Ce chapitre présente la méthode adoptée pour rassembler en une même « famille » les battements cardiaques dont le tracé ECG est de même nature, et distinguer ainsi plusieurs familles[I] de battements.

# I Motivations et objectifs

Dans ce chapitre, nous abordons le problème de la reconnaissance des battements cardiaques, en séparant les battements d'origine supraventriculaire des battements d'origine ventriculaire. Pour réaliser une telle séparation, nous pourrions modéliser un à un les battements à l'aide des bosses présentées dans le chapitre précédent, puis analyser la position et la forme des bosses caractéristiques. Bien que la décomposition en bosses d'un battement soit obtenue par une procédure de calcul rapide (moins de 500ms sous Matlab), le nombre total de battements d'un enregistrement Holter est tel (plus de 100 000 battements) que le temps de calcul d'un modèle pour chaque battement est beaucoup trop important[II]. De plus les battements normaux (qui sont, en général, très largement les plus fréquents) sont très semblables, voire quasi identiques, ce qui rend inutile la modélisation de chacun d'eux.

Cette remarque nous a amené à effectuer une agrégation des battements *préalable* à la modélisation en bosses. Cette opération a pour objectif de regrouper entre eux les battements de même nature ; en particulier, nous regrouperons les battements d'origine supraventriculaire[III] entre eux, ceux de type extrasystole ventriculaire entre eux (en distinguant les extrasystoles de foyers ectopiques distincts), les battements anormalement larges (de type bloc de branche), etc… qui formeront ainsi autant de familles que nécessaire.

[I] Les termes « famille » et « classe » seront indifféremment utilisés dans la suite pour exprimer l'ensemble regroupant les battements de même nature. Le terme « classe » étant celui couramment utilisé en classification et le terme « famille » étant le terme consacré en cardiologie.

[II] Ce qui correspond à environ 13mn sous Matlab pour les 100 000 battements pour cette seule partie de l'analyse. Une autre solution plus efficace pourrait être de ne modéliser *que* l'onde R. Ce n'est pas la solution que nous avons retenue ici.

[III] On verra que le regroupement en familles permet de distinguer les familles de battements dont le foyer initiateur est le sinus de celles dont le foyer est soit auriculaire autre que le sinus, soit nodal.

On peut alors se contenter de ne modéliser en bosses qu'un seul représentant par famille, et l'analyse des bosses obtenues permet d'obtenir des informations valables pour tous les battements appartenant à cette famille.

La méthode de regroupement en familles retenue ici est du type *classification non supervisée*, c'est-à-dire *que l'on ne dispose pas d'une base de référence étiquetée* pour effectuer un apprentissage ; de plus, cette classification se fera sans connaissance a priori *du nombre de classes* : les classes se créent au fur et à mesure de la procédure de classification.
De nombreux algorithmes permettent de traiter ce type de problème et conduisent à de bons résultats ; c'est le cas notamment des algorithmes de quantification vectorielle comme celui des K-moyennes, de classification hiérarchique, et des cartes auto organisatrices par exemple. Ces trois algorithmes sont présentés au paragraphe suivant mais nous verrons en quoi ils ne sont pas adaptés à notre problème. Nous avons donc développé un classifieur ad hoc présenté dans la suite.

# II Présentation d'algorithmes existants

## II.1.1 L'algorithme des K-moyennes

L'algorithme des K-moyennes est un algorithme classique de quantification vectorielle. Son principe est le suivant [Dreyfus, 2002] : on dispose de points de l'espace des observations que l'on souhaite rassembler en classes, sans que l'on dispose de connaissance a priori de propriété(s) particulière(s) sur ces classes ; seul leur nombre $p$ est fixé a priori.
L'algorithme des K-moyennes est itératif ; chaque itération est composée de deux étapes :

- Recherche, pour chaque point d'observation, de son meilleur représentant parmi $p$ référents, où chaque référent représente une classe ;
- Optimisation de chacun de ces référents pour qu'ils représentent au mieux les points d'observations en $p$ classes.

Il existe une preuve de convergence pour cet algorithme. Trois inconvénients ne nous permettent pas de l'utiliser. Le premier est qu'il est nécessaire de connaître le nombre de classes avant de commencer la classification. Or, dans notre cas, nous ne savons pas quel nombre de types de battements différents nous rencontrerons dans un ECG donné. Cependant ce problème pourrait être résolu en imaginant une classification en deux classes : 1)

battements d'origine supraventriculaire et 2) battements d'origine ventriculaire. Un deuxième inconvénient est la grande sensibilité aux conditions initiales, qui se traduit ici par le choix des $p$ référents initiaux. En effet, s'ils sont choisis de manière aléatoire, la convergence de l'algorithme vers un minimum « satisfaisant » n'est pas assurée, ce qui impose, dans la pratique, de multiplier les initialisations, et augmente d'autant le temps de calcul. Enfin l'inconvénient majeur de cette méthode est le suivant : en étudiant l'interprétation probabiliste de cet algorithme, on constate qu'il suppose que les classes suivent des lois de distribution normales réduites, autrement dit, avec la même importance dans toutes les directions de l'espace. Pour nous, les directions de cette espace représentent des grandeurs très différentes, difficilement comparables entre elles comme nous le verrons par la suite (intervalle RR du battement avec le suivant, angle que forme l'axe principal du battement avec une des voies d'enregistrement, produit des valeurs propres issues de l'analyse en composantes principales, etc.).

### II.1.2 Classification hiérarchique

La classification hiérarchique [Everitt, 1974] constitue une autre approche de la classification non supervisée. Elle consiste à calculer une matrice exprimant les distances mutuelles entre les points à classer, puis, en se fondant sur cette matrice, à regrouper entre eux les points les plus proches. Cette méthode permet la construction d'un arbre hiérarchique, qui révèle plusieurs partitions possibles, où chaque point est attribué à l'un des groupes d'une partition donnée. Le choix de la meilleure partition s'effectue une fois la classification hiérarchique terminée ; elle peut être fondé sur différents critères, l'un des plus classiques étant lié à la mesure de l'inertie intergroupe [Thorndike, 1953].

L'avantage de cette méthode est qu'elle n'est soumise à aucune initialisation particulière de paramètre(s) ce qui la rend déterministe, et en outre, que le nombre de classe n'a pas à être fixé a priori. Cependant, ce type de méthode impose le calcul de la matrice des distances de tous les points d'observation avec tous les autres, et cette masse de calculs est beaucoup trop importante compte tenu du temps que nous voulons consacrer à cette étape.

### II.1.3 Cartes auto organisatrices

La méthode des cartes auto organisatrices, développée par T. Kohonen au début des années 1980 [Kohonen, 1984], constitue un compromis entre les deux algorithmes précédents. Cette méthode est plus robuste aux conditions initiales que ne l'est l'algorithme des K-moyennes ; elle ne suppose pas que les directions de l'espace des observations ont la même importance (lois normales réduites pour chacune des classes), et ne nécessite pas la définition a priori du nombre de classes. Elle consiste à représenter, dans un espace de faible dimension (1, 2 ou 3) appelé « carte » des référents ; ces derniers ne sont pas ici des représentants de classes, mais constituent un petit nombre de points abstraits de l'espace des observations. La propriété essentielle des cartes de Kohonen réside dans le fait que la topologie de l'espace initial est conservée dans cette projection dans un espace de faible dimension. Un algorithme itératif, proche de celui des K-moyennes, permet d'aboutir à une distribution des référents sur la carte qui constitue une caricature à faible dimension de l'ensemble des points d'observation. On fait généralement suivre la construction de cette carte par une classification hiérarchique sur les référents.
Deux inconvénients apparaissent ici : le temps de calcul associé aux itérations qui permettent la construction de la carte est important, et surtout, ce type de méthode statistique n'est pas adapté à notre classification. En effet, comme nous le verrons par la suite, nous serons amenés à classer des paquets de 1200 battements en au moins deux groupes qui seront respectivement identifiés comme les battements d'origine supraventriculaire et les battements d'origine ventriculaire. Chez les patients en bonne santé, un grand déséquilibre entre les cardinaux de ces classes est généralement observé : typiquement un patient peut présenter un unique battement ventriculaire pour 1200 battements normaux. Ce type de déséquilibre est mal géré par les méthodes statistiques qui auront tendance à « considérer » ce battement différent comme une réalisation peu probable de loi de probabilité associée à la classe des battements normaux, plutôt que comme le représentant d'une classe à part entière constituée d'un unique élément.

Ces méthodes étant peu adaptées à notre problème, nous avons été amenés à développer un algorithme simple, donc rapide, et bien adapté à la classification qui nous intéresse. Cet algorithme est présenté dans la section suivante.

# III Principe général de l'algorithme

Dans un premier temps, on sépare les battements de l'ECG en trois groupes, en fonction du nombre de voies utilisées lors du calcul en composantes principales : rappelons que le critère de choix du nombre de voies pour le calcul de la voie principale est fondé sur les niveaux de bruits de haute et basse fréquences qui affectent ces voies. On effectue ainsi trois classifications différentes : une pour les battements issus d'une voie principale calculée à partir de 3 voies, une pour ceux qui sont calculés sur 2 voies, et une pour les battements qui ne sont pas exprimés sur une voie principale car seule une voie est valide. Les trois algorithmes sont très proches, mais n'utilisent pas tout à fait les mêmes paramètres, comme nous le verrons dans la suite.

Ainsi on initie l'algorithme en attribuant le premier battement à une famille, qui devient référence pour tout battement construit à partir du même nombre de voies que ce premier battement ; un battement issu d'un nombre différent de voies que ce premier battement conduit à la création d'une nouvelle famille, qui devient lui aussi référence pour ce nombre de voies.

Les battements sont alors traités un à un, dans l'ordre de leur apparition temporelle. Pour un battement donné, si celui-ci a des paramètres proches de ceux d'une famille existante, il est associé à cette famille ; sinon, l'algorithme crée une nouvelle famille qui deviendra référence pour tous les battements à même nombre de voies.

## III.1 Caractérisation d'une famille

Chaque famille est caractérisée par:

- le modèle en bosses du battement qui lui a donné naissance : ce modèle sera appelé *prototype* de la famille[IV],
- des paramètres qui sont *fixés* au moment de la création de la famille,
- d'autres paramètres qui sont *réestimés* chaque fois qu'un battement est ajouté à la famille.

[IV] On remarque ici que l'on modélise le premier battement à l'origine de la famille, qui n'est pas forcément le battement le plus représentatif de cette famille ce point pourrait constituer une évolution eventuelle de l'algorithme.

### III.1.1 Le prototype

Le prototype de la famille est construit à partir du premier battement qui a donné naissance à celle-ci. On applique à ce battement, exprimé sur sa voie principale (donc sur une piste unique), l'algorithme de décomposition en bosses présenté dans le chapitre précédent ; à l'issue de cet algorithme, on obtient un modèle analytique du battement sous la forme d'une somme de 6 bosses (Figure 1) : ce modèle est le *prototype* de la famille.

Chaque battement ultérieur est comparé, par un calcul convenable, au prototype de chaque famille déjà créée, ce qui fournit un critère de décision permettant d'associer ce battement à une des familles, ou de créer une nouvelle famille pour laquelle le modèle de ce battement devient le prototype.

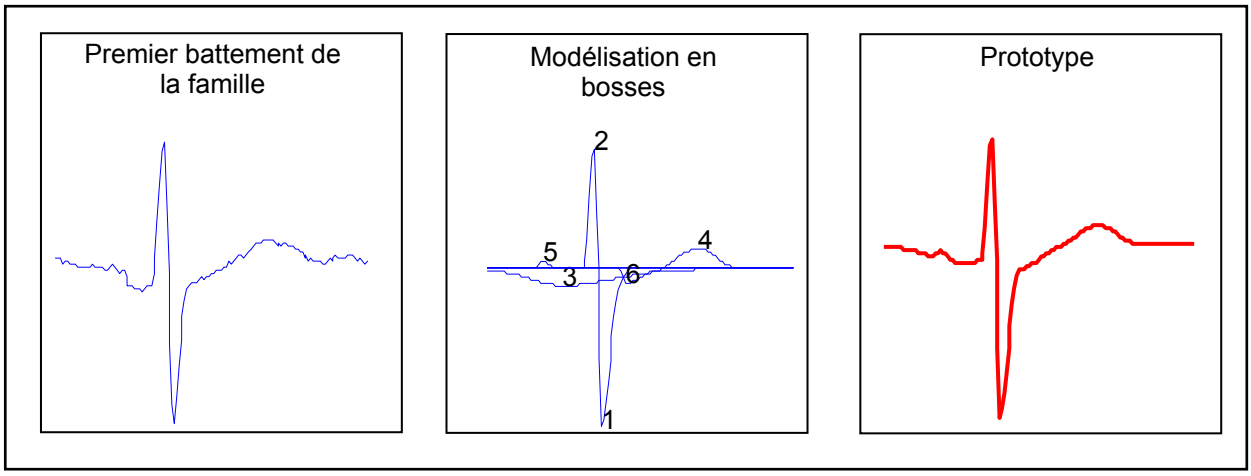


***Figure 1 : A la création d'une nouvelle famille, on construit le prototype de la famille par une modélisation en 6 bosses du battement qui lui a donné naissance.***

### III.1.2 Paramètres fixes

En plus du prototype, des descripteurs permettent de caractériser la famille. Les premiers descripteurs présentés ici sont fixés à la création de la famille, et restent fixes tout au long de l'analyse.

### III.1.2.a Rapport des intervalles RR

Appelons *RRs* la distance entre l'onde R du battement qui engendre la famille et l'onde R du battement suivant, et appelons *RRp* la distance entre l'onde R du battement qui engendre la famille et l'onde R du battement précédent (Figure 2). Le rapport *RRs*/*RRp* est un paramètre qui caractérise la famille. Dans le cas d'un rythme régulier, ce rapport est voisin de 1, mais il peut largement dépasser cette valeur dans le cas d'extrasystole avec repos compensatoire (cf. Chapitre 2). La valeur de ce rapport est fixée à la création de la famille.

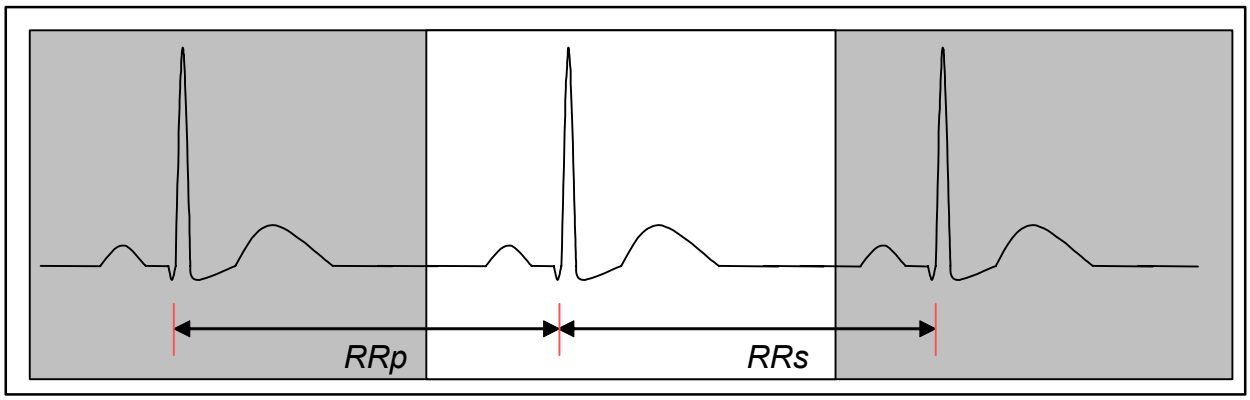


***Figure 2 : Le rapport de RRs/RRp caractérise la famille. Sa valeur est fixée à la création de celle-ci, et reste fixe tout au long de l'analyse.***

L'avantage d'utiliser le *rapport* *RRs*/*RRp* plutôt que l'une ou l'autre des distances qui le définissent est qu'il est *indépendant de la fréquence cardiaque* : ainsi, sa valeur pour un battement normal d'un rythme rapide sera très voisine de celle d'un battement normal d'un rythme lent par exemple, et ce paramètre ne sera alors pas éliminatoire pour le regroupement de battements provenant d'épisodes d'activité cardiaque où seule la fréquence diffère.

### III.1.2.b Intervalles RR avec le battement suivant (RRs)

L'étude de différentes bases de données nous a permis de mettre en évidence le fait que *la distance* *RRs* était un élément parfois utile à la caractérisation de la famille, afin d'éviter des erreurs de classification notamment en présence d'extrasystoles ventriculaires à complexes peu élargis et interpolées : c'est-à-dire des battements ventriculaire qui s'inscrivent dans un rythme régulier pour lequel le rapport précédent vaut 1. Cette distance constitue donc

également un paramètre caractéristique de la famille. L'utilisation de cette grandeur *RRs* et la précédente (le rapport *RRs*/*RRp*) est plus pertinente que l'utilisation séparée des deux grandeurs *RRs* et *RRp* car dans ce dernier cas, lors d'accélération du rythme, les deux paramètres (*RRs* et *RRp)* sont différents d'un battement à l'autre ce qui conduirait à la création d'une famille supplémentaire et nous ne le souhaitons pas, alors qu'avec les paramètres retenues (*RRs* et *RRs*/*RRp)* seule *RRs* est différent et ça ne suffit pas, en général, à créer une nouvelle famille comme nous le verrons dans la suite.

#### III.1.2.c Amplitude du battement initial

Depuis l'étape de segmentation de l'ECG en fenêtres, l'amplitude maximale des battements est normalisée à 1, c'est-à-dire que les complexes ont tous une dynamique de 1, quelle que soit leur taille originale. Pour éviter de confondre des battements de tailles trop différentes, les amplitudes de chacune des voies valides sont retenues comme descripteurs de la famille[V].

### III.1.3 Paramètres réestimés

En plus des paramètres fixes, on caractérise les familles avec des *paramètres qui sont réestimés dès qu'un battement est ajouté à la famille*. Ces paramètres subissent en général une légère dérive due notamment à des changements de position du patient, qui, sur de longues durées, peuvent être importantes. De plus, l'amplitude de variation de ces paramètres dépend fortement du patient; ainsi, une procédure de réestimation lors de l'ajout de chaque nouveau battement permet d'une part de suivre une éventuelle dérive temporelle, et, d'autre part, de *s'adapter spécifiquement au patient*.

#### III.1.3.a L'axe principal

Comme nous l'avons indiqué précédemment, l'axe principal issu de l'analyse en composantes principales est un paramètre particulièrement important pour la caractérisation de la famille : il est parfois un des seuls paramètres qui permette la distinction entre un battement normal et

[V] On note ici que deux battements bi-phasiques très proches peuvent être à l'origine de la création de deux familles différentes.

un battement anormal (chapitre 5). Les informations de l'axe principal sont caractérisées par les deux groupes de valeurs suivantes :

- La moyenne des angles ACP sur les vingt derniers battements[VI] inclus dans la famille. Dans le cas d'un enregistrement à 2 voies, il s'agit donc de *l'angle* $\theta$ défini précédemment (cf. chapitre 5.IV Résultats de l'analyse en composantes principales) entre l'axe principal issu de l'analyse en composantes principales et un axe fixe (Figure 3). Pour les enregistrements sur 3 voies, c'est le couple formé par *l'angle* $\theta$ et *l'angle* $\varphi$ provenant des coordonnées sphériques de l'axe principal qui est retenu. Dans le cas d'un enregistrement n'ayant qu'une seule voie, cette notion d'angle ACP n'a pas de signification : ce paramètre n'est donc pas considéré.
- L'écart type de la distribution de l'angle $\theta$ dans le cas de 2 voies, ou la matrice de covariance 2x2 (pour les angles $\theta$ et $\varphi$) pour les enregistrements 3 voies, ces deux grandeurs traduisant la distribution des valeurs autour de la valeur moyenne.

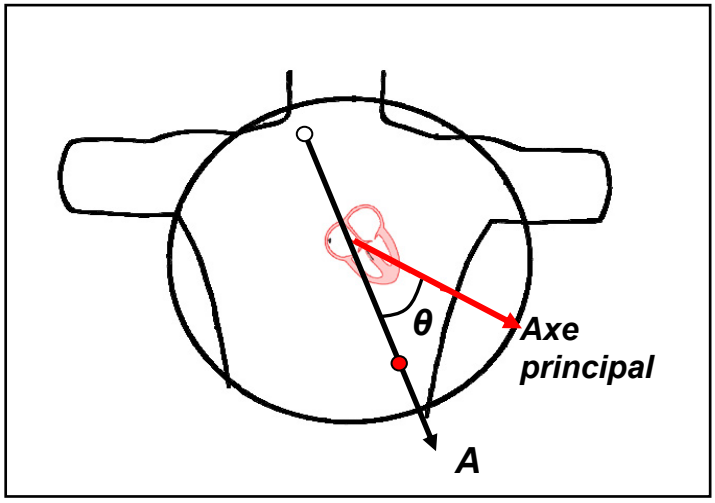


***Figure 3 : l'angle θ correspond à l'angle que forme l'axe principal cardiaque repéré par la méthode d'analyse en composantes principales détaillée au chapitre 5 et un axe fixe, ici l'axe formé par la première voie d'enregistrement A. Dans le cas d'un repérage en trois dimensions de l'axe principal, il est repéré par ses coordonnées sphériques (angle θ et φ) qu'il forme avec ce même axe fixe.***

Le calcul de l'écart-type de la distribution (ou de la matrice de covariance) est imposé par la nécessité de s'adapter au patient ; en effet, certains patients présentent des variations d'angles de quelques centièmes de radians entre deux battements normaux alors que cette variation

[VI] Lorsque la famille n'est pas suffisamment importante (moins de 20 éléments) l'ensemble des battements la constituant est utilisé pour l'estimation des paramètres.

atteint plusieurs dixièmes chez d'autres.L'écart-type (ou la matrice de covariance) permettra de calculer, pour la classification d'un battement, des distances de Mahalanobis [Everitt, 1974] (c'est-à-dire des distances *normalisées par l'écart type* de la famille) plutôt que des distances euclidiennes.

### III.1.3.b Le produit des deux premières valeurs propres de la matrice de covariance

*Le produit des deux plus grandes valeurs propres de la matrice de covariance présentée pour l'analyse en composantes principales* (cf. Chapitre 5, eq. 2) constitue un troisième paramètre pertinent pour la classification des battements en familles. Ce produit est une caractéristique de la forme du vectocardiogramme. Rappelons que les valeurs propres de la matrice de covariance correspondent aux longueurs des axes principaux du vectocardiogramme : le *produit des deux valeurs propres est donc proportionnel à la surface de la courbe.*
De même que pour l'axe principal, nous disposons de deux types de valeurs pour caractériser le produit des deux premières valeurs propres : sa valeur moyenne, qui est recalculée sur les 20 derniers battements de la famille, et son écart-type. Mais, contrairement au cas précédent, on peut éviter la réestimation systématique de l'écart-type. En effet, contrairement à l'écart-type associé à l'axe principal, qui varie en fonction de la position du patient, donc au cours de l'enregistrement, la variation du produit des deux premières valeurs propres est dominée par des caractéristiques physiologiques inhérentes au patient et au positionnement des électrodes. On calcule donc cet écart-type dès le début de l'algorithme, sur l'ensemble des battements de l'ECG.

Ainsi chaque famille créée est définie par 5 paramètres et un prototype continu. Chaque nouveau battement est alors analysé ; il est soit associé à une famille existante, soit à l'origine d'une nouvelle famille.

## III.2 Principe de la classification non supervisée

Comme nous l'avons indiqué plus haut, les battements sont traités un à un dans l'ordre de leur apparition temporelle. On remarque que, au cours du temps, pour un même type de battement, l'axe principal change légèrement de direction, notamment à cause des changements de position du patient, et des mouvement de la cage thoracique pendant la respiration : ainsi, on peut observer que, à un instant donné, l'axe principal d'un battement normal coïncide avec un axe principal d'un battement anormal (typiquement une extrasystole ventriculaire) enregistré peu de temps auparavant, alors que le patient adoptait vraisemblablement une autre position. Or l'adaptation de certains paramètres ne permet pas, à elle seule, d'éviter le regroupement de battements qui doivent être séparés ; on complète donc cette approche adaptative en y ajoutant une découpe de l'ECG en *séries* de 1200 battements ce qui correspond environ à 20 mm d'ECG ; on redéfinit donc les familles tous les 1200 battements. Le regroupement des familles considérées comme identiques des séries temporelles successives n'est effectué qu'une fois que ces familles ont été identifiées, c'est-à-dire une fois que les ondes de leurs prototypes respectifs ont été étiquetées[VII].

Pour chaque série de 1200 battements, la classification s'effectue schématiquement de la manière suivante : pour un battement donné, on calcule une *distance globale* entre le battement et chaque famille déjà créée dans la série ; cette distance, qui sera définie plus précisément dans le paragraphe suivant, est une *combinaison linéaire des distances sur les différents paramètres caractéristiques et d'une valeur de « ressemblance » de l'onde R à celle des battement de la famille calculée en référence à son prototype*. La famille qui représente au mieux le battement est celle pour laquelle la distance est la plus faible. Dans le cas où cette distance est supérieure à un seuil prédéfini, aucune famille n'est satisfaisante et le battement étudié donne naissance à une nouvelle famille.

[VII] Ce deuxième regroupement n'a pas été effectué dans cette étude. Il est principalement nécessaire à l'affichage des données dans le logiciel pour eviter un trop grand nombre de famille ; et ELA Medical possède un tel outil de rassemblement, c'est ce dernier qui sera utilisé.

### III.2.1 Calcul des distances

Les paragraphes qui suivent présentent en détail :

- le calcul de la distance entre la forme de l'onde R d'un battement et celle du prototype de la famille,
- le calcul des distances entre les paramètres caractéristiques,
- le calcul de la *distance globale*, somme pondérée de ces différentes distances.

#### III.2.1.a Distance entre les formes de deux ondes R

Nous définissons une distance normalisée entre la forme de l'onde R du prototype d'une famille de battements et celle du battement considéré exprimé sur sa voie principale. L'emplacement de l'onde R est défini à partir de l'emplacement détecté par l'algorithme de détection des R (cf. chapitre 3). La distance entre l'onde R et l'onde prototype est estimée sur un intervalle de 160 ms qui enferme l'onde R, ce qui correspond à 32 points lors d'un échantillonnage à 200Hz.

Pour ce calcul de distance, le signal et le prototype sont alignés sur le point précis qui repère l'emplacement de leur onde R. Comme cet emplacement peut être légèrement biaisé (de quelques millisecondes) d'un battement à l'autre, on calcule en réalité 7 valeurs de distances (Figure 4), chacune d'elle étant obtenue en décalant légèrement les deux signaux l'un par rapport à l'autre – comme on le fait habituellement pour le calcul de corrélations. On appellera *distance* la plus petite des valeurs calculées.Les décalages relatifs considérés entre les deux signaux à comparer sont de -4 points, -2 points, -1 point, 0 points, 1 point, 2 points et 4 points, ce qui correspond respectivement à des décalages temporels de -20ms, -10ms, -5ms et +5ms +10ms +20ms.

Soit $P^i$ le signal temporel du prototype de la famille $i$, et $S^j$ le signal temporel du battement $j$ ; la distance entre les deux formes, est calculée de la manière suivante :

$$\varepsilon_{FR} = \min_{\tau}\left(\frac{1}{N_p}\sqrt{\sum_{k=1}^{N_p}(P^i(k) - S^j(k-\tau))^2}\right) \qquad \textit{Eq. 1}$$

où $\tau \in [-4,-2,-1,0,1,2,4]$ et $N_p = 32$ points autour de l'onde R.

Remarquons que la valeur du décalage qui minimise la distance pour chaque famille compte parmi les paramètres qui servent à choisir, parmi plusieurs familles, celle qui convient le mieux : si plusieurs d'entre elles sont susceptibles de se voir attribuer le battement considéré, on attribue le signal à la famille pour laquelle cette valeur de décalage est minimum.

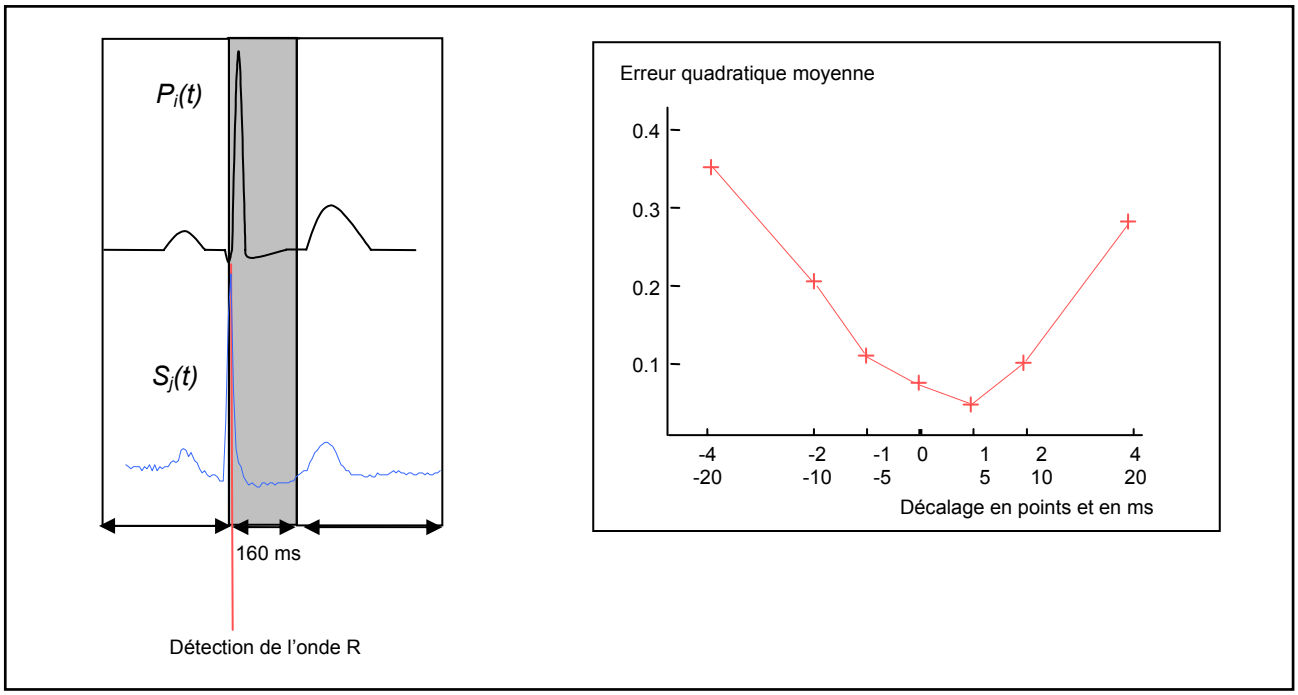


***Figure 4 : L'algorithme de détection des ondes R peut introduire un décalage de quelques ms d'un battement à l'autre dans la position de l'onde. On calcule donc 7 distances obtenues en introduisant 7 décalages différents entre le battement et le prototype autour de la position de l'onde R. Ici la meilleure distance est obtenue pour un décalage de 1 point. La zone grisée sur le battement représente la partie du signal qui est prise en considération dans le calcul, c'est la zone la plus modifiée en cas d'ESV.***

### III.2.1.b Distance sur l'onde P

Afin de séparer dès à présent les battements dont l'onde P est particulière, nous calculons, en plus de la distance sur l'onde R, la distance normalisée sur cette onde. Les points utilisés pour ce calcul sont ceux qui vont du début du segment qui enferme le battement (Figure 4 zone de gauche) jusqu'au point de repérage de l'onde R ; seul un décalage de quelques millisecondes est considéré ici : celui qui résulte du calcul de la distance de forme sur l'onde R (décalage qui minimise cette distance).

Cette distance est un des paramètres contribuant à la décision finale d'intégration où non du battement étudié à la famille.

### III.2.1.c Distance sur les paramètres caractéristiques

Cinq « autres distances» sont calculées, qui sont relatives aux paramètres décrits plus haut :

- La distance sur le *rapport RR*, qui correspond à la différence en valeur absolue entre le rapport RR associé à la famille (exposant $B$) et celui associé au battement (exposant $R$) :

$$\varepsilon_{RR} = \left| \frac{RR_s^B}{RR_p^B} - \frac{RR_s^R}{RR_p^R} \right| \qquad \textbf{Eq. 2}$$

- La distance sur le *RR suivant*, qui correspond ici encore à la différence en valeur absolue entre les deux valeurs :

$$\varepsilon_{RRs} = \left| RR_s^B - RR_s^R \right| \qquad \textbf{Eq. 3}$$

- La distance sur *l'amplitude* du complexe, calculée comme la somme des différences des amplitudes de chacune des voies valides :

$$\varepsilon_A = \sum_{voies\ valides} \left| A^B - A^R \right| \qquad \textbf{Eq. 4}$$

- La distance sur *l'angle ACP*, qui est la différence en valeur absolue entre l'angle ACP associé au battement et l'angle associé à la famille. Cette quantité est normalisée par l'écart-type. Pour les familles calculées sur deux voies, on a la relation suivante :

$$\varepsilon_{angle} = \frac{\left| \theta^B - \theta^R \right|}{Std(\theta^B)} \qquad \textbf{Eq. 5}$$

Pour celles qui sont exprimées sur trois voies, cette différence entre « angles » ACP correspond à la distance de Mahalanobis :

$$\varepsilon_{angle} = \left[ \left( \begin{bmatrix} \theta^B \\ \varphi^B \end{bmatrix} - \begin{bmatrix} \theta^R \\ \varphi^R \end{bmatrix} \right)^T Cov(\theta^B, \varphi^B)^{-1} \left( \begin{bmatrix} \theta^B \\ \varphi^B \end{bmatrix} - \begin{bmatrix} \theta^R \\ \varphi^R \end{bmatrix} \right) \right]^{1/2} \qquad \textbf{Eq. 6}$$

- La distance sur le *produit des deux premières valeurs propres*, qui est également la différence en valeur absolue entre les deux produits, normalisées par l'écart type de la distribution :

$$\varepsilon_{vp} = \frac{\left|\lambda_1^B \lambda_2^B - \lambda_1^R \lambda_2^R\right|}{Std(\lambda_1^B \lambda_2^B)} \qquad \textit{Eq. 7}$$

Ces cinq distances avec la distance sur la forme de l'onde R et celle sur la forme de l'onde P sont autant de mesures de « distances » entre le battement considéré et des familles déjà créées : elles permettent de construire une *distance globale* définie au paragraphe suivant.

### III.2.2 Calcul de la distance globale

La *distance globale* $D$ entre le battement et la famille est la *somme pondérée* des distances présentées ci-dessus, excepté la distance mesurée sur l'onde P qui sera prise en considération de manière différente.

$$D = \alpha_{RR}\varepsilon_{RR} + \alpha_{RR_s}\varepsilon_{RR_s} + \alpha_A\varepsilon_A + \alpha_{angle}\varepsilon_{angle} + \alpha_{vp}\varepsilon_{vp} + \alpha_{vp}\varepsilon_{vp} \qquad \textit{Eq. 8}$$

où les $\alpha$ sont les pondérations associées aux différentes distances partielles. Elles ont été fixées empiriquement à partir de l'analyse d'une base d'apprentissage ; leurs valeurs numériques sont indiquées dans le tableau 1.

| Type de distance | Forme $\varepsilon_{FR}$ | rapport RR $\varepsilon_{RR}$ | Amplitude $\varepsilon_A$ | RR suivant $\varepsilon_{RR_s}$ | rapport des valeurs propres[VIII] $\varepsilon_{vp}$ | angle ACP[VI] $\varepsilon_{angle}$ |
|---|---|---|---|---|---|---|
| Pondération de la distance | $\alpha_{FR}$ 20 | $\alpha_{RR}$ 0.16 | $\alpha_A$ $410^{-4}$ | $\alpha_{RRs}$ 0.0025 | $\alpha_{vp}$ 0.04 | $\alpha_{angle}$ 0.08 |

***Tableau 1 : Valeurs numériques des pondérations des distances dans la distance globale D.***

[VIII] Pour mémoire, cette distance est normalisée par l'écart type de la distribution (ou la matrice de covariance).

Les différentes distances portant sur des caractéristiques très différentes les unes des autres, il est difficile de comparer les pondérations entre elles, puisqu'elles tiennent compte à la fois des différences d'échelles entre les différents paramètres et de l'importance relative du rôle de ces paramètres pour discriminer des battements provenant de processus physiologiques distincts. Cependant afin de mieux saisir la pertinence du choix réalisé, nous présentons ci-dessous un exemple, où la valeur de chaque distance apparaît avec sa pondération (Figure 5): on constate que la pondération permet d'obtenir des distances dont les ordres de grandeur sont tout à fait comparables, lorsque prototype et battement à l'étude sont proches.

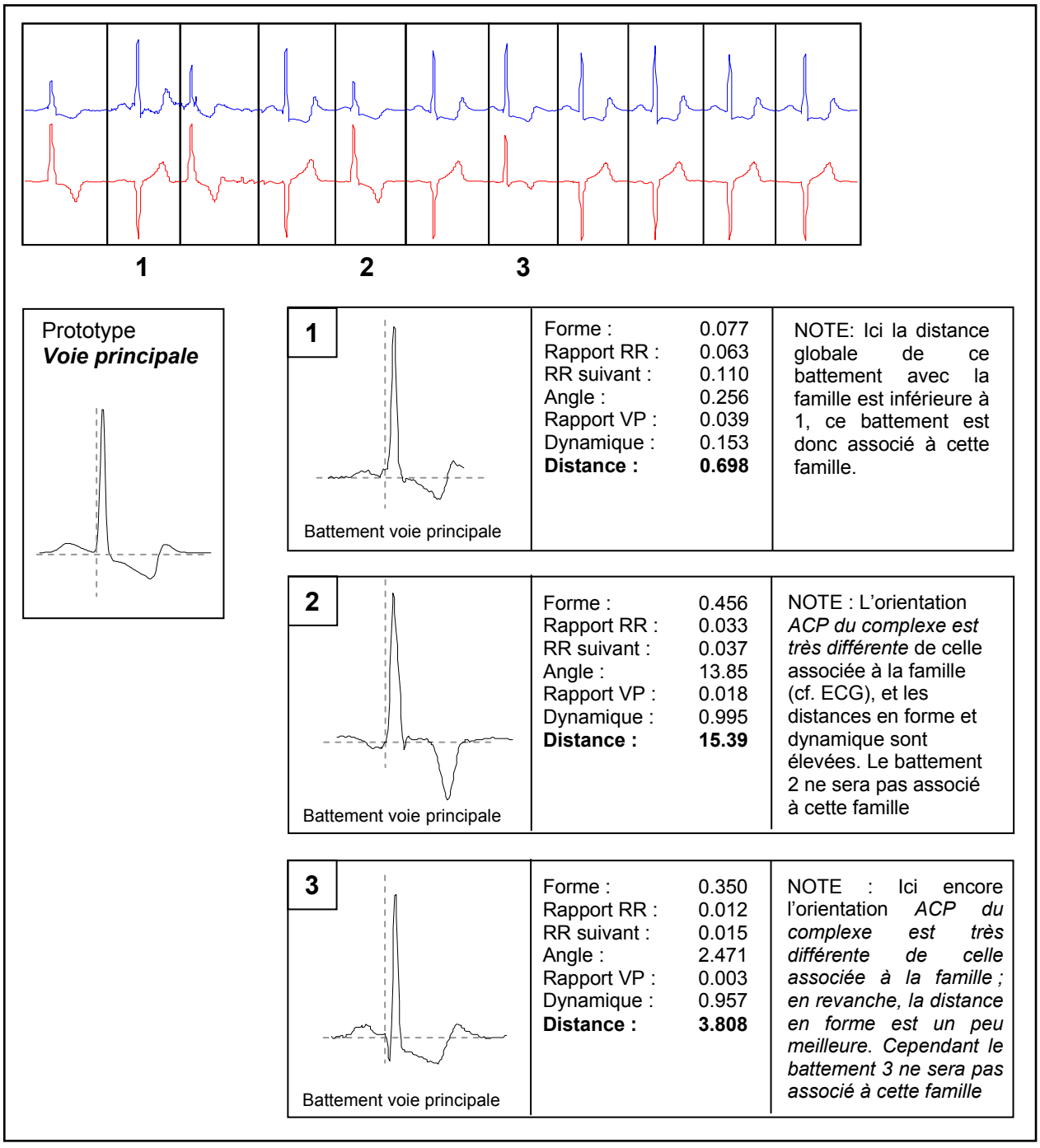


***Figure 5 : les battements 1, 2 et 3 sont comparés à une famille existante dont le prototype est présenté à gauche. Pour chacun de ces battements, on retrouve la valeur pondérée de chaque distance. Seul le battement 1 dans ce cas est associé à la famille.***

### III.2.3 Décision

Les pondérations ci-dessus ont été calculées en cherchant à définir un seuil d'intégration d'un battement dans une famille fixé à la valeur 1. Ainsi, pour qu'un battement donné soit accepté dans une famille, deux conditions doivent être vérifiées :

- la distance globale entre le battement et la famille est inférieure à 1,
- la distance sur l'onde P est inférieure à un seuil prédéfini.

Ainsi, pour chaque battement, un des trois cas suivants se présente :

- une seule famille satisfait les deux conditions : le battement est alors associé à cette famille ;
- aucune des familles existantes ne satisfait simultanément les deux conditions : le battement donne alors naissance à une nouvelle famille ; un prototype de la famille est construit par une modélisation en bosses, et les différents paramètres caractérisant la famille sont alors initialisés ;
- plusieurs familles vérifient les deux conditions : le battement est associé à la famille dont le prototype est le plus proche de ce battement[IX].

Un inconvénient de cette approche réside dans le fait que le nombre de familles n'est pas borné (si ce n'est par le nombre de battements !) ce qui, on l'imagine, pourrait entraîner des divergences, c'est-à-dire la création d'un très grand nombre de famille pour chaque série de 20 mn : on peut imaginer le cas extrême où le nombre de famille créé est égal au nombre de battements soit de l'ordre de 100 000 pour 24 heures, ce qui est considérable en terme d'espace mémoire. Une solution pour remédier à ce problème est le contrôle systématique, au moment de la création d'une nouvelle famille, du nombre total de familles : pour les premiers battements de la série, on autorise la création d'un grand nombre de familles, ce qui permet de s'adapter à toutes les nouvelles formes rencontrées, puis au fur et à mesure, la création de familles est limitée suivant le profil présenté en Figure 6. Donc, si, à un instant donné, un battement nécessite la création d'une famille, mais que le seuil de création de familles est dépassé, on regarde tout d'abord si on peut associer de battement à une des familles existantes en relevant le seuil d'acceptation à 5 (au lieu de 1). Si aucune famille n'est acceptable pour ce nouveau seuil, on crée finalement une nouvelle famille. Le nombre de familles n'est donc pas théoriquement borné, mais en pratique il est fortement limité.

[IX] Le décalage entre l'onde R du battement et celle de la famille est pris en considération à ce niveau.

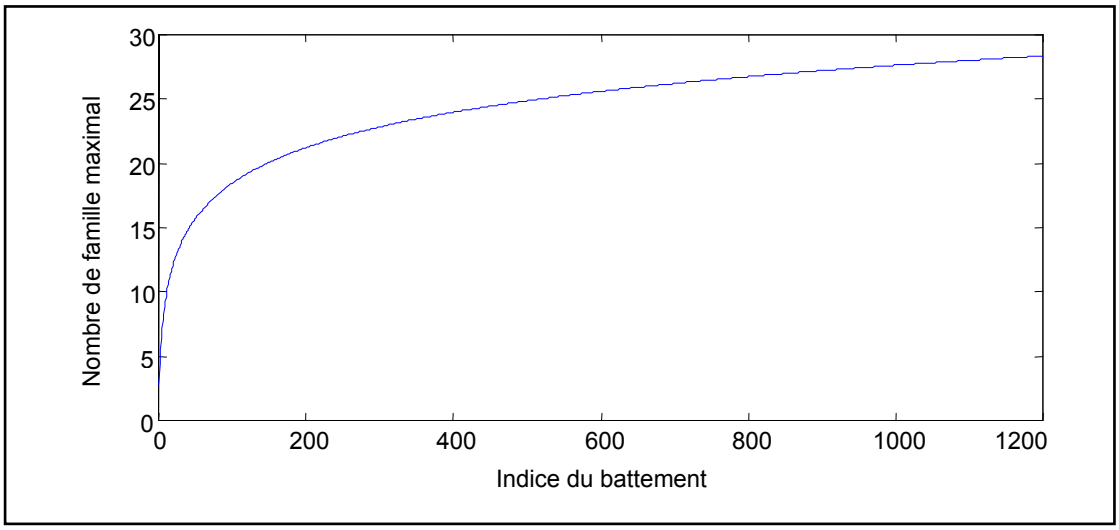


***Figure 6 : Le nombre de familles à chaque instant est contrôlé suivant le profil présenté ci-dessus.Le nombre maximal de familles est limité en fonction de l'indice du battement traité (dans la série de 1200 battements). Si, pour un battement donné, aucune famille existante n'est satisfaisante, et si le nombre maximal de familles à cet instant est atteint, le « seuil d'acceptation » en dessous duquel un battement peut être intégré à une famille est relevé. Suivant ce profil, on constate que l'on autorise au début de la série de nombreuses créations de famille, puis, au fur et à mesure du traitement du signal, on limite de plus en plus ces créations, avec un maximum de 28 familles pour 1200 battements.***

En résumé, cet algorithme est donc un *algorithme non supervisé*, sans aucune connaissance *a priori* du nombre de classes nécessaires pour regrouper les battements de même origine physiologique et surtout pour différencier ceux dont qui résultent de différents processus.

# IV Résultats

La validation de l'algorithme a été effectuée sur la base MIT.

Le premier indice de qualité est l'*homogénéité des familles*. Chaque famille créée par l'algorithme doit être composée de battements tous de même type (en particulier aucune famille ne doit confondre des battements d'origine supraventriculaire avec ceux d'origine

ventriculaire vraisemblable[X]). Le deuxième indice de qualité de l'analyse repose sur le nombre total de famille créé.

Ces deux critères sont opposés : en effet, plus le nombre de familles créées par l'algorithme est important, plus l'homogénéité de chacune d'elle est probable. Inversement, si l'on pénalise la création de nouvelles familles, la taille des familles créées augmente, avec un risque accru d'introduire des éléments qui altèrent l'homogénéité des familles. Les pondérations choisies précédemment permettent un compromis entre ces deux critères, le but étant de ne pas mélanger des battements d'aspect ventriculaire avec des battements d'aspect supraventriculaire lorsque ceux-ci sont clairement distinguables visuellement. Les résultats sont présentés ci-dessous.

## IV.1 Homogénéité des familles

Le principal critère d'homogénéité de la famille est la bonne séparation des battements ventriculaires des battements d'origine sinusale. Sur l'ensemble de la base MIT, et par rapport aux labels de celle-ci, le taux d'erreur est de 1,9%, c'est-à-dire qu'environ 2 battements sur 100 ne sont pas dans des familles homogènes. Il faut cependant noter que les étiquettes de la base distinguent les *prématurités auriculaires* des battements normaux. Or, la forme de l'onde R de tels battements est identique à celle de battements sinusaux (l'impulsion électrique emprunte les mêmes voies de conduction à partir du nœud jonctionnel), et notre classification ne distingue ce type de battement que lorsque l'onde P est vraiment différente.

Ainsi, en ne comptant pas de telles erreurs, le taux d'erreur devient : 0,5%. Le détail des résultats sur l'ensemble de la base MIT est présenté en Annexe D.

[X] À ce niveau de l'analyse, un battement ventriculaire peut être confondu avec un battement d'origine supraventriculaire, nous adressons ici un remerciement au Docteur André Gouérou, du servide de cardiologie de Morlaix, qui a attiré notre attention sur l'existence de complexes QRS qui ressemblent fortement à des ESV mais qui, précédés d'une onde P, peuvent être des complexes d'origine supraventriculaire avec aberration de conduction.

## IV.2 Nombre final de familles

Le nombre de familles par enregistrement de la base MIT est en moyenne de 17 familles créées sur 1 voie et 24 familles créées sur 2 voies. Il est intéressant de ramener ce nombre au nombre de séries de 1200 battements de chaque enregistrement. Ainsi en moyenne pour 1200 battements, on compte 7,3 familles construites à partir d'une voie, et 10,3 familles construites à partir de 2 voies. En moyenne, 3 familles regroupent 90% des battements.
Le détail de cette étude par enregistrement se trouve également en Annexe D.

## IV.3 Perspectives d'amélioration

La classification réalisée ici sépare bien les battements supraventriculaires des battements vraisemblablement ventriculaires, mais ceci parfois au prix de la création de quelques familles inutiles. Ces dernières sont souvent constituées d'un unique battement bruité qui possède soit une forme, soit un autre paramètre, éloigné des familles déjà créées.
Comme nous disposons du niveau des bruits HF et BF de chaque battement, une amélioration envisageable serait d'empêcher la création d'une famille lorsque le battement est trop bruité. Une alternative plus indirecte consisterait à faire varier le seuil de décision sur la distance globale (fixé à 1 ici) en fonction des bruits qui entachent les battements : on fera en sorte que le seuil de décision soit d'autant plus haut que le battement est bruité.

Enfin, compte tenu de l'évolution continue des puissances de calculs disponibles, il est possible que l'étape de classification définie ici devienne inutile, car chaque battement pourra être modélisé par 6 fonctions bosses et, une fois labélissées, ces bosses permettent une meilleure caractérisation de chaque battement et donc un regroupement ultérieur plus fiable.

Résumé :

Ce chapitre présente la classification non supervisée des battements cardiaques. En effet, beaucoup de battements cardiaques proviennent de processus physiologiques identiques et se traduisent par des tracés électrocardiographiques très proches ; il est alors intéressant de rassembler de tels battements pour ne *modéliser qu'un représentant de la classe* ainsi constituée.
Cette classification est effectuée de manière *non supervisée* ; un algorithme spécifique a été développé qui traite les battements par séries de 1200 battements.

On obtient en moyenne sur la base MIT, qui contient des enregistrements d'une demi-heure environ, une vingtaine de classes pour chaque enregistrement, parmi lesquelles deux classes regroupent près de 90% des 1200 battements.

Cet algorithme joue un rôle important dans la rapidité du traitement du signal cardiaque envisagé, car il nous permet de n'étudier que le représentant de chaque famille pour réaliser l'étiquetage des ondes caractéristiques.