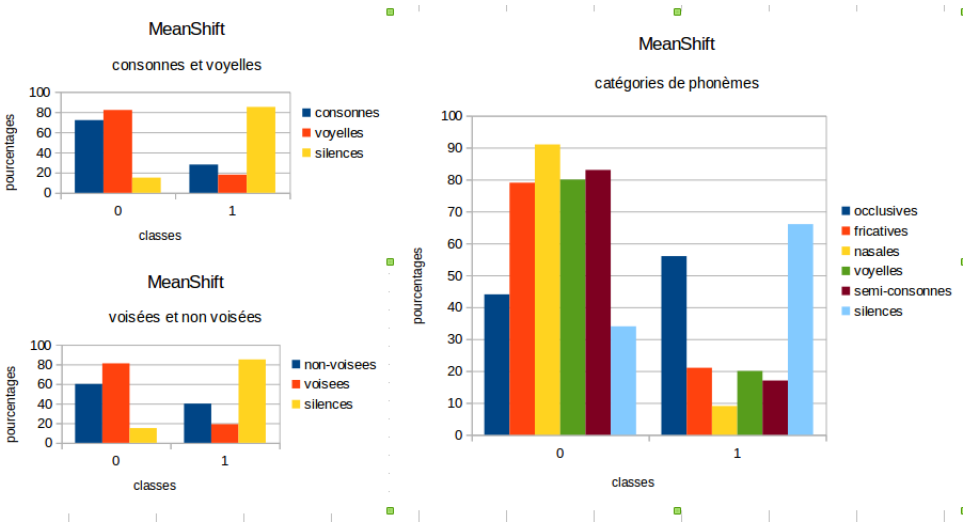

FIGURE 2.6 – histogramme des résultats pour le MeanShift

Voici le résultat du clustering pour chaque phonème (pourcentage de chaque

phonème placé dans chaque cluster pour 6 clusters par les algorithmes de k-means initialisé et non-initialisé, agglomerative clustering, et mean-shift respectivement) :

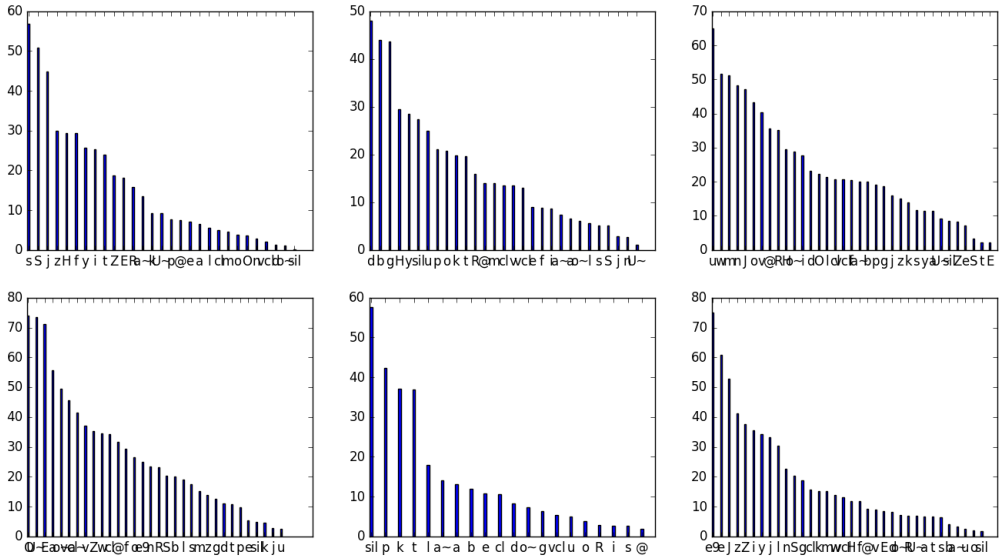

FIGURE 2.7 – kmeans non initialisé pour 6 clusters

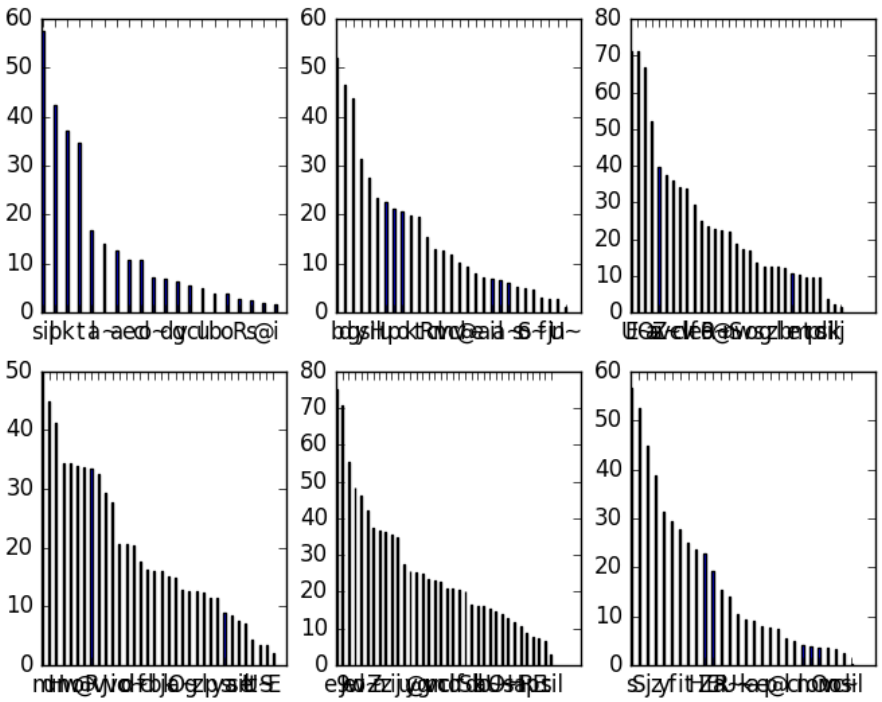

FIGURE 2.8 – kmeans initialisé pour 6 clusters

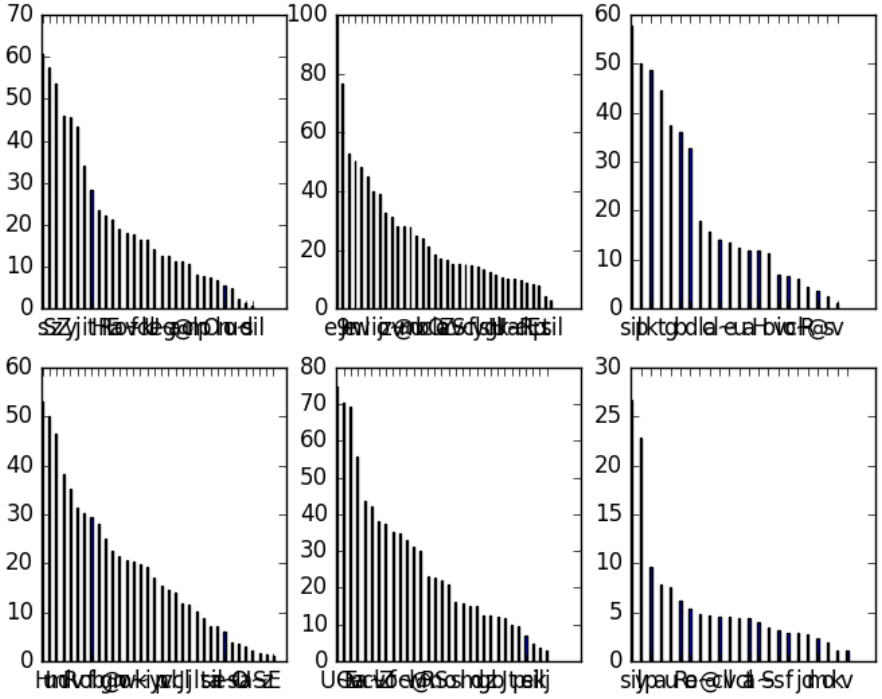

FIGURE 2.9 – agglomerative clustering pour 6 clusters

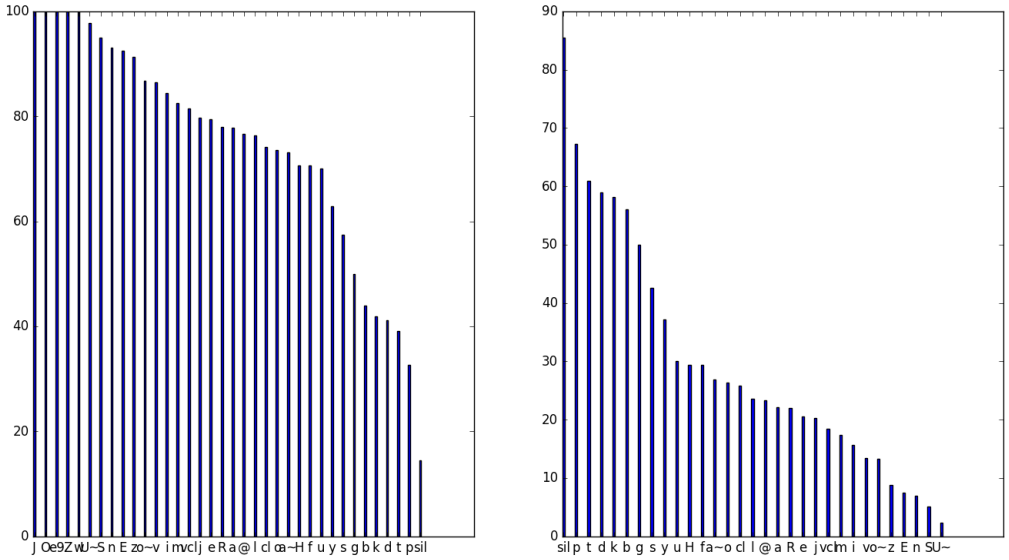

FIGURE 2.10 – meanShift

On observe que certaines fricatives ([s]) et occlusives ([b] et [d]) sont mieux discriminées (clusters 1 et 2), et que d'une façon générales les phonèmes dessinent des paliers dans les clusters, plutôt que des pentes. Le clustering en 6 classes distinctes semble donc donner de légèrement meilleurs résultats lorsqu'on considère les

phonèmes séparément.

3 Analyse avec les coefficients FBANK générés

En ce qui concerne les fbank fait avec python (appliqué sur le fichier Bref80L4M01.wav), le clustering avec kmeans sans plus efficace pour séparer les voisés des non voisés (l'initialisation des centres ne change rien). Le clustering hiérarchique semble mieux pour séparer les consonnes des voyelles. Cependant cette séparation reste assez légère.

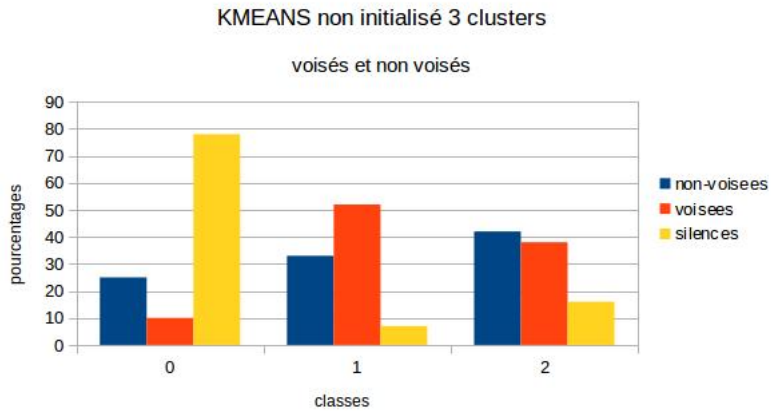

FIGURE 3.1 –

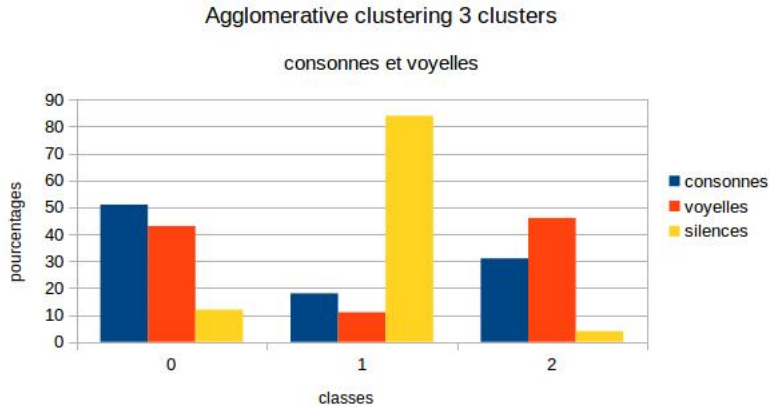

FIGURE 3.2 –

En considérant 6 classes de phonèmes, on n'observe pas vraiment de bons re-

groupements avec un kmeans ou un clustering hiérarchique à 6 classes. L'algorithme meanshift trouve 2 classes et semble associer les silences avec les occlusives, ce qui a un certain sens.

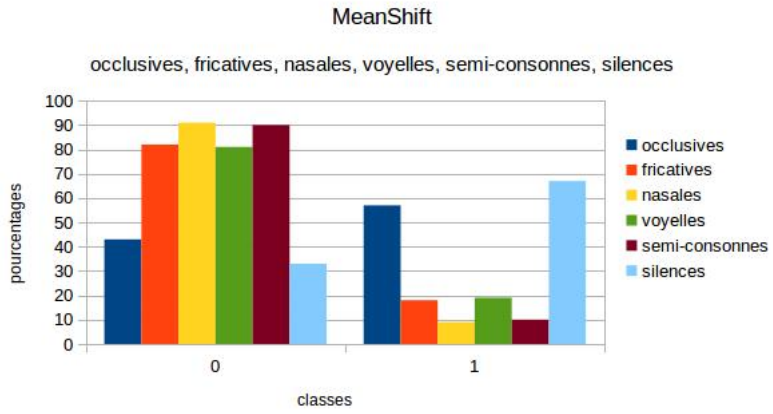

FIGURE 3.3 –

3.1 Travail sur le corpus en japonais :

Clustering (K-means) avec seulement un phonème, mais avec deux locuteurs, un français et un japonais. Voici ici le clustering sur certains phonèmes :

Phonème [b] :
(nombre de phonèmes en japonais : 3, nombre de phonèmes en français : 31)

	classe 0	classe 1
français	51.61	48.38
japonais	100	0

Phonème [p] :
(nombre de phonèmes en japonais : 33, nombre de phonèmes en français : 128)

	classe 0	classe 1
français	29.68	70.31
japonais	36.36	63.63

Phonème [a] :
(nombre de phonèmes en japonais : 63, nombre de phonèmes en français : 315)

	classe 0	classe 1
français	81.58	18.41
japonais	60.31	39.68

Phonème [e] :
(nombre de phonèmes en japonais : 36, nombre de phonèmes en français : 226)

	classe 0	classe 1
français	14.16	85.84
japonais	83.33	16.66

Phonème [i] :
(nombre de phonèmes en japonais : 26, nombre de phonèmes en français : 260)

	classe 0	classe 1
français	68,85	31,15
japonais	7,69	92,30

FIGURE 3.4 –

Pour certains phonèmes (comme le [e] ou le [i]), on distingue plutôt bien les phonèmes prononcés par les français de ceux prononcés par les japonais, ce qui est

très intéressant.

Pour d'autres (le [b] par exemple), les japonais se regroupe bien ensemble mais avec une proportion importante de français. Et enfin, certains phonèmes (comme [a]), un cluster est beaucoup plus important qu'un autre et le résultat n'est pas du tout intéressant.

3.2 Histogrammes des coefficients :

Nous avons fais des histogrammes suivant les bandes de fréquences des coefficient des matrices paramètres. Cependant, sous les différentes transformations, nous n'obtenons pas de distinction entre les phonèmes voisés et non voisés (seul le silence se démarque un peu mieux la plupart du temps). Les histogrammes obtenus sont plus ou moins de ce type :

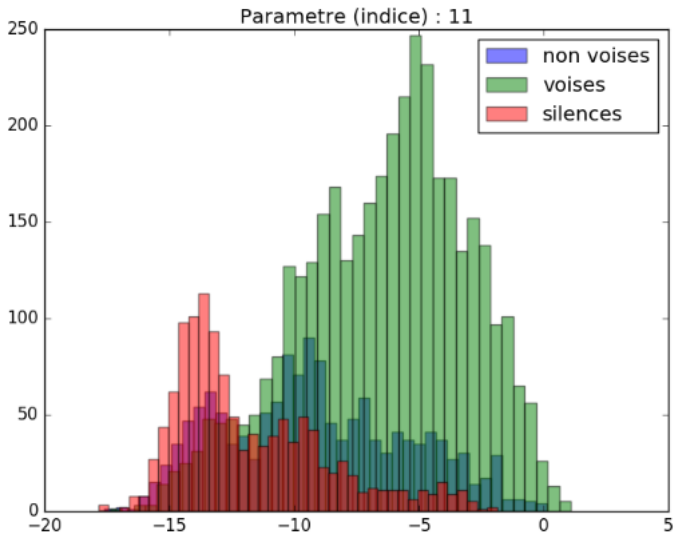

FIGURE 3.5 –

3.3 Utilisation de plus de données

En utilisant deux fichiers sons (Bref80L4M02.wav en plus), c'est à dire en ayant plus de données, les résultats sont légèrement meilleurs : l'écart entre la quantité de voisés et non voisés augmente dans la classe des non voisés et dans la classe des voisés. L'utilisation de plus de deux fichiers sons n'a pas été possible faute de puissance de calcul, mais ces deux fichiers semblent déjà montrer une certaine tendance qui est que les résultats devraient pouvoir être améliorés avec plus de données.

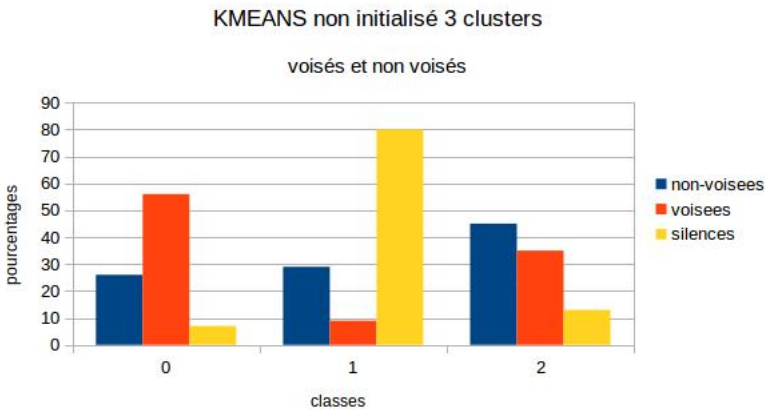
KMEANS non initialisé 3 clusters
voisés et non voisés
pourcentages
90
80
70
60
50
40
30
20
10
0
0
1
2
classes
non-voisees
voisees
silences

FIGURE 3.6 –

4 Analyse avec les coefficients FBANK fournis en format .mat

4.1 k-means non supervisé avec 3 classes

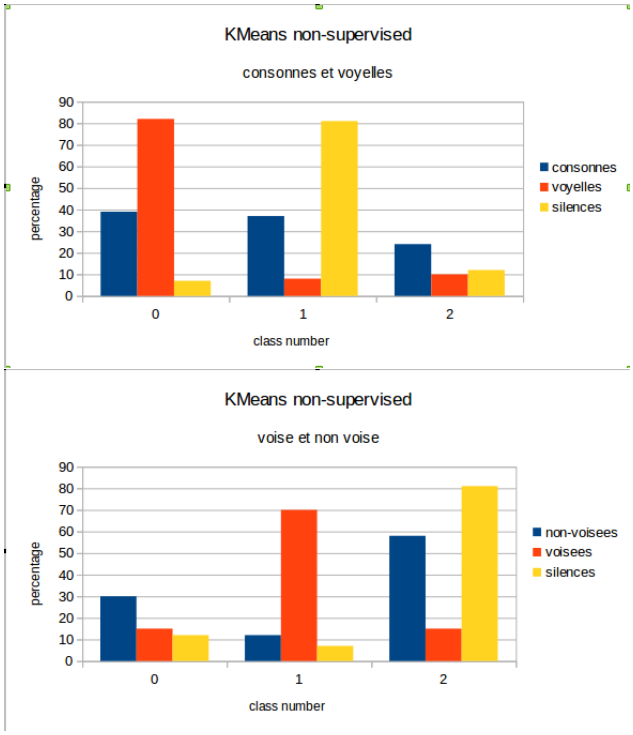

FIGURE 4.1 – K-means 3 classes non supervisé

On remarque qu'avec l'algorithme K-means, on arrive à bien classifer le silence et les sons voisés (ou alors les voyelles).

4.2 k-means non supervisé avec 6 classes

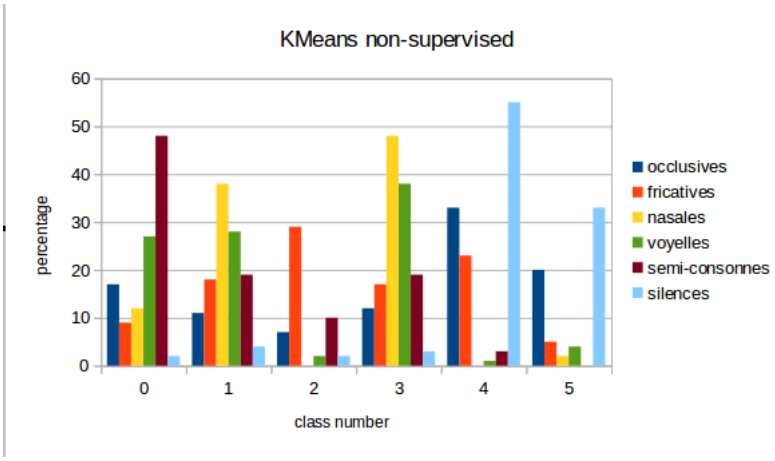

FIGURE 4.2 – K-means 6 classes non supervisé

Pour 6 clusters, on arrive à bien regrouper les semi-consonnes, ainsi que le silence et les consonnes nasales.

k-means non supervisé avec 3 classes :

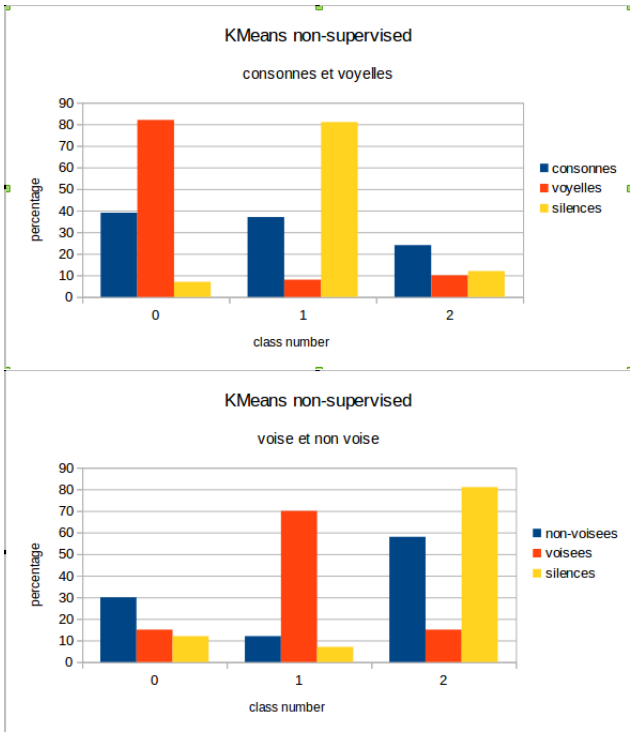

FIGURE 4.3 – K-means 3 classes non supervisé

On remarque qu'avec l'algorithme K-means, on arrive à bien classifer le silence et les sons voisés (ou alors les voyelles).

k-means non supervisé avec 6 classes :

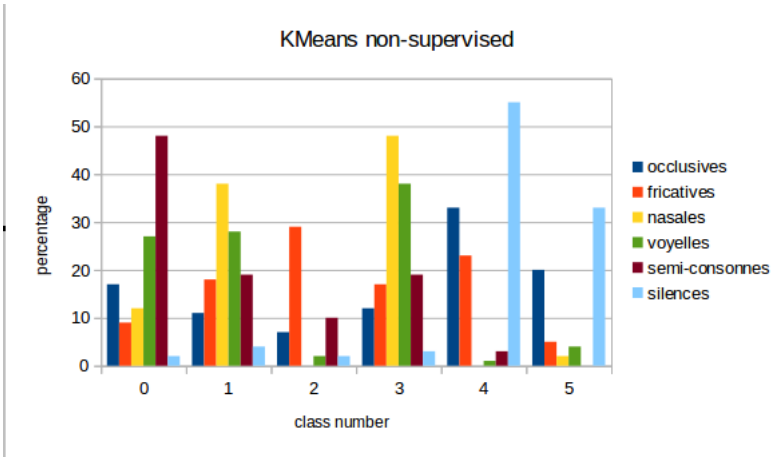

FIGURE 4.4 – K-means 6 classes non supervisé

Pour trouver un meilleur K, on essaye l'algorithme de Gap Statistic. Dans cette algorithme, on calcule deux valeurs qui sont gap et s. Selon l'algorithme, on doit choisir le plus petit K qui satisfait gap(K)> gap(K+1)-s(K+1)

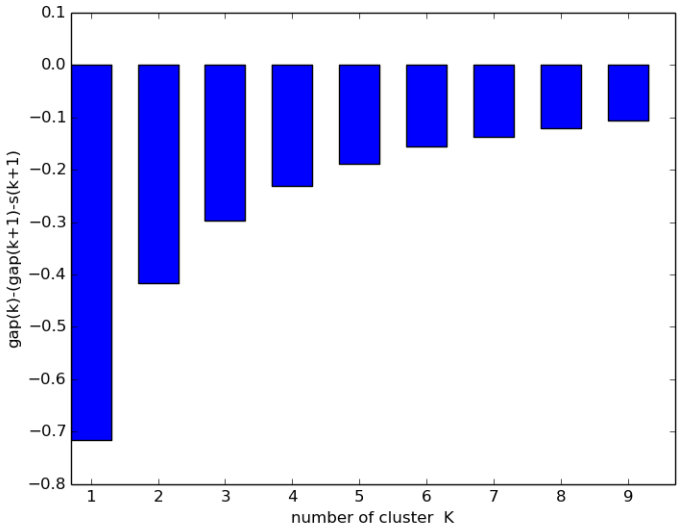

FIGURE 4.5 – Gap Statistic

Mais dans notre cas, on n'a pas trouvé un K qui est assez petit et qui satisfait gap(K) >(gap(K+1)-s(K)).

5 Analyse en ondelettes

5.1 Algorithme de K-MEANS

Les deux figures ci-dessus représentent les résultats de l'algorithme K-means pour la transformée en ondelettes (en utilisant l'ondelette Morlet) du signal. Dans chaque figure, on présente à la fois l'histogramme pour deux matrices :

— La première où l'on a concaténé les lignes correspondant à des fenêtres de 20ms avec un saut de 10ms

— La deuxième où l'on a moyenné ces lignes

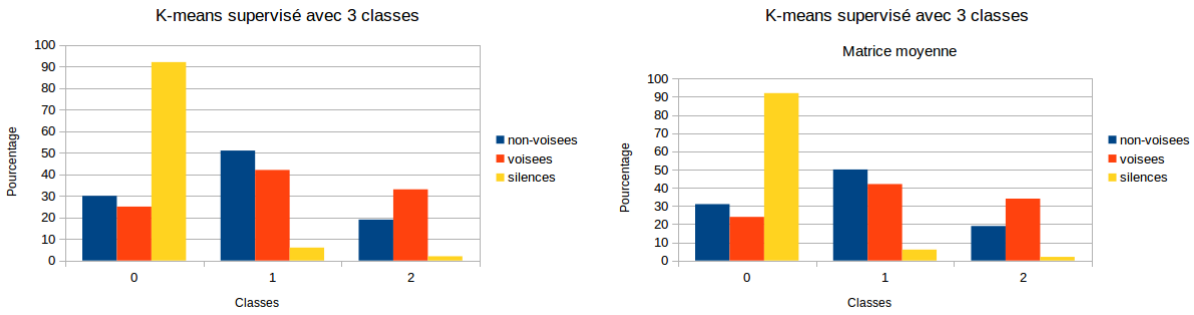

FIGURE 5.1 – K-means 3 classes supervisé

On remarque qu'on a à peu près les mêmes résultats de classification pour les deux types de matrices. On se focalisera donc dans la suite sur la matrice "moyenne" au vu de sa dimension réduite.

5.1.1 K-means supervisé

En gardant la matrice "moyenne", les tests de l'algorithme de classification K-means ont été également faits sur d'autres types d'ondelettes mères, à savoir les ondelettes Paul et dog, les figures ci-dessus montrent les résultats du Kmeans sur les deux types d'ondelettes :

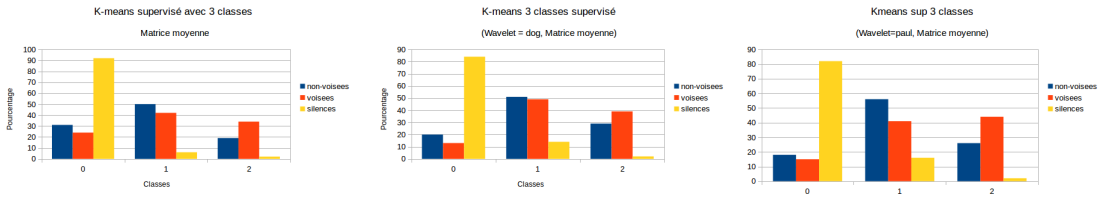

FIGURE 5.2 – Ondelette mère Morlet / Ondelette mère dog / Ondelette mère Paul

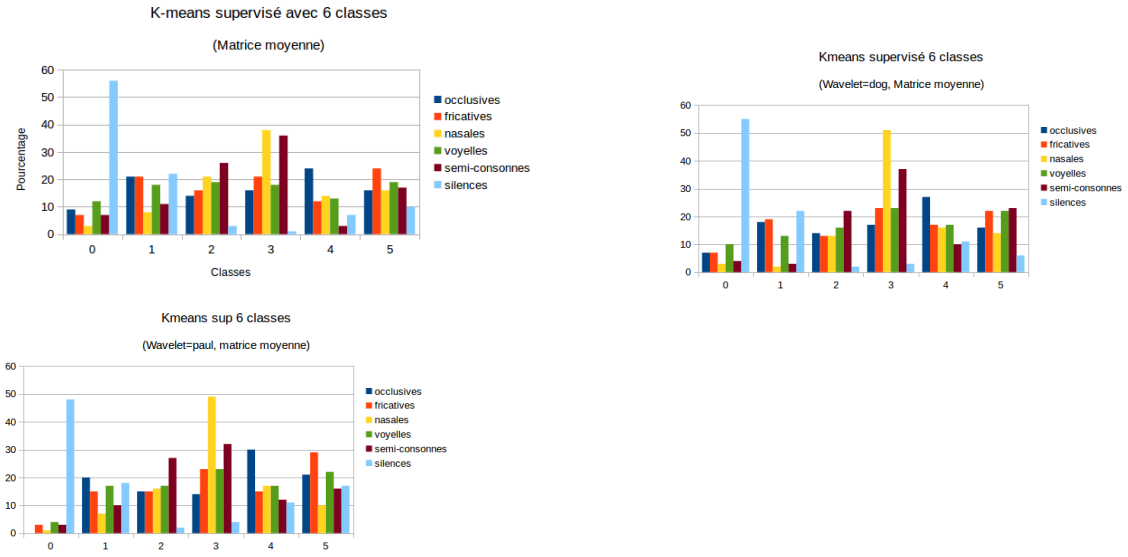

FIGURE 5.3 – Ondelette mère Morlet / Ondelette mère dog / Ondelette mère Paul

Les classes résultantes sont assez mélangées, il n'y a que le silence qui ressort vraiment.Toutefois, on remarque que l'on a une légère meilleure classification quand on utilise l'ondelette mère Paul qui classe un peu mieux les sons non voisés.

5.1.2 K-means non supervisé

En utilisant l'algorithme de Kmeans non supervisé (sans initialisation des centres), on obtient les histogrammes suivants :

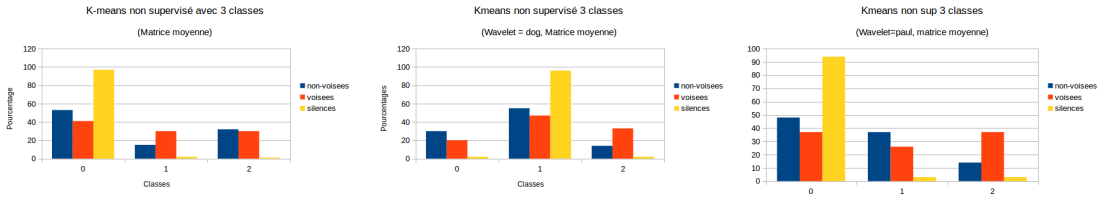

FIGURE 5.4 – Ondelette mère Morlet / Ondelette mère dog / Ondelette mère Paul

Comme pour l'algorithme supervisé, il n'y a pas de séparation distincte des

classes. L'ondelette mère de type Paul donne également les meilleurs résultats de classification.

5.2 Algorithme de l'agglomerative clustering

L'algorithme de l'*Agglomerative Clustering* a été également testé sur la matrice de transformations des ondelettes. Les figures ci-dessous représentent les histogrammes des pourcentages des catégories de sons dans les différentes classes résultantes du clustering.

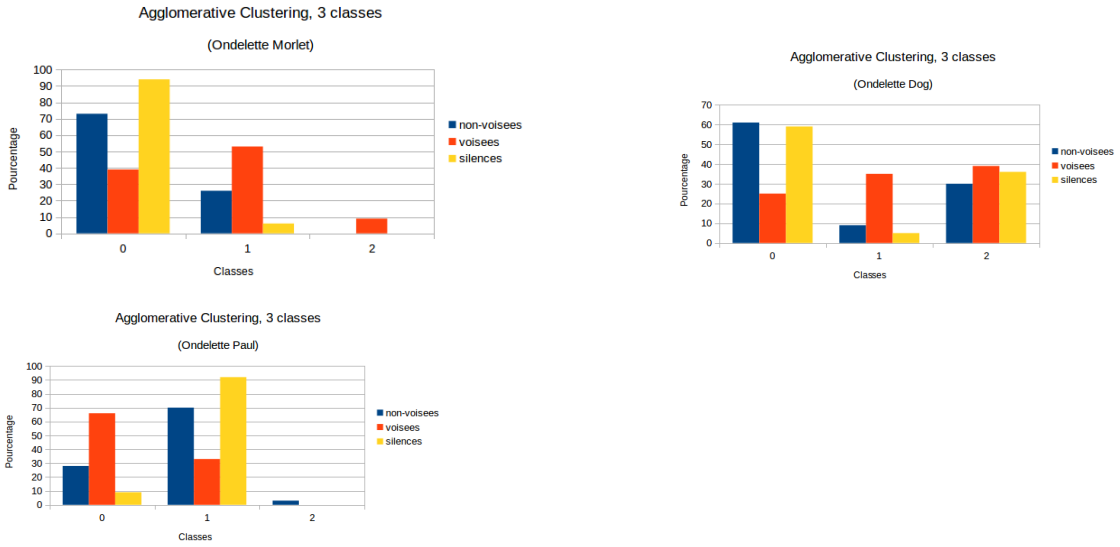

FIGURE 5.5 – Ondelette mère Morlet / Ondelette mère dog / Ondelette mère Paul

Pour la classification en 3 classes, on a les sons voisés qui sont regroupés dans une seule classe en utilisant l'ondelette Paul.

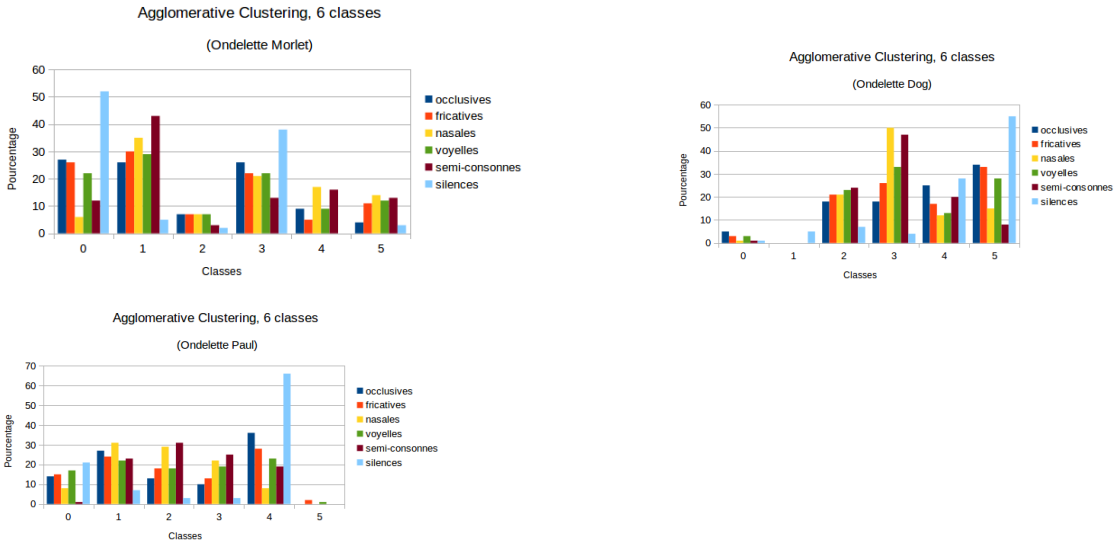

FIGURE 5.6 – Ondelette mère Morlet / Ondelette mère dog / Ondelette mère Paul

Pour la classification en 6 classes, si l'on considère l'ondelette Dog, on a les semi-consonnes, les nasales qui sont dans la même classe, et le silence dans une autre. L'ondelette Paul a regroupé le silence et les occlusives dans la même classe.

Conclusion

Pour les coefficients de mel, le résultat qui ressort c'est la distinction des voisés à 75% pour le clustering en minibatchKmeans supervisé.

Ce qui se dégage du clustering des paramètres mfcc sont les résultats de l'agglomerative clustering : il reconnait très bien les voyelles et les voisées. De plus, pour à peu près tous les algorithmes de clustering, une bonne partie des occlusives sont souvent assimilées à des silences, que ce soit pour 2, 3 ou 6 classes. Les nasales sont également à peu près toujours bien regroupées, mais sont rangées avec les fricatives et les voyelles.

Dans les coefficients FBank générés par Matlab, on arrive à bien classifier le silence et les sons voisés (ou alors les voyelles) en utilisant l'algorithme K-means.

Concernant la transformée en ondelettes du signal, l'utilisation de l'ondelette Paul dans l'algorithme K-means supervisé donne des résultats assez satisfaisants : Pour 3 clusters, on arrive à regrouper le silence et les sons non voisés dans deux classes séparées, et pour 6 clusters, on regroupe le silence et les consonnes nasales. Pour l'algorithme *agglomerative clustering*, c'est l'ondelette Morlet et Dog qui donnent une meilleure classification pour 6 clusters : On arrive à regrouper le silence, ainsi que les consonnes nasales et semi-consonnes. Pour 3 clusters, l'utilisation de l'ondelette Paul est la seule à pouvoir regrouper les sons voisés.