

ARIA



COMPTE-RENDU - PROJET SIGNAL

ÉTUDE ET CARACTÉRISATION DE SIGNAUX D'ACTIVITÉ
HUMAINE

ITAI ZEHAVI - CLARA TROMPETTE

Table des matières

1	Introduction	2
1.1	Objectifs et démarche	2
2	Les données	3
2.1	Présentation des données	3
2.2	Selection des données pertinentes	3
2.3	Limites et problèmes potentiels du dataset	4
2.3.1	Peu d'informations associées au dataset	4
2.3.2	Découpage des données	5
2.4	Pré-traitement des données	5
3	Démarche	5
3.1	Calcul des features	5
3.1.1	Features Classiques	5
3.2	Features basées sur la littérature	7
3.3	Features basées sur les motifs	9
4	Sélection des features pertinentes	16
4.1	Analyse en composante principale	16
4.2	Affichage des features "clusterisantes"	18
4.2.1	Énergie relative dans la bande de 1.5Hz à 2Hz (feature 8)	18
4.2.2	"Energie d'auto-corrélation relative" avec un décalage temporel de $\tau \in [0.22s, 0.25s]$ (feature 25)	19
4.2.3	Distance au motif n°4 (feature 49)	20
5	Discussion	21
6	Conclusion	22

1 Introduction

L'analyse des séries temporelles est une tâche importante dans de nombreux domaines, notamment dans le domaine biomédical. En effet, l'analyse de signaux peut aider à la compréhension et au diagnostic de maladies cérébrales ou cardiaques, avec l'analyse d'électroencéphalogrammes ou d'électrocardiogrammes, mais permet également l'étude de données biomécaniques. En particulier, l'étude biomécanique de la marche humaine est importante pour la compréhension des troubles physiques ou neurologiques affectant la marche, mais est également primordiale dans le développement de prothèses par exemple.

Ce mini-projet se concentre sur l'identification de caractéristiques spécifiques dans des données d'accélérométrie afin de distinguer deux types d'activités physiques : la marche et la montée d'escaliers. L'objectif est de trouver des features pertinentes, c'est-à-dire des caractéristiques mesurables du signal, qui permettent de différencier ces deux actions.

Les accéléromètres, utilisés dans les dispositifs portables et les systèmes de surveillance de l'activité physique, capturent des signaux riches mais souvent complexes à interpréter. La classification de ces signaux en fonction de l'activité physique pratiquée est un défi, car les signaux peuvent être influencés par de nombreux facteurs, tels que la cadence, l'intensité du mouvement ou la posture de la personne.

1.1 Objectifs et démarche

L'objectif de ce mini-projet est d'extraire des features (caractéristiques) permettant de distinguer différents groupes de séries temporelles, dans notre cas, afin d'identifier si une personne est en train de marcher ou de monter des escaliers à partir de données d'accélérométrie.

Pour ce faire, nous allons suivre une démarche structurée en plusieurs étapes. Tout d'abord, il sera essentiel de sélectionner le signal le plus pertinent parmi ceux disponibles dans notre dataset pour cette tâche de classification. Ensuite, nous aborderons le problème en trois principales étapes :

1. **Étude des données** : Nous allons commencer par étudier nos données, c'est à dire comprendre leur nature, les prétraitements qu'elles ont déjà subis et ceux que nous devrons appliquer, puis identifier celles qui sont pertinentes dans le cadre de notre étude.
2. **Calcul des features** : Ensuite, nous allons calculer les features classiquement utilisées dans l'analyse de données, mais également les features intéressantes à étudier d'après la littérature. Enfin, nous calculerons des features à partir d'une détection de motif.
3. **Sélection des features pertinentes** : Enfin, nous procéderons à une sélection des trois features les plus pertinentes parmi l'ensemble de notre jeu de features et nous les utiliserons pour séparer notre dataset en deux groupes de signaux.

2 Les données

2.1 Présentation des données

Le data-set utilisé est tiré de l'étude [Human Activity Recognition using smartphones](#) publiée en 2013, qui étudie les signaux associés aux activités humaines à l'aide de mesures effectuées par un smartphone.

Les expériences de cette étude ont été réalisées sur 30 participants volontaires, âgés de 19 à 48 ans. Les participants, équipés d'un smartphone (Samsung Galaxy SII) fixé au niveau d'une ceinture, devaient réaliser les activités suivantes : marcher, monter les escaliers, descendre les escaliers, être debout, assis et allongé. Le smartphone est équipé de systèmes de mesure intégrés : l'accéléromètre, permettant la mesure triaxiale de l'accélération linéaire , et le gyromètre, permettant la mesure triaxiale de la vitesse angulaire. Les mesures se font avec une fréquence d'échantillonnage de 50Hz. Les mesures ont été réalisées en deux temps, avec la séparation des activités statiques (debout, assis, allongé) et dynamiques (marche, montée et descente des escaliers). Les activités d'un même type sont effectuées consécutivement par les sujets et la mesure se fait de manière continue.

Les signaux ainsi obtenus ont été pré-traitées par les chercheurs. Le premier traitement consistait à réduire le bruit en appliquant un filtre Butterworth passe-bas avec une fréquence de coupure de 20Hz. Ceci permet de capturer la totalité des mouvements, dans la mesure où les mouvements humains se font tous dans des fréquences inférieures ou égales à 15Hz. Dans un second temps, les signaux ont été découpés en fenêtres de 2,56 secondes avec 50 pourcent d'overlap entre les fenêtres. Cette valeur de 2,56s pour le découpage a été choisie de sorte à ce que chaque fenêtre contienne au moins un cycle entier de marche (i.e 2 pas) , sachant que la vitesse minimale de la marche normale est de 1,5 pas/s.

Nous avons donc à notre disposition 9 fichiers de données "brutes" :

- les accélérations selon les axes x (avant-arrière), y (gauche-droite) et z (haut-bas) mesurées par l'accéléromètre (total_acc, unité = m/s²),
- les accélérations du corps selon les axes x,y et z (body_acc, unité m/s²), qui correspondent aux accélérations totales déduites de la gravité,
- les vitesses angulaires selon x, y et z mesurées par le gyromètre (body_gyro, unité rad/s).

Dans chacun de ces fichiers, il y a 7351 lignes correspondant aux signaux, chaque ligne étant composée de 128 éléments. Un signal correspond ainsi à 2.56s de mesure échantillonné à 50Hz.

2.2 Sélection des données pertinentes

Dans le cadre de notre projet, nous ne devons analyser qu'un seul type de signal, il est donc nécessaire de sélectionner le signal le plus pertinent, parmi les signaux d'accélération ou de vitesse angulaire selon l'axe x, y ou z. A priori, tous les signaux ont une certaine pertinence a être utilisés pour différencier la marche à plat et la montée d'escaliers, étant

donné que les accélérations et les vitesses angulaires selon x, y et z varient différemment lors de ces deux activités. En effet, on peut intuitivement remarquer que la structure du pas, les variations de position, d'inclinaison et de vitesse ne sont pas identiques que l'on marche à plat ou en montée d'escalier, et ce dans toutes les directions de l'espace.

Nous avons cependant fait le choix d'étudier les signaux d'accélérations selon l'axe z. Les données de gyrométrie étant plus complexes à comprendre, à relier aux mouvements réels et donc à analyser, elles n'étaient pas les plus pertinentes pour notre étude. Pour ce qui est de l'accélération, l'étude de l'axe z semble la plus adaptée pour discriminer les deux classes, dans la mesure où la montée d'escaliers est associée à une déplacement selon l'axe z importante contrairement à la marche à plat. Les accélérations selon x et y, associées respectivement au déplacement en avant et aux oscillations latérales lors du déplacement, varient probablement différemment lors des deux activités, mais ces variations sont moins intuitivement appréciables que les variations selon l'axe z.

Nous avons ainsi récupéré le fichier *total_acc_z*, contenant donc 7351 signaux de 2.56 secondes (lignes), chacun étant composé de 128 points de mesure (colonnes).

Étant donné que le jeu de données initial contient des signaux correspondant aux 6 activités décrites précédemment, nous avons également dû récupérer les signaux correspondant aux deux activités que nous souhaitons étudier (la marche et la montée d'escaliers).

2.3 Limites et problèmes potentiels du dataset

2.3.1 Peu d'informations associées au dataset

Le premier problème que comporte le dataset est le manque d'information, notamment concernant les sujets participants à l'étude, ce qui nous empêche d'appréhender la variabilité potentielle dans les données. En effet, nous n'avons pas d'informations sur les individus (à part la large fourchette d'âge de 19-48 ans). Hors, l'âge, le sexe, la taille et le poids sont des caractéristiques ayant une influence sur de nombreux paramètres de la marche ([source](#)). En effet, la vitesse, la symétrie et la régularité de la marche sont significativement influencées par les différentes caractéristiques citées précédemment. Ceci a ainsi une influence sur les différentes mesures de notre dataset. Les variations d'accélérations selon l'axe x (avant-arrière) dépendent par exemple de la vitesse de la marche et de taille de l'individu. L'accélération sur l'axe y (axe gauche-droite, caractérisant correspondant aux oscillations latérales lors de la marche) ne sera pas la même selon la position du centre de gravité et la capacité à maintenir l'équilibre des individus, qui sont des paramètres qui varient en fonction du sexe et de la masse corporelle par exemple.

Un autre point notable dans notre cas est la notion de "range of motion" (ROM) ou amplitude des mouvements. Les hommes et les femmes ont en effet des ROM au niveau des hanches, genoux et chevilles différentes lors de la marche. Ces différences d'amplitudes s'observent notamment sur l'axe x (avant-arrière) et sur l'axe z (haut-bas), sur lequel nous nous concentrerons dans notre étude. Il a en effet été établi que, lors de la marche normale, les hommes ont une amplitude de mouvement au niveau des hanches plus importante que les femmes ([source](#)). Cet amplitude est également influencée par d'autres caractéristiques

comme la taille ou la masse corporelle.

Ces différences observables lors de la marche à plat sont également très probablement présentes durant d'autres activités comme la montée des escaliers, bien que ces phénomènes n'aient à priori pas été spécifiquement étudiés.

Finalement, les données d'accélération selon z et les informations que nous allons en extraire sont influencées par un certain nombre de caractéristiques des individus auxquelles n'ont n'avons pas accès. Ceci nous empêche d'anticiper et de prendre en compte les variations intra-classe, ce qui induit une analyse moins éclairée des données.

2.3.2 Découpage des données

Un autre problème potentiel est la méthode de découpage des données.

Dans les expériences réalisées, les individus réalisent les différentes activités les unes après les autres, il n'y a pas des enregistrement distincts pour chaque activité. De ce fait, le découpage réalisé par les chercheurs (qui consiste à découper le signal en fenêtres de 2.56 secondes) ne respecte pas forcément les frontières entre les différentes activités. Ainsi, il y a très probablement un certain nombre de fenêtres qui contiennent une transition entre deux activités, ce qui peut les rendre plus difficile à classer.

2.4 Pré-traitement des données

Le prétraitement réaliser est assez basique. En effet il se base sur l'hypothèse que les informations d'intérêt ont une fréquence comprise entre $f \in [0.2Hz, 14Hz]$. Ainsi, nous considérons toutes les autres fréquences comme étant un bruit de mesure. Ainsi, nous appliquons à tous nos signaux un detrending à l'aide d'un polynome d'ordre 3 et un filtre du butterworth selectif (ordre 5) avec une fréquence de coupure de $15Hz$.

3 Démarche

3.1 Calcul des features

3.1.1 Features Classiques

Dans un premier temps, nous allons nous intéresser aux features classiquement étudiées dans le cas des séries temporelles ainsi que les features qui nous semblent intuitivement pertinentes. Ces caractéristiques et leur pertinences sont listées ci-après. Elles ont également été calculées pour tous les signaux du dataset.

- la **moyenne** (feature 1) : la moyenne du signal correspond à la somme des valeurs d'accélération divisée par le nombre de points de mesure (128). Cette caractéristique pourrait être pertinente pour séparer la marche et la montée d'escalier si une de ces activités est associée à de plus grandes valeurs d'accélération selon z. On peut supposer à priori que la montée d'escaliers, étant associée à des

mouvements plus amples selon l'axe z, est l'activité associée à une moyenne d'accélération la plus importante. Nous l'appliquons sur les signaux dont on a conservé la tendance (signaux non dé-trendés).

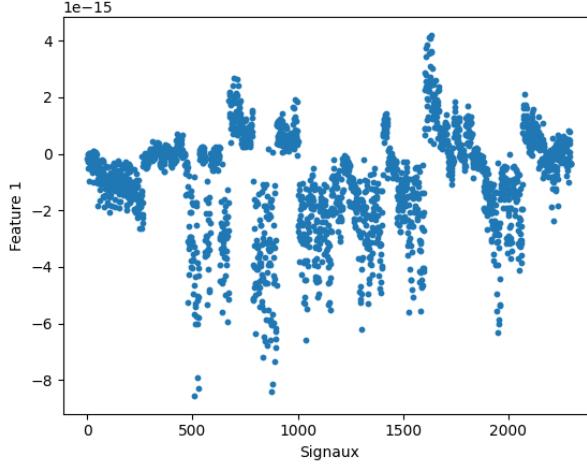


FIGURE 1 – Valeur de la moyenne pour chaque signal du dataset.

- la **variance** (feature 2) : la variance est définie comme la moyenne des carrés des écarts à la moyenne. C'est une caractéristique permettant d'estimer l'intensité des variations du signal autour de sa valeur moyenne. On peut faire l'hypothèse que lors de la montée d'escaliers, les variations de l'accélération selon z sont plus importantes et que cette activité est donc associée à une variance plus forte. Logiquement, les features d'écart-type et du calcul d'amplitude sont très corrélées à la variance et donne donc des résultats très similiaires.

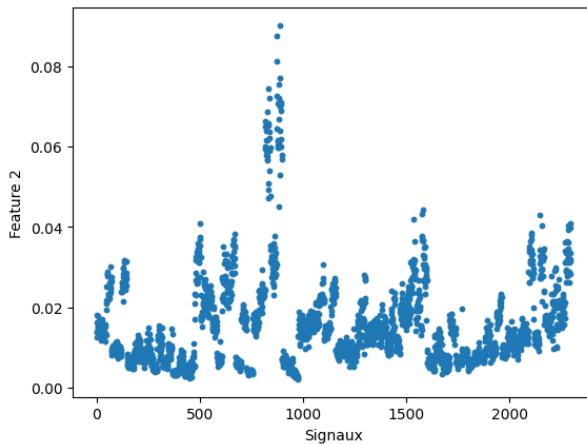


FIGURE 2 – Valeur de la variance pour chaque signal du dataset.

Les plots montrent que les features calculées précédemment, bien que globalement différentes pour chaque signal, ne permettent pas de mettre en évidence deux groupes de signaux distincts et ne sont pas ainsi de bonnes caractéristiques à conserver pour notre étude et pour la séparation des signaux.

3.2 Features basées sur la littérature

Pour identifier d'autres caractéristiques potentiellement pertinentes pour notre projet, nous nous sommes intéressés aux études scientifiques qui explorent les caractéristiques physiques des activités humaines telles que la marche.

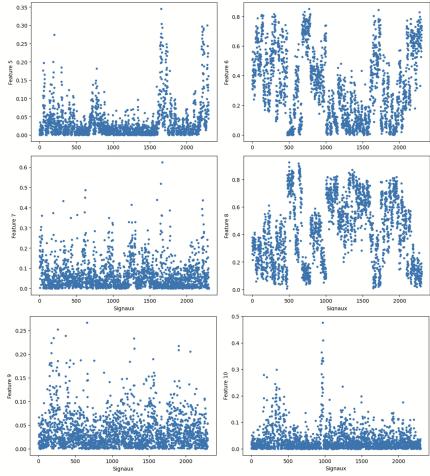
Les articles étudiés sont les suivants :

- (1) Fadel, William F et al. "Differentiating Between Walking and Stair Climbing Using Raw Accelerometry Data." *Statistics in biosciences* vol. 11,2 (2019) : 334-354. doi :10.1007/s12561-019-09241-7. Cette étude se concentre sur l'étude des activités de marche à plat, de montée et descente des escaliers avec des données d'accélérométrie.
- (2) Psaltos DJ, Mamashli F et al. "Wearable-Based Stair Climb Power Estimation and Activity Classification." *Sensors*. 2022 ; 22(17) :6600. <https://doi.org/10.3390/s22176600>. Cet article décrit des méthodes d'estimation de la mesure clinique "stair climb power" et de classification de signaux d'accélérométrie de montée d'escaliers.
- (3) Capela NA, Lemaire ED, Baddour N. "Feature Selection for Wearable Smartphone-Based Human Activity Recognition with Able bodied, Elderly, and Stroke Patients." *PLoS ONE*. 2015 ; 10(4) : e0124414. <https://doi.org/10.1371/journal.pone.0124414>. Cet article se focalise sur l'étude de données d'accéléromètre et de gyroscope mesurée durant la réalisation de nombreuses activités quotidiennes (incluant la marche dans plusieurs contextes) chez différents types de population.

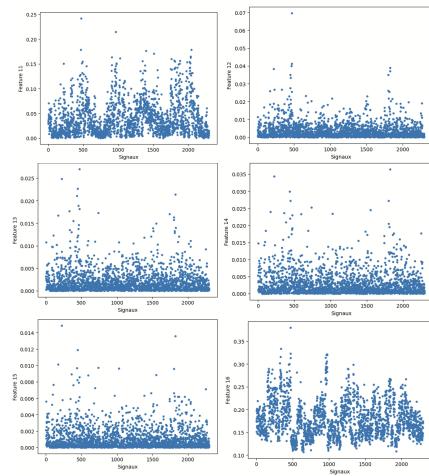
Ces études nous renseignent à la fois sur les notions physiques utiles pour comprendre la marche, mais nous apportent également des informations sur les différentes features plus ou moins pertinentes permettant d'étudier les activités de marche à plat et de montée des escaliers. Parmi les caractéristiques explorées dans ces articles, nous avons choisi de tester les suivantes :

- l'**énergie fréquentielle** (features 5 à 16) : cette énergie, calculée à partir de la Fast Fourier Transform (FFT), permet d'estimer l'amplitude du signal en fonction de la fréquence. Cette caractéristique a été calculée et démontrée comme pertinente dans les 3 articles cités précédemment. Dans le cadre de notre étude, nous nous attendons à des profils différentes d'énergie fréquentielle pour les deux types d'activités étudiées.

Pour notre étude, nous avons découpé les fréquences entre 0 et 6Hz en 12 bandes et calculé l'énergie sur chacune de ces bandes.



(a) Valeur de la densité spectrale pour différents intervalles de fréquence et pour chaque signal du dataset.



(b) Valeur de la densité spectrale pour différents intervalles de fréquence et pour chaque signal du dataset.

Les plots nous permettent de constater que la distribution des signaux est différente selon les bandes de fréquences considérées. Ainsi, pour la deuxième (0.5-1Hz, feature 6) et la quatrième bande (1.5-2Hz, feature 8), la distribution est beaucoup plus large que pour les autres bandes, où l'énergie fréquentielle est très proche de 0 pour la majorité des signaux.

Une grande distribution sur les graphiques, bien qu'elle ne s'accompagne pas nécessairement d'une distinction nette de deux groupes de signaux différents, est l'indicatif de bonnes features permettant de distinguer nos deux types de signaux.

- **l'auto-corrélation** (features 17 à 45) : cette mesure permet d'évaluer la similitude d'un signal avec lui-même à différents décalages temporels. Pour la marche à plat, on peut s'attendre à une auto-corrélation importante, le mouvement étant généralement très régulier. Pour la montée d'escaliers, qui implique plus d'effort et donc potentiellement des petites variations de mouvement à chaque pas, on peut supposer que l'auto-corrélation est plus faible. Cette caractéristique a notamment été qualifiée de pertinente dans l'article (3).

Pour notre étude, nous avons calculé l'autocorrélation pour 50 décalages temporels différents. Puis, nous avons calculé "l'énergie d'auto-corrélation relative" pour 30 intervalles de décalage temporel. "l'énergie d'auto-corrélation relative" correspond au fait de calculer l'intégrale sur un plage de décalage temporelle τ et de diviser l'intégrale total.

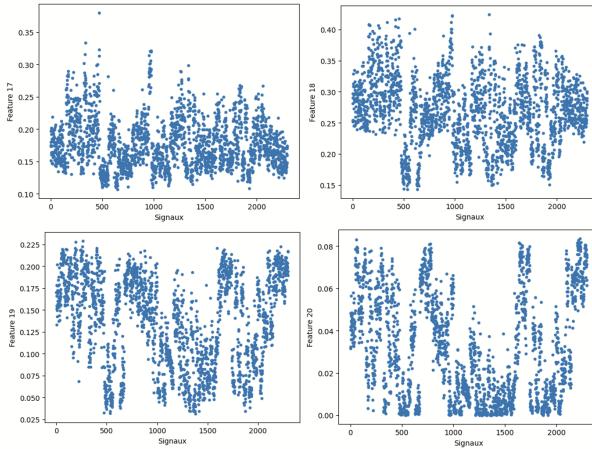


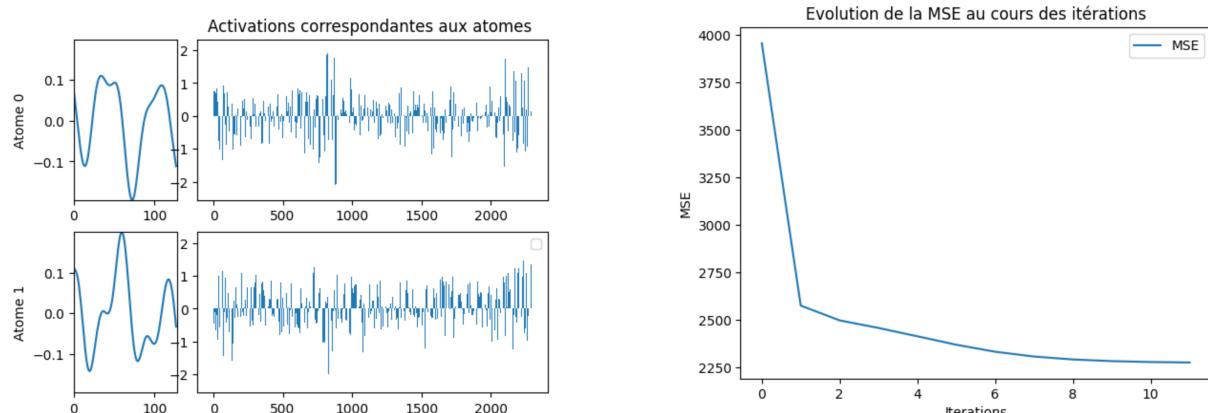
FIGURE 4 – "Energie d'auto-corrélation relative" pour différents intervalles de décalage temporel.

Comme dans le cas de l'énergie par bande de fréquence, les calculs de l'auto-corrélation avec certaines valeurs de décalages temporels sont associées à des distributions relativement larges des données, permettant potentiellement de distinguer différents types de signaux.

3.3 Features basées sur les motifs

Apprentissage de dictionnaire pour la classification :

L'idée de cette sous-partie est de pouvoir distinguer les deux groupes de signaux à l'aide d'une détection de motifs. La première approche est celle de la création de dictionnaire à deux atomes afin de représenter l'ensemble des signaux du dataset. Nous obtenons les résultats suivants :



(a) Atomes et liste des activations correspondantes.

(b) Évolution de l'erreur quadratique au cours des itérations.

FIGURE 5 – Analyse de l'approche par dictionnaire : atomes, activations et erreur quadratique

Nous observons que l'erreur quadratique totale reste très importante à la dernière

itération. En effet, elle prend une valeur de 2276 tandis que l'erreur quadratique pour des atomes et des activations aléatoires est de 4000. L'amélioration liée à l'apprentissage de dictionnaire est donc faible. Nous pouvons d'ailleurs l'observer sur les figures suivantes :

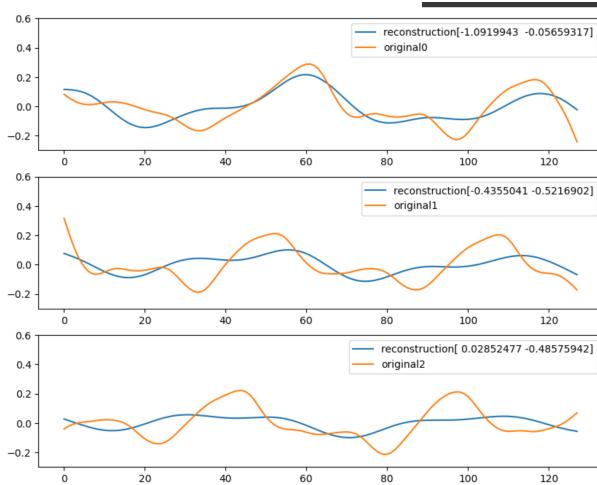


FIGURE 6 – Reconstruction de 3 signaux à l'aide de deux atomes et des activations calculer plus tôt

Sur ces trois figures, nous pouvons observer que la reconstruction paraît pratiquement aléatoire. Ce résultat est pourtant assez cohérent, du fait du nombre très faible d'atome. Bien que les signaux présentent probablement deux motifs majeurs, la reconstruction de ces derniers à l'aide d'atomes ne prend pas en considération le déphasage. Ainsi, même si l'atome présente un des deux motifs majeurs, il sera dans l'incapacité de reconstruire un signal ayant le même motif s'ils ne sont pas en phase. Ainsi, la notion de phase rend l'utilisation de dictionnaire pour la classification inefficace sur notre dataset.

Approche basée sur la détection de motifs :

Nous cherchons donc une méthode permettant l'utilisation des motifs présents dans les signaux pour la classification qui soit insensible au déphasage. L'approche basée sur la détection de motifs paraît être une solution pertinente.

Nous appliquons donc cette méthode à notre dataset. La méthode consiste à sélectionner un intervalle d'une taille L et de calculer sa distance euclidienne normalisée avec le reste du signal, ce qui nous donne une courbe des distances entre notre intervalle et tous les autres sous-intervalles du signal. Ensuite, nous sélectionnons le minimum de cette courbe des distances (en excluant la zone de notre intervalle de référence à plus ou moins L), puis nous ajoutons ce point dans le Matrix profile à la position correspondante à l'intervalle de référence. Ensuite, nous répétons ces étapes pour le sous-intervalle suivant, etc.

Une fois que nous avons balayé tous les sous-intervalles de taille L du signal, nous avons notre matrix profile qui répertorie pour chaque intervalle du signal la distance minimale entre ce dernier et les autres intervalles du signal. Ainsi, une faible valeur dans le matrix profile signifie que l'intervalle correspondant s'est répété au moins une autre dans le signal. Plus cette valeur est faible, plus le motif détecté est proche de la référence.

Nous sélectionnons donc l'intervalle correspondant à la valeur minimale du Matrix Profile et considérons que ce dernier est un motif. Comme nous faisons de la classification, nous comptons toutes les trames contenant ce motif en définissant un seuil de ressemblance, nous supprimons ensuite ces dernières afin de chercher un second motif.

Nous appliquons à présent l'algorithme à nos données :



FIGURE 7 – Temps de calcul estimer : 12h

Le temps de calcul estimé est de 12h sur notre jeu de données pour la détection de seulement un motif, ce qui est conséquent, en particulier si nous devons lancer cet algorithme à plusieurs reprise afin de régler le seuil de ressemblance. Cela est dû à la complexité de calcul du Matrix profile qui est de $O(N^3 \log(N))$.

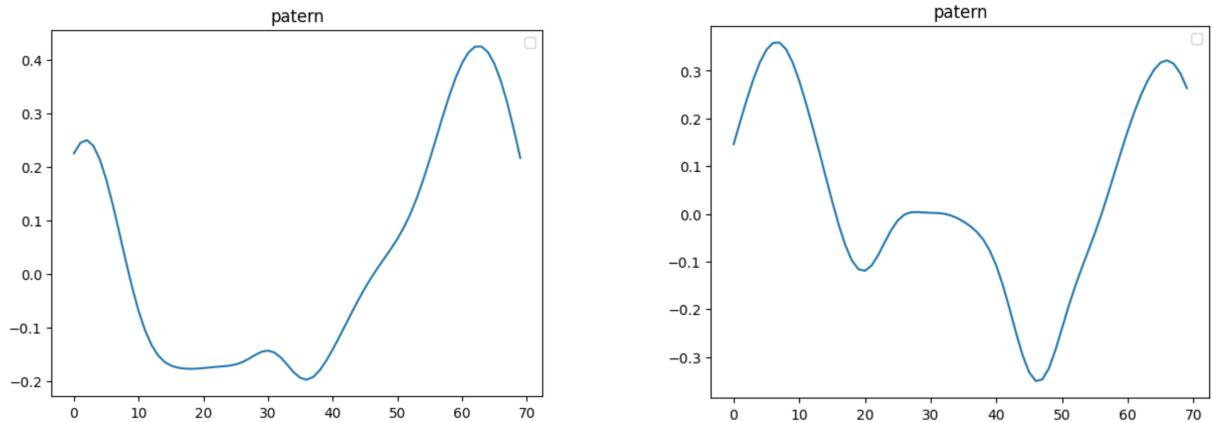
Nous devons donc chercher une solution nous permettant de réduire le temps de calcul de notre algorithme. En faisant l'hypothèse que chaque trame a une grande probabilité de contenir un des deux motifs majeurs, une idée serait de sélectionner un motif sur 10% des trames du dataset. Puis de compter le nombres de signaux contenant ce motif et de les supprimer afin refaire une détection de motif sur le même nombre de signaux et ainsi obtenir les deux motifs majeurs. Cette méthode induit une approximation, dans la mesure où le motif obtenu n'est extrait que sur une partie du dataset.



FIGURE 8 – Temps de calcul avec l'optimisation : 12min

Le temps de calcul pour la détection d'un motif passe à 12min. L'apport de l'optimisation sur le temps de calcul est donc considérable (le temps de calcul est divisé par 50). Cependant, en contrepartie, la qualité du motif détecté diminue. Plus le temps de calcul est réduit (donc moins nous considérons de trames pour le matrix profile), moins la probabilité d'avoir le meilleur motif du dataset est importante.

Les deux premiers motifs détectés sont les suivants :



(a) Premier motif détecté. Le motif se répète 598 fois sur le dataset avec un seuil de ressemblance fixé à 4.

(b) Second motif détecté. Le motif se répète 617 fois sur le reste du dataset avec un seuil de ressemblance fixé à 4.

FIGURE 9 – Motifs détectés par la méthode 1 avec un seuil de ressemblance de 4

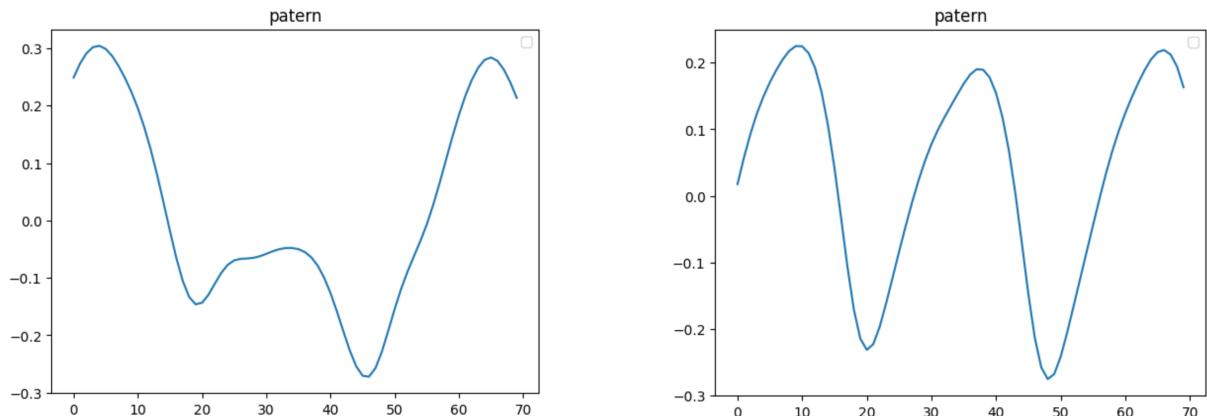
Notre premier motif se répète dans 598 signaux / 2298 signaux au total. En supposant que notre dataset est équilibré, nous devrions avoir environ 1100 signaux possédant ce premier motif. Nous sommes donc loin des chiffres attendus. Ainsi, nos deux motifs sont présents dans 1215 signaux / 2298 signaux aux total. Nous parvenons à classer 53% du dataset avec cette approche.

Les résultats précédents mettent en évidence une limite de cette méthode. En effet, cette méthode sélectionne les motifs qui se sont répétés parfaitement au moins une fois, ce qui n'est pas exactement ce que nous cherchons dans notre cas de figure. En effet, l'objectif est de pouvoir classifier nos signaux en deux groupes à l'aide de la détection de deux motifs majeurs, représentant chacun une classe de signaux.

Une idée serait de modifier le calcul du matrix profil et faire intervenir le nombre de fois où le motif se répète. Ainsi, nous allons ajouter à l'algorithme les étapes suivante :

1. De la même manière que dans l'algorithme précédent nous commençons par calculer pour chaque intervalle, la distance euclidienne normalisée avec le reste du signal.
2. Ensuite, nous comptons tous les minimum locaux inférieurs à un certain seuil.
3. Puis nous attribuons au matrix profile amélioré le minimum global (hors intervalle considéré) divisé par le nombre de minimum local.
4. Enfin, le motif sélectionné est celui présentant la valeur la plus basse dans ce matrix profile amélioré.

En appliquant cette méthode nous obtenons les résultats suivant :



(a) Premier motif. Le motif se répète 945 fois.

(b) Second motif. Ce dernier se répète 347 fois.

FIGURE 10 – Motifs obtenus par l'approche du Matrix profile amélioré

Nous remarquons une augmentation de la répétition de nos motifs détectés pour le même seuil. Cependant, nous observons que les répétitions ne sont pas équilibrées entre les deux motifs : le premier se répète 3x plus que le second. De plus ces deux motifs ne couvrent que 59% du dataset. Le problème vient donc de l'hypothèse que nous avons faite.

En effet, nous avons supposé que deux motifs seraient représentatifs du dataset, dans la mesure où nous voulons séparer nos signaux en 2 catégories (un groupe pour "marcher" et un autre pour "monter les escalier"). Cependant, il est important de noter que ces mesures ont été fait sur plusieurs individus d'âge, de taille, de sexe et de masse corporelle différents, et ainsi probablement de manière de se déplacer différentes. Cela peut donc induire l'existence de plusieurs motifs plus ou moins différents pour la même activité.

Pour prendre en compte cette possibilité, nous allons laisser tourner notre algorithme détecteur de motifs jusqu'à ce qu'il y en ai suffisamment pour représenter 85% du dataset.

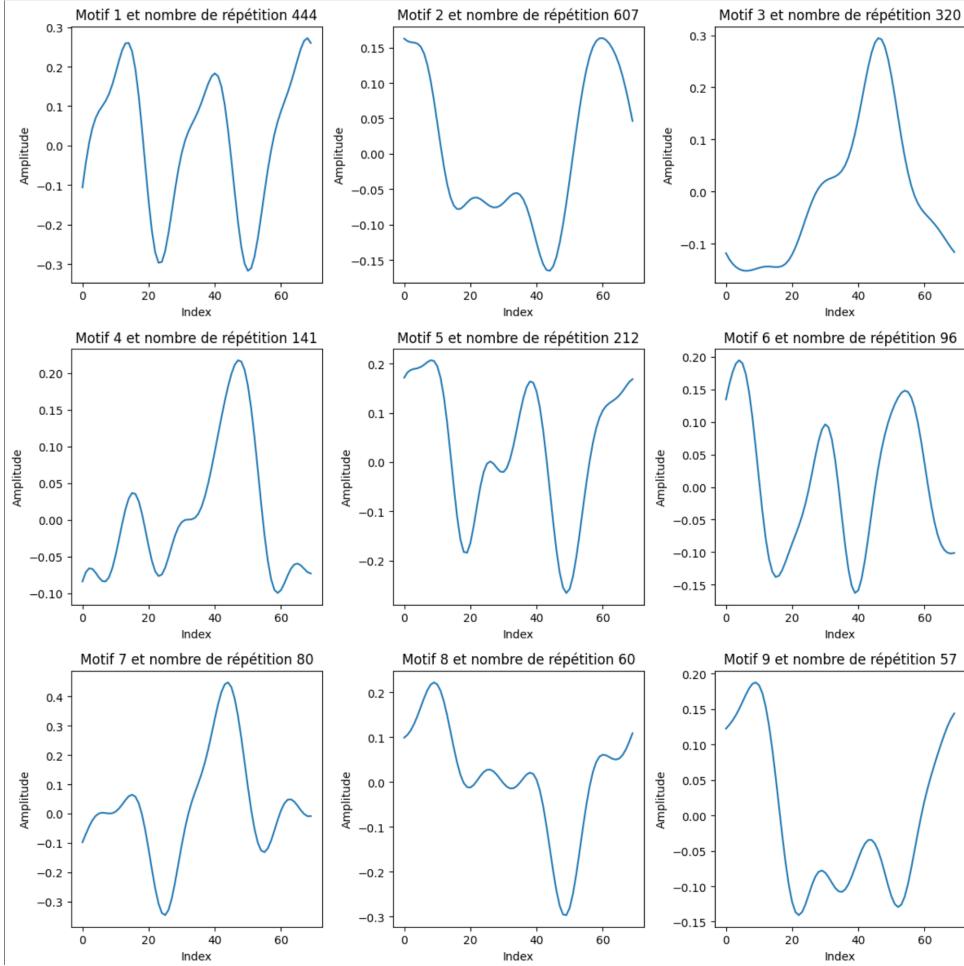


FIGURE 11 – Totalité des motifs présent dans 85% du dataset avec leurs nombre de répétition respectif.

Nous observons sur cette figure une certaine variété : tous les motifs sont différents. On peut cependant observer certaines ressemblances (par exemple entre les motifs 1, 3 et 6, ou entre les motifs 4 et 7 par exemple).

Une idée serait donc de créer plusieurs familles de motifs en calculant la DTW (Dynamic Time Warping), et de regrouper dans la même famille les motifs ayant une distance DTW faible.

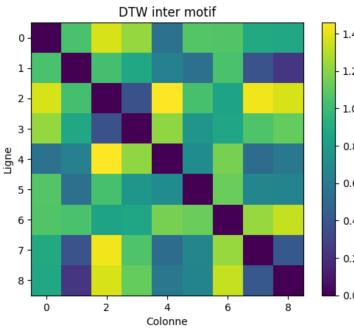


FIGURE 12 – Matrice de la DTW entre tous les motifs

Certains motifs présentent des distances DTW faibles, nous pouvons donc faire l'hypothèse que ces derniers ont un lien et qu'ils appartiennent à la même catégorie. Afin de pouvoir créer ces groupes, nous allons considérer notre problème sous forme de graphes, en considérant que tous les noeuds dont il existe un chemin les reliant appartiennent au même groupe. Nous considérons qu'un lien existe entre deux noeuds lorsque leur distance est inférieure à un certain seuil.

L'algorithme fait par exemple les associations suivantes :

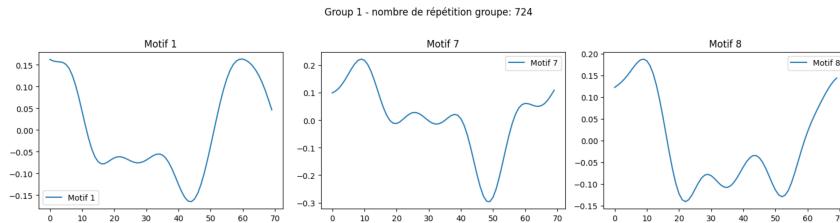


FIGURE 13 – exemple de Groupe de motifs avec leur nombre de répétition dans le dataset.

Cependant le motif 1 et 6 n'ont pas été associé alors qu'ils sont ressemblants. Le seuil d'appariement est donc sûrement trop restrictif. Nous augmentons donc légèrement le seuil.

En augmentant le seuil, les motifs 2, 6, 8, 5 et 9 forment un groupe alors que les motifs 1 et 6 appartiennent au même groupe. Ainsi, en augmentant le seuil nous avons des associations à priori mauvaises et non des fusions des groupes, tandis que des associations évidentes ne se font pas. Nous en déduisons donc qu'il n'existe pas de seuil permettant d'associer nos motifs correctement avec cette démarche.

L'expertise humaine est donc nécessaire pour réunir nos motifs par groupes pertinents. A partir de la forme des signaux, on peut ainsi faire l'hypothèse que les motifs 1, 3 et 6 correspondent à de la marche à plat, avec seulement 1 pic principal. Les motifs 2, 4, 7, 8 et 9 pourraient correspondre à de la montée d'escaliers, étant caractérisés par un grand pic principal et un ou plusieurs autres petits pics, qui pourraient se rapporter au fait que la marche dans les escaliers se fait en 2 temps (dans un premier temps c'est seulement la jambe qui bouge pour se poser sur la marche, puis le reste du corps).

A présent, nous cherchons à extraire de ces motifs des features. Pour cela nous calculons la distance euclidienne centrée réduite entre notre signal et le motif. Nous avons ainsi 9 nouvelles features (features 46 à 54) qui correspondent aux distances du signal à chacun des 9 motifs, que nous calculons sur notre dataset :

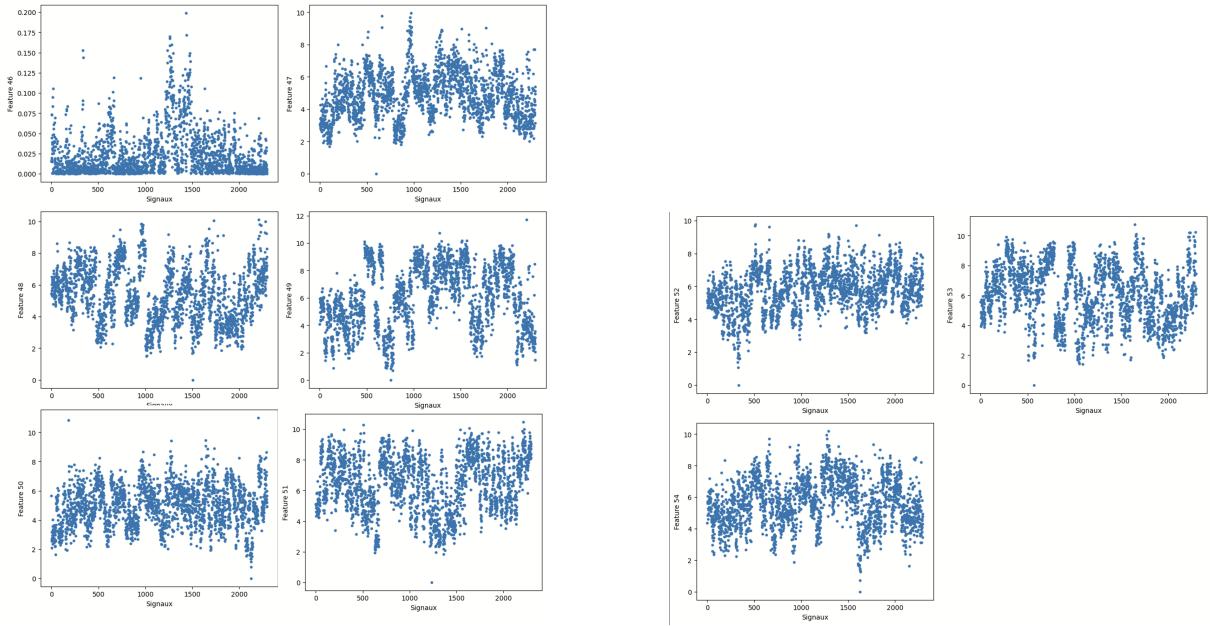


FIGURE 14 – Valeurs des distances minimales entre chaque motif (9 motifs) et chaque signal

Selon les motifs considérés, les features de distance euclidienne permettent plus ou moins de séparer les signaux. Par exemple, les distances des signaux au motif 4 (feature 49) séparent mieux les signaux que les distances des signaux au motif 1 (feature 46).

Ces différentes approches nous ont permis de mettre en évidence quelles sont les caractéristiques les plus pertinentes pour distinguer les signaux.

4 Sélection des features pertinentes

Dans cette partie, nous allons sélectionner et expliquer les 3 caractéristiques nous semblant être les plus pertinentes pour distinguer les signaux associés à la marche à plat et ceux associés à la montée d'escaliers.

4.1 Analyse en composante principale

Une première approche consiste à faire une analyse en composante principale de toutes nos features. Cette analyse est très largement utilisée dans l'analyse de données. Elle consiste à diagonaliser la matrice de covariance des features (centrée réduite) afin d'identifier les composantes principales, qui matérialisent simplement les directions dans l'espace des features qui maximisent la variance des données (ces directions sont orthogonales car la matrice de covariance est symétrique). Nous appliquons donc cette algorithme à l'ensemble de nos features et sélectionnons les deux premières composantes principales. Puis traçons l'ensemble de nos features et signaux projetés sur l'espace vectoriel généré par nos deux composantes principales.

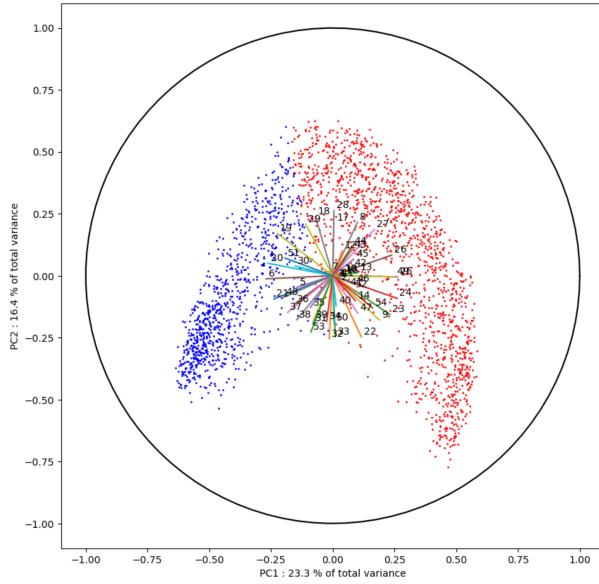


FIGURE 15 – PCA sur l’ensemble des features, les droites corresponds aux features et les points correspondent aux signaux projeter sur les deux composantes principales.

Nous observons sur cette figure deux clusters de signaux, un premier en bleu et un second en rouge. Nous avons obtenues ces deux clusters grâce à l’algorithme "k-mean" appliqué sur l’espace de toutes les features puis projeté sur les deux composantes principales. Au premier abord nous pouvons penser qu’il n’y a pas de cluster. Cela vient du fait que la seconde composante principale "dilue" les groupes formés par la première composante et nous donne l’impression d’un "U" retourné. Cependant, si nous projections toutes les données sur la première composante principale nous observerions deux clusters bien distincts.

Nous cherchons à présent les features nous permettant de séparer au mieux ces deux cluster. Les features : 49, 25, 24, 26, 6, 20, et 21 semblent posséder cette propriété. Nous pouvons aussi séparer les points selon la features 8.

Ensuite, nous affichons les graphiques de toutes les features chaque features sur chaque signal (détail dans la partie suivante), afin de déterminer celles qui semblent séparer aux mieux les données. Les features 8, 25 et 49 semblent être les meilleures.

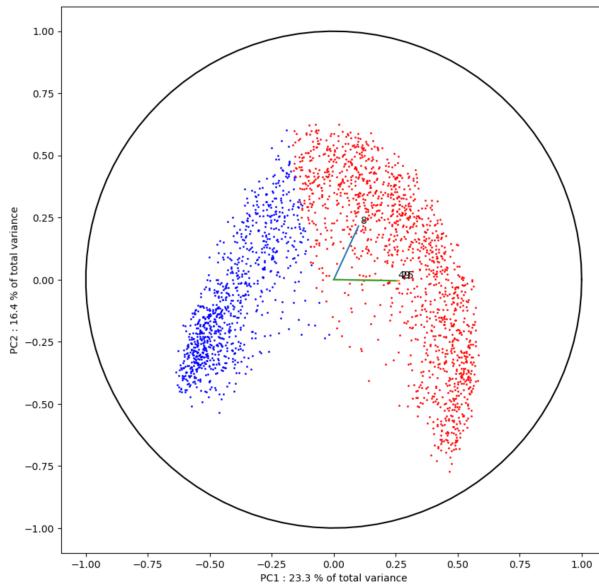


FIGURE 16 – PCA où nous affichons uniquement les 3 features sélectionnées.

Nous remarquons par ailleurs que les features 25 et 49 sont superposées, et donc sont très probablement corrélées.

4.2 Affichage des features "clusterisantes"

Pour identifier puis choisir les features les plus clusterisantes (c'est-à-dire celles qui permettent de séparer le mieux le dataset en 2 groupes), nous avons analysé les 54 figures des valeurs des features pour chaque signal (figures exposées dans la partie précédente). Ces graphiques nous permettent de voir selon quelles features les différents signaux de notre dataset sont le plus largement répartis, et donc selon quelles features ils peuvent être facilement discriminés.

4.2.1 Énergie relative dans la bande de 1.5Hz à 2Hz (feature 8)

L'énergie relative dans la bande 1.5Hz à 2 Hz correspond vraisemblablement à la bande de fréquence dans lequel se trouve le pas (la cadence de marche normale étant de minimum 1.5 pas/seconde). Une forte énergie relative dans cette bande de fréquence indique que l'individu marche à priori entre 1.5 et 2 pas par seconde.

Or, la cadence de la montée d'escalier est plus faible, du fait de l'effort engendré par cette activité. Certaines études ([exemple](#)), ou encore les informations associées aux tapis de course de type escaliers, nous indiquent que la cadence normale de la marche en montée d'escalier se trouve entre 1,2 et 1,7 pas par seconde. Ainsi, les signaux associés à la plus faible énergie dans la bande de fréquence 1.5-2Hz sont vraisemblablement les signaux de montée d'escaliers, dans la mesure où ces signaux ont probablement une énergie plus forte dans la bande de fréquence inférieure.

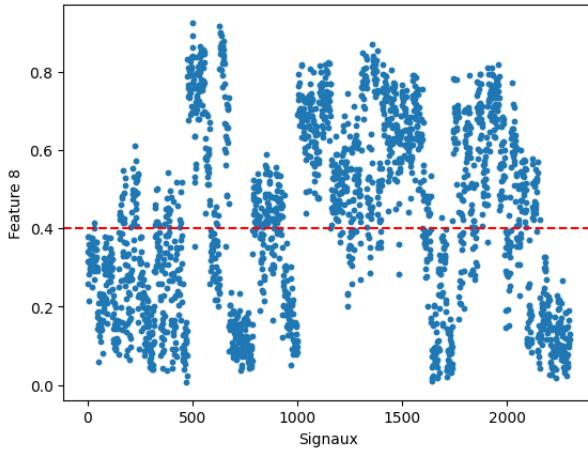


FIGURE 17 – Features Energie relative fréquentielle pour $f \in [1.5Hz, 2Hz]$

Afin de calculer les features d'énergie relative par bande de fréquence, nous appliquons une transformée de fourrier à notre signal. Puis nous séparons notre FFT en 12 intervalles pour des fréquence allant de 0 à 6 Hz. Nous calculons ensuite l'énergie contenu dans chaque intervalle divisée par l'énergie totale du signal. La feature 8 correspond ainsi au 4ème intervalle.

4.2.2 "Energie d'auto-corrélation relative" avec un décalage temporel de $\tau \in [0.22s, 0.25s]$ (feature 25)

L'énergie d'auto-corrélation relative permet d'évaluer la corrélation du signal avec lui-même à différentes valeurs de décalage temporel, ici avec un décalage entre 0.22 et 0.25s. Ce décalage temporel correspond donc à des fréquences entre 4 et 4.5Hz.

L'énergie d'autocorrélation correspond à l'autocorrélation au carré, ainsi une forte énergie d'autocorrélation signifie que l'autocorrélation à ce décalage temporel est soit fortement positive soit fortement négative (à l'inverse, une faible énergie signifie que l'autocorrélation est proche de zéro). Une auto-corrélation fortement positive correspond à un phénomène de superposition parfaite, c'est à dire que le signal se reproduit après le décalage τ (ce qui signifierait que la période du signal est entre 0.22 et 0.25s dans ce cas). Une autocorrélation fortement négative correspondrait à la superposition d'un creux et d'un pic au niveau , donc au niveau d'une demie période (dans ce cas, la demie période est de τ).

Dans notre cas, on peut faire l'hypothèse que la forte énergie d'autocorrélation observée correspond à une autocorrélation fortement négative, et que le décalage de [0.22 s ,0.25 s] correspond à la demie période. Dans ce cas, la période est dans l'intervalle [0.44s, 0.5s], et donc la fréquence associée est entre 2 et 2.27Hz.

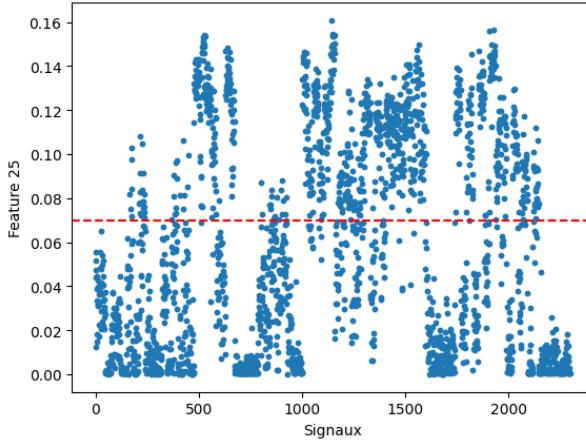


FIGURE 18 – Features énergie relative d’auto-corrélation pour $\tau \in [0.22s, 0.25s]$

Afin de calculer cette feature, nous commençons par calculer la fonction d’auto-corrélation de notre signal pour un décalage temporelle compris entre $\tau \in [-1s, 1s]$. Puis nous séparons l’intervalle $\tau \in [0s, 1s]$ en 30 sous intervalles, puis nous réalisons le calcul d’énergie (de la même manière que pour la feature précédente). Puis nous sélectionnons la valeurs issues du 8e sous-intervalle.

4.2.3 Distance au motif n°4 (feature 49)

Le pattern 4 est un motif qu’on a supposé être associé à la montée d’escalier (voire partie motif). Ainsi, les signaux qui ont une distance faible à ce motif sont probablement ceux correspondant à la montée d’escalier.

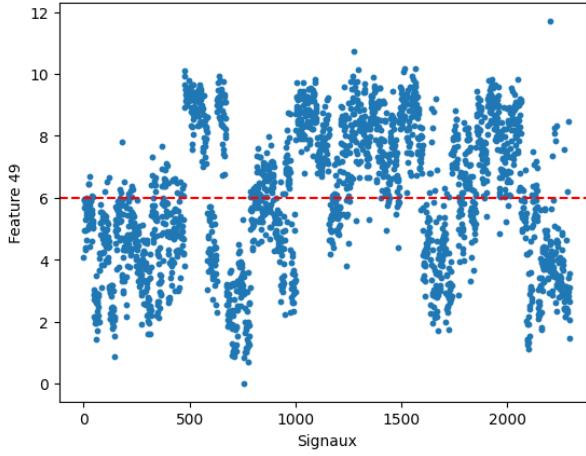


FIGURE 19 – Distance minimal entre les signaux et le motif 4 pour chaque signal.

Afin de calculer cette features, nous importons les paterns calculer dans la partie précédente. Puis, nous calculons la distance euclidienne centrée réduite entre notre 4e motif et un signal. Enfin, la valeurs minimal de cette distance nous donne la features 49.

Comme vu dans la partie précédente, les features associées à une grande distribution des données sont à priori des features plus pertinentes pour séparer les signaux. Dans le

cas des features choisies, bien qu'on ne puisse voir deux groupes de signaux bien distincts, on peut tout de même définir un seuil (ligne rouge en pointillés d'équation) pour séparer grossièrement deux groupes. Nous discuterons dans la discussion de la pertinence de cette classification.

5 Discussion

Nous pouvons à présent nous demander si la manière dont les trois features choisies permettent de séparer les signaux constitue une séparation pertinente ou pas. L'analyse en composante principale suggère que les features séparent plus ou moins de la même manière le dataset. Pour cela, nous allons, pour chacune des features choisies, appliquer sur les données le seuil estimé en partie précédente pour séparer les signaux en deux groupes, puis comparer les groupes formés. Ainsi, si les groupes obtenus pour chaque feature sont similaires ou identiques, cela signifie que les features séparent les signaux de la même manière, ce qui indique une séparation à priori pertinente dans le cadre de notre étude.

Le code permettant de comparer ces groupes opère les étapes suivantes :

- pour chaque feature, les signaux sont séparés en deux groupes selon le seuil choisi et leurs indices regroupés dans deux listes. Les listes a et b correspondent respectivement aux signaux ayant une valeur de feature 8 inférieure et supérieure au seuil. De même pour les listes c et d avec la feature 25, et les listes e et f avec la feature 49.
- on compare, à l'aide de la fonction set() de python, les listes obtenues deux à deux (les listes a et b sont comparées aux listes c et d puis e et f, et ne sont par ailleurs pas comparées entre elles). On obtient ainsi le nombre ou le pourcentage d'éléments en commun entre les listes.

Les résultats obtenus sont les suivants :

```

Nombre et pourcentage d'éléments en commun entre les listes a et c : 1072 soit 99.0 %
Nombre et pourcentage d'éléments en commun entre les listes a et d : 11 soit 1.0 %
Nombre et pourcentage d'éléments en commun entre les listes b et c : 296 soit 24.4 %
Nombre et pourcentage d'éléments en commun entre les listes b et d : 919 soit 75.6 %
Nombre et pourcentage d'éléments en commun entre les listes c et e : 1056 soit 77.2 %
Nombre et pourcentage d'éléments en commun entre les listes c et f : 312 soit 22.8 %
Nombre et pourcentage d'éléments en commun entre les listes d et e : 95 soit 10.2 %
Nombre et pourcentage d'éléments en commun entre les listes d et f : 835 soit 89.8 %
Nombre et pourcentage d'éléments en commun entre les listes a et e : 922 soit 85.1 %
Nombre et pourcentage d'éléments en commun entre les listes a et f : 161 soit 14.9 %
Nombre et pourcentage d'éléments en commun entre les listes b et e : 229 soit 18.8 %
Nombre et pourcentage d'éléments en commun entre les listes b et f : 986 soit 81.2 %

```

FIGURE 20 – Pourcentage de ressemblance entre les groupes

On observe donc des forts pourcentages de similitude entre les listes a et c, c et e, a et e, de même entre les listes b et d, d et f, b et f. On peut ainsi voir que, à plus ou moins d'exceptions près, les 3 features permettent de séparer les signaux de la même manière, et que les listes a, c et e correspondent à une classe de signal et les listes b, d et f correspondent à l'autre.

Ainsi, le premier groupe (listes a, c et e identiques à au moins 75%) correspond à priori à la montée d'escaliers, donc aux signaux ayant une faible énergie dans la bande 1.5Hz à 2 Hz et une faible distance avec le motif 4

Le deuxième groupe (listes b, d et f) correspondrait aux signaux de la marche à plat.

6 Conclusion

L'objectif de notre projet était de trouver des caractéristiques pertinentes permettant la distinction entre deux types de signaux temporels : la marche à plat et la montée d'escaliers.

Nous avons ainsi pu, à partir de données d'accélérométrie selon l'axe z, calculer diverses caractéristiques de ces signaux. Ces caractéristiques, notamment celles d'énergie par bande de fréquence, d'autocorrelation et de distance à des motifs, se sont révélées être informatives quant à la nature des signaux. A partir de ces résultats et de notre analyse (à la fois basée sur la logique et la littérature), nous avons faire l'ébauche d'un classifieur permettant de séparer des signaux de marche et de montée d'escaliers.

Malgré des résultats relativement corrects et concluant, notre étude était limitée, notamment par la structure et la qualité des données, mais aussi par des contraintes matérielles et de temps, notamment en ce qui concerne la partie sur la détection de motifs.