Fouille de dynamiques multivariées, application à des données temporelles en cardiologie.

Jerome Dumont

N° d'ordre: 3682

THÈSE

présentée devant

L'UNIVERSITÉ DE RENNES 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention TRAITEMENT DU SIGNAL ET TÉLÉCOMMUNICATIONS

par

Jérôme DUMONT

Equipe d'accueil :	Laboratoire Traitement du Signal et de l'Image (LTSI)
Ecole doctorale :	MATISSE
Composante universitaire :	UFR Structure et Propriétés de la Matière

Titre de la thèse :

Fouille de dynamiques multivariées,
application à des données temporelles en cardiologie.

A soutenir le 9 octobre 2008 devant la commission d'examen

Rapporteurs :	M. Hervé	RIX	Pr, Université de Nice-Sophia Antipolis
	M. Raimon	JANE	Pr, Universitat Politècnica de Catalunya
Examinateurs :	M. Jean-Marc	BOUCHER	Pr, ENST Bretagne
	M. Philippe	MABO	PU-PH, Université de Rennes 1
Directeur de Thèse :	M. Guy	CARRAULT	Pr, Université de Rennes 1
Co-Directeur de Thèse :	M. Alfredo	HERNÁNDEZ	CR INSERM, Université de Rennes 1

La science est une chose merveilleuse...
tant qu'il ne faut pas en vivre !
Albert Einstein

Table des matières

Introduction

Depuis déjà plus d'une cinquantaine d'années, l'ECG s'est révélé être un outil puissant et irremplaçable d'exploration et de diagnostic. L'intérêt est son acquisition, réalisée à partir d'un appareillage simple, peu coûteux et d'une innocuité totale pour le patient. Ceci a de fait favorisé son utilisation en tant qu'examen de routine en milieu hospitalier mais aussi en dehors, par le biais de systèmes portatifs comme le Holter. Comparé aux autres modalités d'observations de l'activité cardiaque, l'ECG est le premier et parfois le seul témoin de modifications se produisant aux niveaux moléculaire et cellulaire. Il constitue un outil de diagnostic essentiel pour des pathologies fréquentes telles que l'ischémie myocardique [Sleker et al., 1997, Pope et Selker, 2003], les arythmies ou des pathologies plus rares comme les dystrophies musculaires cardiaques ou le syndrome de Brugada [Brugada et Brugada, 1992]. Son intérêt a aussi largement dépassé l'exploration propre du myocarde par l'étude du retentissement du système nerveux autonome sur le système cardio-vasculaire, par exemple pour le suivi du réveil en anesthésie [Wodey et al., 2003] ou encore l'étude de la récupération chez des sportifs de haut niveau [Carré, 2002, Reland et al., 2003].

L'acquisition du signal ECG et son exploitation ont aussi bénéficié des nombreux progrès en électronique, en traitement du signal, en informatique et en intelligence artificielle. Plus récemment, ces progrès ont principalement été orientés vers l'extraction d'informations à caractère physiologique, et donc facilement exploitables par les spécialistes, mais aussi vers le développement d'outils plus avancés d'aide au diagnostic fondé sur la physiologie tels que la modélisation cardiaque [Hernández, 2000, Le Rolle, 2006]. Cependant ces méthodes de détection et d'évaluation de pathologies avec le signal ECG ne sont pas toujours satisfaisantes. Les principales difficultés rencontrées concernent notamment la gestion des divers indicateurs extraits du signal ECG, la gestion du temps et de la dynamique temporelle associée aux indicateurs, et finalement la présentation de l'ensemble de ces données aux spécialistes pour en effectuer une interprétation.

Le problème de l'analyse conjointe de données multivariées et de leur représentation à des fins exploratoires a déjà été résolu par de nombreuses méthodes d'analyses descriptives telles que les Analyses en Composates Principales et les Analyses Factorielles des Correspondances ou bien encore par des méthodes de classification multivariées. Ces méthodes ont été appliquées avec succès dans le domaine de l'électrocardiologie [Wong, 2004]. Cependant, les aspects temporels, et notamment la dynamique de l'ensemble des variables étudiées, ne sont que très rarement intégrés dans ces méthodes. Ceci n'est pas sans poser de problèmes car l'information contenue dans l'ECG ou dans les indicateurs extraits est évolutive et reflète des phénomènes transitoires dont l'importance est avérée.

Ce travail s'inscrit donc dans une optique d'exploitation améliorée des dynamiques d'indicateurs extraits de l'ECG. Au vu des méthodes déjà existantes et des besoins en cardiologie, il

nous est apparu important de développer des outils axés sur la recherche d'un bon compromis entre :

- la prise en compte d'une information la plus complète possible (i.e. multivariée et temporelle dans notre cas),
- la clarté et la simplicité pour la représentation et l'interprétation des résultats,
- la flexibilité pour la réalisation des tâches de diagnostic ou d'extraction de connaissances lors d'études exploratoires.

Ainsi, les contributions de ce travail portent i) sur l'extraction d'indicateurs pertinents du signal ECG et plus particulièrement sur la segmentation précise de chaque battement, ii) sur la modélisation des dynamiques de ces indicateurs et iii) sur l'exploitation de ces modèles dans un cadre de fouille de données, pour la classification, le clustering [1] et la représentation des données.

Le chapitre 1 présente des notions de base de l'électrocardiographie clinique pour mieux aborder les objectifs de ce travail. Après un bref rappel de la base anatomique et biochimique du coeur, l'activité électrique cardiaque est introduite. L'acquisition de cette activité électrique par les dérivations cardiaques est expliquée ainsi que l'origine et la signification des ondes élémentaires constituant le signal ECG.

Le chapitre 2 pose le problème de l'analyse des dynamiques des indicateurs extraits du signal ECG. La détection de l'ischémie est prise en exemple pour illustrer notre propos. Les modifications de l'électrogénèse liées à l'ischémie sont tout d'abord détaillées puis plusieurs méthodes, exploitant l'ECG pour le diagnostic de pathologies coronariennes en test d'effort et pour la détection d'épisodes ischémiques dans des enregistrements Holters, sont commentées. Nous verrons que les aspects multivariables et temporelles sont très importants mais finalement pris en compte de manière peu satisfaisante dans ces méthodes. Ceci nous a conduit à formaliser et à proposer une démarche de résolution qui s'inscrit dans le cadre de la fouille de données temporelles et multivariées.

Le chapitre 3 discute de l'extraction des caractéristiques du signal ECG. Les caractéristiques employées sont majoritairement issues de la segmentation des ondes fondamentales P, Q, R, S et T de chaque battement. La première partie de ce chapitre est donc un état de l'art des méthodes de segmentation de l'ECG. Ceci nous a amené à mettre en place, dans la suite de ce chapitre, un algorithme de segmentation basé sur une décomposition en ondelettes et à résoudre le délicat problème de l'ajustement des paramètres et des seuils de décision par un algorithme évolutionnaire. Les performances obtenues par cet algorithme, avec les paramètres et les seuils optimisés, sont évaluées sur une base de données de signaux ECG segmentés manuellement. Enfin, cet algorithme est intégré dans une station d'analyse de l'ECG directement exploitable par un clinicien dans un cadre de recherche. Cette station permet d'extraire un large ensemble d'indicateurs et de les présenter aux utilisateurs.

Le chapitre 4 est dédié à la modélisation des séries temporelles extraites, en voulant caractériser principalement leur dynamique, et à l'exploitation de modèles pour la fouille de données. Tout d'abord, un état de l'art des différents modèles de représentation des séries temporelles multivariées est présenté. Nous nous sommes ensuite intéressés plus précisément aux modèles

[1] nous différencierons dans ce travail la notion de classification et de clustering. La première est liée à une notion d'apprentissage supervisé tandis que la seconde constitue des classes sans connaissance *à priori*, et relève donc de l'apprentissage non supervisé.

Semi-Markovien Cachés (MSMC) pour représenter, simuler et classer des séries temporelles. Le clustering, plus complexe, a été résolu indépendamment à l'aide d'un algorithme Expectation-Maximisation (EM) flou, toujours basé sur les MSMC. Des résultats sur données simulées sont présentés dans ce chapitre afin de souligner le bien fondé de nos propositions.

Enfin, le chapitre 5 réalise une exploitation des algorithmes pour l'étude de deux pathologies différentes. La classification et le clustering sont appliqués pour détecter des épisodes ischémiques et, dans une seconde section, une étude exploratoire et de classification de patients atteints du syndrome de Brugada est menée à partir d'ECG acquis lors d'une épreuve d'effort.

Bibliographie

[Brugada et Brugada, 1992] Brugada, P. et Brugada, J. (1992). Right bundle branch block, persistent st segment elevation and sudden cardiac death : a distinct clinical and electrocardiographical syndrome. a multicenter report. *J. Am. Coll. Cardiol.*, 20 :1391–6.

[Carré, 2002] Carré, F. (2002). Cardiovascular benefits and hazard of physical practice. *Ann Cardiol Angeiol*, 51(6) :351–6.

[Hernández, 2000] Hernández, A. I. (2000). *Fusion de signaux et de modèles pour la caractérisation d'arythmies cardiaques.* PhD thesis, Université de Rennes 1, France.

[Le Rolle, 2006] Le Rolle, V. (2006). *Modélisation multi-formalisme du système cardiovasculaire associant Bond Graph, équations différentielles et modèles discrets.* PhD thesis, Université de Rennes 1, France.

[Pope et Selker, 2003] Pope, J. et Selker, H. (2003). Diagnosis of acute cardiac ischemia. *Emergency Medicine Clinics of North America*, 21(1) :27–59.

[Reland et al., 2003] Reland, S., Ville, N., Wong, S., Gauvrit, H., Kervio, G., et Carré, F. (2003). Exercice heart rate variability of older women in relation to the level of physical activity. *Journal of Gerontology Series A, Biological Sciences*, 58(7) :585–91.

[Sleker et al., 1997] Sleker, H., Zalenski, R., Antman, E., Aufderheide, T., SA, S. B., Bonow, R., Hagen, W. G. M., Johnson, P., Lau, J., McNutt, R., Ornato, J., Schwartz, J., Scott, J., Tunick, P., et Weaver, W. (1997). An evaluation of technologies for identifying acute cardiac ischemia in the emergency department : A report from a national heart attack alert program working group. *Ann Emerg Med January*, 29(1) :17–20.

[Wodey et al., 2003] Wodey, E., Senhadji, L., Bansard, J. Y., Terrier, A., Carré, F., et Ecoffey, C. (2003). Comparison of heart rate response to an epinephrine test dose and painful stimulus in children during sevoflurane anesthesia : heart rate variability and beat-to-beat analysis. *Regional Anesthesia and Pain Medicine*, 28(5) :439–444.

[Wong, 2004] Wong, S. (2004). *Segmentation de l'intervalle RT et description par analyse factorielle de la variabilité de la fréquence cardiaque et de la repolarisation ventriculaire.* PhD thesis, Université de Rennes 1, France.

Chapitre 1

Electrophysiologie cardiaque

L'objectif de ce travail est de mettre en oeuvre des nouvelles méthodes d'exploitation et de traitement du signal ECG afin de réaliser des outils de diagnostic et d'exploration temporels concernant les pathologies cardiaques évolutives. Ce premier chapitre n'a pas prétention à l'originalité, il a seulement pour ambition de donner les bases physiologiques pour la compréhension de l'origine du signal ECG et des travaux réalisés dans la suite de ce mémoire. Il s'inspire donc fortement des thèses précédentes réalisées au laboratoire [Hernández, 2000, Defontaine, 2006].

1.1 Généralités sur le système cardiovasculaire

La fonction principale du système cardiovasculaire est d'assurer un flux de sang continu aux organes et aux tissus cellulaires du corps, pour *i)* leur fournir de l'oxygène et des nutriments, *ii)* évacuer les produits métaboliques générés pendant leur activité et *iii)* transporter les hormones produites par les glandes endocrines vers les récepteurs. Ce système est constitué d'un organe pompe, le cœur, et d'un réseau continu et fermé de conduits qui permettent le transport du sang, le système vasculaire.

Le cœur est situé vers le front de la cavité thoracique et est légèrement déplacé vers la gauche. Sa forme est similaire à un cône inversé (sa base vers le haut et à droite et son apex en bas et à gauche). L'axe anatomique du cœur (une ligne imaginaire de la base jusqu'à l'apex) est défini entre la partie supérieure et postérieure du thorax droit, jusqu'à la partie basse, antérieure et gauche. La localisation exacte du cœur peut varier d'un individu à un autre par la forme du diaphragme et les différences de taille du cœur.

Les parois du cœur sont constituées par un muscle, le myocarde, qui est composé majoritairement de fibres contractiles disposées de façon spiroïdale autour du cône, les autres étant dirigées linéairement de la base vers l'apex. Le cœur est divisé en quatre cavités (figure 1.1). Les deux supérieures, les oreillettes gauche et droite, sont chargées de recevoir le sang et sont séparées par le septum interauriculaire. Les deux cavités inférieures, les ventricules gauche et droit, sont divisées par le septum interventriculaire et assurent l'expulsion de sang dans le système vasculaire. Les ventricules sont séparés des oreillettes au moyen des valves auriculo-ventriculaires, formées par des ailerons de tissu connectif. La valve tricuspide sépare l'oreillette et le ventricule droit et la valve mitrale sépare l'oreillette gauche du ventricule gauche. La fonction des valves auriculo-ventriculaires est d'éviter une réentrée du sang aux oreillettes une fois qu'il est arrivé aux ventricules tandis que les valves sigmoïdes (pulmonaire et aortique) évitent le retour du sang vers les ventricules, une fois expulsé vers l'artère pulmonaire et vers l'aorte.

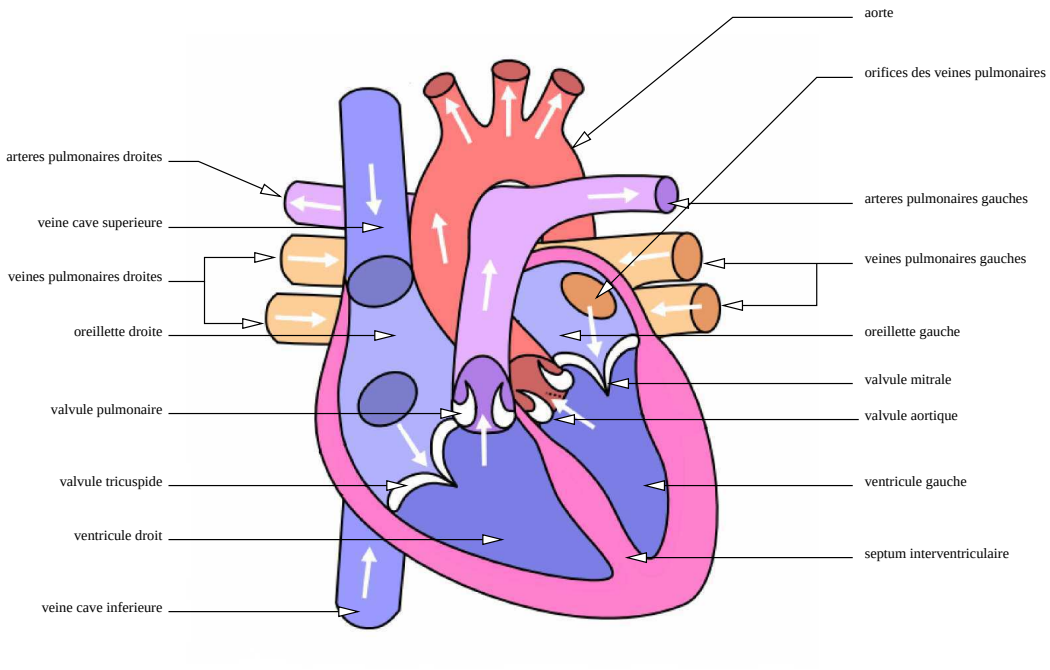


FIG. 1.1 – Structure anatomique du cœur – *image d'après Wikipedia, permission de copier, distribuer et/ou modifier ce document selon les termes de la Licence de Documentation Libre GNU (GFDL).*

Les événements mécaniques successifs qui caractérisent la fonction de pompe du cœur pendant le cycle cardiaque peuvent être divisés en deux périodes : la systole et la diastole. La systole est la période d'éjection sanguine et est composée de trois phases : la systole auriculaire (figure 1.2 a), la contraction ventriculaire isovolumique et la systole ventriculaire (figure 1.2 b). La diastole est la période de relaxation du cœur, pendant laquelle il est rempli de sang. Cette période est composée de deux phases : la relaxation ventriculaire isométrique ou "proto-diastole" et la phase finale de la diastole, ou période de repos du cœur.

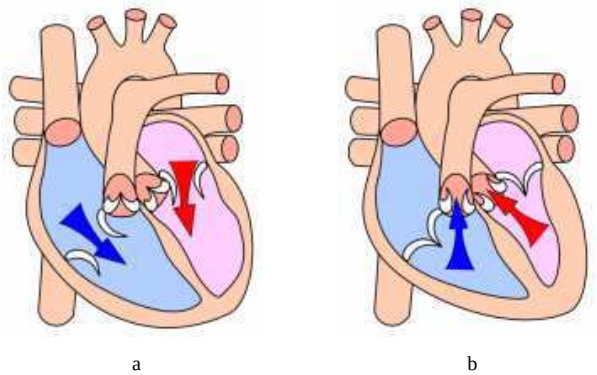


FIG. 1.2 – Systoles *a)* auriculaire et *b)* ventriculaire – *d'après Wikipedia, permission de copier, distribuer et/ou modifier ce document selon les termes de la Licence de Documentation Libre GNU (GFDL).*

Cette fonction mécanique cardiaque est la conséquence d'une activation électrique organisée du myocarde. Pour accomplir cette fonction de pompe, le myocarde est constitué principale-

ment de deux types de tissus :

- **le tissu de conduction ou tissu nodal** : ce tissu est constitué de cellules présentant des propriétés spécialisées d'excitabilité, de conductibilité et d'automaticité. Ces propriétés permettent la génération régulière et spontanée des impulsions électriques et la conduction de ces impulsions d'une manière organisée au travers du myocarde, afin d'assurer une contraction adéquate et un pompage efficace ;
- **le tissu myocardique contractile** : ce type de tissu est largement majoritaire et présente aussi des propriétés d'excitabilité et de conductibilité cellulaire. Cependant, à la différence du tissu nodal, il est constitué de cellules pourvues d'un grand nombre de fibres musculaires capables de se contracter.

1.2 Activité électrique cardiaque

Chaque cellule cardiaque (nodale ou myocardique) est entourée et remplie avec une solution qui contient des ions [Rudy, 2004]. Les trois plus importants sont le sodium (Na^+), le potassium (K^+) et le calcium (Ca^{2+}). Dans la période de repos de la cellule, l'intérieur de la membrane cellulaire est chargé négativement par rapport à l'extérieur, qui est pris comme référence. Cette différence de potentiel, ou potentiel de repos cellulaire, est approximativement de $-85\ mV$ pour les cellules ventriculaires et dépend : *i)* des concentrations ioniques dans les milieux intracellulaire et extracellulaire (équation de Nernst) et *ii)* des protéines, chargées négativement, qui présentent une concentration majeure dans le milieu intracellulaire. Les processus actifs et passifs de mouvement des ions au travers des canaux ioniques traversant la membrane cellulaire, ainsi que la propagation de ces ions de cellule à cellule, constituent les fondements de l'activité électrique cellulaire.

1.2.1 Propriété d'excitation (le potentiel d'action)

Quand une impulsion électrique d'amplitude suffisante (supra-liminaire) arrive à une cellule excitable, l'intérieur de cette cellule (nodale ou myocardique) devient rapidement positif par rapport à l'extérieur. Ce processus est connu comme la dépolarisation cellulaire. Le retour de la cellule cardiaque stimulée à son état de repos est appelé repolarisation. A la fin de cette dernière phase, l'intérieur de la membrane cellulaire récupère sa négativité normale et, dans les cellules myocardiques, reste dans un état de repos jusqu'à l'arrivée d'une nouvelle excitation. L'enregistrement des différences de potentiel mesurées entre les milieux intracellulaire et extracellulaire, pendant les processus de dépolarisation et de repolarisation d'une cellule, correspond au Potentiel d'Action cellulaire (PA). Il est constitué de cinq phases (figure 1.3) :

- **la phase 0 ou dépolarisation rapide** : après une excitation électrique au-dessus du seuil d'activation (ou potentiel liminaire) de la cellule en repos, les potentiels mesurés présentent une inversion rapide de polarité. Cette dépolarisation est générée par l'ouverture de plusieurs canaux ioniques, dépendant de la différence de potentiel transmembranaire et permettant la diffusion passive et généralisée de Na^+ vers le milieu intracellulaire ;
- **la phase 1 ou début de la repolarisation** : elle se caractérise par une repolarisation rapide, de courte durée, due à l'inactivation des canaux Na^+ ;
- **la phase 2 ou plateau** : pendant cette phase, la repolarisation continue mais à un taux très lent. Le plateau est principalement dû à l'ouverture des canaux calciques, permettant

la diffusion lente du Ca^{2+} vers l'intérieur de la cellule ;
- l**a phase 3 ou repolarisation rapide** : elle est caractérisée par une repolarisation majeure, produite par la fermeture des canaux ioniques spécifiques, qui emmène la cellule au potentiel de repos original. Il existe aussi, dans la dernière partie de la phase 3, une activation des canaux de potassium, autorisant l'expulsion de ces ions et facilitant le retour à la négativité originale du potentiel transmembranaire ;
- **la phase** 4 : elle correspond au potentiel de repos, où la cellule devient plus facilement excitable. Les caractéristiques de cette phase dépendent du type de cellule concernée.

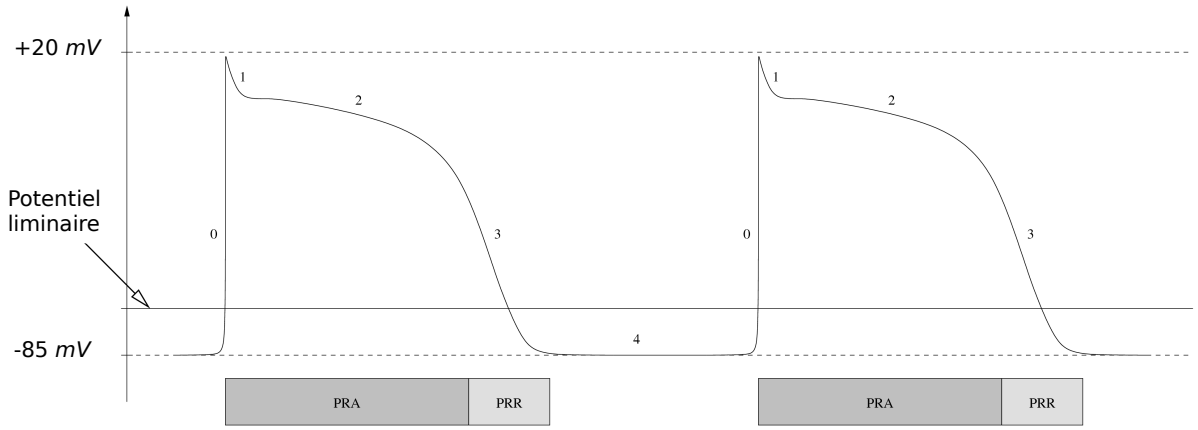


FIG. 1.3 – Potentiel d'action des cellules cardiaques ventriculaires.

La propriété d'excitabilité cellulaire change pendant les différentes phases du PA. Pendant les phases 0, 1, 2 et la première partie de la phase 3 (généralement jusqu'aux alentours de -50 mV), une stimulation externe sera incapable de provoquer un nouveau PA (figure 1.3). Cette période est appelée la Période Réfractaire Absolue (PRA). La durée de la PRA dépend de la fréquence à laquelle la cellule est stimulée, plus longue pour des fréquences plus basses et inversement pour des fréquences de stimulation plus élevées. La Période Réfractaire Relative (PRR) est associée à la dernière partie de la phase 3 (souvent pour des potentiels inférieurs à -50 mV). Pendant la PRR, une stimulation d'amplitude supérieure à la normale (supra-stimulus) peut provoquer un nouveau PA, qui présentera une durée de la phase 0 rallongée en fonction de la prématurité de la stimulation.

Outre les variations morphologiques des PA dérivées des propriétés d'adaptation à la fréquence de stimulation, communes à toutes les cellules cardiaques, la durée et la morphologie de chacune des phases du PA dépendent *i)* du type de cellule considérée (nodale ou musculaire), *ii)* de la localisation de la cellule dans le myocarde et *iii)* de l'état physiopathologique du patient. Dans tous les cas, l'activation électrique des cellules cardiaques est due à un enchaînement organisé d'activations de canaux ioniques, ce qui permet une diffusion sélective de certains ions. Dans les cellules cardiaques, les ions les plus importants sont les ions Na^+, K^+ et Ca^{2+}. Les diffusions de ces ions au travers de la membrane cellulaire créent des courants ioniques qui s'additionnent pour générer le potentiel d'action. La figure 1.4 présente un exemple de PA de myocyte ventriculaire et les principaux courants associés.

La sous-section suivante explique comment ces potentiels d'actions se propagent de cellules en cellules dans le tissu myocardique et le tissu nodal.

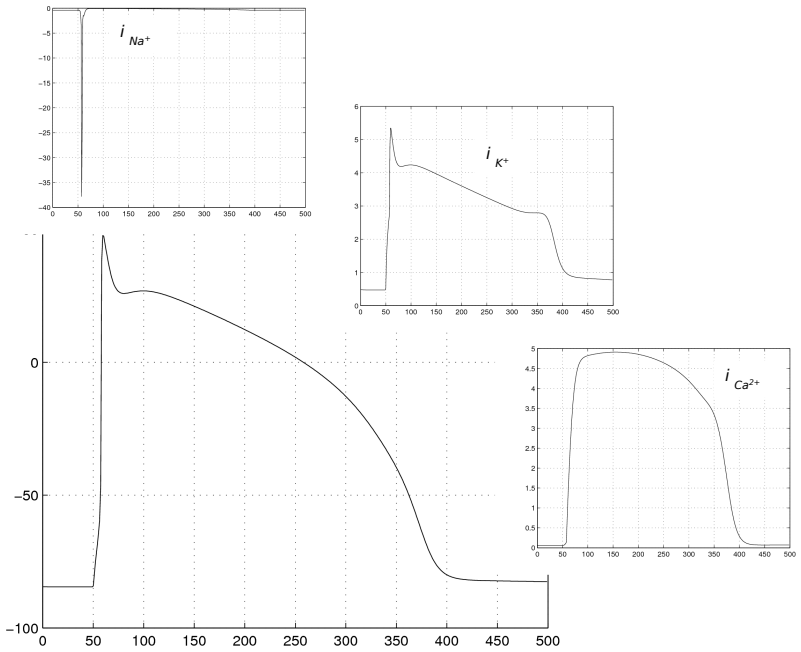


FIG. 1.4 – Exemple d'un potentiel d'action ventriculaire et principaux courants ioniques associés (ions Na^+, K^+ et Ca^{2+}).

1.2.2 Propriété de conductibilité

L'excitation électrique d'une cellule cardiaque (nodale ou myocardique) génère un "front de dépolarisation" se propageant le long de la cellule. Cette propagation intracellulaire est assurée par la diffusion des ions aux surfaces interne et externe de la membrane cellulaire, due au gradient électrique créé sur les frontières du front d'activation (figure 1.5). A cause de cette diffusion ionique, les zones de la cellule les plus proches du front d'activation voient leur différence de potentiel transmembranaire réduite et atteignent leur potentiel liminaire. Le PA est donc propagé de façon unidirectionnelle, à partir du point d'excitation vers les extrémités de la cellule. La direction de cette propagation, dans une cellule, ne peut pas être modifiée à cause de l'impossibilité d'exciter les zones réfractaires.

Les milieux intracellulaires de deux cellules voisines sont connectés physiquement (cytoplasme à cytoplasme) au moyen des petites anastomoses présentes dans les membranes cellulaires, appelées disques intercalaires. Ces derniers représentent des passages intercellulaires permettant la diffusion des ions entre les deux cellules. Quand une cellule est dépolarisée et que son front de dépolarisation arrive à une extrémité, un gradient électrique est créé entre les milieux intracellulaires de la cellule dépolarisée et ses voisines en état de repos. Ce gradient provoque la diffusion des ions au travers des disques intercalaires, générant une dépolarisation progressive des cellules voisines pour atteindre leurs potentiels liminaires. Ainsi, le front d'activation est transmis aux cellules voisines (conduction intercellulaire).

1.2.3 Propriétés spécifiques des cellules nodales

Les cellules nodales présentent des propriétés d'excitation et de conduction électrique spécialisées se traduisant par une conduction plus rapide du front de dépolarisation et une exci-

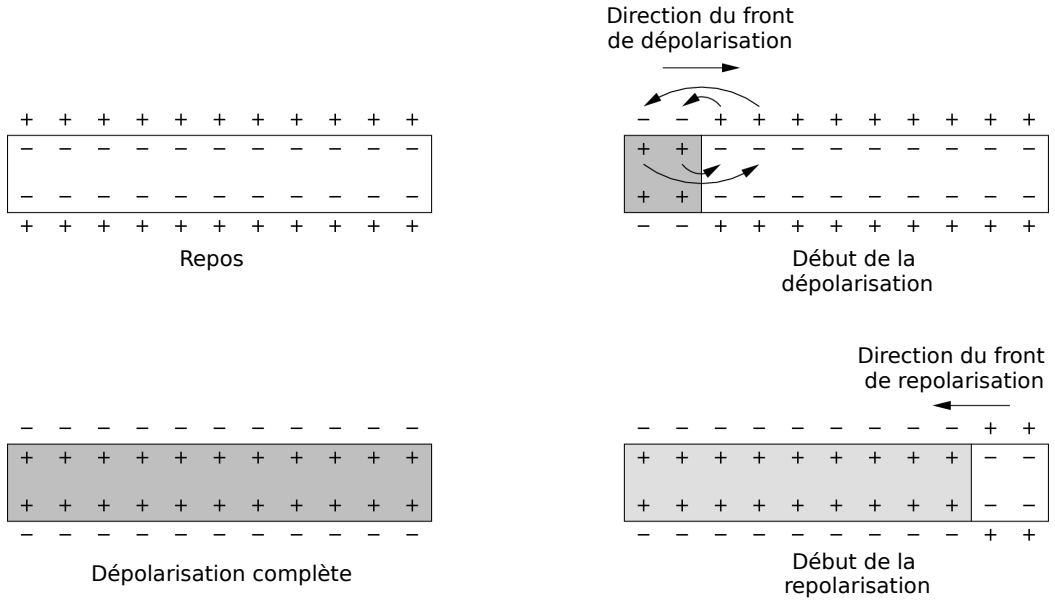


FIG. 1.5 – Cycle de dépolarisation - repolarisation.

tabilité facilitée. De plus, ces cellules ont la propriété d'automatisme, c'est-à-dire, la capacité de se dépolariser lentement et spontanément pendant la phase 4, pour atteindre le potentiel liminaire et produire un nouveau potentiel d'action sans être stimulées extérieurement. Le tissu nodal est disposé dans un système spécialisé d'excitation et de conduction, responsable de la génération et de la conduction organisées du front d'activation cardiaque, afin de produire une contraction myocardique efficace. Le système spécialisé d'excitation-conduction comprend : le nœud sinusal, le nœud auriculo-ventriculaire (NAV), le faisceau de His, avec les branches gauche et droite, et les fibres de Purkinje, localisés sur la figure 1.6.

Les cellules associées à chacune des parties du système d'excitation-conduction présentent une pente de dépolarisation diastolique lente (phase 4) différente. Dans le cas physiologique, cette pente est plus prononcée sur le nœud sinusal. Ceci implique que le seuil liminaire est atteint et qu'un potentiel d'action est généré plus rapidement dans le nœud sinusal que dans les autres parties du système spécialisé. Ainsi, le nœud sinusal est appelé le pacemaker dominant du cœur et les autres centres du tissu d'excitation contraction sont considérés comme des pacemakers latents ou subsidiaires. Dans le cas normal, l'activité électrique du cœur suit la séquence d'activation suivante :

- **le nœud sinusal (NS)** : l'activité électrique est générée spontanément dans le nœud sinusal. Il est situé dans la partie haute de la paroi intérieure de l'oreillette droite, au niveau où débouche la veine cave supérieure ;
- **les oreillettes** : l'impulsion cardiaque initiée dans le nœud sinusal est transmise aux deux oreillettes ;
- **le nœud auriculo-ventriculaire (NAV)** : il est situé en bas de l'oreillette droite et est constitué de cellules qui présentent une conduction électrique lente. L'activation électrique qui arrive au NAV est ralentie (approximativement 100 ms) avant d'arriver au faisceau de His. Cette propriété physiologique du NAV permet de protéger les ventricules

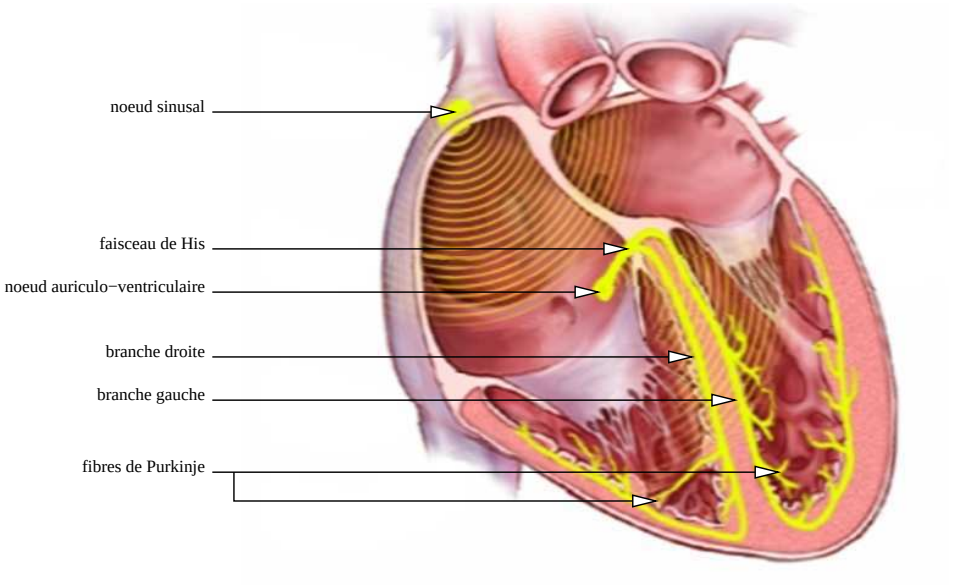


FIG. 1.6 – Localisation du système spécialisé de conduction.

d'un nombre excessif d'activations auriculaires et concède aux oreillettes un temps de vidange plus grand, optimisant ainsi la contraction ventriculaire ;

- **le faisceau de His** : il est situé dans la partie haute du septum interventriculaire et ses fibres traversent le tissu connectif (non excitable) qui sépare électriquement les oreillettes des ventricules. Dans les cas normaux, le NAV et le faisceau de His constituent la seule voie de propagation de l'activité électrique cardiaque entre les oreillettes et les ventricules. L'ensemble de ces deux structures est souvent appelé la jonction auriculo-ventriculaire. Une fois dans l'étage ventriculaire, le faisceau de His se divise en deux branches, la gauche et la droite ;
- **la branche droite** : elle est située le long de la partie droite du septum interventriculaire et facilite la conduction de l'activité électrique dans le ventricule droit ;
- **la branche gauche** : cette branche se subdivise en deux hémibranches, l'une antérieure gauche et l'autre postérieure gauche ;
- **les fibres de Purkinje** : les branches du faisceau de His finissent dans un réseau de fibres qui arrivent dans les parois ventriculaires. Les fibres de Purkinje terminent en anastomoses avec les fibres myocardiques musculaires, facilitant leur excitation.

1.3 Observation de l'activité électrique cardiaque

La direction du flux et l'amplitude des courants électriques générés par les processus de dépolarisation et de repolarisation de l'ensemble des cellules myocardiques peuvent être détectées par des électrodes disposées sur la surface du thorax. L'analyse de cette activité électrique s'est révélée comme une technique primordiale pour le diagnostic des maladies cardiovasculaires et constitue un outil fondamental dans le monitoring cardiaque. Cette section explique brièvement la relation entre les potentiels observés à la surface du corps et les potentiels d'action cellulaire, ainsi que les moyens d'observation de cette activité.

1.3.1 Electrogenèse cardiaque et origine du vectocardiogramme (VCG)

Dans la surface externe de la membrane de chaque fibre cardiaque excitée, deux zones de polarités différentes peuvent être observées de chaque côté du front d'activation (figure 1.5) : une zone dépolarisée, chargée négativement, et une zone positive adjacente (prête à être dépolarisée). Ces deux zones constituent un dipôle qui peut être représenté, à chaque instant t, par un vecteur de champ électrique $\vec{v}$ (ou vecteur dipolaire élémentaire instantané). La direction de ce vecteur est confondue avec l'axe anatomique de la fibre cardiaque et son amplitude dépend du PA associé à la cellule concernée. Ces concepts sont aussi applicables au processus de repolarisation, pendant lequel le dipôle présente une amplitude et une direction similaire à celles de la dépolarisation, mais de sens inverse. L'ensemble de ces dipôles équivalents qui correspondent à toutes les cellules actives (en dépolarisation ou repolarisation) à un instant t, peut être enregistré de façon invasive sur la surface du cœur ou dans ses chambres. Les signaux ainsi acquis sont désignés par "électrogrammes".

Le thorax étant un volume conducteur, les potentiels sont propagés de la surface du cœur à la surface du corps. Cependant, le thorax est un volume conducteur anisotropique, irrégulier et dynamique, ce qui empêche une propagation uniforme et produit des distorsions dans le champ électrique superficiel. L'approche d'électrogenèse de l'électrocardiogramme, proposée par Einthoven au début du siècle, représente une simplification de ce problème. Cette approche, qui s'est révélée très utile en pratique clinique, est basée sur la notion du dipôle cardiaque équivalent et suppose qu'un vecteur cardiaque instantané peut être estimé, pour tout instant t, par la somme vectorielle exacte de tous les vecteurs élémentaires instantanés associés aux cellules actives dans le myocarde et que ce vecteur coïncide avec l'observation sur la surface thoracique. La direction et le sens du vecteur cardiaque instantané représentent ceux du front d'activation cardiaque et son module dépend de la quantité et du type de cellules actives à l'instant t.

Dans le même sens, une représentation de l'activité électrique moyenne du cœur dans un intervalle temporel $[t_1\ t_2]$ (ou vecteur cardiaque moyen) peut être calculée par une simple somme vectorielle des dipôles cardiaques instantanés correspondant à l'intervalle étudié. Ainsi, les différents étages de l'activation électrique cardiaque peuvent être caractérisés par une séquence de vecteurs cardiaques moyens, calculés sur les intervalles temporels correspondants. On distingue (figure 1.7) :

i) **Le vecteur cardiaque moyen d'activation auriculaire** : il représente l'excitation auriculaire et correspond au processus de dépolarisation du nœud sinusal, des voies internodales et des oreillettes.

ii) **Le vecteur cardiaque moyen d'activation ventriculaire** : il est associé à l'excitation globale ventriculaire et sa direction, appelée aussi axe électrique du cœur, est un indicateur clinique utile pour le diagnostic de quelques pathologies (comme la cardio-mégalie). De plus, il explique l'interrelation morphologique entre les différentes voies de l'ECG de surface. L'axe électrique du cœur présente une variation importante interindividus, mais aussi, pour le même individu, notamment au cours du temps. Le vecteur moyen ventriculaire peut être décomposé en trois vecteurs successifs :

- **le vecteur d'activation septale** : il représente le début de la dépolarisation ventriculaire, qui survient dans la partie centrale du septum interventriculaire. Il est de petite amplitude et dirigé vers la droite, en bas et avant ;

- **le vecteur d'activation ventriculaire** : il représente la dépolarisation des parois ventriculaires. C'est le vecteur de plus grande amplitude et il est dirigé vers la gauche (parce que le ventricule gauche présente une masse musculaire plus importante), vers le bas et vers le dos ;
- **le vecteur d'activation basale** : il représente la dépolarisation des portions basses des ventricules. Il présente une amplitude réduite et est dirigé vers le haut et vers le dos.

iii) **Le vecteur cardiaque moyen de repolarisation ventriculaire** : il correspond à la repolarisation des ventricules. Il est d'amplitude moyenne et de sens inverse au vecteur de dépolarisation ventriculaire.

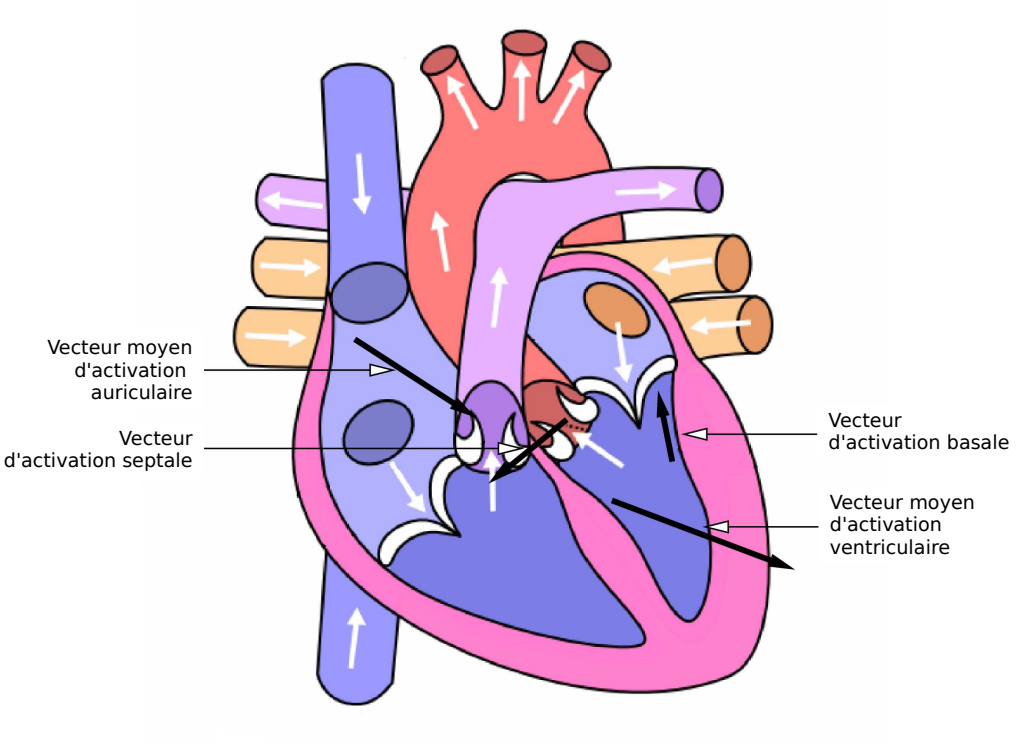


FIG. 1.7 – Représentation vectorielle du processus de dépolarisation sur le plan frontal.

Le vectocardiogramme (VCG) est l'enregistrement continu des vecteurs cardiaques instantanés, obtenu au moyen de l'application d'un ensemble d'électrodes sur la surface du thorax. Il est caractérisé par une séquence de boucles décrites dans l'espace, associée à la dépolarisation et repolarisation du tissu myocardique (figure 1.8). Bien que l'étude du VCG ait eu une application clinique importante pendant les années 50 et 60, il est aujourd'hui largement remplacé par l'analyse des 12 dérivations standard de l'électrocardiogramme.

1.3.2 L'électrocardiographie 12 dérivations

A la différence du VCG, qui représente l'activité électrique cardiaque dans l'espace, l'électrocardiogramme (ECG) est l'enregistrement des mêmes potentiels électriques cardiaques, projetés sur un axe spécifique, en fonction du temps (figure 1.8). Chacun de ces axes, ou dérivations, est capable d'observer, avec une résolution propre, les phénomènes électriques

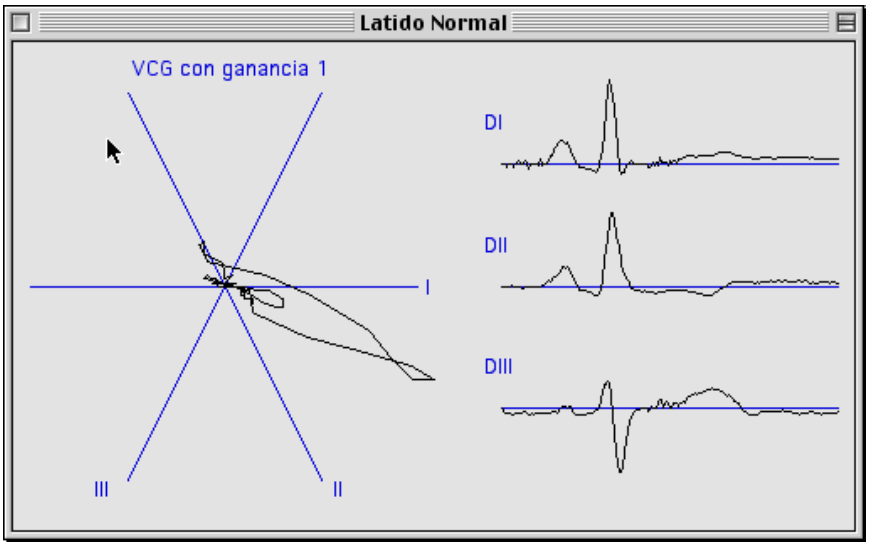


FIG. 1.8 – Vectocardiogramme sur le plan frontal et sa projection sur trois dérivations standard ECG.

cardiaques. L'électrocardiographie moderne s'appuie sur l'étude de 12 dérivations (ou dérivations standard). Parmi les 12 dérivations, trois sont appelées bipolaires, trois unipolaires augmentées et six précordiales. Différentes méthodes de placement d'électrodes ont été proposées pour l'électrocardiographie de 12 dérivations. La méthode classique consiste à placer les électrodes sur les deux bras et la jambe gauche du patient pour obtenir les trois dérivations bipolaires et les trois unipolaires augmentées (figure 1.9). Une quatrième électrode est placée sur la jambe droite afin de réduire le bruit de mode commun dans l'étape d'amplification, mais cette électrode ne contribue pas à la formation des dérivations.

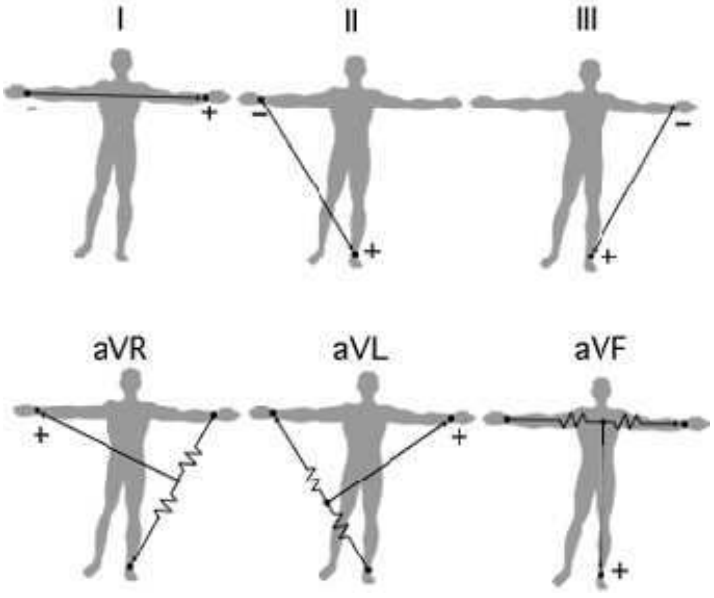


FIG. 1.9 – Dérivations bipolaires et triangle d'Einthoven (haut) ; dérivations unipolaires augmentées (bas).

Les trois dérivations bipolaires ont été introduites par Einthoven et déterminent la différence de potentiel entre les couples d'électrodes disposés sur les membres (figure 1.9) : la dérivation *DI* représente la différence du potentiel mesurée entre le bras droit (électrode négative) et le bras gauche (électrode positive), la dérivation *DII* entre le bras droit (électrode négative) et la jambe gauche (électrode positive) et la dérivation *DIII* entre le bras gauche (électrode négative) et la jambe gauche (électrode positive).

Ces trois dérivations forment un triangle sur le corps (ou triangle d'Einthoven) et, selon l'hypothèse du dipôle cardiaque équivalent, suivent une relation simple : l'amplitude du potentiel enregistré dans la dérivation *DI* plus celle de la dérivation *DIII* est égale à l'amplitude des potentiels dans la dérivation *DII* ($DI + DIII = DII$).

Les dérivations unipolaires ont été introduites initialement par Wilson en mesurant les potentiels de chaque membre du triangle d'Einthoven par rapport à une référence, appelée borne centrale de Wilson, construite en appliquant une résistance de 5000 Ω aux trois électrodes des membres (figure 1.9). Golberger a introduit plus tard le concept de dérivations unipolaires augmentées. Les dérivations unipolaires des membres sont acquises en plaçant l'électrode positive sur : le bras droit (aVR), le bras gauche (aVL) et la jambe gauche (aVF) (figure 1.9).

Les dérivations précordiales, notées $V1$ à $V6$, sont aussi des dérivations unipolaires, qui mesurent la différence de potentiel entre la borne centrale de Wilson et l'ensemble de positions spécifiques de la surface thoracique montrées dans la figure 1.10.

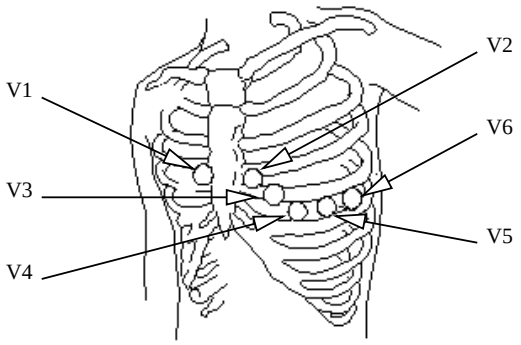


FIG. 1.10 – Dérivations précordiales (V_1 à V_6).

La relation entre l'ECG de 12 dérivations et le VCG, ainsi que la transformation entre les différents schémas d'application d'électrodes, ont été des sujets de recherche importants dans les années 70 et 80. Les contributions les plus importantes sont dues à Dower, qui a proposé une matrice de transformation linéaire pour la génération des 12 dérivations ECG à partir des voies X, Y et Z du VCG [Dower et al., 1980, Dower, 1984] et à Levkov, qui a traité la synthèse du VCG à partir des potentiels mesurés par les 12 dérivations standard d'ECG, en s'appuyant aussi sur une méthode linéaire [Levkov, 1987].

1.3.3 Configuration de l'ECG physiologique

Les processus de dépolarisation et repolarisation des structures myocardiques se présentent dans l'ECG comme une séquence de déflections ou ondes superposées à une ligne de potentiel zéro, appelée ligne isoélectrique (figure 1.11). L'ordre et la morphologie de ces ondes dépendent de deux aspects fondamentaux : *i)* la structure anatomique d'initiation de l'impulsion électrique (*i.e.* le nœud sinusal, une structure jonctionelle, ...) et *ii)* la séquence de conduction au travers du myocarde.

Dans le cas physiologique, comme il a déjà été présenté, l'impulsion est initiée dans le nœud sinusal. Le front de dépolarisation auriculaire résultant est représenté dans l'ECG par l'onde P. Cette onde se caractérise au niveau spectral par une composante basse fréquence de faible énergie, qui limite souvent son observation dans plusieurs dérivations ECG, spécialement dans des conditions de bruit. La repolarisation auriculaire est représentée par l'onde Ta et sa direction est opposée à celle de l'onde P. Généralement l'onde Ta n'est pas visible dans l'ECG car elle coïncide avec le complexe QRS d'amplitude plus importante. Ce dernier correspond à la dépolarisation ventriculaire et représente la déflection de plus grande amplitude de l'ECG. Il est constitué de trois ondes consécutives (les ondes Q, R et S) qui sont associées respectivement aux vecteurs moyens d'activations septale, ventriculaire et basale, présentés auparavant. Le processus de repolarisation ventriculaire est reflété par l'onde T. Dans certaines occasions, une onde, dite onde U, de très basse amplitude peut être observée après l'onde T. Bien que son origine physiologique n'ait pas encore été démontrée, l'onde U (fréquemment observée chez les athlètes) est souvent associée aux processus de repolarisation ventriculaire tardive.

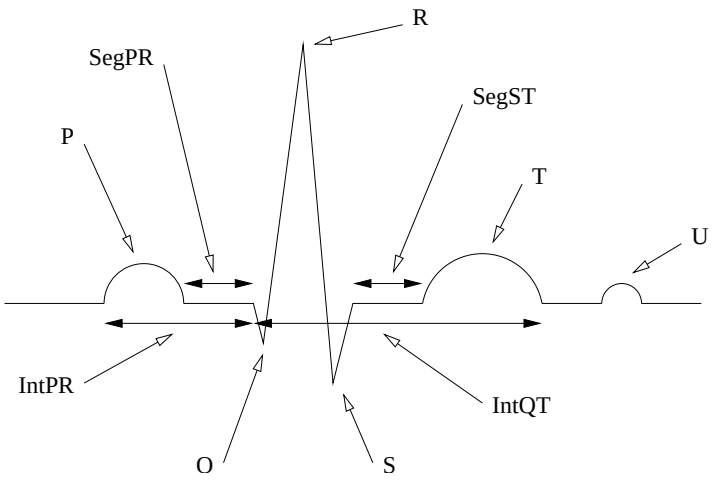


FIG. 1.11 – Ondes, intervalles et segments dans l'ECG pour un battement physiologique.

Outre les formes d'ondes, un battement cardiaque est aussi caractérisé par plusieurs segments et intervalles (figure 1.11) :

- **l'intervalle** PR : il est mesuré entre le début de l'onde P et le début du complexe QRS. Cet intervalle représente la dépolarisation des oreillettes et du nœud A-V. Sa durée normale est de 0.14 à 0.2 s ;
- **le segment** PR : c'est la période temporelle comprise entre la fin de l'onde P et le début du complexe QRS. Il représente le temps de transmission du front de dépolarisation par

le nœud A-V ;

- **le segment** ST : il est compris entre la fin du complexe QRS (ou point J) et le début de la phase ascendante de l'onde T. Ce segment correspond au temps pendant lequel l'ensemble des cellules myocardiques sont dépolarisées (phase de plateau) et donc, dans le cas normal, doit être isoélectrique. En cas contraire, le niveau d'amplitude et la pente de ce segment sont des indicateurs de l'état ischémique du myocarde ;
- **l'intervalle** QT : le temps entre le début du complexe QRS et la fin de l'onde T. Il représente une indication de la longueur des phases de dépolarisation et repolarisation ventriculaire (longueur moyenne d'un PA ventriculaire). Sa durée varie avec la fréquence cardiaque entre 0.3 et 0.38 s.

Bibliographie

[Defontaine, 2006] Defontaine, A. (2006). *Modélisation multirésolution et multiformalisme de l'activité électrique cardiaque.* PhD thesis, Universtié de Rennes 1, France.

[Dower, 1984] Dower, G. E. (1984). The ECGD : A derivation of the ECG from VCG leads. *Journal of Electrocardiology*, 17(2) :189–192.

[Dower et al., 1980] Dower, G. E., Machado, H. B., et Osborne, J. A. (1980). On deriving the electrocardiogram from the vectorcardiographic leads. *Clinical Cardiology*, 3 :87–95.

[Hernández, 2000] Hernández, A. I. (2000). *Fusion de signaux et de modèles pour la caractérisation d'arythmies cardiaques.* PhD thesis, Université de Rennes 1, France.

[Levkov, 1987] Levkov, C. L. (1987). Orthogonal electrocardiogram derived from the limb and chest electrodes of the conventional 12-lead system. *Medical and Biological Engineering and Computing*, 25 :155–164.

[Rudy, 2004] Rudy, Y. (2004). *Cardiac electrophysiology : from cell to bedside – Edition 4*, chapter 28 – Ionic mechanisms of cardiac electrical activity : a theoretical approach, pages 255–266. W B Saunders Company.

Chapitre 2

Analyse de dynamiques temporelles en cardiologie : l'exemple de l'ischémie

Le chapitre précèdent s'est concentré sur l'activité électrophysiologique normale aux niveaux cellulaire, tissulaire (conduction entre cellules voisines) et organe (activation ordonnée des différentes structures du myocarde). Cette activité électrique peut cependant être fortement altérée dans le cas pathologique et ces modifications sont directement traduites sur l'ECG. Ce chapitre expose, sur la détection de l'ischémie, comment ces modifications peuvent être détectées et souligne aussi pourquoi l'analyse des dynamiques du signal ou des variables qui y sont extraites est fondamentale.

La première section aborde les modifications électriques survenant lors de l'ischémie. La seconde présente différentes méthodologies pour la détection des épisodes ischémiques et conclut sur le fait que l'information temporelle est importante mais peu exploitée. La troisième section formalise le problème et le pose dans le cadre méthologique de la fouille de données qui sera adopté dans la suite de ce mémoire.

2.1 L'activité électrique cardiaque pendant l'ischémie

L'ischémie myocardique est définie par un déséquilibre de la balance apports/besoins en oxygène du myocarde et est la conséquence d'un arrêt temporaire ou définitif de la circulation coronaire (qui alimente le cœur), avec accumulation des produits du métabolisme cellulaire (acides organiques, radicaux libres, ...). L'ischémie peut survenir au repos, à cause d'une chute brutale du débit coronaire, sans augmentation obligatoire des besoins. Elle peut aussi survenir à l'effort, lorsque l'augmentation du débit coronaire est insuffisante pour équilibrer l'augmentation considérable de la consommation d'oxygène. Les principales conséquences de l'ischémie sont [Daubert, 1998] :

- **d'un point de vue électrique** : modification de l'électrogenèse, abaissement du seuil fibrillatoire et émergence d'une hyperexcitabilité à l'étage ventriculaire ;
- **d'un point de vue mécanique** : perte de la relaxation et de la contraction dans le territoire ischémié ;
- **d'un point de vue métabolique** : dette en oxygène et en radicaux énergétiques avec acidose et accumulation de radicaux libres ;
- **d'un point de vue clinique** : apparition d'une douleur d'angine de poitrine.

Les principales conséquences de l'ischémie myocardique au niveau cellulaire sont liées aux perturbations de la respiration aérobie et de la production d'ATP. Les modifications de l'ATP intracellulaire altèrent le fonctionnement des canaux actifs, modifiant la distribution de certains ions critiques (notamment Na^+ et K^+), qui sont en grande partie responsables de la génération et de la propagation du potentiel d'action (PA). Ces modifications concernent particulièrement : une augmentation du potassium extracellulaire, $[K^+]_o$ (ou hypercalémie) ; une diminution du pH intracellulaire et une réduction de l'ATP intracellulaire. Les effets de ces différentes manifestations sur l'excitabilité ont été largement étudiés, expérimentalement et théoriquement.

L'augmentation de la concentration de potassium extracellulaire ($[K^+]_o$) pendant l'ischémie implique une dépolarisation du potentiel de repos cellulaire, qui modifie l'excitabilité cellulaire de deux façons opposées : *i)* pour des augmentations faibles de $[K^+]_o$, l'excitabilité cellulaire est augmentée car le potentiel de repos est plus proche du seuil liminaire et *ii)* à partir d'une certaine valeur d'augmentation du $[K^+]_o$, le potentiel de repos dépolarisé réduit la disponibilité des canaux sodiques, limitant l'excitabilité cardiaque. La diminution du pH intracellulaire réduit la conductance sodique et la pente maximale de la phase 0, ce qui provoque également une réduction de l'excitabilité et une propagation plus lente du PA.

Mais l'ischémie affecte également la repolarisation cellulaire. En effet, le déficit d'ATP et les modifications de la $[K^+]_o$ accélèrent la sortie du K^+ pendant les phases 1 et 2, réduisant la durée du PA d'environ 60% [Shaw et Rudy, 1997]. Les inhomogénéités de durée du PA qui peuvent survenir, par exemple, aux bords d'une zone ischémique, ont été identifiées comme l'une des principales causes des arythmies ventriculaires létales.

Les modifications du potentiel d'action induites par l'ischémie peuvent entraîner, en fonction de la taille et de la localisation de la zone affectée, une modification de l'électrocardiogramme de surface et notamment du segment ST [Yan et Antzelevitch, 1999]. Comme indiqué précédemment, le segment ST est normalement isoélectrique car la plupart des cellules ventriculaires se trouvent dans la phase de plateau. La différence de durée des PA des cellules saines et ischémiques crée une différence de potentiel, qui est mesurable au niveau de l'ECG de surface au cours du segment ST. Le déplacement vers le haut ("sus-décalage" ou "sus-dénivellation") ou vers le bas ("sous-décalage" ou "sous-dénivellation") de l'amplitude du segment ST, par rapport à la ligne isoélectrique, indique généralement un tel état pathologique.

En pratique clinique, le suivi des patients suspectés d'ischémie est recommandé dans [Gibbons, 2000] et suit le cheminement présenté figure 2.1 pour l'insuffisance coronarienne aiguë (ICA). L'ECG n'est donc pas l'unique source d'information utilisée par les médecins. Dans sa phase précoce, le diagnostic se base aussi sur l'étude des symptômes décrits par le patient (et aussi sur son historique et les facteurs de risques) ou sur la concentration de marqueurs biologiques tels que la troponine, la créatine-kinase ou la myoglobine. Dans une seconde phase, lorsque l'ICA est très fortement suspectée, une angiographie est appliquée pour localiser l'artère coronaire affectée et l'importance du thrombus.

2.2 Indicateurs ECG de l'ischémie

Comme expliqué précédemment, l'ischémie engendre des modifications de l'électrogenèse qui affectent principalement le segment ST. Cependant, l'ischémie peut avoir d'autres manifestations électrocardiographiques dont l'intérêt n'est pas à négliger [Roy et al., 2005]. Les

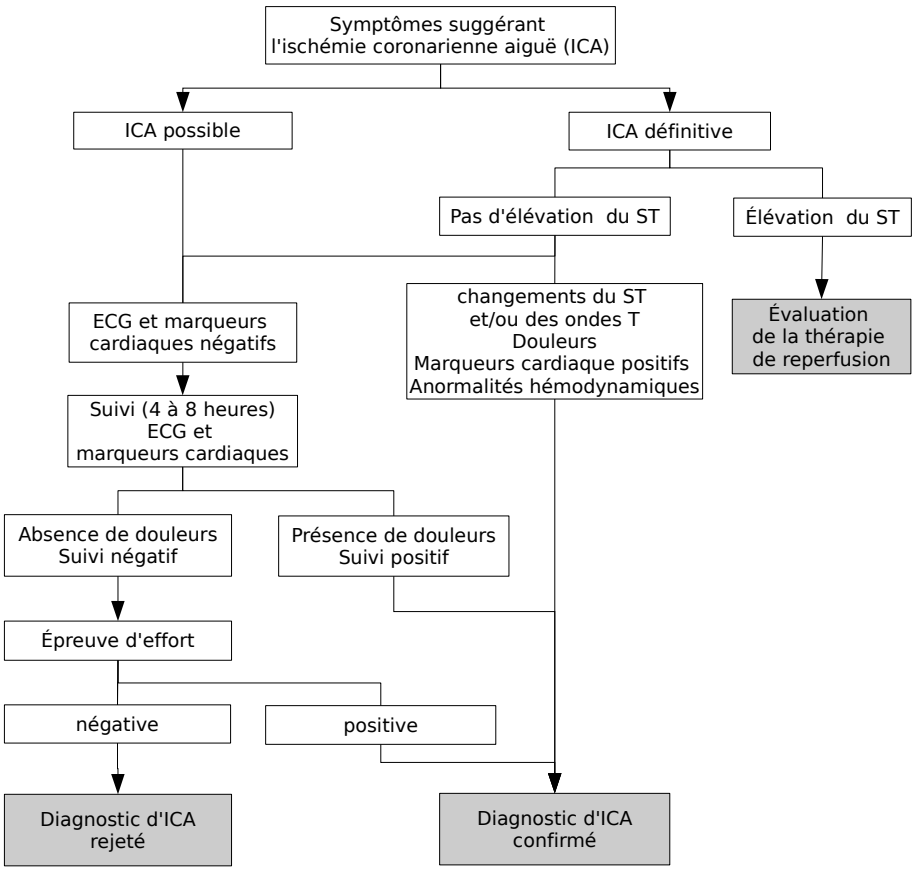


FIG. 2.1 – Evaluation et suivi des patients suspectés d'insuffisance coronarienne aiguë (ICA).

différents indicateurs analysés dans la littérature sont reportés ici, suivant qu'ils concernent la dépolarisation ventriculaire, la repolarisation ventriculaire ou le système nerveux autonome (SNA) :

- **Indicateurs de la dépolarisation ventriculaire** : plusieurs indicateurs sont extraits du complexe QRS, ils se basent principalement sur une analyse des amplitudes des ondes ou sur les hautes fréquences du complexe QRS. Des mesures des amplitudes des ondes du QRS [Michaelides et al., 1995, Toth et al., 2001] ainsi que des pentes du QRS [Pueyo et al., 2008] ont été proposées avec des résultats concluants sur des épreuves d'effort. De manière générale, les modifications du complexe QRS sont souvent dynamiques et réversibles. Il peut s'agir d'une augmentation de l'onde R, de la distortion de la fin du QRS ou de l'apparition ou disparition de l'onde Q. Les hautes fréquences du QRS ont été analysées dans [Pettersson et al., 2000] et une diminution de cette activité haute fréquence est observée. Cependant une forte variabilité inter-patients remet en cause ces résultats dans [Trägårdh et al., 2004] et l'intérêt de cet indicateur est toujours controversé [Abboud et Zlochiver, 2006].
- **Indicateurs de la repolarisation ventriculaire** : plusieurs études montrent des allongements significatifs du QT, par exemple ceux observés au début des ischémies aiguës provoquées par angioplastie [Arini et al., 2006, Kenigsberg et al., 2007]. Les changements de morphologies portant sur l'ensemble $ST-T$ ont aussi été étudiés, à l'aide des coefficients issus de la transformée de Karhunen-Loève (KLT) et des modifications significatives sont relevées [Garcia et al., 1998, Schindler et al., 2007]. Sur l'onde T, on relève principalement une symétrisation, un pic plus prononcé et parfois des changements de signe.
- **Indicateurs du SNA** : Airaksinen [Airaksinen et al., 1987] a montré que l'installation de l'ischémie s'accompagne de la réduction des capacités régulatrices du système nerveux autonome. Une baisse de la variabilité est observée et ceci parfois même avant l'apparition de symptômes provoqués par l'ischémie. Ceci est confirmé dans [Huikuri et Mäkikallio, 2001] qui précise que la baisse de variabilité est surtout visible dans la réduction des hautes fréquences.

L'ensemble de ces études montrent que l'ischémie ne se limite pas à une modification du segment ST mais que plusieurs indicateurs supplémentaires peuvent être extraits de l'ECG et apporter des informations complémentaires. Cependant il est important de noter que le diagnostic de l'ischémie est très dépendant du contexte clinique et que l'ECG seul sert souvent plus à la suggestion de l'ischémie qu'à sa confirmation. La section suivante montre comment les diagnostics d'ischémie aiguë ou d'ischémie transitoire sont posés à partir de ces indicateurs.

2.3 Diagnostic de l'ischémie

L'ischémie apparaît sous différents syndromes avec des routines de diagnostics propres et dépendant des symptômes présentés et de facteurs de risques. On distingue notamment :

1. **l'ischémie myocardique silencieuse (IMS)**. Elle est la plus difficile à détecter car elle ne s'accompagne pas de douleurs et sa manifestation est transitoire. Ces accès ischémiques correspondent le plus fréquemment à des vasospasmes. Elle est dépistée généralement par un test d'effort et parfois par l'enregistrement Holter qui est moins spécifique mais qui permet d'évaluer la fréquence de ces accès ischémiques. En effet, il a été montré que ces épisodes sont souvent précurseurs d'infarctus [Touzé et al., 2005,

Lovett et al., 2003] et la mesure de leur fréquence d'apparition est donc un indicateur utile pour la stratification du risque ;

2. **l'angor vasospastique**, qui s'exprime avec des douleurs mais est, comme pour l'IMS, transitoire. Son analyse est souvent réalisée à l'aide de l'enregistrement Holter et permet la stratification du risque ;
3. **l'angor chronique stable**, qui est la forme clinique la plus fréquente et qui provient de sténoses fixes. Elle se manifeste par de l'angor d'effort et est d'ailleurs diagnostiquée à l'aide du test d'effort ;
4. **l'angor instable**, qui est un syndrome coronarien aigu répétitif avec absence d'élévation du segment ST. Ce type d'angor provient généralement des thrombus partiellement occlusifs. L'identification de ce syndrome est important car elle évite à court terme des accidents coronariens aigus. Par contre, elle est relativement facile à détecter car l'ECG montre des anomalies à la fois transitoires et permanentes ;
5. **l'infarctus**, qui survient lors d'occlusions thrombotiques et dont les anomalies observées sur l'ECG évolue en fonction de la progression de la nécrose. L'infarctus et l'angor instable font parties des ICA (insuffisance cardiaque aiguë).

La sous-section suivante montre comment une aide logicielle au diagnostic de l'ICA est réalisée. Nous verrons que la méthode proposée intègre déjà des aspects multi-variables mais l'information temporelle. Une seconde sous-section explique comment le test d'effort peut être utilisé, par exemple, pour la détection d'angor chronique stable, et en intégrant une information temporelle simple. Enfin nous discuterons de la détection d'épisodes ischémiques dans des enregistrements Holter, avec l'utilisation d'une information temporelle plus importante puis nous examinerons les limitations présentent dans ces différentes méthodes de dépistage de l'ischémie.

2.3.1 Détection de l'ischémie aiguë : exemples d'outils de diagnostic multivariable

En pratique clinique, un enjeu important en électrocardiographie est le triage dans les urgences des patients suspectés d'être atteints d'une ischémie aiguë en raison d'angor de poitrine. Un des outils d'aide au diagnostique et utilisé en pratique clinique est l'ACI-TIPI (Acute Coronary Ischemia - Time Insensitive Predictive Instrument), [Selker et al., 1997]. Cet outil est présenté ici car il reflète parfaitement l'intérêt de faire des analyses multivariables et multimodalités dans le cas de l'ischémie. En effet, l'ACI-TIPI réalise le calcul d'une probabilité, exprimée en pourcent, de présence d'ICA par régression logistique des variables suivantes : (1) l'age, (2) le sexe, (3) la présence ou l'absence *a)* de douleur de poitrine, *b)* de douleur dans le bras gauche, (4) de la prédominance de la douleur de poitrine ou pression, (5) la présence ou l'absence d'ondes Q dans l'ECG, (6) la présence et le degré dans l'ECG d'élévation ou de dépression du segment ST et (7) la présence et le degré d'élévation ou d'inversion de l'onde T. La prise en compte de ces différentes variables donne des scores de classification intéressants : sensibilité de 95% et spécificité de 73%, et une aire sous la courbe récepteur-opérateur de 0.85. Ceci a favorisé son intégration dans les outils d'acquisition et de visualisation de l'ECG.

L'information utilisée est donc multi-modale et instantanée. Ceci est adapté pour la détection des ACI dans les urgences, lorsque les patients présentent des symptômes suggérant cette ischémie. Cependant, il est à noter que dans un contexte de monitoring ou de détection précoce, l'information temporelle prend une toute autre importance.

2.3.2 Epreuve d'effort : méthodologie basée sur l'analyse de l'hystérésis ST/HR

Les déviations du segment ST peuvent avoir une cause coronarienne mais aussi être dues à une accélération du rythme cardiaque provoquant aussi une chute du rapport apport/besoin en oxygène des cellules myocardiques. Il convient alors d'analyser non seulement la déviation du segment ST mais aussi de prendre en compte le rythme cardiaque. C'est ce qui est réalisé sur les analyses de type hystérésis ST/HR. Pour la détection de maladies coronariennes en épreuve d'effort, l'hystérésis ST/HR est tracée au cours du temps, avec en abscisse le rythme cardiaque et en ordonnée la dépression du segment ST (figure 2.2). Différents critères à but diagnostique sont proposés dans la littérature :

- le sens de l'hystérésis, proposé par [Okin et al., 1989] est un indicateur issu directement de la boucle ST/HR et qui donne des résultats quantitatif sur la détection de pathologies coronariennes (SE = 72%, PP = 85%). Une boucle qui évolue dans le sens horaire indique un sujet sain alors qu'une boucle qui évolue dans le sens anti-horaire est associée à un patient avec une pathologie coronarienne.
- l'aire de l'hystérésis, proposée dans [Lehtinen et al.,] présente un indicateur de pathologies coronariennes plus performant que le sens de la boucle, avec une sensibilité et une spécificité pouvant être tout deux supérieurs à 80% pour le seuil correspondant.

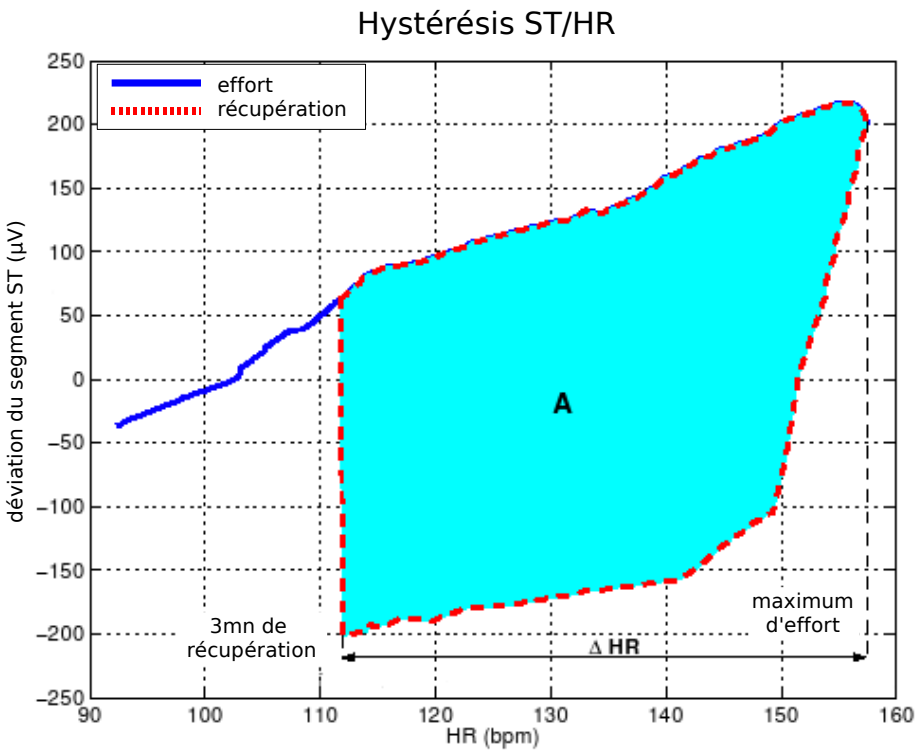


FIG. 2.2 – Hystérésis ST/HR issue d'une épreuve d'effort.

En dehors du cadre de l'analyse de l'hystérésis, [Bailon et al., 2003] exposent un grand nombre d'indicateurs extraits de l'épreuve d'effort, classés suivant qu'ils caractérisent la répolarisation, la dépolarisation ou la variabilité cardiaque, ainsi que leur efficacité à discriminer des patients à faibles risques ou à fort risque d'ICA (ces derniers étants soumis à une angiogra-

phie). Ces indicateurs sont ensuite exploités dans une analyse discriminante et la combinaison des indicateurs de la variabilité cardiaque fournit les meilleurs résultats avec une sensibilité de 94% pour une spécificité de 92%.

2.3.3 Détection des épisodes ischémiques : vers une utilisation de l'information temporelle

Le décalage du segment ST précédemment évoqué est un indicateur de présence d'épisodes ischémiques très sensible mais peu spécifique. En effet, les déviations du segment ST peuvent survenir lors de changements du rythme cardiaque, de position du patients ou à cause du bruit dans l'ECG. La littérature présente deux axes de recherche sur ce problème [Moody et Jager, 2003, Pueyo et al., 2004] :

- l'extraction de caractéristiques plus pertinentes de l'ECG afin de mieux caractériser l'ischémie,
- la prise en compte des dynamiques des indicateurs observés pendant les épisodes ischémiques.

La plupart des indicateurs extraits se présentent sous la forme de séries temporelles. Par exemple les durées de QRS, l'intervalle QT ou l'amplitude du segment ST sont relevés pour chaque battement. Il faut alors determiner comment utiliser ces séries temporelles pour identifier les épisodes ischémiques. Les paragraphes suivants présentent deux méthodologies très différentes dédiées à ce problème.

2.3.3.1 Méthodologie basée uniquement sur la série temporelle de l'amplitude du segment ST

Langley [Langley et al., 2003] propose un algorithme simple pour détecter les épisodes ischémiques uniquement à partir de la série temporelle de la déviation du segment ST. Il suppose que les décalages sont plus importants et perdurent plus longtemps en cas d'épisodes ischémiques que pour les épisodes non-ischémiques. Pour être considéré comme ischémique, un épisode doit donc avoir une déviation supérieur à V_{min} pendant une durée d'au moins T_{min}. La figure 2.3 donne un aperçu du fonctionnement de cette méthode et de l'utilisation des seuils et intervalles temporels.

Les résultats de détection d'épisodes ischémiques sur un ensemble de test et présentés dans [Langley et al., 2003] donnent une précision de 81.4% et sont obtenus pour des paramètres V_{min} et T_{min} déterminés de façon empirique, alors qu'une recherche plus approfondie sur ces paramètres pourrait donner de meilleurs résultats. Cet algorithme est aussi limité à l'utilisation du décalage du segment ST alors que d'autres variables peuvent apporter de l'information supplémentaire, comme le paragraphe suivant l'expose.

2.3.3.2 Méthologie basée sur une classification des battements

Cette approche se base sur le fait qu'on puisse labéliser les battements en deux groupes : "normaux" et "ischémiques". La détection d'épisodes ischémiques est alors réalisée en deux étapes : la labélisation de chacun des battements puis la détection des épisodes en considérant les périodes où le nombre de battements labellisés "ischémiques" est important. Par exemple Papaloukas [Papaloukas, 2002] effectue une analyse en composantes principales des battements puis utilise les vecteurs propres avec un réseau de neurones pour classer les battements. L'information temporelle est ensuite prise en compte de manière très simple pour décider de la présence d'un épisode ischémique : dans un intervalle de 30 secondes, il faut que 75% des battements soient classifiés comme ischémiques. Une sensibilité de 86% et une spécificité de

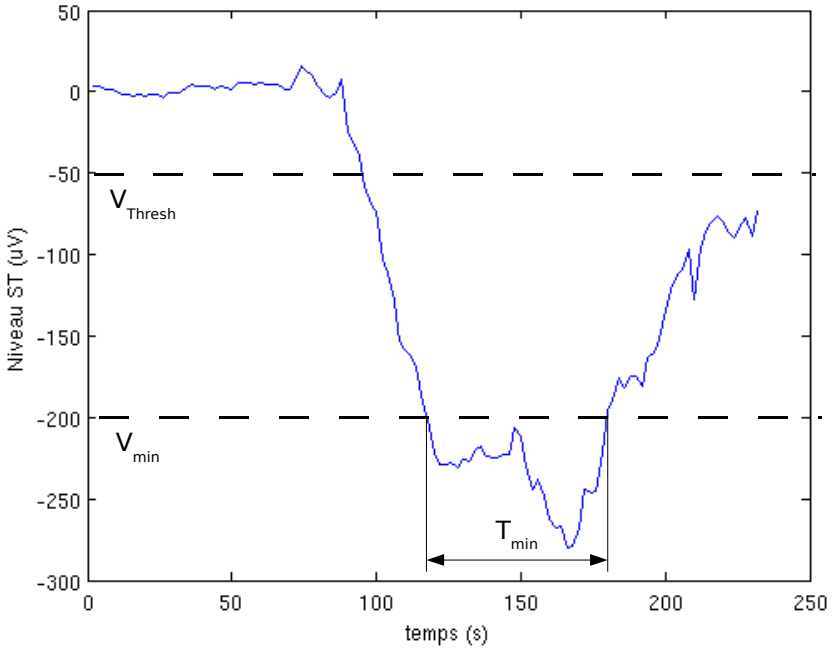


FIG. 2.3 – Illustration des seuils et intervalles temporels de l'algorithme de base de [Langley et al., 2003].

87% sont reportées sur une base de données annotées. Il est cependant important de noter que la classification des battements consiste à faire une très forte discrétisation du problème qui n'est pas justifiée étant donnée que l'apparition des conséquences de l'ischémie sur l'ECG n'est pas instantanée.

2.3.4 Limitation des méthodes existantes

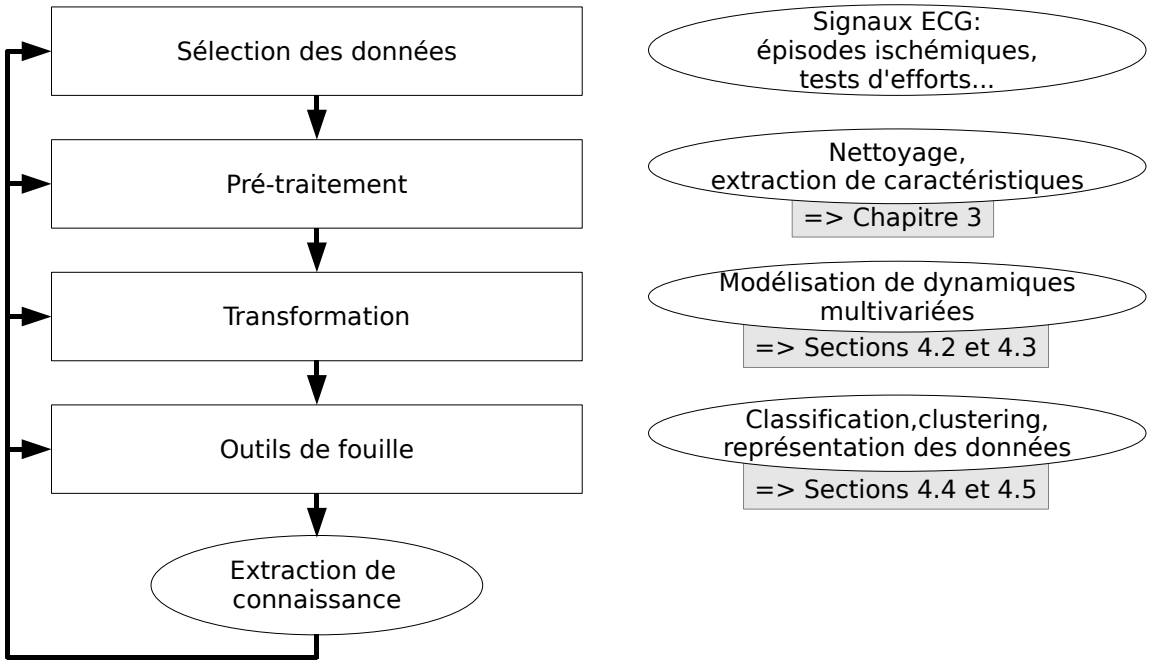
Le problème de la détection des épisodes ischémiques est donc double : i) il faut être capable d'isoler un certain nombre de caractéristiques qui différencient l'état ischémique de l'état sain, ii) il faut être capable d'exploiter ces caractéristiques et notamment prendre en compte leur évolution temporelle. Les méthodes proposées dans les paragraphes précédents ne sont pas réellement adaptés à la prise en compte des dynamiques : par exemple, le calcul de l'aire sous la courbe d'hystérésis peut donner le même résultat si le patient a fait peu d'effort au début (beaucoup de points au début de l'hystérésis) et un effort violent ensuite (peu de points autour du maximum d'effort) que l'inverse. Bien qu'une information temporelle soit conservée sous forme de séquence de points, l'information de durée n'est quant à elle pas exploitée. Cette information de durée est intégrée de manière très simple dans plusieurs approches, comme par exemple celle de Langley [Langley et al., 2003] ou celle de Papaloukas [Papaloukas, 2002] avec la définition d'intervalles de temps qui sont ajustés de manière empirique.

2.4 Positionnement du problème dans le cadre de la fouille de données temporelles multivariées

A partir de l'exemple de l'ischémie décrit, deux points particulièrement importants, sur lesquels reposent ce travail, peuvent êtres soulevés :

- On a montré qu'une quantité importante d'informations peut être extraite de l'ECG et que cette information est potentiellement utile pour fournir des indications sur le diagnostic, sur la thérapie à adapter ou bien encore pour réaliser une stratification du risque (d'infarctus par exemple). Dans des cadres de recherches exploratoires, elle peut aussi contribuer à l'apport de nouvelles connaissances.
- On a montré que nous sommes confrontés à des phénomènes transitoires et dont les différentes dynamiques reflètent des comportements particuliers n'ayant pas les mêmes significations physiologiques. Hors la prise en compte de ces dynamiques représente un problème complexe, qui est très souvent négligé, alors que son importance est pourtant évidente.

Ces deux constats expliquent que notre travail trouve une formalisation particulièrement adaptée dans le cadre de la fouille de données temporelles et multivariées. La figure 2.4 schématise ce processus de fouille de données et associe à chacune de ses étapes une correspondance dans le cadre de ce travail.



Etapes du processus de fouille de données et leur correspondance dans ce mémoire.

Sélection des données Les données sont donc des signaux ECGs ou des portions de signaux ECG récupérés sur des ensembles de patients. Il est à noter que cette étape est déjà importante. Procéder à une sélection des données permet de diminuer la quantité d'informations à traiter, par exemple en cas d'analyse de l'ischémie et à partir d'enregistrements de type Holter, il est plus important de se concentrer sur les périodes où l'ischémie est suspectée plutôt que sur les périodes où l'ECG est normal. Il est aussi important de réaliser des groupes de patients les plus homogènes possibles et de taille significative.

Pré-traitement Le pré-traitement des données consiste à nettoyer au mieux le signal ECG puis à en extraire l'information utile. Cette information correspond aux indicateurs utilisés par les cardiologues pour caractériser le cycle cardiaque (notamment les intervalles et amplitudes des ondes présentées figure 1.11) ainsi que le rythme cardiaque. Cette étape d'extraction est fondamentale car elle permet de réduire la dimension des données et de conserver une information interprétable par les spécialistes. Le troisième chapitre de ce mémoire est dédié à cette étape de nettoyage/extraction. Une des difficultés rencontrées est le choix de seuils de décisions optimaux et ce chapitre propose donc une nouvelle méthode d'optimisation des seuils. Une autre difficulté est de proposer une plateforme directement exploitable par un clinicien. Une station d'analyse multi-variée a donc aussi été mise au point.

Transformation A l'issu de l'étape d'extraction de caractéristiques nous disposons de séries temporelles multivariées, constituées d'une succession de vecteurs d'observations :

$$O = \{\overline{o}_t : t \in T\}$$

où $\overline{o}_t$ est, le vecteur d'observation, regroupant les différents indicateurs x à chaque instant et défini dans $\mathbb{R}^n$:

$$\overline{o}_t = \{x_{(1,t)}, x_{(2,t)}, ..., x_{(n,t)}\}$$

L'objectif à ce stade est de prendre en compte l'information présente dans la dynamique de ces observations. Plutôt que de transformation, nous parlerons de caractérisation de la dynamique, et ceci sera typiquement réalisé à travers un modèle M de la forme :

$$\overline{o}_{t+1} = M(o_t, o_{t-1}, ..., o_{t-d}) \text{ avec } d \in \mathbb{N}$$

Le problème de la caractérisation des dynamiques est abordé au quatrième chapitre de ce mémoire et résolu à l'aide de modèles Semi-Markovien Cachés.

Outils de fouille Nous verrons que les modèles de type semi-markovien se prêtent particulièrement bien aux tâches de fouille de données et qu'il est aisé d'apprendre des dynamiques spécifiques, de classer de nouveaux individus ou bien de les représenter suivant leurs dynamiques. Cependant le problème du clustering, consistant à réaliser des groupes naturels d'individus à partir des observations extraites, est plus complexe. Il nécessite, par exemple, la mise au point d'algorithmes d'optimisation de type Estimation-Maximisation (EM) et sera traité de manière plus précise à la fin du chapitre 4.

Bibliographie

[Abboud et Zlochiver, 2006] Abboud, S. et Zlochiver, S. (2006). High-frequency QRS electrocardiogram for diagnosing and monitoring ischemic heart disease. *Journal of Electrocardiology*, 39(1) :82.

[Airaksinen et al., 1987] Airaksinen, K., Ikäheimo, M., Linnaluoto, M., Niemelä, M., et Takkunen, J. (1987). Impaired vagal heart rate control in coronary artery disease. *British Heart Journal*, 58 :592.

[Arini et al., 2006] Arini, P., Martinez, J., et P., L. (2006). Evolution of T wave width during severe ischemia generated by percutaneous transluminar coronary angioplasty. *Computers in Cardiology 2006*, 33 :713.

[Bailon et al., 2003] Bailon, R., Mateo, J., Olmos, S., Serrano, P., Garcia, J., del Rio, A., Ferreira, I. J., et Laguna, P. (2003). Coronary artery disease diagnosis based on exercise electrocardiogram indexes from repolarisation, depolarisation and heart rate variability. *Medical and Biological Engineering and Computing*, 41(5) :561–571.

[Daubert, 1998] Daubert, J. C. (1998). Cardiopathies ischémiques – ischémies myocardiques transitoires. Technical report, Département de cardiologie et maladies vasculaires – CHU de Rennes.

[Garcia et al., 1998] Garcia, J., Lander, P., Sörnmo, L., Olmos, S., Wagner, G., et Laguna, P. (1998). Comparative study of local and Karhunen-Loève-Based ST-T Indexes in recordings from human subjects with induced myocardial ischemia. *Computers and Biomedical Research*, 31(4) :271.

[Gibbons, 2000] Gibbons, R. J. (2000). Guidelines for the management of patients with unstable angina and non-st-segment elevation myocardial infarction : Executive summary and recommendations. *Circulation*, 102 :1193–1209.

[Huikuri et Mäkikallio, 2001] Huikuri, H. et Mäkikallio, T. (2001). Heart rate variability in ischemic heart disease. *Auton Neurosci.*, 90(1-2) :95.

[Kenigsberg et al., 2007] Kenigsberg, D. N., Khanal, S., Kowalski, M., et Krishnan, S. C. (2007). Prolongation of the QTc interval is seen uniformly during early transmural ischemia. *Journal of the American College of Cardiology*, 49(12) :1299.

[Langley et al., 2003] Langley, P., Bowers, E. J., Wild, J., Drinnan, M. J., Allen, J., Sims, A. J., Brown, N., et Murray, A. (2003). An algorithm to distinguish ischaemic and non-ischaemic st changes in the holter ecg. *Computers in Cardiology*, 30 :239–242.

[Lehtinen et al.,] Lehtinen, R., Sievänen, H., Viik, J., Turjanmaa, V., Niemelï¿½, K., et Malmivuo, J. Accurate detection of coronary artery disease by integrated analysis of the st-segment depression/heart rate patterns during the exercise and recovery phases of the exercise ecg test. *American Journal of Cardiology*, 78(9).

[Lovett et al., 2003] Lovett, J., Dennis, M., Sandercock, P., Bamford, J., Warlow, C., et Rothwell, P. (2003). Very early risk of stroke after a first transient ischemic attack. *Stroke, American Heart Association, Inc*, 34 :138.

[Michaelides et al., 1995] Michaelides, A., Ryan, J., Bacon, J., Pozderac, R., Toutouzas, P., et Boudoulas, H. (1995). Exercise-induced QRS changes (Athens QRS score) in patients with coronary artery disease : a marker of myocardial ischemia. *Journal of Electrocardiology*, 26(5) :263.

[Moody et Jager, 2003] Moody, G. et Jager, F. (2003). Distinguishing ischemic from non-ischemic ST changes : The physionet/computers in cardiology challenge 2003. *Computers in Cardiology*, 30 :235–237.

[Okin et al., 1989] Okin, P. M., Ameisen, O., et Kligfield, P. (1989). Recovery-phase patterns of st segment depression in the heart rate domain. identification of coronary artery disease by the rate-recovery loop. *Circulation*, 80 :533.

[Papaloukas, 2002] Papaloukas, C. (2002). An ischemia detection method based on artificial neural networks. *Artificial Intelligence in Medicine*, 24(2) :167.

[Pettersson et al., 2000] Pettersson, J., Pahlm, O., Carro, E., Edenbrandt, L., Ringborn, M., et Sörnmo, L. (2000). Changes in high-frequency QRS components are more sensitive than ST-segment deviation for detecting acute coronary artery occlusion. *Journal of the American College of Cardiology*, 36(6) :1827.

[Pueyo et al., 2004] Pueyo, E., Garcia, J., Wagner, G., Bailon, R., Sornmo, L., et Laguna, P. (2004). Time course of ecg depolarization and repolarization changes during ischemia in ptca recordings. *Methods of Information in Medicine*, 43(1) :43–46.

[Pueyo et al., 2008] Pueyo, E., Sörnmo, L., et Laguna, P. (2008). QRS slopes for detection and characterization of myocardial ischemia. *Biomedical Engineering, IEEE Transactions on*, 55(2) :468.

[Roy et al., 2005] Roy, L. D., Brohet, C., et Renard, M. (2005). *ECG pathologique*. Masson.

[Schindler et al., 2007] Schindler, D., Lux, R., Shusterman, V., et Drew, B. (2007). Karhunen-Loëve representation distinguishes ST-T wave morphology differences in emergency department chest pain patients with non-ST-elevation myocardial infarction versus nonacute coronary syndrome. *Journal of Electrocardiology*, 40(6) :145.

[Selker et al., 1997] Selker, H. P., Zalenski, R. J., Antman, E. M., Aufderheide, T. P., Bernard, S. A., Bonow, R., Gibler, W. B., Hagen, M. D., Johnson, P., Lau, J., McNutt, R. A., Ornato, J., Schwartz, J. S., Scott, J. D., Tunick, P. A., Weaver, W. D., et on "Evaluation of Technologies for Identifying Acute Cardiac Ischemia in the Emergency Department", N. H. A. A. P. C. C. W. G. (1997). Aci-tipi. *Annals of Emergency Medicine*, 29(1) :43–50.

[Shaw et Rudy, 1997] Shaw, R. M. et Rudy, Y. (1997). Electrophysiologic effects of acute myocardial ischemia : a theoretical study of altered cell excitability and action potential duration. *Cardiovascular Research*, 35 :256–272.

[Toth et al., 2001] Toth, A., Marton, Z., Czopf, L., Kesmarky, G., Halmosi, R., Juricskay, I., Habon, T., et Toth, K. (2001). QRS Score : a composite index of exercise-induced changes in the Q, R, and S waves during exercise stress testing in patients with ischemic heart disease. *Ann Noninvasive Electrocardiol*, 6(4) :310.

[Touzé et al., 2005] Touzé, E., Varenne, O., Chatellier, G., Peyrard, S., Rothwell, P. M., et Mas, J.-L. (2005). Risk of myocardial infarction and vascular death after transient ischemic attack and ischemic stroke. *Stroke, American Heart Association, Inc*, 36 :2748.

[Trägårdh et al., 2004] Trägårdh, E., Pahlm, O., Wagner, G., et Pettersson, J. (2004). Reduced high-frequency QRS components in patients with ischemic heart disease compared to normal subjects. *Journal of Electrocardiology*, 37(3) :157.

[Yan et Antzelevitch, 1999] Yan, G. X. et Antzelevitch, C. (1999). Cellular basis for the Brugada syndrome and other mechanisms of arrhythmogenesis associated with ST-segment elevation. *Circulation*, 100(15) :1660–1666.

Chapitre 3

Extraction des caractéristiques du signal ECG

Comme dans la majorité des applications de fouille de données, une étape de pré-traitement est nécessaire à l'analyse des dynamiques de l'activité électrique cardiaque. Cette étape de pré-traitement va notamment permettre d'extraire l'information utile du signal ECG et de la rendre aisément exploitable, à la fois par les cliniciens ou par des algorithmes automatiques de traitement de données. Cette extraction est typiquement réalisée par le clinicien dans toute analyse ECG d'un patient. Son automatisation est cependant importante face à de gros volumes de données, ce qui permet un gain de temps ainsi qu'une meilleure répétabilité des analyses effectuées. Au sein de cette procédure de traitement de l'ECG, l'étape de segmentation battement à battement demeure vraisemblablement le problème le plus complexe, en particulier pour les ondes P et T.

L'objet de ce chapitre est justement d'y apporter des solutions originales. Après une courte bibliographie sur les méthodes appliquées à la segmentation des ondes du signal ECG, la seconde section détaille la chaîne de traitement du signal proposée et les modifications apportées à un algorithme à base d'ondelettes issu de la littérature. Bien que quelques améliorations aient été apportées sur l'algorithme de segmentation, la contribution majeure de ce chapitre est de définir une procédure permettant d'optimiser les paramètres de l'algorithme. Cette procédure est décrite en détail section 3. Les performances de l'algorithme, utilisé avec des paramètres optimisés, sont évaluées sur une base de données de battements annotés et les résultats obtenus sont discutés section 4. Enfin, la dernière section nous ramène au coeur du sujet délicat de transfert et d'exploitation en clinique des algorithmes développés. A cet effet, une station d'analyse du signal ECG, qui intègre l'algorithme de segmentation à base d'ondelettes ainsi que d'autres modules (extraction des paramètres liés à la variabilité cardiaque, visualisation de l'évolution temporelle des indicateurs, corrections battements à battements, segmentation semi-automatiques des battements) nécessaires à une exploitation directement par le clinicien, est décrite.

3.1 Segmentation des ondes : position du problème et bibliographie

L'objectif principal de l'étape de segmentation est d'identifier les instants de début, de fin et les maxima des ondes P, Q, R, S et T de chaque battement du signal ECG. Nous ne nous intéresserons pas ici à l'onde U, qui a une signification physiologique encore controversée [Wu et al., 2002]. Une fois cette segmentation effectuée, il est aisé d'extraire d'autres indica-

teurs utiles, tels que les amplitudes des ondes, le niveau et la pente du segment ST ou les pentes du complexe QRS.

De manière générale, les outils de segmentation automatique de l'ECG peuvent être décomposés en quatre étapes présentées figure 3.1.

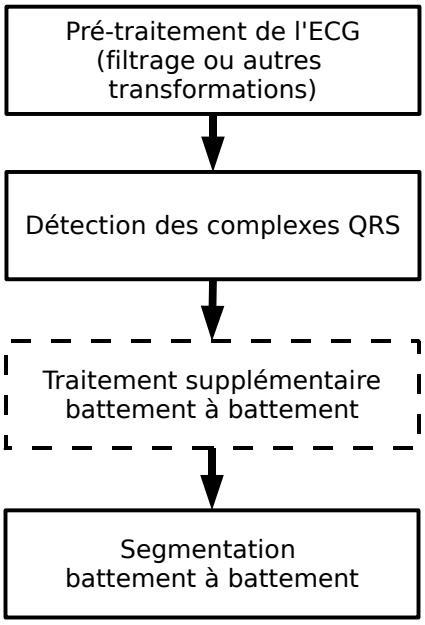


FIG. 3.1 – Les 4 étapes de la segmentation du signal ECG. La 3ième étape est facultative (dépend de l'application).

La première étape consiste principalement à éliminer certains bruits qui entachent le signal ECG (interférence à 50Hz avec l'alimentation des appareils de mesures, artefacts liés aux électrodes, déviation de la ligne de base liée aux mouvements du coeur et du patient etc...). Différentes méthodes reportées dans la littérature proposent déjà des performances très satisfaisantes. Elles sont basées sur des filtres simples ou des filtres adaptatifs [Thakor et Zhu, 1991, Ziarani et Konrad, 2002]. Dans cette étape de pré-traitement, on peut aussi inclure les méthodes qui utilisent l'information contenue dans les différentes voies de l'ECG pour créer des dérivations factorielles et ainsi réduire la présence de bruits sur ces dérivations. C'est par exemple l'intérêt du VCG [Edenbrandt et Pahlm, 1988, Kachenoura et al., 2007].

La seconde étape 2, associée à la détection des complexes QRS, peut aussi être considérée comme un problème déjà résolu de manière satisfaisante. Une bibliographie est reportée dans [Köhler et al., 2002] et une étude réalisée au sein de notre laboratoire montre que chaque détecteur n'atteint pas les mêmes performances suivant les conditions d'utilisation [Portet et al., 2005].

La troisième étape est facultative. Elle est généralement fondée sur la pseudo-périodicité du cycle cardiaque : il est par exemple possible de réduire le bruit des battements en effectuant un moyennage sur une fenêtre temporelle et en isolant les battements ectopiques. Cette étape dépend de la pathologie étudiée : en cas de changements brusques intervenant d'un battement

à l'autre le moyennage ne peut être employé. Par contre si la pathologie se manifeste par des changements s'installant progressivement, sur plusieurs battements, alors cette méthode devient intéressante.

La quatrième étape, la segmentation elle-même, est la plus complexe. Effectuer une délimitation précise des ondes est une tâche difficile pour plusieurs raisons : i) malgré les étapes de pré-traitement, les battements ECG peuvent présenter un rapport signal-bruit faible, ii) une grande variété de morphologies d'ondes existe, même parmi les sujets sains, iii) il n'y a pas de définition universelle sur les positions des bornes des ondes (en particulier pour les ondes Q et T), iv) les composantes spectrales des ondes se chevauchent comme l'illustre la figure 3.2. L'ensemble de ces points témoigne de la difficulté de cette tâche et explique que la littérature sur les techniques de segmentation de l'ECG est vaste. Il est d'ailleurs à noter que, même entre cardiologues, la variabilité portant sur la segmentation peut être élevée [CSE, 1985].

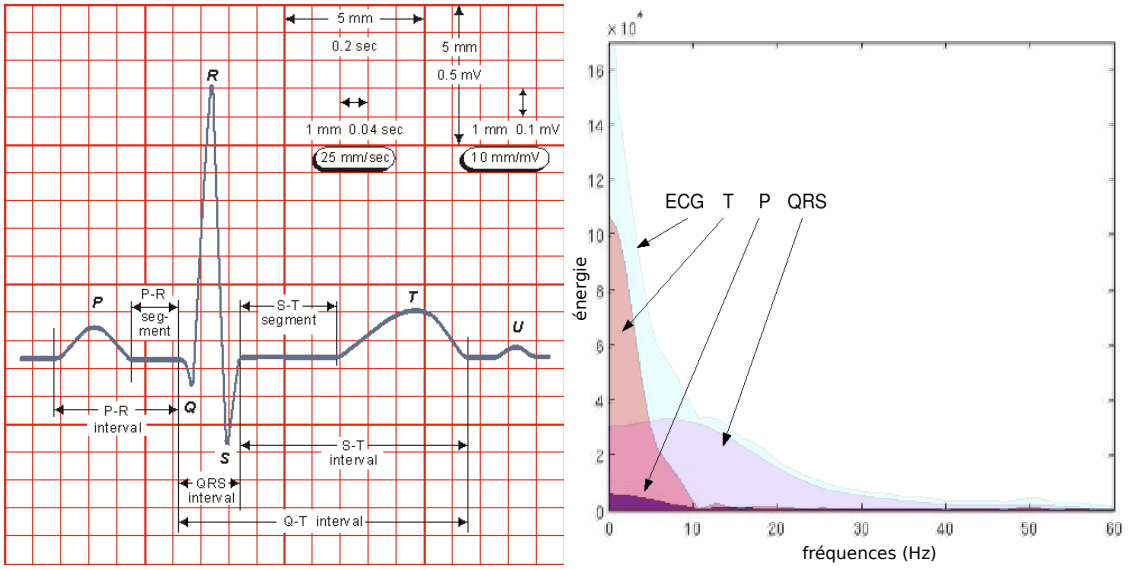


FIG. 3.2 – Représentation d'un battement ECG dans le domaine temporel et dans le domaine fréquentiel.

Les outils de segmentation automatique (ou semi-automatique) de l'ECG intègrent différentes étapes de transformation ou de modélisation de l'ECG. Afin d'expliquer l'intérêt de notre approche et de nos choix, les bases des différentes méthodes issues de la littérature ainsi que leurs résultats sont brièvement commentés.

3.1.1 Filtrage et dérivation

La méthode proposée par Laguna [Laguna et al., 1994] fait office de référence de par les résultats satisfaisants qu'elle produit et de par sa simplicité. Après avoir éliminé les déviations de la ligne de base, un filtre passe-bas dérivateur est appliqué. Une technique simple à base de seuils et de fenêtres temporelles est alors utilisée pour déterminer les pentes significatives caractérisant les complexes QRS. Pour les ondes P et T, un deuxième filtrage passe-bas est employé. De même que pour le QRS, des fenêtres temporelles sont fixées pour détecter les pentes significatives, avec, pour l'onde T, une fenêtre relative à la fréquence cardiaque. Les pics sont définis comme les points de croisements par zéro des signaux dérivés-filtrés puis les limites

sont recherchées en définissant des seuils relatifs aux amplitudes des pentes significatives. Cette approche montre qu'il est intéressant d'analyser les ondes du complexe QRS et les ondes P, T avec au moins deux niveaux de filtrage distincts. Ceci constitue une prise en compte très simple de la non-stationnarité présente dans les battements. De plus, les résultats obtenus sur la base QTDB [Laguna et al., 1994], présentés table 3.1 en terme d'erreur moyenne et d'écart type, montrent que la segmentation basée sur des fenêtres et des seuils est une solution envisageable.

	Pon	$Poff$	$QRSon$	$QRSoff$	$Tpeak$	$Toff$
$mean(ms)$	10.4	-3.7	-4.1	-1.0	-8.0	11.6
$\sigma(ms)$	12.5	11.4	9.0	8.4	15.0	28.2

TAB. 3.1 – Résultats de segmentation obtenus par Laguna, sur la base QTDB de Physionet. Pon, $Poff$, $QRSon$, $QRSoff$ sont respectivement les débuts et fins de l'onde P et du complexe QRS tandis que $Tpeak$ et $Toff$ désignent le pic et la fin de l'onde T.

3.1.2 Filtrage Adaptatif

La méthode suggérée par Soria-Olivas [Soria-Olivas et al., 1998] combine un filtre simple et un filtre adaptatif. Elle inclut, d'une certaine manière une étape d'apprentissage puisque des cardiologues réalisent une détection de l'onde T suivant leur critère personnel. Ceci permet l'estimation d'une constante d'adaptation, introduite dans le filtre adaptatif. Par la suite, l'ECG est filtré par un filtre classique puis par le filtre adaptatif. Un minimum local de l'ECG filtré qui apparaît à la fin de chaque onde T est détecté et dénote toutes les fins d'ondes T. Les résultats de cet algorithme sont jugés satisfaisants mais il est important de rappeler qu'il ne peut être employé sans une pré-détection par un cardiologue. De plus les auteurs n'ont pas réalisés de tests sur des bases de données. Il est donc difficile d'apprécier exactement les différences de performance avec une méthode totalement automatique.

3.1.3 Réalignement temporel par programmation dynamique

Le réalignement temporel par programmation dynamique ou dynamic time warping (DTW) est initialement une mesure entre séries temporelles qui est insensible aux dilatations et aux compressions temporelles. Cette méthode consiste à réaligner les points d'une série temporelle un à un, sur une autre série prise comme référence, de manière à minimiser une distance cumulée sur l'ensemble des points. Dans le cadre d'une segmentation de battements ECG, c'est principalement cette propriété de réalignement qui est utilisée : le battement à segmenter est comparé et ré-aligné avec une base de données de battements préalablement segmentés (figure 3.3). Les différentes méthodologies à base de DTW se distinguent principalement sur le pré-traitement des battements. Cela peut être une approximation linéaire par morceau [Vullings et al., 1998] ou approximation constante adaptative par morceaux (Adaptive Piecewise Constant Approximation) [Zifan et al., 2006].

Cet algorithme est naturellement extrêmement dépendant de l'ensemble des battements de référence. Il est également difficile en temps réel de comparer tous les battements annotés au battement à traiter. Une solution est alors de créer une matrice de distance entre les battements annotés, puis de l'organiser pour privilégier les comparaisons avec les battements voisins, si la distance courante est déjà faible, ou pour aller explorer des battements très différents, si la distance courante est importante. Par exemple [Zifan et al., 2006] réalise des clusters de battements annotés à partir d'une base de données. Enfin, la variabilité battement

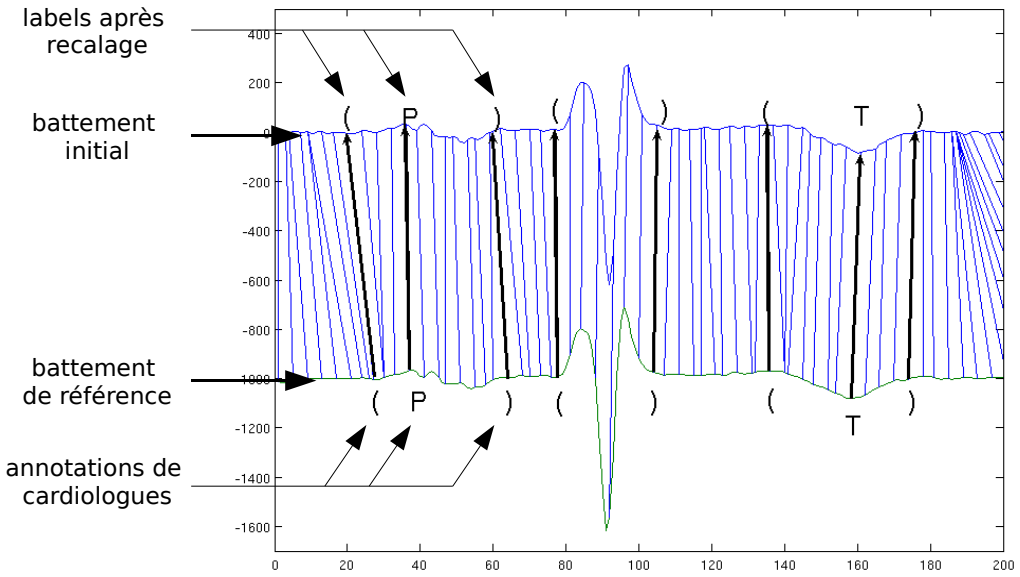


FIG. 3.3 – Segmentation par recalage avec le Dynamic Time Warping.

à battement de la segmentation automatique comparée à des annotations manuelles est importante ([Vullings et al., 1998] et [Zifan et al., 2006]). Par exemple, dans [Vullings et al., 1998], elle est doublée pour les indicateurs des ondes P et du QRS et l'est presque pour les indicateurs de l'onde T, lorsqu'une comparaison est effectuée avec les résultats de Laguna [Laguna et al., 1994]. Cette variabilité peut être expliquée par deux raisons : i) en cas de problème de ré-alignement, l'erreur commise peut-être très grande ; ii) lorsque le battement de référence change, une variation significative et instantanée dans la position des points labelisés sur le battement courant est inévitablement introduite.

3.1.4 Modélisation markovienne

Au cours des années 80-90 les modèles de Markov ont reçu une attention particulière pour la modélisation de séries temporelles, notamment en traitement de la parole. Les premières applications en modélisation du signal ECG ont tout d'abord concerné la simulation d'arythmies et ont été introduites par Doerschuk [Doerschuck, 1985]. Plus tard, Coast [Coast, 1990] a proposé un modèle de Markov pour la détection des séquences d'ondes de l'ECG, notamment la présence ou l'absence d'ondes P pour la détection d'arythmies supra-venticulaires. La première application à la segmentation des complexes QRS est présenté par [Coast, 1993] et est généralisée un peu plus tard par Thoraval [Thoraval, 1995] avec une segmentation de l'ensemble des ondes. Les premiers modèles, par exemple celui de Koski [Koski, 1996], pour la simulation d'ECG, et celui de Clavier [Clavier et Boucher, 1996], pour la détection précoce de fibrillation auriculaire après une segmentation et une classification des ondes P, intègrent dans les états des information de pentes, d'amplitudes ou de durées pour différencier les ondes. Dans [Hughes et al., 2003], deux modèles sont créés : le premier fonctionnant de manière similaire à ceux de Koski et Clavier, avec des paramètres extraits suite à une régression linéaire sur le signal ; le deuxième exploitant des coefficients d'une décomposition en ondelettes. Des résultats significativement meilleurs sont obtenus avec la décomposition en ondelettes (par exemple un taux de détection d'onde P plus grand). Ces mêmes résultats sont aussi comparés à ceux

reportés dans [Graja et Boucher, 2003] et [Laguna et al., 1994] et sont plus compétitifs. Les décompositions en ondelettes ont aussi été utilisées par Thoraval [Thoraval, 1995], combinés avec d'autres paramètres (fréquentiels, d'amplitudes...) dans des modèles semi-Markoviens ou par Lepage [Lepage et al., 2001], avec des modèles de Markov par niveau ou encore plus récemment dans [Graja et Boucher, 2005] par un modèle de Markov en arbre, où la connectivité entre tous les états n'est pas complète, mais seulement entre les coefficients d'ondelettes de niveau voisin, et où chaque état émet, comme observation, un coefficient d'ondelette particulier. Cet structure est testée avec succès mais sur seulement quelques enregistrements, et n'a pas été évaluée sur de larges bases de données.

Les modèles de Markov sont particulièrement adaptés pour la décomposition de la séquence ECG, par exemple pour retrouver des épisodes d'arythmie ou des battements anormaux. Bien que des raffinements aient été apportés sur les premiers modèles proposés de Coast et de Doerschuk pour effectuer des tâches de segmentation, leur application sur de larges bases de données est peu fréquente. Le choix des caractéristiques dont la dynamique doit être modélisée demeure un problème, ainsi que le grand nombre de données nécessaires à l'apprentissage.

3.1.5 Modélisation physiologique

L'algorithme de Vila [Vila et al., 2000] modélise des potentiels d'action de manière à ajuster puis segmenter l'onde T : deux potentiels d'action sont modélisés puis sommés au sein d'une équation à 7 degrés de libertés, qui constitue le modèle physiologique. Les paramètres sont estimés par un algorithme d'optimisation non-linéaire de manière à minimiser l'erreur quadratique moyenne entre le signal observé et le modèle. L'onde T est ensuite segmentée en analysant les dérivées première et seconde de l'équation obtenue. Ces dérivées peuvent être interprétées comme des versions débruitées des dérivées issues du signal initial et la difficulté de la segmentation est ainsi reportée sur l'estimation des paramètres du modèle. Les auteurs reportent des améliorations de segmentation comparé à l'algorithme de filtrage/dérivation de Laguna [Laguna et al., 1994], par exemple le taux de pics d'ondes T bien classés (moyenne inférieure à 15 ms et écart type inférieur à 30.6 ms) est de 82%, contre 72% dans [Laguna et al., 1994]. Il a aussi été montré par Wong [Wong, 2004] qu'en présence de bruit cette approche devient intéressante comparée à celle de Laguna. Cependant, une limitation majeure actuelle est qu'elle ne permet que la segmentation de l'onde T.

3.1.6 Approche temps-échelles par décomposition en ondelettes

Compte-tenu de la non-stationnarité du signal ECG et du recouvrement spectrale des ondes à extraire (figure 3.2), les approches temps-échelles, qui permettent la réduction des bruits aux échelles larges, puis la redéfinition plus précise des positions aux échelles fines, fournissent une alternative intéressante comparée aux autres méthodes. La transformée en ondelettes décompose un signal $f(t)$ en une somme pondérée de fonctions de bases qui sont des versions dilatées et translatées d'une fonction ψ appelée ondelette mère :

$$f(t) = \sum_{a \in \mathbf{Z}} \sum_{k \in \mathbf{Z}} c_{ak} \psi(s^a t - k) \tag{3.1}$$

Pour un scalaire s fixé, le coefficient a contrôle l'échelle à laquelle le signal est analysé par ψ. Si a est large, les caractéristiques temporelles longues (i.e. les basses fréquences) sont analysées, si a est faible les caractéristiques courtes (les hautes fréquences) sont analysées. Cette décomposition permet donc d'étudier le signal dans différentes bandes de fréquence. Une fois l'ECG décomposé en plusieurs niveaux, les ondes sont recherchées de manière séquentielle,

en utilisant les minima et maxima locaux trouvés sur les niveaux qui caractérisent le mieux chaque onde individuellement. La recherche de ces minima et maxima passe par la définition de fenêtre temporelle, limitant la zone de recherche, ainsi que la définition de seuils, permettant uniquement la conservation des minima et maxima significatifs. Pour chaque onde, plusieurs niveaux peuvent être utilisés, en recherchant d'abord les pentes, qui nécessitent plus de lissage et sont déterminées sur des niveaux de basses fréquences, puis en analysant les niveaux de fréquences plus élevées, notamment pour trouver les pics des ondes, une fois que les deux pentes correspondant à chaque pic ont été trouvées. Un exemple simple, avec la délimitation d'une onde T à l'aide d'une décomposition en deux niveaux, est présenté figure 3.4.

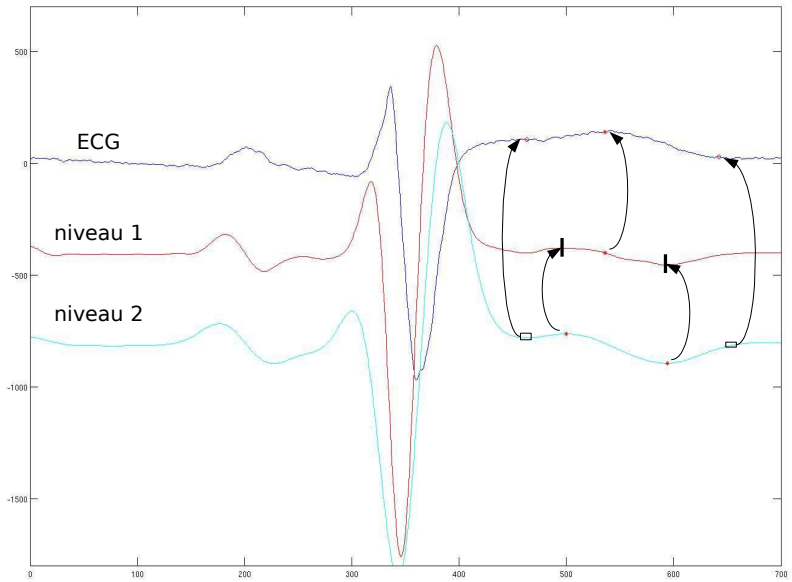


FIG. 3.4 – Detection d'une onde T et de ses bornes à l'aide d'une décomposition en ondelettes à deux niveaux.

Les travaux de Li [Li et al., 1995] appliquent cette méthode notamment pour la détection des complexes QRS et pour la segmentation des ondes P et T. Une généralisation a été effectuée par Lhotska [Lhotska et al., 2003] et aussi par Martinez [Martinez et al., 2004], avec une segmentation du QRS et surtout une évaluation rigoureuse sur plusieurs bases de données. Ces travaux ont souligné que la décomposition en ondelettes se révèle être une approche intéressante, à la fois pour détecter et segmenter des ondes. Par exemple, Li obtient un taux de détection correct des QRS de 99.8% tandis que Martinez, qui compare la détection de la fin de l'onde T avec Vila [Vila et al., 2000] et Laguna [Laguna et al., 1994], montre que l'erreur moyenne est de 1.6 ms contre 13.5ms et 0.8ms et surtout que l'écart type est réduit à 18.1ms contre 27 et 30.3ms dans les travaux de Vila et Laguna respectivement. Cependant, la principale difficulté de cette méthode est liée à la définition des nombreux paramètres, les fenêtres temporelles et les seuils, qui sont ajustés empiriquement.

3.1.7 Bilan de l'état de l'art

Cette revue bibliographique a mis en avant les caractéristiques particulières des différentes classes de méthodes de segmentation du signal ECG ainsi que la comparaison de résultats

lorsque celle-ci est possible. Ceci nous a permis de mettre en exergue que les méthodes fondées sur la décomposition en ondelettes ont prouvé leur utilité et ont montré, sur des bases données annotées, des résultats supérieurs aux autres classes de méthodes. De ce fait, bien que le nombre de paramètres à ajuster puisse être grand, une approche à base d'ondelettes est privilégiée dans la procédure de segmentation.

3.2 Méthode de segmentation d'ECG proposée

La chaîne de traitement appliquée pour segmenter l'ECG se décompose en 4 étapes principales, dont les objectifs respectifs ont déjà été abordés dans la section précédente (figure 3.1). Un exemple représentatif d'application de ces quatre étapes sur un ECG réel est présenté figure 3.5. Des détails sur les algorithmes employés sont donnés dans les sous-sections suivantes, avec notamment une description plus précise de l'étape de segmentation, sur laquelle des améliorations ont été apportées par rapport aux méthodes relevées dans la littérature.

3.2.1 Etape 1a : Suppression des déviations de la ligne de base

Les déviations de la ligne de base sont des déviations basses-fréquences de l'amplitude de l'ECG qui sont liées principalement aux mouvements du patient et à la respiration. De nombreuses méthodes ont été proposées, avec notamment des approches par filtrage ou par approximation par splines. Cependant les approches de filtrage sont préférées car les approximations par splines peuvent génèrer des distortions [Jané et al., 1992]. Un algorithme à base de bancs de filtres multicadences inspiré de [Shusterman et al., 2000] est appliqué ici. Ces étapes de filtrage/sous-échantillonnage permettent de réaliser un filtre dont la fréquence de coupure est bien maîtrisée, tout en évitant les déphasages et les temps de calcul plus longs qui seraient introduits par un filtre unique, mais d'ordre plus élevé.

3.2.2 Etape 1b : Suppression des interférences à 50hz

Après la détection des battements, un recalage puis un moyennage de ces battements sont effectués. Ces deux étapes peuvent être biaisées en présence d'une composante à 50Hz, provenant des matériels électroniques, et qu'il convient d'éliminer.

La suppression de l'interférence à 50Hz a fait l'objet de nombreuses publications et la revue de Levkov [Levkov et al., 2005] recense un grand nombre de méthodes. Les méthodes à base de filtrage adaptatif sont très souvent retenues car elles permettent de suivre les changements de phases et d'amplitudes de l'interférence. Ces filtres sont très usités en électocardiographie pour l'élimination de la ligne de base, la suppression de l'interférence 50Hz, l'annulation du complexe $QRS - T$ [Thakor et Zhu, 1991, Vasquez et al., 2001]. Par contre il est remarqué [Levkov et al., 2005, Thakor et Zhu, 1991], que les filtres LMS introduisent des perturbations, notamment après les portions du signal ECG qui ont une forte amplitude (i.e. le complexe QRS). En effet les fortes amplitudes de l'ECG vont accélérer l'adaptation du filtre mais le rapport signal/interférence étant élevé l'interférence n'est pas correctement annulée.

La figure 3.6 décrit le fonctionnement du filtre employé et notamment l'intégration d'une normalisation du pas d'adaptation en fonction de l'énergie instantanée $Ei_x(t)$ du signal de l'ECG.

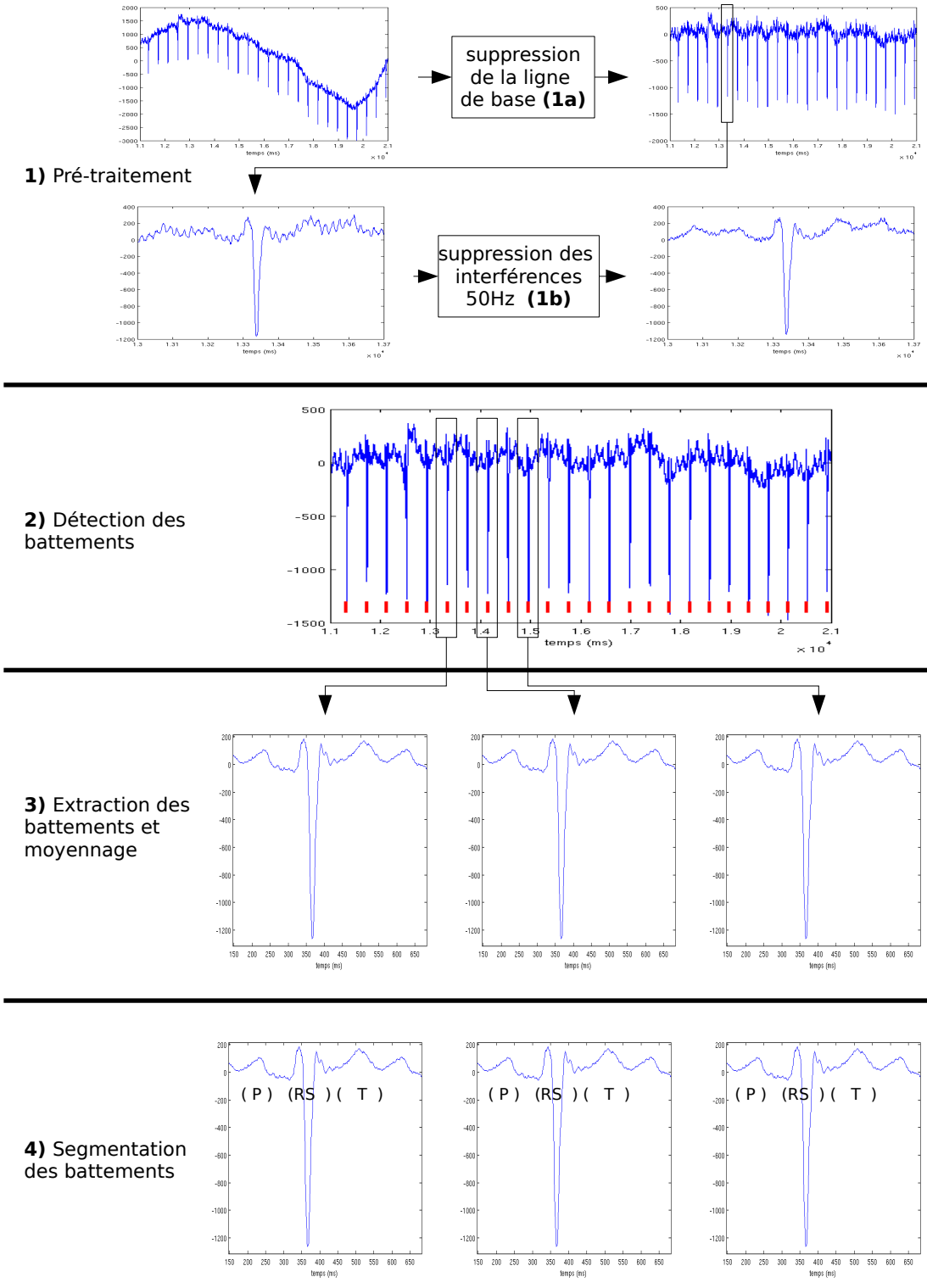


FIG. 3.5 – Chaîne de traitement de l'ECG pour la segmentation.

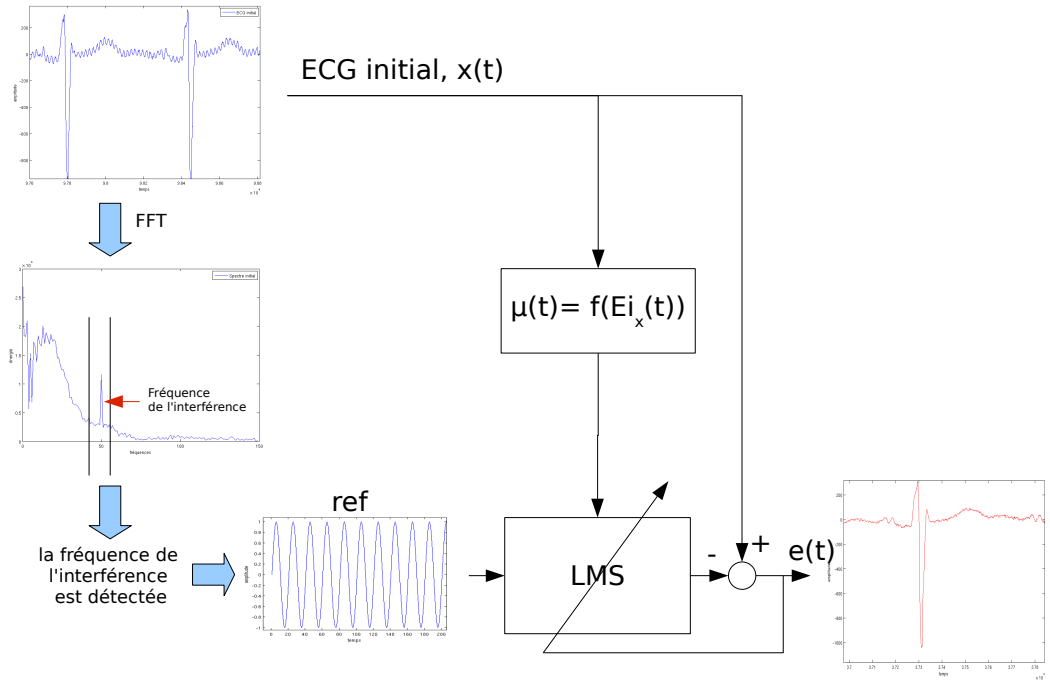


FIG. 3.6 – Principe de fonctionnement d'un filtre adaptatif NLMS pour supprimer les interférences 50Hz.

La première étape est la recherche de la fréquence exacte de l'interférence, celle-ci n'étant pas toujours exactement à 50Hz. Une fois cette fréquence déterminée, un signal d'interférence est généré. Ceci peut être fait aussi bien hors ligne qu'en ligne. Le pas d'adaptation $\mu(t)$, défini en fonction de l'énergie instantanée du signal d'entrée $Ei_x(t)$, s'écrit :

$$\mu(t) = \mu_0 * (1 - min(1, \frac{Ei_x(t)}{E_{max}}))$$

où $Ei_x(t)$ est évaluée sur un intervalle de temps de la taille du filtre. E_{max} est l'énergie de l'ECG au delà de laquelle le filtre n'est plus adapté et est fixé au 85ième percentile de l'énergie de l'ECG mesurée sur un intervalle ΔT de 5 secondes. Les paramètres ΔT et le 85ième percentile sont choisis de manière expérimentale et ceci pour obtenir un pas d'adaptation faible pendant les périodes d'énergie importante de l'ECG. La référence du bruit est filtrée puis soustraite à l'ECG, ce qui donne en $e(t)$ un ECG débruité. Les poids du filtre sont mis à jour avec le pas d'adaptation $\mu(t)$ intégré dans l'algorithme LMS standard.

Ce type de filtrage adaptatif, proche du filtre adaptatif normalisé des moindres carrés (NLMS), fournit une très bonne solution, avec des propriétés satisfaisantes :

- l'adaptativité du filtre permet de bien supprimer l'interférence, en s'adaptant aux distorsions qui peuvent l'affecter.
- l'algorithme NLMS procure une adaptation modulée en fonction du rapport signal à bruit (SNR) : elle est rapide entre les battements (SNR faible, bruit bien adapté) et lente au niveau des complexes QRS (SNR élevé, bruit peu adapté). La composante à 50Hz est alors bien réduite, même après les QRS.

3.2.3 Étape 2 : Détection des battements

L'algorithme de Pan et Tompkins [Pan et Tompkins, 1985], largement reporté dans la littérature, a été choisi pour détecter les battements et appelle à peu de commentaires. Il retourne la position d'un point, appelé point fiducial pour chaque complexe détecté. Ce point se situe au niveau où la dérivée du battement présente l'énergie la plus élevée.

3.2.4 Étape 3 : Extraction des battements et moyennage

Une fois les battements détectés, une fenêtre temporelle, dont la taille est fixée de manière à contenir un battement entier, est ouverte autour de chacun d'entre eux. Ces battements (associés à leur fenêtre temporelle) sont ensuite recalés un par un sur des battements battements prototypes, qui représentent des types de morphologies différentes et qui sont des battements moyennés. La mesure de corrélation utilisée pour le recalage sert aussi aussi à affecter le battement analysé au prototype le plus ressemblant, suivant le maximum de corrélation entre les prototypes. Le battement analysé participe ensuite au moyennage du prototype. Ceci est fait pour chaque battement, excepté lorsqu'aucun battement prototype ne donne une mesure de corrélation supérieure à un seuil donné et dans ce cas le battement courant devient alors lui-même un prototype.

3.2.5 Étape 4 : Application de l'algorithme d'ondelettes et segmentation

L'ondelette mère $\psi(t)$, à partir de laquelle sont dérivés le filtre d'approximation $L(z)$ et le filtre de détail $H(z)$, est une spline quadratique, dont la transformée de Fourier est :

$$\psi(\omega) = j\omega \left(\frac{sin(\frac{\omega}{4})}{\frac{\omega}{4}} \right)^4$$

Cette ondelette peut être considérée comme la dérivée d'une fonction de type passe-bas. Elle est ainsi très utile pour analyser les pentes des différentes ondes de l'ECG et a déjà été appliquée avec succès dans plusieurs travaux [Li et al., 1995, Bahoura et al., 1997, Martinez et al., 2004]. Comme dans les travaux de [Li et al., 1995] et [Martinez et al., 2004], cinq niveaux d'ondelettes, $W2^1$ à $W2^5$, ont été utilisés, avec une décomposition dyadique. Le niveau $W2^0$ fait référence au signal initial. Ces cinq niveaux de décomposition peuvent donc être réalisés par une cascade de filtre (en octave) présentée figure 3.7 et dont les transformées de Fourier sont présentées figure 3.8

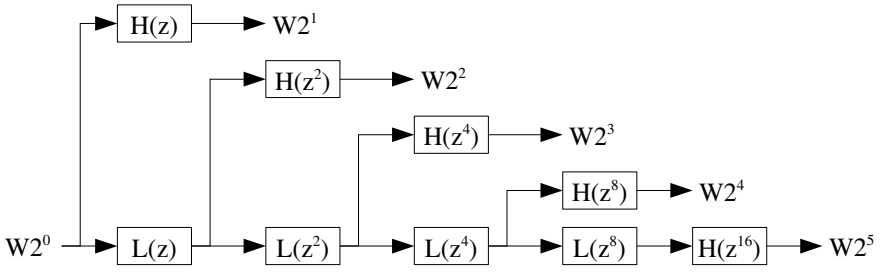


FIG. 3.7 – Banc de filtres de la décomposition en ondelettes. $L(z)$ et $H(z)$ sont respectivement les filtres d'approximation et de détails. $W2^k$ sont les sorties du filtre aux echelles 2^k (k = 1 à 5), $W2^0$ désigne le battement initial.

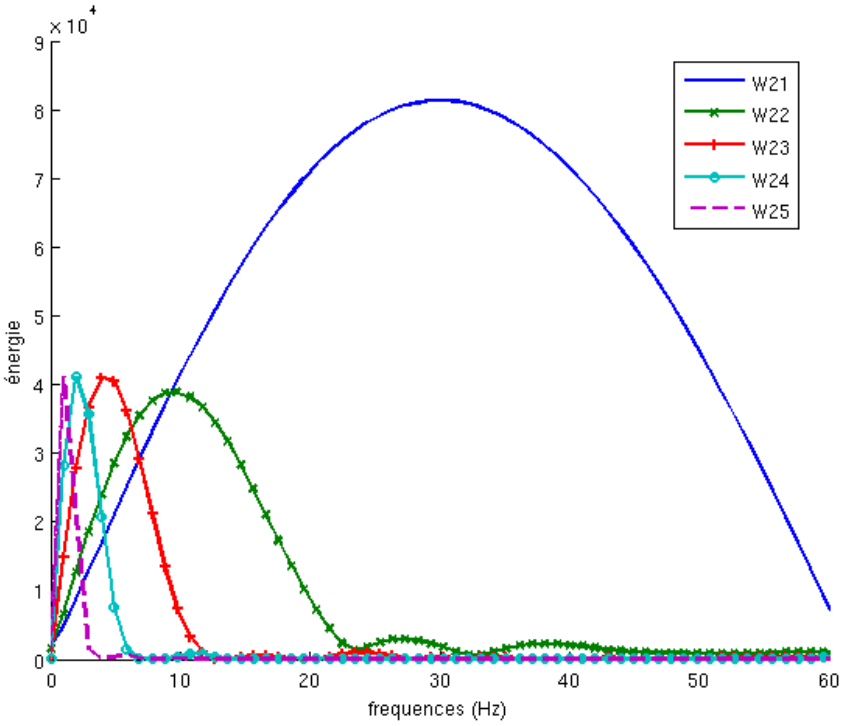


FIG. 3.8 – Transformée de Fourier des filtres.

Certains indicateurs sont directement extraits des échelles 1 à 5, sans remonter au signal initial. Ne pas décimer entre chaque décomposition favorise alors la résolution temporelle pour l'extraction de ces indicateurs, ceci explique que les décimateurs usuellement présents dans les transformations en ondelettes n'ont pas été conservés dans ce travail.

Une fois la décomposition effectuée, l'onde R, puis les ondes Q et S et enfin les ondes P et T sont successivement extraites. La procédure globale est la suivante :

- L'onde R est détectée en analysant $W2^2$, en estimant les deux plus grands extrema de signes opposés et ensuite en cherchant le croisement par zéro, situé sur $W2^1$ entre ces deux extrema. Si plusieurs croisements par zéro apparaissent, celui associé avec la plus grande amplitude de $W2^0$ est retenu. Les paramètres R1 et R2 définissent le support temporel utilisé pour la recherche de l'onde R.
- Le niveau iso-électrique est déterminé à l'aide de la méthode décrite par Smrdel et Jager [Smrdel et Jager, 2004].
- Les ondes Q et S, ainsi que le début et la fin du QRS (QRSon et le QRSoff), sont estimés en analysant les maxima locaux sur les deux premières échelles, $W2^1$ and $W2^2$, comme dans Martinez [Martinez et al., 2004]. Les seuils $\gamma_{QRSpre,QRSpost}$ sont proportionnels à la pente maximale de l'onde R et sont utilisés pour identifier les autres pentes significatives des ondes Q et S. Les supports temporels pour rechercher ces ondes sont définis par les paramètres $QRS_{Qlim,Slim}$.
- Les ondes P et T sont délimités en analysant d'abord les échelles $W2^4$ et $W2^5$: si ces ondes ne sont pas détectées sur $W2^4$, elles sont aussi recherchées sur $W2^5$. Les extrema sont trouvés sur les échelles juste précédentes (i.e. $W2^3$ ou $W2^4$). Les seuils pour les détections sont $\epsilon_{(P,T)}$ et sont relatifs aux énergies des échelles $W2^4$ ou $W2^5$.

Les pentes significatives sont identifiées suivant les seuils $\gamma_{T,P}$. Les fenêtres temporelles pour les ondes P et T dépendent de l'intervalle RR précédant (seuils P_RR et T_RR). T1•/P1• sont les bornes de gauche de ces fenêtres et T2•/P2• les bornes de droite, • désigne l'index qui change en fonction de l'intervalle RR.

- Les débuts et fin des ondes sont détectés suivant les seuils $\xi_{(QRSon,QRSend,Ton,Tend,Pon,Pend)}$ relatifs aux amplitudes des premières ou dernières pentes significatives.
- Comparé aux travaux de Martinez [Martinez et al., 2004], un test supplémentaire est effectué pour vérifier si les pics des ondes sont plus éloignés du niveau iso-électrique que les début et fin des ondes. Si ce n'est pas le cas, il est considéré que la délimitation effectuée n'est pas correcte et le support temporel est redéfini. Ce test est appliqué successivement, en partant d'un support temporel large et en le réduisant.

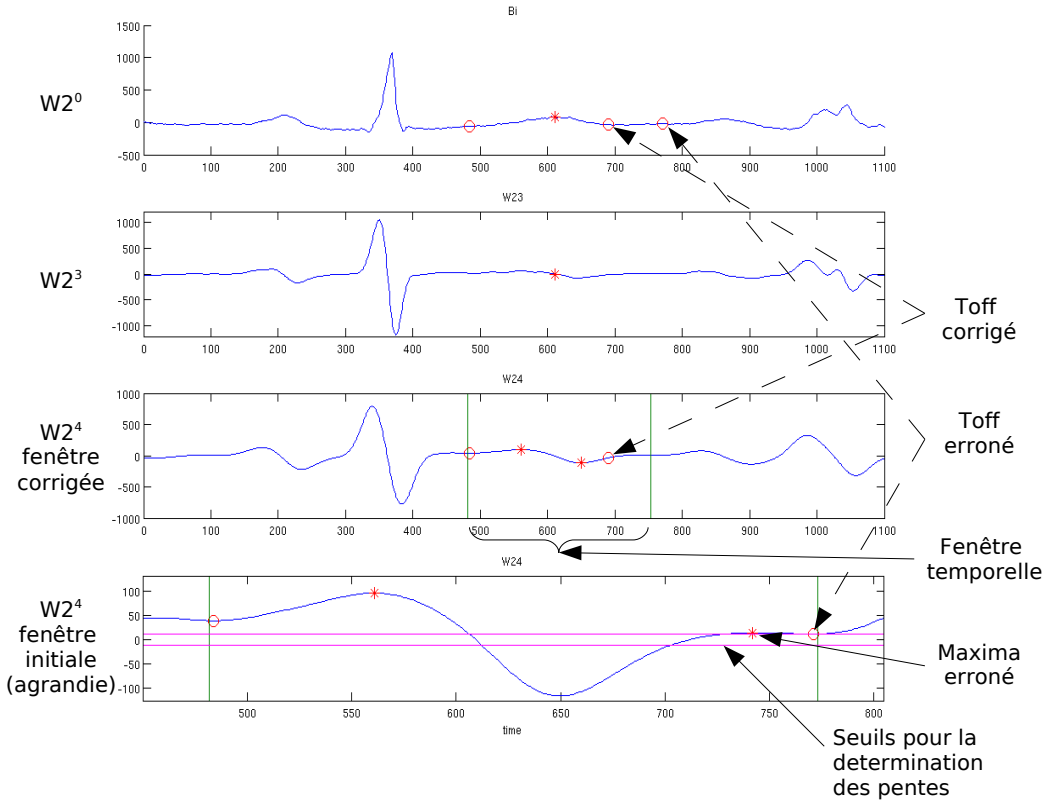


FIG. 3.9 – Exemple de segmentation d'une onde T nécessitant un ré-ajustement du Toff.

Une segmentation effectuée sur une onde T d'un battement sinusal et en utilisant les échelles 0, 3 et 4 est illustrée sur la figure 3.9. Le premier tracé présente le battement original ($W2^0$), le second est la décomposition au niveau 3 et les deux derniers tracés sont des décompositions au niveau 4, avec des supports temporels différents. Les lignes verticales représentent les limites de ces supports temporels. Sur le dernier tracé, un maximum local de faible amplitude (l'étoile sur la droite) est détecté dans cette fenêtre temporelle et se trouve juste au dessus du seuil de décision pour l'identification des pentes significatives. Prendre en compte ce maximum conduit à une position de Toff qui ne correspond pas à la véritable fin de l'onde T. L'étape de vérification identifie ce problème et réajuste le support temporel de telle sorte que ce maximum

ne soit plus pris en compte. Les limites finales sont présentées sur $W2^0$ et définissent bien le début et la fin de l'onde T.

Les performances de cette approche dépendent naturellement de la bonne définition des fenêtres d'analyse temporelle, mais aussi des seuils de décision qui spécifient les pentes significatives et les limites des ondes. Dans la littérature [Li et al., 1995, Lhotska et al., 2003, Martinez et al., 2004], ces seuils sont souvent ajustés empiriquement sans savoir si leurs multiples interactions conduisent à une performance (en terme de segmentation) optimale. Cette question a accompagné notre réflexion et explique qu'une méthodologie numérique d'optimisation de seuils est décrite dans le paragraphe suivant.

3.3 Procédure d'optimisation des paramètres de la chaîne de traitement

Les paramètres de l'algorithme de segmentation doivent être ajustés de manière à réaliser une segmentation similaire à celle des cardiologues. La procédure d'optimisation a pour objectif d'apprendre ces paramètres pour que les segmentations effectuées soient le plus proche possible de celles enregistrées dans une base de donnée. Cette procédure, présentée figure 3.10, n'utilise aucune connaissance *a priori* autre que les annotations manuelles enregistrées et est capable de trouver une solution globale dans une domaine de recherche de grande dimension.

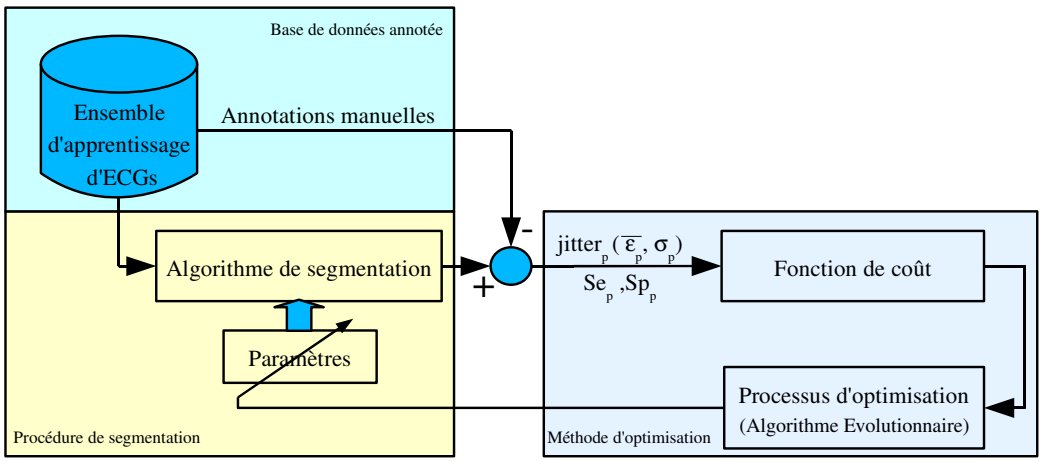


FIG. 3.10 – Procédure d'optimisation des paramètres de l'algorithme de segmentation à partir d'une base de donnée annotée et d'un algorithme évolutionnaire.

Plusieurs contraintes doivent être prises en compte dans le choix de l'algorithme d'optimisation :

- étant donné la complexité de l'espace de recherche des paramètres et les interactions qui existent entre eux, la fonction de coût risque de contenir des minima locaux ;
- l'algorithme de segmentation est composé de différents seuils, ce qui signifie que le critère évaluant les performances sera très certainement discontinue dans l'espace des paramètres.

Ces deux contraintes ne sont pas compatibles avec l'utilisation des méthodes d'optimisation multivariables basées sur les gradients (méthodes de Gauss-Newton) ou des méthodes

du simplex, qui risquent de tomber dans des minima locaux. D'un autre coté les méthodes d'optimisation stochastique, telles que les algorithmes évolutionnaires, sont particulièrement bien adaptées pour ce type problème.

Les algorithmes évolutionnaires sont des techniques d'optimisation, inspirées des théories de l'évolution et de la sélection naturelle, qui peuvent être employées pour trouver des configurations optimales pour un système donné et avec des contraintes spécifiques [Michalewicz, 1996]. Dans ces algorithmes, chaque individu (ou chromosome) d'une population représente une configuration donnée de l'ensemble de paramètres à optimiser. Une population initiale est créée, usuellement il s'agit d'un ensemble aléatoire de chromosomes et cette population va évoluer, en améliorant sa performance globale, à l'aide d'un processus itératif. Au cours de ce processus, la performance de chaque individu pour le problème posé est évaluée à l'aide d'une fonction de coût. Une nouvelle génération est produite en appliquant des opérateurs de mutation et de cross-over en priorité sur les individus qui présentent de faibles valeurs pour la fonction de coût. La figure 3.11 résume ce fonctionnement. Les algorithmes évolutionnaires ont été utilisés dans différentes applications biomédicales, pour estimer de larges ensembles de paramètres, avec des résultats satisfaisants [Hernãndez et al., 2002, R.M. Eichler West et Wilcox, 1998].

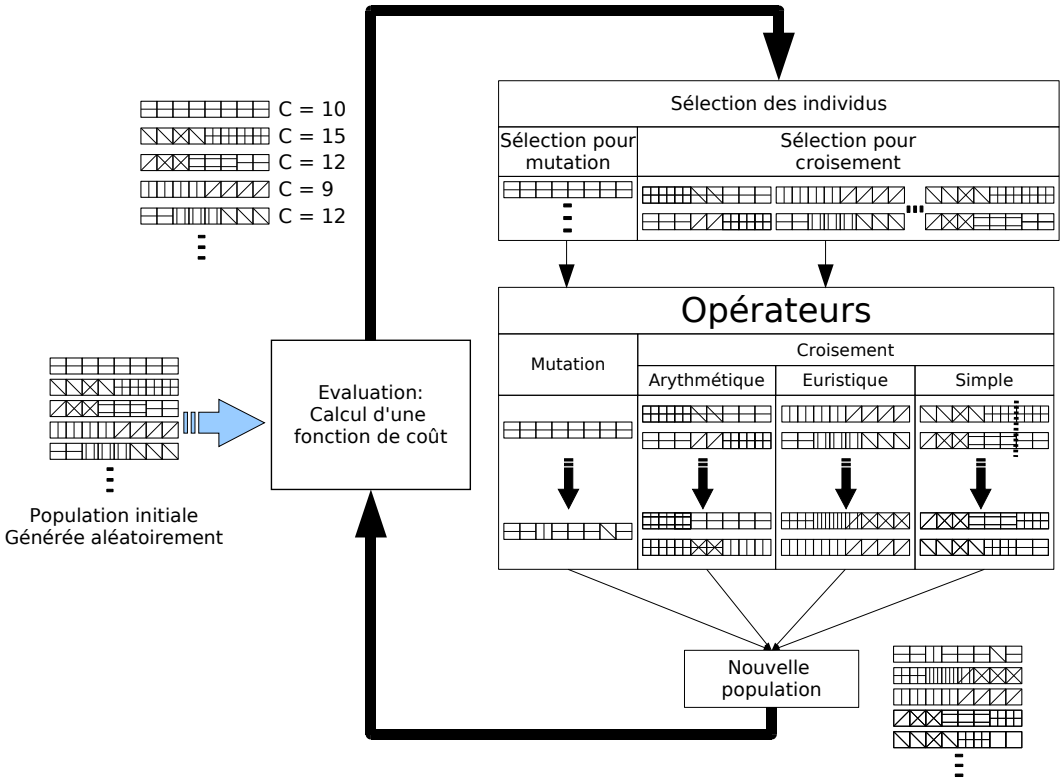


FIG. 3.11 – Principe de fonctionnement des algorithmes évolutionnaires, avec : la définition d'une population, le calcul d'une fonction de coût, la sélection des individus suivant leur coût et l'application d'opérateurs pour créer une nouvelle population.

Les différentes composantes de l'algorithme évolutionnaire spécifique mis en oeuvre sont détaillées dans la sous-section suivante. Une seconde sous-section commente l'introduction de cet algorithme dans notre procédure d'optimisation.

3.3.1 Fonctionnement de l'algorithme évolutionnaire

Comme shématisé figure 3.11, un algorithme évolutionnaire se décompose en trois fonctions principales : la fonction de coût qui évalue la performance des individus de la population courante ; la méthode de sélection qui tire de manière aléatoire, à partir du résultat de la fonction de coût, les individus qui servent à produire la nouvelle population ; les opérateurs qui sont appliqués sur les individus tirés pour générer la nouvelle population. Ces trois fonctions, réalisées pour répondre au problème d'optimisation posé, sont décrites dans les sous-sections suivantes.

3.3.1.1 Fonction de coût

L'algorithme de segmentation fourni la position des 7 indicateurs suivants : Pon, Ppeak, Poff, QRSon, QRoff, Tpeak, Toff. La fonction de coût évalue plusieurs critères sur chacun de ces indicateurs. Ces critères sont :

- La probabilité d'erreur de détection, définie comme :

$$Perr_p = \sqrt{(1 - Se_p)^2 + (1 - Sp_p)^2}$$

  où Se_p et Sp_p sont, respectivement, les sensibilités et spécificités obtenues un indicateurs p ($p \in \{Pon,\ Ppeak,\ldots\ Tpeak,\ Toff\}$). La probabilité $Perr$ est calculée comme la distance entre le point (Se_p, Sp_p) et le point de détection ($Se = 1$ et $Sp = 1$).
- L'erreur moyenne du jitter de segmentation $\overline{\varepsilon_p}$, commise sur les M enregistrements et sur l'ensemble des N_m battements de chaque enregistrement où l'indicateur p est à la fois détecté et aussi présent dans les annotations manuelles :

$$\overline{\varepsilon_p} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N_m} annoté_p(m,n) - détecté_p(m,n)}{\sum_{m=1}^{M} N_m} \tag{3.2}$$

- L'écart type du jitter de segmentation σ_p. Il est calculé comme la moyenne des écarts types du jitter de segmentation pour chaque enregistrement (σ_p^m) qui sont pondérés par leur nombre de battements :

$$\sigma_p = \frac{\sum_m (N_m \sigma_p^m)}{\sum_m N_m} \tag{3.3}$$

Ces trois critères sont évalués pour l'ensemble des points caractéristiques p. Ce type d'optimisation multi-objectif peut être ramené à un problème d'optimisation simple, avec une fonction de coût définie comme une somme pondérée des différents critères :

$$O = \sum_{p=1}^{P}\left(\frac{Perr_p}{a_p} + \frac{\overline{\varepsilon_p}}{b_p} + \frac{\sigma_p}{c_p}\right) \tag{3.4}$$

Cette fonction de coût évalue l'amélioration entre nos résultats ($Perr_p, \overline{\varepsilon_p}, \sigma_p$) et les meilleurs résultats reportés dans la littérature [Martinez et al., 2004, Laguna et al., 1994] appelés ici a_p, b_p, and c_p. C'est le résultat de cette fonction de coût qui est minimisé par l'algorithme d'optimisation.

3.3.1.2 Codage des individus

Dans la définition originale des AG par Holland [Holland, 1992], les chromosomes ont été conçus comme des chaînes de G paramètres (G gènes), chacun codé sur b bits. L'extension des concepts des AG au codage par valeurs réelles a été proposée par Michalewiks [Michalewicz, 1996]. Cette représentation est souvent accompagnée d'une matrice de valeurs maximum et minimum pour chaque gène, afin de restreindre l'espace de recherche aux solutions valides et aussi d'accélérer la convergence vers la solution optimale. Le codage réel a été comparé expérimentalement avec les représentations binaires et a démontré une rapidité de calcul, une précision et une reproductibilité plus grande des solutions estimées [Janikow et Michalewicz, 1991]. En outre, les paramètres à coder dans AE1 et AE2 sont tous à valeurs réelles. Un codage réel a donc été choisi. Les bornes hautes et basses de chaque gène, respectivement nommées a_i et b_i, sont définies de la manière suivante : les fenêtres temporelles sont issues des positions et des durées extrêmes de chaque onde tandis que les limites des autres seuils sont déterminées en augmentant (limite haute) ou diminuant (limite basse) largement les paramètres trouvés dans [Martinez et al., 2004]. Ces intervalles sont employés par les AEs lors de la construction de la population initiale.

3.3.1.3 Méthode de sélection

Chaque nouvelle génération d'individus est créée aléatoirement à partir des individus de la génération courante, au moyen d'un processus de sélection qui favorise l'évolution des individus les plus adaptés. En général, les méthodes de sélection associent à chaque individu (l) une probabilité de sélection Ps, basée sur sa performance Fl, calculée préalablement. Plusieurs critères d'assignation des probabilités de sélection ont été proposés dans la littérature. Les plus courantes sont la méthode de la roulette (introduite par Holland [Holland, 1992]) et la méthode de classement. Dans le cas de la méthode par classement, les valeurs des probabilités Ps(l) sont calculées en fonction du rang de l'individu l, lorsque les individus sont triées dans l'ordre décroissant de leur performance Fl. Les probabilités sont alors données par :

$$p_l = \frac{q*(1-q)^{r_l-1}}{1-(1-q)^P}$$

où q est la probabilité de tirer l'individu de rang 1, r_l est le rang de l'individu l, P est le nombre total d'individus.

D'autres méthodes de sélection, comme les méthodes de "tournoi" ou élitiste peuvent être utilisées, mais elles ne sont que des cas particuliers de la méthode de classement. La méthode de sélection par classement a été employée, avec un paramètre q fixé à 0.08, comme proposé dans [Houck et al., 1995].

3.3.1.4 Les opérateurs de cross-over et de mutation

Plusieurs opérateurs de recombinaison ont été proposés dans la littérature, d'abord pour la représentation génétique binaire puis étendus à la représentation réelle. Des cross-over simples, arithmétiques et heuristiques sont appliqués :

Cross-over simple Le cross-over simple appliqué aux chromosomes à valeurs réelles est identique à celui présenté dans la version binaire : les parents (notés I_1 et I_2) sont scindés en deux parties qui sont ensuite recombinées entre elles pour donner les enfants (notés E_1 et E_2) :

$$E_1 = [I_1\{1..p\}I_2\{p+1...N_p\}]$$
$$E_2 = [I_2\{1..p\}I_1\{p+1...N_p\}]$$

où N_p, est la taille du chromosome et p est une variable aléatoire suivant une loi uniforme sur l'intervalle $[1, N_p - 1]$.

Les cross-over arithmétiques et heuristiques, spécifiques aux chromosomes à valeurs réelles, ont été introduits par Michalewiks [Michalewicz, 1996].

Cross-over arithmétique Deux combinaisons linéaires complémentaires des parents sont générées, pondérées par une valeur aléatoire ρ repartie uniformément dans l'intervalle $[0, 1]$:

$$E_1 = \rho I_1 + (1 - \rho) I_2$$
$$E_2 = (1 - \rho) I_1 + \rho I_2$$

Cross-over heuristique Cet opérateur profite de l'information de performance des parents. Les enfants sont calculés au moyen des relations :

$$E_1 = I_1 + \rho * (I_1 - I_2)$$
$$E_2 = I_1$$

où I_1 a une performance plus grande que I_2 et ρ est toujours une variable aléatoire de distribution uniforme dans l'intervalle $[0, 1]$.

Les opérateurs de mutation n'affectent qu'un seul individu, sur un ou plusieurs de ces gènes. Trois types de mutation ont été employés :

Mutation uniforme Un gène, tiré aléatoirement est remplacé par une valeur aléatoire issue d'une distribution uniforme sur l'intervalle possible de cette variable :

$$E = \begin{cases} I(j) \text{ pour tout } j \neq i \\ U[a_i \; b_i] \text{ pour le gène } i \end{cases} \tag{3.5}$$

où i est une variable aléatoire uniforme tirée sur l'ensemble des gènes et a_i, b_i sont les bornes du gène i.

Mutation non-uniforme La valeur d'un gène est aléatoirement changée dans la direction de l'une des bornes supérieure ou inférieure de ce gène :

$$E = \begin{cases} I(j) \text{ pour tout } j \neq i \\ \begin{cases} E(i) + (b_i - E(i)) * f(gen) \text{ si } r1 = 0 \\ E(i) - (a_i + E(i)) * f(gen) \text{ si } r1 = 1 \end{cases} \text{ pour le gène i} \end{cases} \tag{3.6}$$

où $r1$ est une expérience de Bernouilli, et f(gen) retourne un coefficient dépendant du numéro de la génération courante :

$$f(gen) = (r2(1 - \tfrac{gen}{Maxgen}))^b$$

r2 est une variable suivant une loi uniforme sur $[0, 1]$ et b est un paramètre de forme.

Mutation non-uniforme multiple Il s'agit d'une mutation uniforme appliquée sur l'ensemble des gènes des chromosomes.

3.3.2 Application de l'algorithme évolutionnaire

Afin de réduire la dimension de l'espace de recherche, et en conséquence le nombre d'individus et le nombre d'itérations de l'algorithme évolutionnaire, la procédure d'optimisation globale est décomposée en deux étapes :

- Dans la première étape, appelée AE1, les paramètres jouant un rôle dans la segmentation de l'onde P et du complexe QRS sont ajustés conjointement.
- Dans une seconde étape, AE2, les paramètres associés à l'onde T sont optimisés, en réutilisant les meilleurs paramètres obtenus à la fin de AE1.

Cette décomposition est possible car le support temporel de recherche de l'onde T est lié à la détection de l'onde S et il est considéré que la détection de l'onde T ne sera optimale que lorsque l'onde S sera au préalable bien détecté.

Pour les deux processus AE1 et AE2, les populations sont entraînées sur 80 générations, avec 60 individus. La probabilité de cross-over, p_c, est fixée à 0.7. Dans le but d'obtenir des solutions plus stables et plus fiables, la probabilité de mutation, p_m, est adaptée au cours de l'apprentissage, en commençant avec des valeurs élevées durant les premières générations, pour assurer une recherche dans l'ensemble de l'espace d'état et en diminuant vers la fin, pour faciliter la convergence vers un minimum fiable. La solution de Bäck [Bäck et Schütz, 1996] a été retenue en raison des résultats intéressants qui sont reportés dans la littérature [Sebag et al., 1997, Madeline, 2002, Thierens, 2002] :

$$p_m = (2 + \frac{(N_p - 2)}{Maxgen - 1} * gen)^{-1}$$

avec N_p le nombre de paramètres, $Maxgen$ le nombre maximum de générations et gen le numéro de la génération actuelle.

3.3.3 Conclusion

Les algorithmes évolutionnaires permettent, au cours d'un processsus itératif, d'obtenir un vecteur de paramètres minimisant la fonction de coût fournie. Dans le cadre de notre vecteur de paramètres de grande taille (19 paramètres pour la première partie de l'apprentissage et 11 pour la seconde) et de la fonction de coût discontinue cette approche stochastique est donc privilégiée. La section suivante présente les résultats de l'apprentissage des paramètres et de la segmentation de battements obtenus à partir d'une base de données de signaux ECG.

3.4 Résultats de l'apprentissage et de la segmentation

Les résultats sont présentés en deux étapes, tout d'abord une analyse des paramètres obtenus après l'étape d'apprentissage est effectuée. Ensuite, les performances de segmentation, obtenues avec les paramètres optimisés, sont présentées et comparées avec d'autres résultats extraits de la littérature. La base de données utilisée pour cette application est la base QTDB de physionet [Laguna et al., 1994]. La base QTDB fournit une large variété de pathologies, avec un total de 105 enregistrements contenant deux voies ECG échantillonnées à 250Hz. Comparée aux autres bases de données, elle contient aussi un large nombre de battements annotés par enregistrement : 30 battements annotés au lieu d'un seul dans la base de données CSE. Les annotations sont aussi très complètes, avec toutes les positions des points Pon, Ppeak, Poff, QRSon, QRSoff, Tpeak, Toff. Cette base de données a donc été utilisée pour entraîner notre algorithme de segmentation avec les AEs et, dans un second temps, pour valider les résultats.

3.4.1 Apprentissage et paramètres optimaux

Dans les travaux précédents sur les méthodes de segmentation à base d'ondelettes, une définition manuelle de l'ensemble des paramètres était effectuée. Dans [Martinez et al., 2004], une base de données spécifique était employée pour ajuster les paramètres mais ceux-ci étaient définis manuellement, comme dans [Li et al., 1995] et [Lhotska et al., 2003]. En considérant les problèmes dus aux nombreuses morphologies, le nombre de paramètres, et les objectifs opposés (par exemple une erreur moyenne de segmentation plus faible, au détriment d'une erreur de détection plus élevée) il apparaît difficile d'obtenir des résultats fiables avec de telles approches. Utiliser la procédure d'optimisation sur une base de données manuellement annotée, telle que la base QTDB, permet de résoudre ce problème.

Pour comparer nos résultats avec ceux obtenus avec les autres méthodes [Martinez et al., 2004] et [Laguna et al., 1994], il est requis d'effectuer la procédure de test, sur tous les battements de la base de données. Puisque l'étape d'entraînement est aussi effectuée sur la même base, un ensemble d'apprentissage L et un ensemble de test T, doivent être définis. Tous les enregistrements sont d'abord divisés en trois parties équivalentes, appelées sous-enregistrements. Deux tiers de l'ensemble des sous-enregistrements sont choisis aléatoirement et affectés à l'ensemble d'apprentissage. Les sous-enregistrements restants sont alloués à l'ensemble de test. Un total de 13 ensembles apprentissage/test sont générés de la même manière pour réaliser une étude en cross-validation de l'algorithme de segmentation et évaluer la sensibilité des paramètres optimaux obtenus pour une instance donnée de L et T. L{1-13} et T{1-13} sont donc les ensembles d'apprentissage et de test et $OP_{L\{1-13\}}$ sont les paramètres optimisés.

Un résumé de tous les paramètres, optimisés avec les AE appliqués sur les différents ensembles d'apprentissage, est présenté tableau 3.2, avec leurs moyennes et écarts types pour les 13 ensembles d'apprentissage.

EA1				EA2	
P11	278±31	γ_{QRSpre}	0.09±0.03	ϵ_T	0.24±0.06
P12	240±17	$\gamma_{QRSpost}$	0.11±0.03	γ_T	0.28±0.07
P21	88±14	$\xi_{QRSonpos}$	0.07±0.04	ξ_{Ton}	0.17±0.09
P22	99±27	$\xi_{QRSonneg}$	0.07±0.04	ξ_{Tend}	0.36±0.07
R1	118±34	$\xi_{QRSendpos}$	0.21±0.12	T11	111±24
R2	111±37	$\xi_{QRSendneg}$	0.23±0.11	T21	441±75
P_{RR}	664±182	QRS_{Qlim}	88±22	T12	90±16
ϵ_P	0.12±0.05	QRS_{Slim}	154±32	T22	0.6±0.08
γ_P	0.4±0.09			T23	581±94
ξ_{Pon}	0.41±0.08			T_{RR1}	705±155
ξ_{Pend}	0.76±0.05			T_{RR2}	1231±70

TAB. 3.2 – Paramètres utilisés par l'algorithme de segmentation (sous-section 3.2.5) avec leurs valeurs optimales, représentés par moyenne ± écart type, obtenus par le processus d'optimisation.

Les valeurs de certains paramètres présentent peu de variations, par exemple les fenêtres temporelles pour les ondes P, Q, S et T ou les seuils qui définissent les débuts et fins des ondes P et T, ce qui indique une forte sensibilité de l'algorithme de segmentation à ces paramètres. D'un autre coté, les fenêtres temporelles utilisées pour rechercher l'onde R, ou les seuils qui définissent les débuts et fins des QRS montrent de larges variations parmi les différents en-

sembles d'apprentissage. Ces résultats révèlent la sensibilité moindre de la segmentation des QRS à ces paramètres. En effet, l'onde R présente généralement un rapport signal sur bruit élevé et est toujours proche du point fiducial. Les pentes du début et de la fin du QRS sont aussi plus nettes que les pentes des ondes P et T, donc une large gamme de paramètres donne approximativement les mêmes résultats. Il est important de souligner que l'approche proposée n'est pas seulement utile pour ajuster les paramètres mais aussi pour analyser la sensibilité de chacun des paramètres vis à vis des résultats de la segmentation.

3.4.2 Résultats de la segmentation

Jané [Jané et al., 1997] a proposé un cadre d'analyse pour les algorithmes de segmentation évalués sur des bases de données semblables à la QTDB. Afin de valider le notre et la méthode d'optimisation, le même cadre d'analyse est appliqué. Une comparaison avec les algorithmes de Martinez [Martinez et al., 2004] et Laguna [Laguna et al., 1994] est alors possible.

Les enregistrements ECG de la base QTDB sont constitués de deux voies. Pour annoter les battements, les cardiologues identifient tout d'abord quelle voie permet la meilleure estimation d'un indicateur donnée et cette voie est alors utilisée pour tous les battements de l'enregistrement. Dans le cas d'une segmentation automatique indépendante entre les deux voies, la meilleure voie est choisie *a posteriori*, en analysant l'erreur commise par rapport aux annotations manuelles.

Les erreurs moyennes et écarts types sont pondérées par le nombre de battements par enregistrements et moyennés sur les enregistrements présents dans l'ensemble de test (équations 3.2 et 3.3). Les probabilités d'erreur de détection ($Perr_p$) sont dérivées à partir de la sensibilité et de la prédictivité, calculées comme dans [Martinez et al., 2004]. Ces 3 critères, erreur moyenne, écart type et probabilité d'erreur de détection, sont évalués pour tous les indicateurs. La figure 3.12 montre les distributions de nos résultats sur les différents ensembles de tests, sous forme de boîtes à moustaches. Les croix (+) et les étoiles (*) sont respectivement les résultats de [Martinez et al., 2004] et [Laguna et al., 1994] présentés dans ces références. Sur la colonne de droite, qui présente la moyenne sur les indicateurs, le rond (o) est notre résultat.

Quelques commentaires peuvent être faits sur chaque critère :
- Pour la déviation moyenne, la médiane des résultats est plus proche de zéro que pour les deux autres méthodes, pour les points Pon, Ppeak, et QRSon. Elle est aussi meilleure que [Martinez et al., 2004] pour Poff et meilleure que [Laguna et al., 1994] pour QRSoff, Tpeak et Toff.
- Pour l'écart type, la médiane est plus faible pour Pon, Peak, Poff, QRSon et Tpeak, plus élevée que [Martinez et al., 2004] pour Toff et plus élevée que [Martinez et al., 2004] et [Laguna et al., 1994] pour QRSoff.
- La probabilité d'erreur de détection est plus basse pour l'onde P, mais pas l'onde T.

La caractéristique de la probabilité d'erreur de détection est plus difficile à analyser que les autres critères car tous les battements automatiquement segmentés n'ont pas été manuellement annotés, ce qui biaise la mesure de la spécificité. Pour vérifier que la probabilité d'erreur de détection est bien optimisée, un test supplémentaire a été effectué : les courbes de caractéristiques opérationnelles du récepteur (COR) pour les ondes P et T sont tracées, pour un ensemble de test donné, et il est vérifié que les paramètres (ϵ_P et ϵ_T) qui minimisent les probabilités d'erreurs sur les courbes COR sont proches de ceux obtenus par le processus d'op-

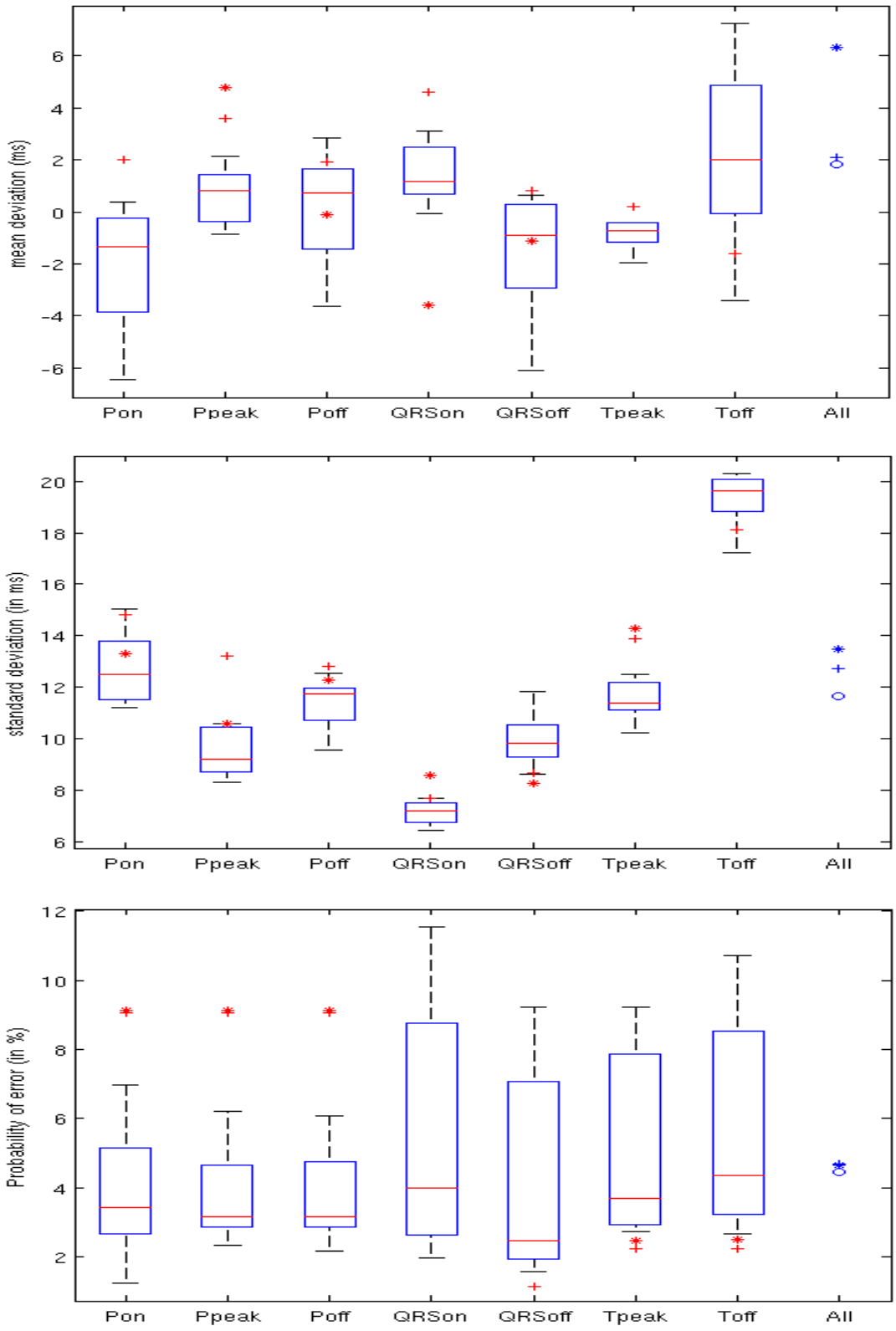


FIG. 3.12 – Boites à moustaches des résultats de segmentation associés aux trois critères évalués (moyenne et écart type du jitter et probabilité d'erreur de détection) et évalués sur chacun des ensembles de tests, pour tous les indicateurs extraits.

timisation, sur l'ensemble d'apprentissage correspondant.

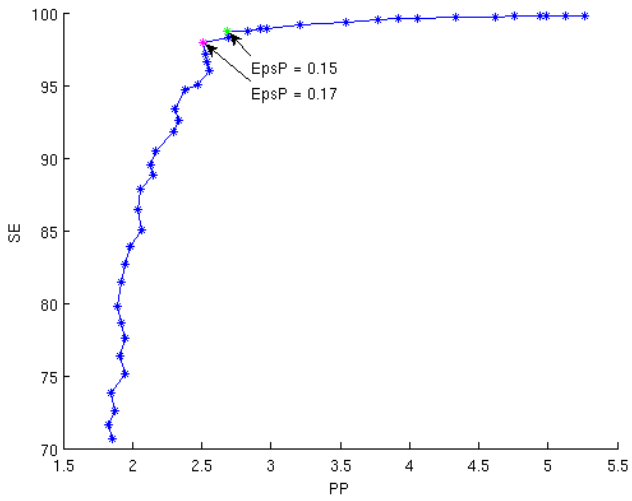


FIG. 3.13 – Courbe COR pour l'évaluation de la détection de l'onde P suivant le paramètre ϵ_P.

La figure 3.13 montre la courbe COR concernant la détection de l'onde P, qui varie suivant le paramètre ϵ_P (variation de 0 à 0.43 avec un pas de 0.01). Le point minimisant cette probabilité de détection est trouvé pour $\epsilon_P = 0.17$ avec la courbe COR tandis que l'algorithme d'optimisation a fixé ce paramètre à 0.15 lors de la procédure d'apprentissage.

D'autres indicateurs de la qualité de l'optimisation sont les scores globaux et scores partiels, calculés dans la fonction de coût, pour les 13 ensembles de tests générés. Ces scores sont présentés table 3.3. Ils correspondent aux ratios des trois critères (erreur moyenne, écart type, et probabilité d'erreur de détection) obtenus par la segmentation et ces mêmes critères issus de [Laguna et al., 1994] et [Martinez et al., 2004] (le meilleur des deux), comme dans la définition de la fonction de cout (équation 3.3.1.1). L'ensemble est sommé pour les 7 indicateurs p, ce qui produit des scores inférieurs à 7 lorsque des améliorations ont été obtenues et supérieurs à 7 sinon.

Pop.	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	$\overline{T_i} \pm \sigma(T_i)$
$\overline{\varepsilon}$	2.66	4.05	7.7	3.72	5.75	7.04	5.97	6.56	5.2	4.07	3.42	4.02	5.22	5.03 ± 1.52
σ	6.69	6.59	6.71	6.20	6.65	6.7	5.97	7.33	6.82	7.21	6.73	7.25	6.78	6.74 ± 0.38
$Perr$	2.88	5.95	12.69	5.26	14.14	6.92	5.49	6.32	8.81	13	2.98	4.5	10.92	7.68 ± 3.85
cost O	12.23	16.60	27.11	15.19	26.52	18.66	18.50	20.20	20.83	24.27	13.13	15.78	22.92	19.38 ± 4.83

TAB. 3.3 – Coûts obtenus pour les différentes populations de test

Plusieurs constatations peuvent être faites de ce tableau :

- une majorité des coûts partiels sont inférieurs à 7. Ceci indique que des améliorations sont obtenues. Le coût moyen calculé sur l'ensemble des populations de test est de 19.38 et révèle une amélioration globale de 7.7% sur les meilleurs résultats de [Martinez et al., 2004] et [Laguna et al., 1994].
- il n'y a pas un seul critère optimisé au détriment des autres. En effet, on observe qu'aucun

d'entre eux n'est toujours inférieur (ou supérieur) à 7 pour tous les ensembles de test. Ceci confirme que les améliorations sont générales et portent sur tous les critères.

– l'écart type calculé sur l'ensemble des populations de test est élevé.

Le dernier point doit être examiné plus en détail. Cet écart élevé peut provenir soit d'un mauvais apprentissage des paramètres pour des populations d'apprentissage spécifiques, soit de populations de test constituées d'enregistrements ECG qui sont difficiles à segmenter. Les tableaux 3.4 et 3.5 donnent des éléments de réponse sur ce sujet.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	$\overline{T_i} \pm \sigma(T_i)$
param. OP_{L1}	12.2	14.3	22.5	15.9	25.6	14.4	15.9	17.1	20.8	22.3	14.6	14.0	23.0	17.9 ± 4.35
param. OP_{L3}	12.6	17.5	27.1	15.9	26.9	18.2	15.9	20.8	20.0	26.2	16.8	14.1	24.3	19.7 ± 5.49

TAB. 3.4 – Coût obtenu en appliquant les meilleurs et pires paramètres optimisés à toutes les populations de test.

OP	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$	$L7$	$L8$	$L9$	$L10$	$L11$	$L12$	$L13$	$\overline{OP_{Li}} \pm \sigma(OP_{Li})$
pop. T1	12.2	14.3	12.6	12.4	14.3	13.1	15.4	19.3	14.2	13.6	11.5	11.7	13.2	13.69 ± 2.03
pop. T3	22.5	23.5	27.1	26.2	25.9	29.7	21.2	23.1	19.0	23.2	20.8	22.7	20.1	23.45 ± 3.04

TAB. 3.5 – Coûts obtenus en segmentant la meilleure et la pire population de test avec l'ensemble des paramètres optimisés.

Quand le meilleur et le pire des ensembles de paramètres (respectivement OP_{L1} et OP_{L3}) sont appliqués pour segmenter toutes les populations de tests, une forte variance est obtenue : respectivement 4.35 pour les meilleurs paramètres et 5.49 pour les pires (3.4). D'un autre coté, la meilleure (T1) et la pire (T3) population de test produisent une variance plus faible (1.98 et 1.85) lorsque les différents ensembles de paramètres sont appliqués pour les segmenter (3.5). A partir de ces deux tests, l'écart type élevé du tableau 3.3 peut être principalement expliqué par le fait que quelques enregistrements, ceux présentant des pathologies particulières, peuvent provoquer des coûts très élevés pour les ensembles de tests qui les contiennent (par exemple T3 et T5). La phase d'apprentissage a moins d'effet sur cette variance élevée. Cependant, sur l'ensemble des populations de test, il est possible de voir que les résultats ont été améliorés.

3.5 Intégration des algorithmes dans une station d'analyse d'ECG

Si les précédentes sections de ce chapitre ont visé à produire un algorithme de segmentation robuste et une procédure autonome pour le réglage des seuils, l'objectif de celle-ci est de montrer leur intégration dans une station directement exploitable par le clinicien afin de permettre l'extraction et la présentation cohérente d'un grand nombre d'indicateurs de l'ECG.

La figure 3.14 présente l'interface utilisateur de la station où il est par exemple possible de retrouver les résultats de la segmentation pour chacun des battements (coin supérieur droit), la mesure des indicateurs du battement courant (coin inférieur droit), ou encore les séries temporelles des indicateurs qui sont extraits sur la durée totale de l'ECG analysé (coin inférieur gauche).

Cette station est constituée de plusieurs modules, dont le principal effectue une segmentation complètement automatique de l'ensemble des battements avec la chaîne de traitement précédemment proposée. Les modules supplémentaires permettent l'exploitation de cette segmentation, avec plusieurs possibilités de corrections des erreurs de segmentation, ainsi que le calcul des indicateurs et leur affichage. L'ensemble de ces modules est représenté figure 3.15.

Ces deux problématiques, de correction des erreurs de segmentation et de calculs des indicateurs, qui sont particulièrement importantes dans le cadre d'exploitation clinique de l'outil d'analyse de signaux ECG, sont traitées dans les deux sous-sections suivantes.

3.5.1 Corrections des erreurs : segmentation semi-automatique par recalage dynamique

Les résultats de la segmentation automatique peuvent présenter quelques défauts, qui apparaissent par exemple dans les ECG d'effort, lorsque le rythme cardiaque devient élevé et que les ondes sont plus difficiles à délimiter. Cette déficience est résolue à deux niveaux dans la station :

- en cas de défauts ponctuels, sur quelques battements, des corrections manuelles peuvent être effectuées par les cardiologues pour combler les lacunes de la segmentation automatique ;
- en cas de défauts fréquents, une méthode de segmentation semi-automatique est utilisée. Elle requiert l'enregistrement de quelques battements correctement segmentés, de manière manuelle ou automatique, puis un algorithme basé sur un réalignement avec le DTW est appliqué pour segmenter le reste des battements.

3.5.1.1 Principe de la segmentation semi-automatique

La segmentation proposée est effectuée sur l'ensemble des battements, exceptés les battements de référence qui ont été manuellement segmentés et qui sont enregistrés à part dans un premier temps. La segmentation des autres battements se déroule ensuite en 4 étapes :

1. filtrage et dérivation du battement courant : le recalage par DTW est insensible aux décalages temporelles mais est sensible aux décalages d'amplitudes. Pour éviter des perturbations, les battements sont donc dérivés et filtrés. Par raison de simplicité le filtre employé correspond au deuxième niveau de détail de la décomposition en ondelette effectuée précédemment (3.2.5).

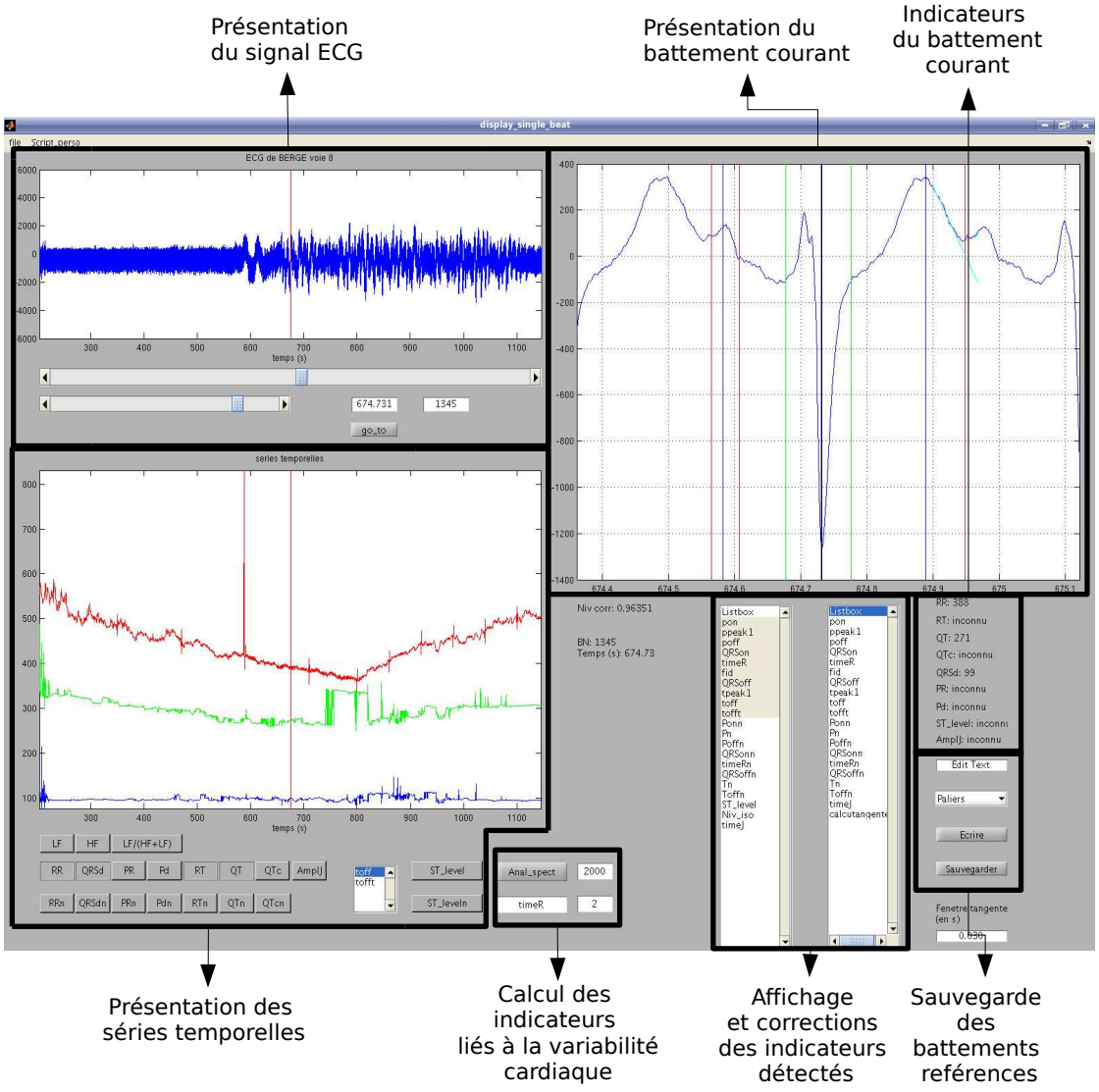


FIG. 3.14 – Interface utilisateur principale de la station d'analyse de signaux ECG.

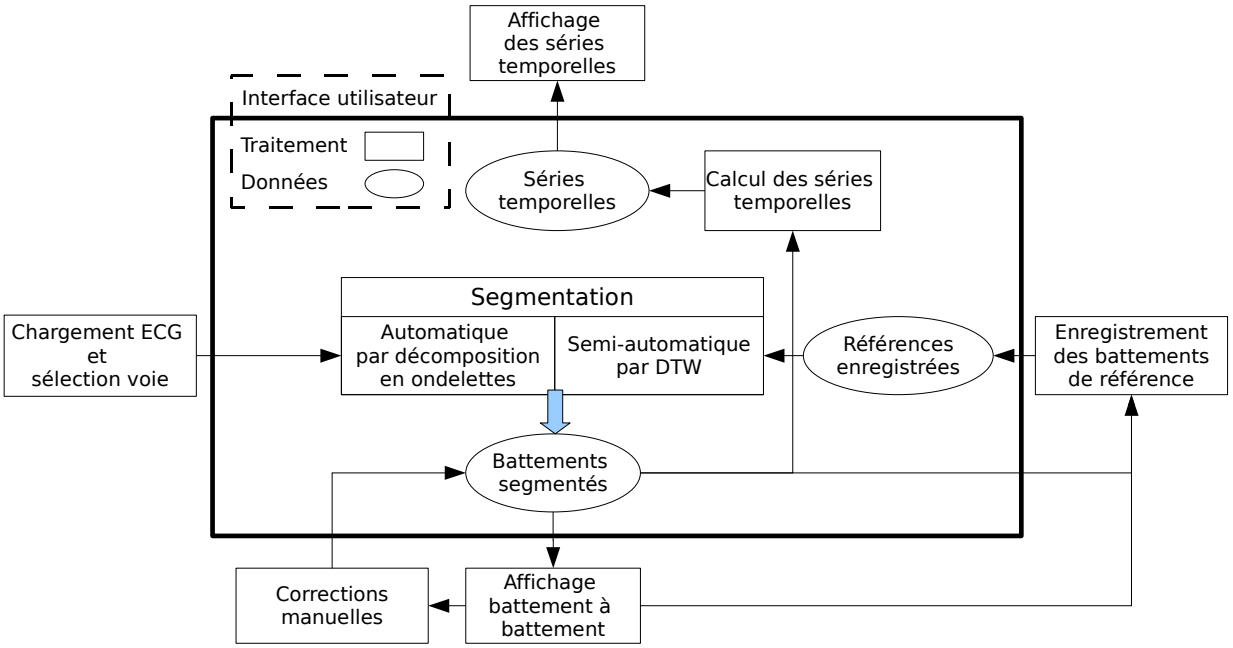


FIG. 3.15 – Représentation schématique des principaux modules de la station et de leur interactivité.

2. sous-échantillonnage du battement : pour optimiser les temps de calcul les battements sont éventuellement sous-échantillonés avant le recalage.
3. recalage du battement : le battement est à la fois recalé avec le battement de référence antérieur et le battement de référence postérieur. Celui dont la distance fournie par le DTW est la plus faible est conservé pour la segmentation.
4. segmentation du battement : la segmentation du battement de référence permet de segmenter le battement qui vient d'être recalé (figure 3.3).

Dans le cadre d'épreuves d'efforts, où la segmentation automatique est difficile, l'approche semi-automatique est idéale pour l'obtention de séries temporelles fiables : il suffit de sélectionner quelques battements bien segmentés, avec donc une interaction avec le clinicien minime, et ceci permet d'obtenir rapidement une segmentation sur l'ensemble de l'ECG avec un nombre d'erreurs plus faible. Des exemples concrets de segmentation permettent d'évaluer l'intérêt de cette méthode alternative, en complément de la segmentation automatique.

3.5.1.2 Exemples de segmentation sur des signaux réels

Segmentation d'intervalle QT La figure 3.16 présente l'extraction de l'intervalle temporel QT lors d'une épreuve d'effort. Sur la série temporelle, plusieurs pics peuvent être observés, avec en plus, à environ 680 secondes, une élévation anormale de l'intervalle QT mesurée sur plusieurs battements. L'analyse battement à battement permet d'identifier le problème : les ondes P et T se chevauchent lorsque l'effort est important, vers le milieu de l'enregistrement. Au niveau du battement **a** les ondes sont bien séparées par l'algorithme de segmentation automatique. Au niveau des battements **b** et **c**, l'onde P et l'onde T se confondent, ce qui trompe le détecteur concernant la position de la fin de l'onde T. Il est à rappeler ici que l'algorithme de segmentation a été optimisé avec un grand nombre d'enregistrements mais pas

avec des enregistrements d'efforts, où le rythme cardiaque est beaucoup plus élevé qu'au repos.

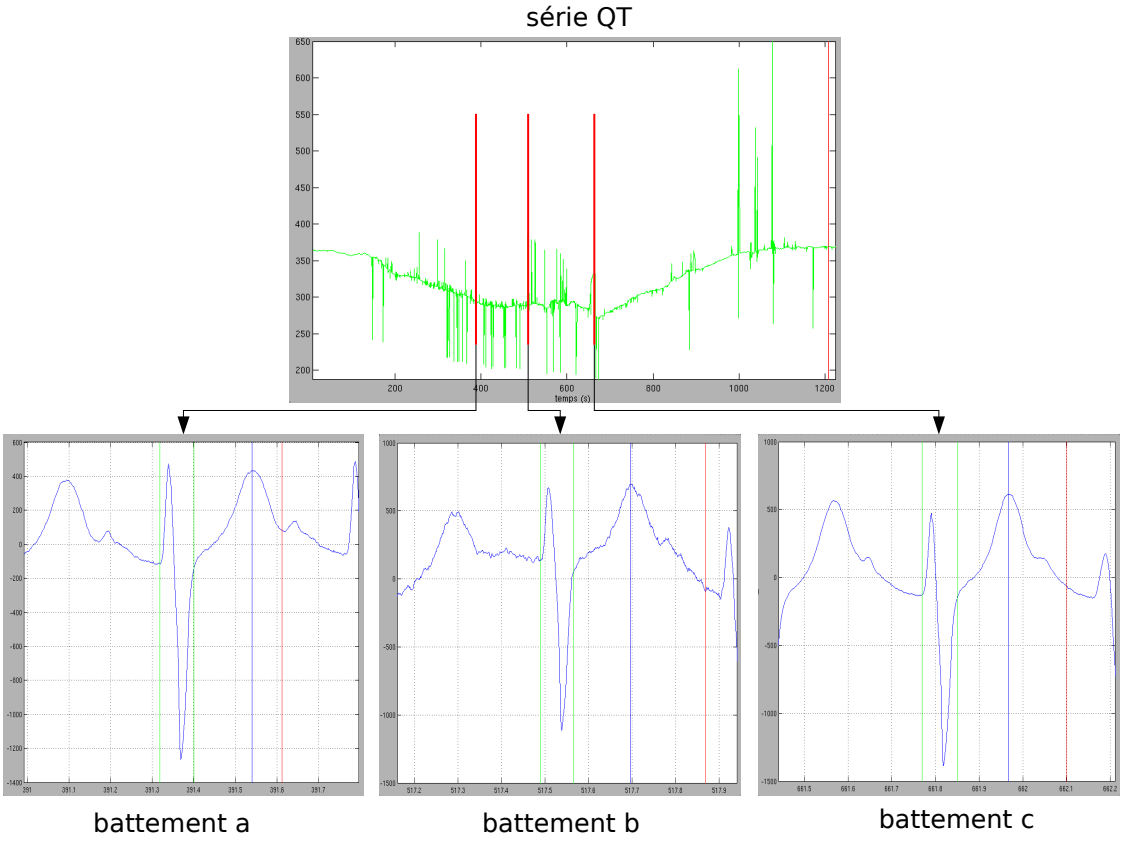


FIG. 3.16 – Résultats d'une segmentation automatique de l'onde T : série temporelle de l'intervalle QT obtenue et 3 battements extraits. Il est à noter que la série QT est bruitée et que les fins d'onde T sont mal positionnées pour les battements b et c.

Avec la segmentation semi-automatique, figure 3.17, l'onde T est mieux segmentée. La fin de l'onde T sur les battements **b** et **c** n'est plus confondue avec la fin de l'onde P. La série temporelle de l'intervalle QT ne présente plus de déviations aussi grandes qu'initialement, notamment sur la première moitié de la série temporelle. Cependant il est à relever que des déviations, inexistante initialement, sont aussi apparues. Elles correspondent à des petites variations dans le recalage (sensibilité au bruit) et aussi au fait que le battement courant est segmenté tantôt à partir du battement de référence précédent, tantôt à partir du battement de référence suivant, avec donc l'introduction d'une discontinuité lors du passage de l'un à l'autre. Ces déviations sont cependant d'amplitudes plus faibles que celles observées initialement.

Segmentation du complexe QRS De même sur les complexes QRS, des petites déflections entraînent des positions variables pour le début ou la fin du QRS. L'exemple figure 3.18 montre la segmentation d'un complexe QRS. La série temporelle $QRSd$ obtenue par segmentation automatique apparaît comme étant très bruitée, avec notamment des raccourcissements qui sont dus à une fin d'onde S peu marquée par rapport au début de l'onde T d'une part, et avec des déflections d'importance variable d'autre part. La segmentation automatique commet donc une erreur sur le battement **b**, alors que le battement **c** est bien segmenté. Ceci produit donc une série temporelle avec des discontinuités (**a**). Avec la segmentation semi-automatique

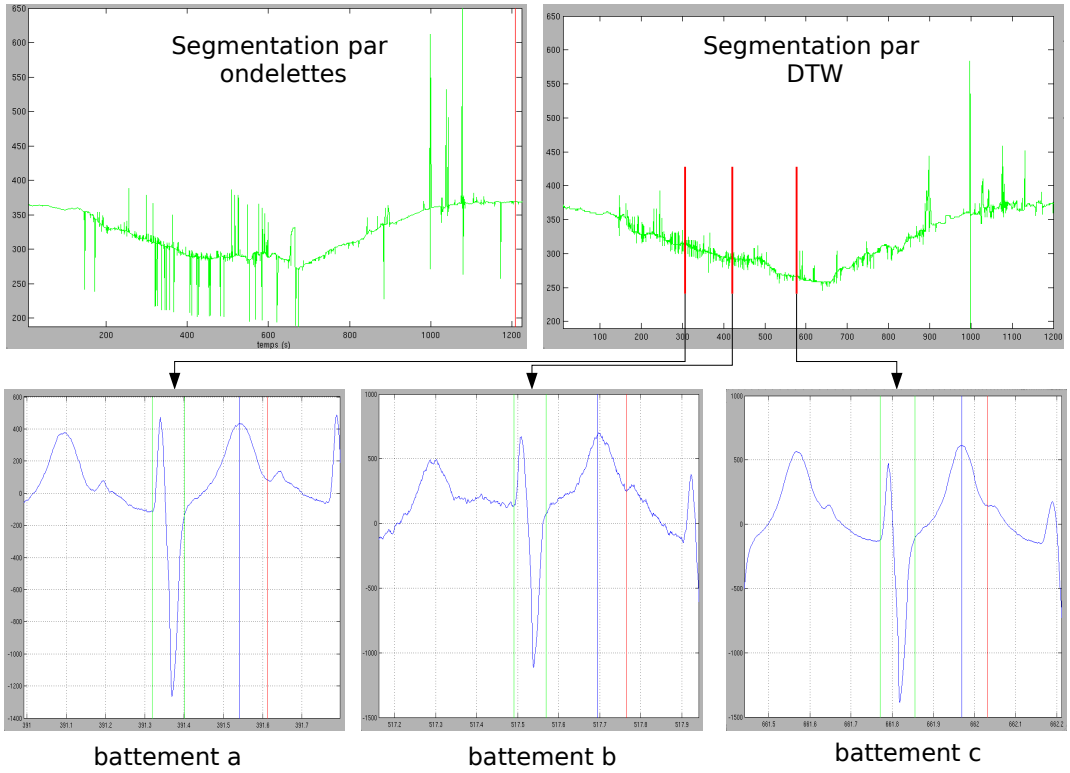


FIG. 3.17 – Résultats d'une segmentation semi-automatique de l'onde T : série temporelle de l'intervalle QT obtenue et 3 battements extraits. Il est à noter que la série QT est moins bruitée que précédemment (celle de droite par rapport à celle de gauche) et que les fins d'ondes T sont maintenant correctement positionnées sur les battements b et c.

les deux battements sont bien segmentés (**c** et **d**) et la série temporelle $QRSd$ présente moins de discontinuités.

3.5.2 Calcul des séries temporelles

Le module d'estimation des séries temporelles calcule certains indicateurs à partir des battements segmentés et génère les séries correspondantes sur la durée totale de l'enregistrement ECG. Les séries sont ensuite sauvegardées et affichées pour être soumises à l'analyse des experts.

Ce module permet d'extraire les indicateurs listés table 3.6.

La majorité des séries temporelles caractérisant les intervalles temporels et les amplitudes est aisément déduite de la segmentation proposée dans la section 3.2. Cette sous-section détaille donc uniquement les indicateurs nécessitant un traitement supplémentaire. Ces indicateurs sont ceux qui caractérisent :

- la variabilité de la fréquence cardiaque (VFC),
- le segment ST,
- l'intervalle QT corrigé en fonction de la fréquence cardiaque.

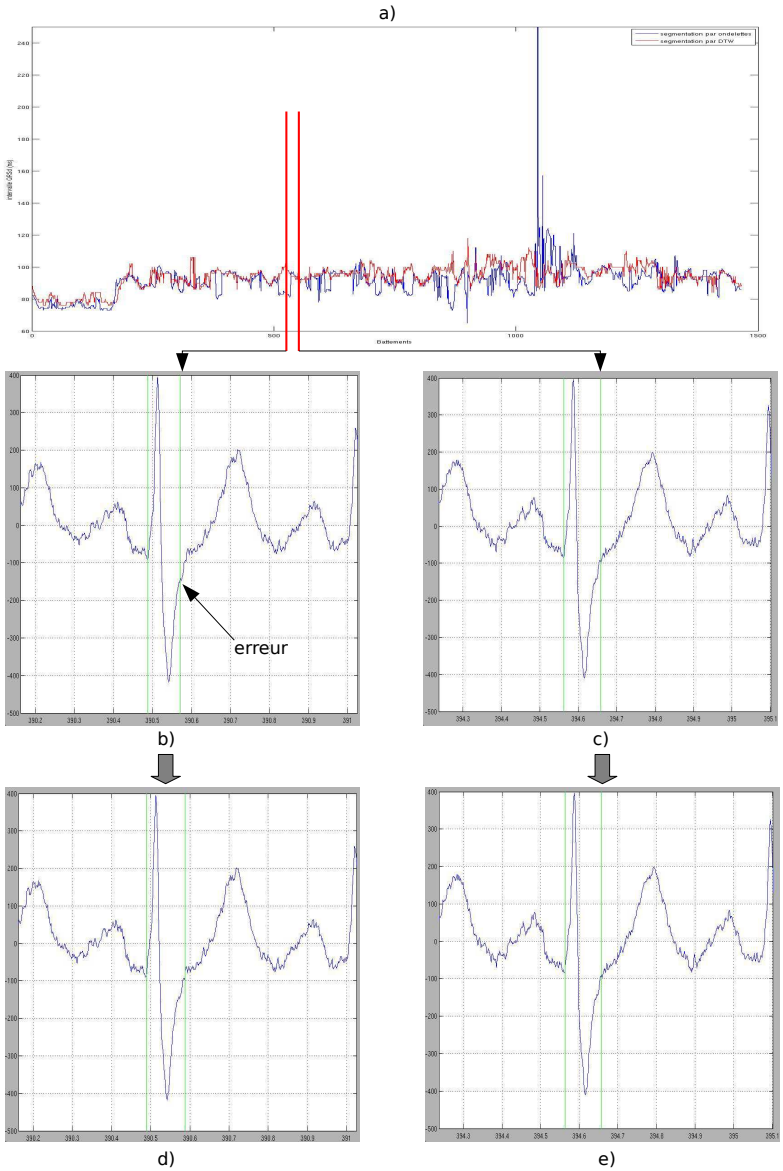


FIG. 3.18 – Segmentation automatique et semi-automatique des complexes QRS.

3.5.2.1 Indicateurs caractérisant la variabilité de la fréquence cardiaque

Les indicateurs caractérisant la variabilité du rythme cardiaque sont très souvent utilisés comme indicateurs de l'activité du système nerveux autonome. Différentes méthodes temporelles ou fréquentielles ont été appliquées à l'analyse du rythme cardiaque [TA-, 1996]. La méthode standard d'analyse spectrale fournit de l'information sur l'énergie en fonction de la fréquence. Deux bandes fréquentielles sont particulièrement utiles :

- la bande LF : comprise dans l'intervalle [0.04-0.15Hz], elle est considérée comme un marqueur de l'activité sympathique [Malliani et al., 1994, Montano et al., 1994] ou un marqueur mixte vagal et sympathique [Akselrod et al., 1981].
- la bande HF : comprise dans l'intervalle [0.15-0.50Hz], elle est due aux variations de fréquence cardiaque en réponse aux influence de la respiration sous le contrôle parasym-

intervalles temporels	RR	intervalle de temps entre deux battements : $R(i) - R(i-1)$
	$QRSd$	dépolarisation ventriculaire : $QRSoff(i) - QRSon(i)$
	PR	activité électrique auriculaire : $QRSon(i) - P(i)$
	Pd	dépolarisation auriculaire : $Poff(i) - Pon(i)$
	RT	$Toff(i) - R(i)$
	QT	activité électrique ventriculaire : $Toff(i) - QRSon(i)$
	QTc	intervalle QT corrigé
amplitudes	$niveau_ST$	amplitude mesurée sur le segment ST
	$ampl_P$	amplitude du pic de l'onde P
	ampl_R	amplitude du pic de l'onde R
	$ampl_T$	amplitude du pic de l'onde T
VFC	LF	Énergie dans la bande spectrale basse des fréquences de la VFC
	HF	Énergie dans la bande spectrale haute des fréquences de la VFC
	LF/HF	balance sympatho/vagale

TAB. 3.6 – Séries temporelles disponibles pour l'analyse. Le terme VFC désigne la Variabilité de la Fréquence Cardiaque.

pathique. La bande HF est reconnue comme un marqueur de l'activité vagale.

Les séries temporelles LF, HF, ainsi que le rapport LF/HF qui reflète la balance sympatho/vagal, sont calculés par une analyse spectrale de la série temporelle RR en fenêtre glissante : pour chaque position de la fenêtre, le spectre est estimé et l'énergie dans les bandes LF et HF est calculée. Ceci permet l'analyse de l'évolution temporelle de ces indicateurs. Le problème des méthodes d'analyse spectrale est qu'elles assument que le signal est stationnaire, ce qui n'est pas le cas du rythme cardiaque. Les non-stationnarités tels que des tendances linéaires ou plus complexes peuvent causer des distorsions dans l'analyse spectrale. Les origines de ces non-stationnarités sont discutées dans [Berntson et al., 1997].

Suppression des non-stationarités Deux approches pour supprimer les non-stationnarités se distinguent dans la littérature :

- soit les non-stationnarités sont détectés et uniquement les portions stationnaires sont conservées pour l'analyse du spectre comme dans [Weber et al., 1992],
- soit les non-stationnarités sont d'abord enlevées du signal, ce qui donne un signal filtré avant l'analyse spectrale. Les tendances non-stationnaires sont modélisés par des polynômes d'ordre 1 [Litvack et al., 1995] ou d'ordre supérieurs [Mitov, 1998].

Dans Grossman [Grossman, 1992], il est noté que les segments sélectionnés par l'approche de Weber [Weber et al., 1992] ne sont pas forcément très représentatifs de l'ensemble de l'enregistrement. Une suppression complète des non-stationnarités est donc préférable. Pour les approches de modélisation polynomiale, le choix de l'ordre du polynôme est problématique. Certaines portions du signal avec peu de non-stationnarités peuvent très bien être modélisées par un polynôme du premier ordre alors que d'autres portions nécessitent un ordre plus élevé. L'approche des *a priori* de lissage (smoothness priors), telle que présentée par Tarvainen [Tarvainen et al., 2002], est plus souple et avec moins de paramètres à ajuster.

Cette approche sépare le signal RR en deux composantes :

$$RR = RR_{stat} + RR_{tend} \tag{3.7}$$

où RR_{stat} est la composante stationnaire et RR_{tend} est la tendance non-stationnaire.

Cette tendance non-stationnaire est modélisée à partir d'un modèle linéaire :

$$RR_{tend} = \widehat{RR_{tend}} + \varepsilon = H\theta + \varepsilon. \tag{3.8}$$

où $\widehat{RR_{tend}}$ est la tendance estimée, ε est l'erreur résultante, H est la matrice d'observation et θ sont les paramètres à ajuster.

Ces paramètres sont ajustés par l'approche des *a priori* de lissage (prior smoothing). Il s'agit d'une minimisation des moindres carrés standard à laquelle une contrainte sur la continuité est rajoutée :

$$\hat{\theta_\lambda} = argmin_\theta\{\|H\theta - z\|^2 + \lambda^2\|D_d(H\theta)\|^2\} \tag{3.9}$$

où λ est le paramètre de lissage et D_d indique l'approximation discrète de la dérivation d'ordre d.

Les détails de l'implémentation sont donnés dans [Tarvainen et al., 2002]. Une dérivation du second ordre pour D_d est choisie, ceci favorise une solution localement linéaire pour la tendance estimée. Le seul paramètre à déterminer, λ est fixé en utilisant les abaques de [Tarvainen et al., 2002] : pour une fréquence d'échantillonnage du signal $RR(t)$ de 2Hz, un paramètre λ de 150 assure que la fréquence de coupure équivalente au filtre réalisé sera comprise entre 0.04 et 0.022Hz. La matrice H est choisie comme étant une matrice identité.

La figure 3.19 présente un exemple de suppression des non-stationnarités d'une courte série RR. En a) les séries temporelles sont présentées et en b) leur spectre, obtenu par une modélisation auto-regressive. Le spectre de la série initiale contient une forte composante basse-fréquences qui masque la bande LF. Par contre cette bande se distingue nettement sur le spectre de la série dont les non-stationnarités ont été enlevées.

Estimation de la densité spectrale L'estimation du spectre du signal $RR(t)$ désaisonalisé peut-être réalisée, soit par une transformée de Fourier standard, soit pas une analyse paramétrique. Une analyse paramètrique est choisie car elle permet de lisser le spectre, tout en minimisant l'erreur quadratique. Un modèle auto-régressif (AR), dont l'ordre est fixé à 16, à partir des recommandations de [Boardman et al., 2002], est donc ajusté par une méthode des moindres carrés et utilisé pour modéliser la série temporelle. Le spectre peut alors directement s'écrire en fonction des coefficients du modèle AR et de la variance de l'erreur de prédiction.

Calcul des séries LF et HF Pour chaque position de la fenêtre glissante, le signal RR est désaisonnalisé, le spectre est estimé et l'énergie dans les bandes LF et HF est simplement calculée par intégration sur les bandes de fréquences correspondantes. La figure 3.20 présente un exemple de calcul de ces séries.

3.5.2.2 Indicateurs caractérisant le segment ST

L'amplitude du segment ST est défini à l'aide d'une méthode inspirée de [Smrdel et Jager, 2004]. Il est calculé en partant de la position du point J et en y ajoutant un délai D, variant en fonction du rythme cardiaque.

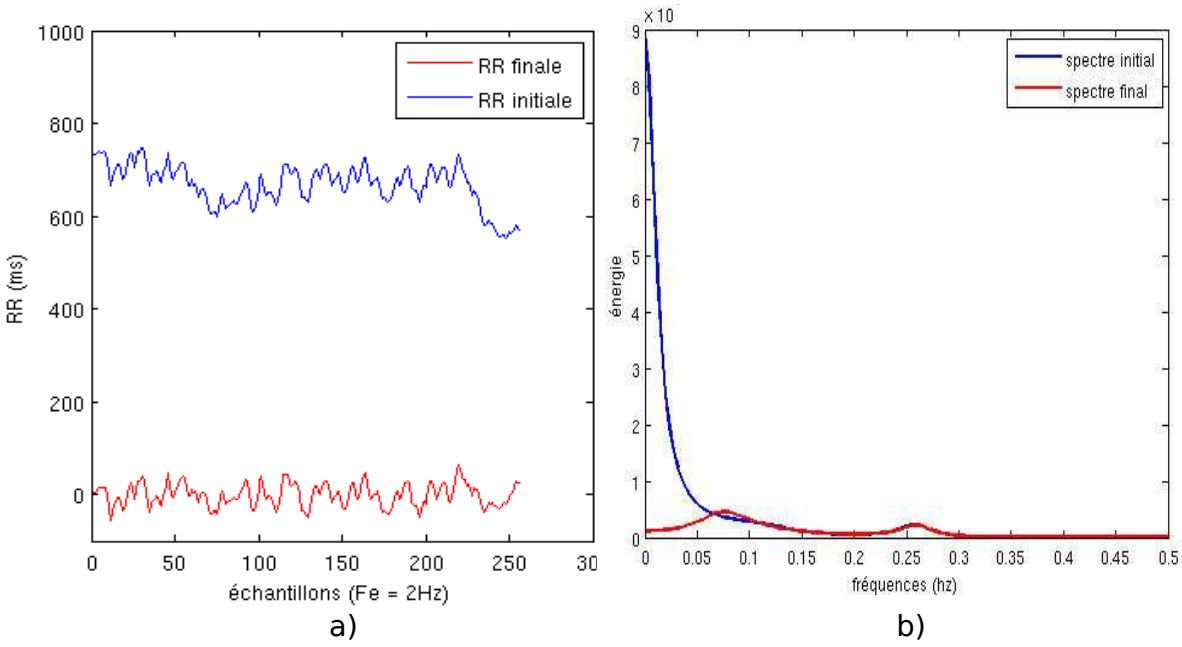


FIG. 3.19 – Suppression des non-stationnarités d'une série d'intervalles *RR*.

HR	100	110	120	130	140	150	160	170
D	80	72	64	60	56	52	48	44	40

TAB. 3.7 – Délai D, choisit en fonction du rythme cardiaque.

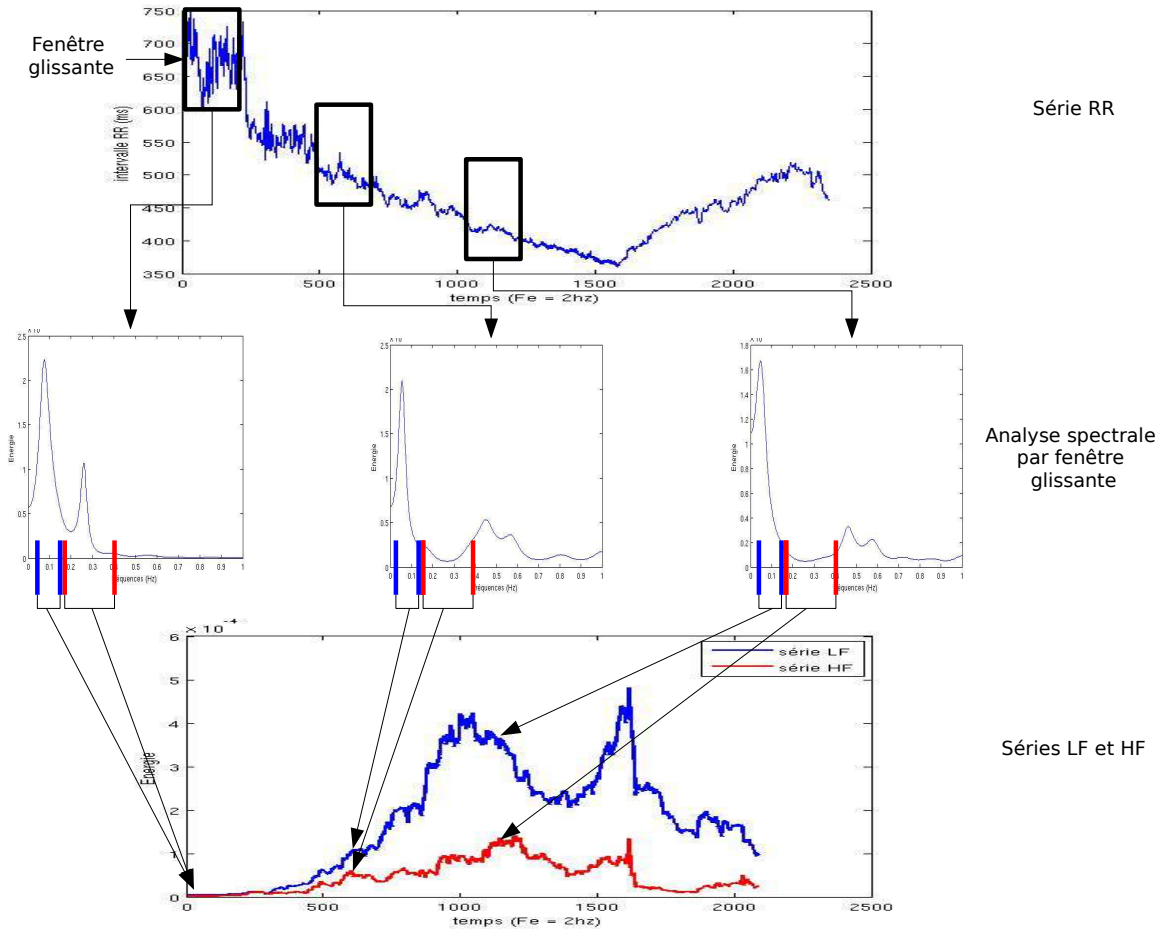


FIG. 3.20 – Extraction des séries LF et HF à partir de la série RR.

La pente du segment ST est calculée par interpolation linéaire sur un intervalle de temps borné par QRSoff + 30ms et QRSoff + 70 ms. La figure 3.21 représente le calcul de l'amplitude et de la pente du segment ST.

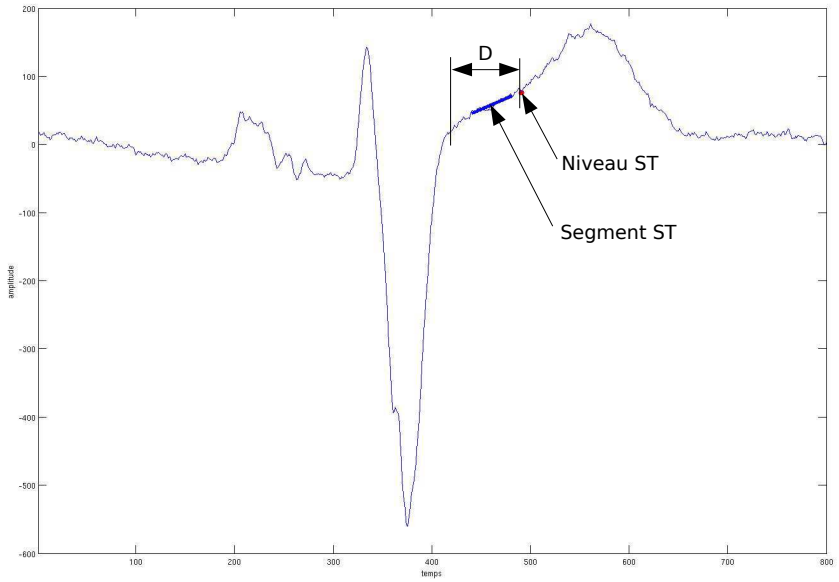


FIG. 3.21 – Extraction de l'amplitude et de la pente du segment ST à partir d'un battement ECG.

3.5.2.3 Indicateurs d'intervalle QT corrigé

L'indicateur QT, qui est un marqueur de la durée de repolarisation ventriculaire, est aussi très affecté de manière physiologique par le rythme cardiaque. Le but de la correction est de s'affranchir de cette source de variation pour mieux isoler les variations uniquement liées à des troubles de la repolarisation ventriculaire. De nombreuses méthodes ont été proposées pour résoudre ce problème (une bibliographie est présentée dans [Wong, 2004]). La correction de Bazett [Bazett, 1920] fait simplement intervenir la racine carrée de l'intervalle RR :

$$QTc = QT/\sqrt{RR} \tag{3.10}$$

Bien qu'imparfaite à cause d'un intervalle QTc trouvé anormalement long pour des intervalles RR faibles, cette correction demeure privilégiée [ICH :E14, 2004] et a donc été choisie.

3.6 Conclusion

Une chaine complète de traitement du signal, terminée par un algorithme de segmentation à base d'ondelettes issu de la littérature, est proposée afin d'extraire plusieurs indicateurs du signal ECG. Cette approche de décomposition en ondelettes est efficace car elle prend bien en compte les non-stationnarités présentent dans le signal. En revanche, la détection multi-échelle entraîne l'ajustement d'un grand nombre de seuils et de fenêtres temporelles d'analyse. Ces problèmes ont été résolus à l'aide d'un algorithme évolutionnaire, qui minimise une fonction de coût défini de manière à évaluer la performance de la segmentation. La validation de cette

approche sur une base de données montre des résultats intéressants comparés à d'autres méthodes, témoignant ainsi l'intérêt d'effectuer un ajustement rigoureux et global des paramètres. L'exploitation de la chaîne de traitement réalisée, et notamment l'extraction de caractéristiques du signal ECG, est concrétisée par la réalisation d'une station qui met en oeuvre les avantages de la segmentation automatique, tout en permettant de corriger les erreurs éventuelles par intervention manuelle de l'utilisateur.

Différentes extensions à la méthodologie présentée dans ce travail peuvent être envisagées. L'algorithme de segmentation traite les voies de manière indépendante, mais une version multi-voies peut être envisagée. Des travaux suivant cette orientation ont déjà été reportés dans la littérature mais au prix d'une complexité accrue et d'une efficacité parfois moindre, notamment en présence de bruit et d'artefacts sur certaines voies. Une extension possible concerne aussi l'exploitation de l'apprentissage par l'algorithme évolutionnaire : examiner plus précisément les individus présents dans la population finale peut aider à améliorer la procédure de segmentation. Par exemple, il est possible que les chromosomes montrent une tendance à se diviser en deux populations distinctes et à converger vers deux minimas locaux différenciés par un ou deux paramètres. L'ensemble de paramètres associé à un de ces minimum locaux peut produire une bonne segmentation pour une morphologie de battement donnée, l'autre ensemble étant plus approprié pour les morphologies restantes. Ainsi, en fonction des paramètres en question, des améliorations pourraient être réalisées sur l'algorithme de segmentation en rajoutant des règles heuristiques qui permettraient de prendre en compte l'ensemble des morphologies.

Bibliographie

[TA-, 1996] (1996). Heart rate variability : standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*, 93(5) :1043–1065.

[Akselrod et al., 1981] Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Berger, A. C., et Cohen, R. J. (1981). Power spectrum analysis of heart rate fluctuation : a quantitative probe of beat-to-beat cardiovascular control. *Science*, 213(4504) :220–222.

[Bahoura et al., 1997] Bahoura, M., Hassani, M., et Hubin, M. (1997). DSP implementation of wavelet transform for real time ecg wave forms detection and heart rate analysis. *Comput Methods Programs Biomed*, 52(1) :35–44.

[Bazett, 1920] Bazett, J. (1920). An analysis of time relations of electrocardiograms. *Heart*.

[Berntson et al., 1997] Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., et van der Molen, M. W. (1997). Heart rate variability : origins, methods, and interpretive caveats. *Psychophysiology*, 34(6) :623–648.

[Boardman et al., 2002] Boardman, A. ; Schlindwein, F. S., Rocha, A. P., et Leite, A. (2002). A study on the optimum order of autoregressive models for heart rate variability. *physiological measurement*.

[Bäck et Schütz, 1996] Bäck, T. et Schütz, M. (1996). Intelligent mutation rate control in canonical genetic algorithms. *Proc of the International Symposium on Methodologies for Intelligent Systems*, pages 158–167.

[Clavier et Boucher, 1996] Clavier, L. et Boucher, J. (1996). Segmentation of electrocardiograms using a hidden Markov model. In *Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE*, volume 4, pages 1409–1410vol.4.

[Coast, 1993] Coast, D. (1993). Segmentation of high-resolution ECGs using hidden Markov models. *Acoustic, Speech, and Signal Processing, ICASSP-93*, 1 :67–70.

[Coast, 1990] Coast, D.A. ; Stern, R. C. G. B. S. (1990). An approach to cardiac arrhythmia analysis using hidden markovmodels. *Biomedical Engineering, IEEE Transactions on*, 37(9) :826–836.

[CSE, 1985] CSE (1985). Recommandations for measurement standards in quantitative electrocardiography. *European Heart Journal : The CSE working Party*, 6 :815–825.

[Doerschuck, 1985] Doerschuck, P. (1985). *A markov chain approach to electrocardiogram modeling and analysis*. PhD thesis, MIT.

[Edenbrandt et Pahlm, 1988] Edenbrandt, L. et Pahlm, O. (1988). Vectorcardiogram synthetized from a 12-lead ecg : superiority of the inverse dover matrix. *J Electrocardiol*, 21 :361.

[Graja et Boucher, 2003] Graja, S. et Boucher, J.-M. (2003). Multiscale hidden Markov model applied to ECG segmentation. In *Intelligent Signal Processing, 2003 IEEE International Symposium on*, pages 105–109.

[Graja et Boucher, 2005] Graja, S. et Boucher, J.-M. (2005). Hidden Markov tree model applied to ECG delineation. In *IEEE Transactions on instrumentation and measurement*, number 6, pages 2163–2166.

[Grossman, 1992] Grossman, P. (1992). Breathing rhythms of the heart in a world of no steady state : a comment on Weber, Molenaar, and van der Molen. *Psychophysiology*, 29(1) :66–72 ; discussion 73–5.

[Hernãndez et al., 2002] Hernãndez, A. I., Carrault, G., Mora, F., et Bardou, A. (2002). Model-based interpretation of cardiac beats by evolutionary algorithms : signal and model interaction. *Artif Intell Med*, 26(3) :211–235.

[Holland, 1992] Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267-1 :66–72.

[Houck et al., 1995] Houck, C., Joines, J., et Kay, M. (1995). A Genetic Algorithm for function optimization : A Matlab implementation. *NCSU-IE TR 95-09.*

[Hughes et al., 2003] Hughes, N., Tarassenko, L., et Roberts, S. (2003). Markov models for automated ECG interval analysis. *NIPS*, 16.

[ICH :E14, 2004] ICH :E14 (2004). The clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. *CHMP/ICH/2/04, ICH Step 2.*

[Janikow et Michalewicz, 1991] Janikow, C. Z. et Michalewicz, Z. (1991). An experimental comparison of binary and floating point representations in genetic algorithms. pages 31–36.

[Jané et al., 1997] Jané, R., Blasi, A., Garcia, J., et Laguna, P. (1997). Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database. In *Computers in Cardiology 1997*, pages 295–298.

[Jané et al., 1992] Jané, R., Laguna, P., Thakor, N. V., et Caminal, P. (1992). Adaptive baseline wander removal in the ECG : comparative analysis with cubic spline technique. In *Computers in Cardiology 1992.*

[Kachenoura et al., 2007] Kachenoura, A., Poree, F., Hernandez, A., et Carrault, G. (2007). Surface ECG reconstruction from intracardiac EGM : a PCA-vectorcardiogarm method. *Signals, Systems and Computers, 2007 (ACSSC)*, pages 761–764.

[Koski, 1996] Koski, A. (1996). Modelling ECG signals with hidden Markov models. *Artif Intell Med*, 8(5) :453–471.

[Köhler et al., 2002] Köhler, B.-U., Hennig, C., et Orglmeister, R. (2002). The principles of software QRS detection. *IEEE Engineering in Medecine and Biology.*

[Laguna et al., 1994] Laguna, P., Jané, R., et Caminal, P. (1994). Automatic detection of wave boundaries in multilead ECG signals : validation with the CSE database. *Comput Biomed Res*, 27(1) :45–60.

[Lepage et al., 2001] Lepage, R., Boucher, J.-M., Blanc, J.-J., et Cornilly, J.-C. (2001). ECG segmentation and P-wave feature extraction : application to patients prone to atrial fibrillation. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 1, pages 298–301vol.1.

[Levkov et al., 2005] Levkov, C., Mihov, G., Ivanov, R., Daskalov, I., Ivaylo, C., et Dotsinsky, I. (2005). Removal of power-line interference from the ECG : a review of the subtraction procedure. *Biomed Eng Online*, 4 :50.

[Lhotska et al., 2003] Lhotska, L., Fejtova, M., Macek, J., et Novak, D. (2003). Biological data preprocessing : A case study. *In Intelligent and Adaptive Systems in Medicine, EUNITE Workshop.*

[Li et al., 1995] Li, C., Zheng, C., et Tai, C. (1995). Detection of ECG characteristic points using wavelet transforms. *IEEE Trans Biomed Eng*, 42(1) :21–28.

[Litvack et al., 1995] Litvack, D. A., Oberlander, T. F., Carney, L. H., et Saul, J. P. (1995). Time and frequency domain methods for heart rate variability analysis : a methodological comparison. *Psychophysiology*, 32(5) :492–504.

[Madeline, 2002] Madeline, B. (2002). New low cost and undedicated genetic operators. Research Report 4573, INRIA.

[Malliani et al., 1994] Malliani, A., Pagani, M., et Lombardi, F. (1994). Physiology and clinical implications of variability of cardiovascular parameters with focus on heart rate and blood pressure. *Am J Cardiol*, 73(10) :3C–9C.

[Martinez et al., 2004] Martinez, J. P., Almeida, R., Olmos, S., Rocha, A. P., et Laguna, P. (2004). A wavelet-based ECG delineator : evaluation on standard databases. *IEEE Trans Biomed Eng*, 51(4) :570–581.

[Michalewicz, 1996] Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Berlin and Heidelberg, 3rd edition.

[Mitov, 1998] Mitov, I. P. (1998). A method for assessment and processing of biomedical signals containing trend and periodic components. *Med Eng Phys*, 20(9) :660–668.

[Montano et al., 1994] Montano, N., Ruscone, T. G., Porta, A., Lombardi, F., Pagani, M., et Malliani, A. (1994). Power spectrum analysis of heart rate variability to assess the changes in sympathovagal balance during graded orthostatic tilt. *Circulation*, 90(4) :1826–1831.

[Pan et Tompkins, 1985] Pan, J. et Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*, 32(3) :230–236.

[Portet et al., 2005] Portet, F., andez, A. H., et Carrault, G. (2005). Evaluation of real-time QRS detection algorithms in variable contexts. *Med Biol Eng Comput*, 43(3) :379–385.

[R.M. Eichler West et Wilcox, 1998] R.M. Eichler West, E. D. S. et Wilcox, G. (1998). "Using Evolutionary Algorithms to search for control parameters in a nonlinear partial differential equation". Technical report, University of Minnesota Supercomputer Institute Research, Report UMSI 97/61.

[Sebag et al., 1997] Sebag, M., Schoenauer, M., et Ravise, C. (1997). Toward civilized evolution : Developing inhibitions. In Bäck, T., editor, *Proc. of the Seventh Int. Conf. on Genetic Algorithms*, pages 291–298, San Francisco, CA. Morgan Kaufmann.

[Shusterman et al., 2000] Shusterman, V., Shah, S. I., Beigel, A., et Anderson, K. P. (2000). Enhancing the precision of ECG baseline correction : selective filtering and removal of residual error. *Comput Biomed Res*, 33(2) :144–160.

[Smrdel et Jager, 2004] Smrdel, A. et Jager, F. (2004). Automated detection of transient ST-segment episodes in 24h electrocardiograms. *Med Biol Eng Comput*, 42(3) :303–311.

[Soria-Olivas et al., 1998] Soria-Olivas, E., Martínez-Sober, M., Calpe-Maravilla, J., Guerrero-Martínez, J. F., Chorro-Gascó, J., et Espí-López, J. (1998). Application of adaptive signal processing for determining the limits of P and T waves in an ECG. *IEEE Trans Biomed Eng*, 45(8) :1077–1080.

[Tarvainen et al., 2002] Tarvainen, M. P., Ranta-Aho, P. O., et Karjalainen, P. A. (2002). An advanced detrending method with application to HRV analysis. *IEEE Trans Biomed Eng*, 49(2) :172–175.

[Thakor et Zhu, 1991] Thakor, N. V. et Zhu, Y. S. (1991). Applications of adaptive filtering to ECG analysis : noise cancellation and arrhythmia detection. *IEEE Trans Biomed Eng*, 38(8) :785–794.

[Thierens, 2002] Thierens, D. (2002). Adaptive mutation rate control schemes in genetic algorithms. *Evolutionary Computation, Proceedings of the 2002 Congress on*, 1 :980–985.

[Thoraval, 1995] Thoraval, L. (1995). *Analyse statistique de signaux électrocardiographiques par modèles de markov cachés*. PhD thesis, Université de Rennes.

[Vasquez et al., 2001] Vasquez, C., Hernandez, A., Mora, F., Carrault, G., et Passariello, G. (2001). Atrial activity enhancement by wiener filtering using an artificial neural network. *Biomedical Engineering, IEEE Transactions on*, 4 :940–944.

[Vila et al., 2000] Vila, J. A., Gang, Y., Presedo, J. M. R., andez Delgado, M. F., Barro, S., et Malik, M. (2000). A new approach for TU complex characterization. *IEEE Trans Biomed Eng*, 47(6) :764–772.

[Vullings et al., 1998] Vullings, H., Verhaegen, M., et Verbruggen, H. (1998). Automated ECG segmentation with dynamic time warping. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, volume 1, pages 163–166.

[Weber et al., 1992] Weber, E. J., Molenaar, P. C., et van der Molen, M. W. (1992). A nonstationarity test for the spectral analysis of physiological time series with an application to respiratory sinus arrhythmia. *Psychophysiology*, 29(1) :55–65.

[Wong, 2004] Wong, S. (2004). *Segmentation de l'intervalle RT et description par analyse factorielle de la variabilité de la fréquence cardiaque et de la repolarisation ventriculaire.* PhD thesis, Université de Rennes 1.

[Wu et al., 2002] Wu, J., Wu, J., et Zipes, D. (2002). Early afterdepolarizations, u waves, and torsades de pointes. *Circulation*, 105(6) :675–676.

[Ziarani et Konrad, 2002] Ziarani, A. K. et Konrad, A. (2002). A nonlinear adaptive method of elimination of power line interference in ECG signals. *IEEE Trans Biomed Eng*, 49(6) :540–547.

[Zifan et al., 2006] Zifan, A., Saberi, S., Moradi, M. H., et Towhidkhah, F. (2006). Automated ECG segmentation using piecewise derivative dynamic time warping. *International Journal of Biomedical Sciences*, 3-1 :181–185.

Chapitre 4

Caractérisation et clustering de dynamiques

4.1 Introduction

Ce chapitre concerne l'exploitation des indicateurs dont l'extraction de l'ECG vient d'être détaillée. Il a été mentionné au cours des chapitres précédents qu'une analyse des dynamiques de ces indicateurs peut faciliter l'appréhension de certains problèmes cliniques rencontrés en cardiologie ; on pense en particulier à des pathologies fréquentes telles que l'ischémie ou les conséquences du diabète, ou des pathologies plus rares telles que le Brugada (cf chapitre 1).

La fouille de données désigne l'extraction de nouvelles informations non-triviales contenues dans de larges bases de données. Pour le cas particulier de données temporelles, les différentes tâches à accomplir sont généralement : (i) la prédiction, (ii) la classification, (iii) le clustering, (iv) la recherche par contenu, (v) la découverte de motifs, (vi) la visualisation et (vii) la détection d'anomalies, [Antunes et Oliviera, 2001, Mörchen, 2006, Laxman et Sastry, 2006]. Dans le domaine biomédical, l'aspect "larges bases de données" est souvent présent mais demeure facultatif. Par contre, les objectifs à atteindre sont aisément mis en relation avec les tâches caractéristiques de la fouille de données. Le paragraphe suivant résume cette relation étroite.

Lors d'une étude d'un ensemble de sujets, il est par exemple utile d'identifier les groupes naturels formés (**clustering**), pour isoler des comportements particuliers ou prévoir différents traitements suivant ces groupes. Ces groupes peuvent aussi révéler des tendances spécifiques (**recherche de motifs**) ou encore des **anomalies** présentes chez quelques individus, apportant ici des informations complémentaires pour mieux appréhender les pathologies étudiées. Pour évaluer ces tendances spécifiques ou ces anomalies et les présenter de manière compréhensible aux experts, des contraintes de **visualisation** doivent aussi être respectées. Ces contraintes de visualisation sont aussi présentes dans l'analyse des groupes de patients : il est important de visualiser la position des patients les uns par rapport aux autres pour évaluer leur proximité et analyser comment les groupes se distinguent. Les objectifs de prise de décision dans le cadre hospitalier (diagnostic, choix de traitements) correspondent naturellement à un problème de **classification**. Enfin, il est intéressant de pouvoir **prédire** l'évolution de l'état ou des caractéristiques d'un patient, en fonction des observations courantes ou du traitement qui va lui être appliqué.

Rien que pour le domaine du biomédical, les applications possibles de la fouille de données

temporelles sont tout aussi nombreuses que diverses. Cependant, son application concrète est finalement rare, comparée à la fouille de données statiques, pour des raisons de complexité : la dimension des données est beaucoup plus grande, les relations temporelles sont souvent complexes et donc difficiles à exprimer de manière simple. Ceci se traduit par un compromis récurrent dans la littérature : d'un coté l'importance d'exploiter de grandes quantités de données, de modéliser des relations complexes (les relations temporelles, mais aussi l'interaction entre les différentes variables) et d'un autre coté le besoin de simplicité inhérent aux tâches de fouille de données.

En conséquence, les approches proposées sont systématiquement dépendantes des données à traiter, du système dont elles sont observées et des objectifs à atteindre. Dans le deuxième chapitre, nous avons relevé l'importance de la prise en compte des dynamiques en cardiologie, pour comprendre le fonctionnement de certaines pathologies et permettre un diagnostic. Le problème maintenant posé concerne l'analyse de ces dynamiques, issues de séries temporelles à valeurs réelles et multivariées, et leur utilisation dans un cadre de fouille de données.

Dans la littérature, la caractérisation des dynamiques est usuellement décomposé en deux problèmes distincts : la définition de similarité et la représentation des séries temporelles. La revue de Mörchen [Mörchen, 2006] présente une classification des nombreuses méthodes appliquées à la résolution de ces deux problèmes. Plusieurs points clés montrent l'intérêt des approches basées modèles, comparées aux autres méthodes, pour effectuer une analyse des dynamiques dans le cadre d'une fouille de données :

- les mesures de similarités basées "formes" (distance Euclidienne, Dynamic Time Warping (DTW) [Berndt et Clifford, 1994], Longest Common Subsequence [Das et al., 1997]) nécessitent peu de connaissances *a priori* et sont bien adaptées pour les problèmes de clustering. Par contre la résolution des problèmes de classification ou de prédiction n'est pas directe, de même pour la prise en compte de données multivariées.
- les mesures de similarités basées "caractéristiques" nécessitent plus de connaissances *a priori* des données : il faut connaître les caractéristiques importantes. Ceci explique que les étapes d'extraction de caractéristiques (ou de transformation dans un autre espace) soient, comme le souligne Mörchen [Mörchen, 2006], très dépendantes du domaine d'application et des données à traiter. Ceci est d'autant plus vrai que chaque caractéristique présente différents inconvénients : par exemple les approches fréquentielles supposent la stationnarité des signaux tandis que les approches temps-fréquences sont difficiles à employer dans le cas de signaux de différentes tailles. Enfin, la définition d'une distance dans l'espace des caractéristiques n'est pas évidente, qui plus est dans le cas de données multivariées.
- les approches basées modèles existantes permettent, pour la plupart, la prise en compte de séries temporelles multidimensionnelles et de différentes tailles. L'introduction de connaissances *a priori* peut éventuellement servir à raffiner le modèle suivant l'application. Les modèles permettent aussi de mesurer l'adéquation d'un individu au modèle (tâche de classification) ou de réaliser des prédictions, par simulation.

Cette comparaison succincte explique pourquoi la modélisation de dynamiques de signaux est un vaste domaine de recherche, qui s'est développé considérablement depuis les années 80. Une littérature ciblée sur la modélisation des dynamiques est donc présentée dans la section suivante. Dans la section 3, une approche basée sur les modèles semi-Markoviens est proposée et une comparaison avec d'autres approches de caractérisation de dynamiques est effectuée sur différentes tâches de fouilles de données. La section 4 s'attache à la résolution du clustering. Un

nouvel algorithme de type "clustering descendant", basé sur les modèles semi-markoviens, est proposé. Ce dernier est évalué dans la section 5 sur différents ensembles de données simulées.

4.2 Etat de l'art des modèles pour la caractérisation de dynamiques multivariées

Cet état de l'art se focalise sur quatre classes de modèles adaptés aux séries temporelles multivariées. Ces quatre modèles sont : (i) les modèles dans les espaces des phases reconstruits ; (ii) les filtres de Kalman ; (iii) les réseaux de neurones pour données dynamiques ; (iv) les modèles de Markov cachés. Une cinquième sous-section discute aussi des systèmes combinant à la fois les réseaux de neurones et les modèles de Markov de cachés.

4.2.1 Modélisation de trajectoires dans les espaces des phases reconstruits

Les espaces des phases sont définis comme un espace de représentation utilisé principalement en mathématiques et en physique pour démontrer et visualiser les changements dans les variables dynamiques d'un système. Chacune de ces variables est représentée par un axe de l'espace des phases et, pour chaque état possible du système, un point est alors tracé dans cet espace. La succession de ces points représente l'évolution dynamique du système et la forme obtenue permet d'exhiber des propriétés du système qui ne sont pas visibles autrement. De plus, la trajectoire formée dans cet espace est constante dans le temps ; il est ainsi possible de calculer l'état du système dans le futur (ou dans le passé) à l'aide de la fonction de transfert f :

$$X(t+1) = f(X(t))$$

où $X(t)$ désigne l'état du système à l'instant t. Enfin, comme chaque point de l'espace des phases n'a qu'un successeur temporel, deux trajectoires différentes ne peuvent s'intersecter.

Lorsque ce sont les variables observées et non les variables d'états qui sont analysées, un Espace des Phases Reconstruit (EPR, appelé aussi pseudo-espace des phases) est alors créé. Dans ce cas, les différents axes intègrent des versions retardées de la même variable :

$$[x(t-\tau), x(t-2\tau), ..., x(t-M\tau)]$$

En théorie, le théorème de Takens [Takens, 1980] spécifie qu'en choisissant correctement ces variables retardées et dans le cas d'observations non-bruités de systèmes déterministes, les trajectoires intégrées à l'EPR ont les mêmes propriétés, d'un point de vue topologique, que celles des véritables attracteurs représentés dans l'espace des phases. La caractérisation de la dynamique est alors effectuée en deux étapes :

1. intégration de la série dans un EPR particulier.
2. modélisation de la trajectoire parcourue dans cet EPR.

Intégration de la série dans l'EPR : La figure 4.1 illustre l'intégration d'un attracteur de Lorenz dans un EPR de dimension 3 réalisé à partir de la variable x retardée. La création de l'EPR nécessite notamment la détermination du délai τ et de la dimension M de l'espace. Dans le cas théorique de séries non-bruitées et de durées infinies, Takens a montré qu'une dimension supérieure à deux fois la dimension fractale assure une trajectoire dans le pseudo espace des phases avec une topologie équivalente à celle de l'espace des phases initial, et

ce quel que soit le délai τ. Hors de ce cadre théorique, Buzug [Buzug et Pfister, 1992] ou Celluci [Cellucci et al., 2003] révèlent l'importance du choix des paramètres. Par exemple, un délai trop faible va donner des points extrêmement corrélés entre les différents axes et les corrélations réellement importantes ne seront pas visibles, tandis qu'un délai trop grand donnera des points qui sembleront distribués aléatoirement dans l'EPR. Pour la dimension M, une sur-évaluation complexifiera de manière inutile l'exploitation de la trajectoire tandis qu'une sous-évaluation créera une trajectoire repliée sur elle même. Une position à un instant t ne donnera pas forcément une position unique à l'instant t+1. Une minimisation de l'information mutuelle est fréquemment employée pour déterminer le délai et donne des résultats satisfaisants [Fraser, 1989, Radoji et al., 2002, Kraskov, 2004] tandis que l'algorithme des faux plus proches voisins [Kennel et al., 1992] détermine de manière itérative la dimension suffisante pour éviter les repliements non-significatifs.

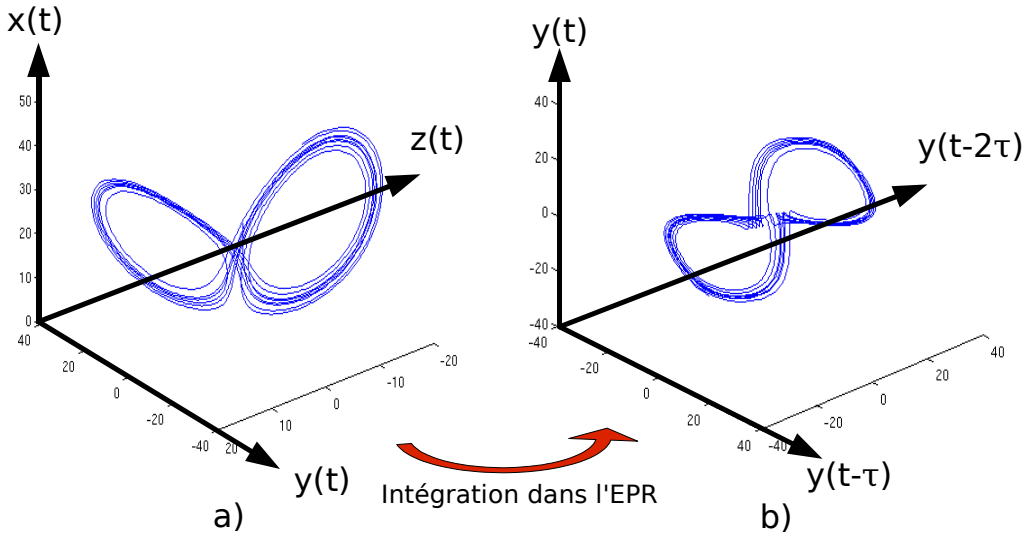


FIG. 4.1 – Intégration des variables d'états du système de Lorenz dans un EPR.

Modélisation de la trajectoire parcourue dans le pseudo-espace des phases :

Cette trajectoire est représentée par la fonction f de $\mathbb{R}^M \rightarrow \mathbb{R}$:

$$x(t+1) = f(X)$$

où X est le vecteur de retard de dimension M : $X = [x(t)x(t-\tau)...x(t-(M-1)\tau)]$.

Parmi la multitude d'approches de modélisation, deux groupes sont distingués : les modélisations locales et les modélisations globales.

Les approches locales partitionnent l'EPR en sections dans lesquelles les trajectoires sont généralement décrites par des relations linéaires. L'approche locale la plus courante est celle des plus proches voisins [Farmer et Sidorowich, 1987, Liu et al., 1998]. Elle est limitée par le besoin de faire un partitionnement adéquat, nécessitant d'évaluer les proximités entre chaque point mais aussi par le fait que le modèle créé aura un nombre de paramètres extrêmement large, avec un sous-ensemble de paramètres par section constituée. De plus, la connexion entre

deux partitions introduit des discontinuités, ce qui, pour certaines applications pose problème [Casdagli, 1989].

Les modélisations globales permettent une représentation compacte de l'ensemble de la trajectoire, dans l'ensemble de l'EPR, avec généralement un ajustement des paramètres par régression. Les modèles polynomiaux (simples ou rationnels) ont été très largement appliqués et des résultats intéressants ont été reportés [Abarbanel et al., 1994, Bezruchko et al., 2001]. Les approches par réseaux de neurones sont également employées dans [Weigend et al., 1990]. L'approche de modélisation par noyaux gaussiens (`RBF`) est aussi utilisée et présente une technique d'interpolation globale avec de bonnes propriétés de localisation, utile pour des données bruitées. Par contre, les performances dépendent fortement de la position des centres. Par exemple, Povinelly [Povinelly, 2005] utilise des mixtures de gaussiennes pour modéliser les portions de l'espace de phases parcourues par les séries temporelles.

Les méthodes d'analyse de dynamiques par les modélisations dans les EPR sont plus pertinentes que les modèles auto-régressifs : i) les points voisins analysés pour faire la prédiction sont voisins dans le sens des dynamiques et non pas dans le sens temporel, ii) les modélisations des trajectoires dans l'EPR permettent d'exhiber des non-linéarités, ce qui n'est autrement fait qu'avec les modèles NARMAX, iii) les modélisations probabilistes telles que les mixtures de gaussiennes permettent la définition d'une distance modèle-série probabiliste, au lieu d'une simple distance euclidienne sur l'erreur de prédiction.

Aspect multivariables : La majorité des applications utilisant les espaces de phases reconstruits sont univariables mais il existe quelques applications en multivariable. Par exemple, dans [Suzuki et al., 2003] plusieurs types d'attracteurs sont construits en utilisant une, deux ou trois des variables observées. Cependant, afin de limiter la complexité, les dimensions et les délais sont fixés à une constante donnée. Le modèle est local, similaire à celui des plus proches voisins et deux indicateurs évaluant la performance de la prédiction sont analysés en fonction de l'attracteur construit. Les résultats de [Suzuki et al., 2003] montrent que les attracteurs multivariés sont plus performants que les attracteurs univariés et leur utilisation se retrouve par exemple dans [Garcia et Almeida, 2005] et [Wan et Han, 2007].

4.2.2 Filtre de Kalman

Le filtre de Kalman [Kalman, 1960] est un estimateur récursif dont le principe de fonctionnement est représenté figure 4.2. Le filtre permet dans une première étape d'estimer l'état futur (à t+1) à partir de l'état courant (à t). Une fois l'observation à t+1 relevée, l'état correspondant est corrigée. L'historique des observations et des estimations n'est ainsi pas requis, ce qui correspond donc à la propriété markovienne. Dans le cas d'un système sans commande, le modèle linéaire, à états continus et sur lequel est appliqué le filtre de Kalman, s'écrit :

$$\begin{cases} x_t = M_t x_{t-1} + w_t & \text{(équation d'état)} \\ y_t = \phi x_t + v_t & \text{(équation de mesure)} \end{cases} \tag{4.1}$$

où x_t et y_t sont respectivement le vecteur d'états et le vecteur d'observations, M_t est la matrice de transitions d'états, ϕ est la transformation entre les états et les observations, v_t est le vecteur de signaux aléatoires qui polluent les mesures et w_t est un vecteur de signaux aléatoires qui vient perturber l'équation d'état du système.

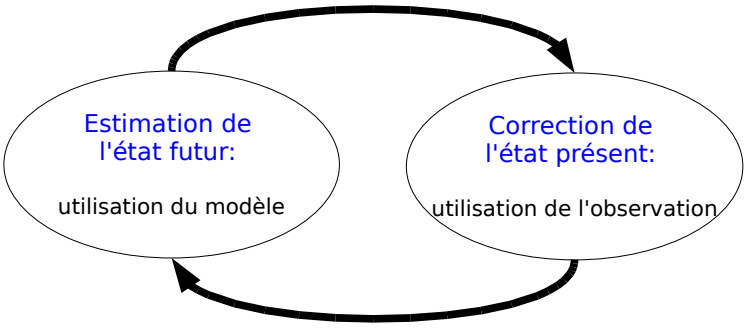


FIG. 4.2 – Principe du filtre de Kalman.

Les filtres de Kalman sont principalement utilisés pour effectuer du lissage (détermination des états x(t)) ou de la prédiction [Shumway et Stoffer, 1982]. Cependant, d'un point de vu plus générique, ce filtre est adapté pour la modélisation des dynamiques des états d'un système (équation d'état) à partir des observations effectuées (équation de mesure), puisqu'il va permettre d'estimer les états au cours du temps et avec la motivation d'intégrer aussi l'incertitude du modèle (w_t) et des mesures (v_t).

Pour utiliser le filtre de Kalman, la covariance du bruit de mesure (R) et la covariance du bruit du système (Q) doivent être évaluées à l'avance. La détermination de la covariance du bruit du système est généralement plus difficile parce que l'observation directe de l'état du système est impossible.

Dans le cadre du modèle linéaire, l'identification des paramètres du modèle (la matrice de transition M_t, la matrice d'observation ϕ, ainsi que les covariances des bruits w_t et v_t) est basée sur la maximisation de la vraisemblance des observations, qui est donnée par :

$$L = \begin{cases} -1/2log|\Sigma| - 1/2(x_0 - \mu)'\Sigma^{-1}(x_0 - \mu) \\ -n/2log|Q| - 1/2\sum_{t=1}^{n}(x_t - \phi x_{t-1})'Q^{-1}(x_t - \phi x_{t-1}) \\ -n/2log|R| - 1/2\sum_{t=1}^{n}(y_t - M_t x_{t-1})'R^{-1}(y_t - M_t x_t) \end{cases} \tag{4.2}$$

Comme dans [Shumway et Stoffer, 1982] et [Ghahramani et Hinton, 1996], cette vraisemblance est fréquemment optimisée par l'algorithme d'Expectation-Maximisation (EM) de Dempster [Dempster et al., 1977], qui permet d'estimer les variables cachées dans l'étape E (l'étape de lissage du filtre de Kalman) et d'adapter les paramètres du modèle pour maximiser la vraisemblance dans l'étape M. L'étape E est réalisée par une méthode forward-backward qui calcule l'état x_t à l'aide des observations précédentes y_1 à y_t dans la phase forward et à l'aide des observations y_{t+1} à y_T dans la phase backward. L'étape M est constituée de l'ensemble des équations de mise à jour des paramètres et appliquées avec les états estimés dans l'étape E.

L'extension du filtre de Kalman simple à des équations de dynamiques et d'observations non-linéaires est appelée le filtre de Kalman étendu :

$$\begin{cases} x_t = f(x_{t-1}) + w_t & \text{(équation d'état)} \\ y_t = g(x_t) + v_t & \text{(équation de mesure)} \end{cases} \tag{4.3}$$

Ces équations peuvent être utilisées pour l'étape de prédiction de l'état mais pas pour

la prédiction de la covariance de l'erreur. Lorsque les fonctions f et g sont différentiables et que le bruit est gaussien, une approche consiste alors à linéariser localement le système pour prédire cette erreur. La fonction réalisée approxime alors le système dynamique stationnaire non-linéaire par un système non-stationnaire linéaire. A chaque point du système d'état, les dérivées des fonctions f et g sont linéarisées et représentées par les matrices jacobiennes $A_{\tilde{x}}$ et $C_{\tilde{x}}$:

$$A_{\tilde{x}} \equiv \frac{\partial f}{\partial x}|_{x=\tilde{x}} \text{ et } C_{\tilde{x}} \equiv \frac{\partial g}{\partial x}|_{x=\tilde{x}}$$

qui permettent d'approximer la covariance des états. Une modification de l'algorithme initial d'apprentissage EM pour le filtre de Kalman étendu est proposée par Roweis [Roweis, 2000]. Cependant en cas d'une estimation erronée de l'état initial ou bien d'une fonction de transition fortement non-linéaire, le filtre peut diverger rapidement. De plus il est montré que la matrice de covariance a tendance à sous-estimer la covariance réelle. Pour compenser ces défauts une alternative est proposée avec le filtre de Kalman dit sans odeur [Julier et Uhlmann, 1997]. Ce filtre, au lieu de réactualiser la matrice de covariance à l'aide d'une approximation linéaire des fonctions f et g, estime la nouvelle distribution des états en prenant en compte un certain nombre de points autour de la moyenne estimée et en leur appliquant la relation non-linéaire. La matrice de covariance complète est alors reconstruite en prenant en compte les quelques points tirés.

Le filtre de Kalman a de nombreuses applications, en suivi de formes sur des séquences vidéos : en imagerie générique [Kim et Woods, 1998] ou plus spécifiquement en suivi de profil [Jebara et Pentland, 1997], en économie [Wells, 1995], en navigation et localisation (de GPS [Ramjattan et Cross, 1995], de téléphones cellulaires [Zainab et Mark, 2005]), ou encore dans le domaine de la vision assistée par ordinateur, par exemple en détection de profondeurs à partir de séquences d'images [Matthies et al., 1989, Hung et Ho, 1999].

4.2.3 Réseaux de neurones pour données temporelles

Les réseaux de neurones ont connu un essor très important pour de nombreuses tâches de décision ou de classification à partir de données structurées durant les deux dernières décennies. Le perceptron simple couche de Rosenblatt [Rosenblatt, 1958] (figure 4.3 a) représentait un premier système artificiel, capable d'apprendre par expérience en respectant plusieurs analogies avec le neurone biologique. Les entrées, assimilées aux synapses, sont pondérées par des poids et sommées, puis une fonction d'activation non-linéaire est appliquée pour générer une activité O (initialement binaire). Cependant c'est principalement le Perceptron Multi-Couche (PMC) de Rumelhart [Rumelhart et al., 1986] qui a marqué une grande avancée vers la modélisation de systèmes non-linéaires complexes.

Les recherches portant sur la prise en compte d'information temporelle avec les réseaux de neurones ont débuté principalement dans le domaine du traitement de la parole. Les travaux de Waibel [Waibel, 1989] ont donné naissance au réseau de neurones temps-retard (Time-Delayed Neural Networks, TDNN). Ce type de réseau de neurones intègre le temps comme une dimension supplémentaire dans chaque couche du PMC. La figure 4.4 schématise l'architecture d'un TDNN. La première couche effectue une mémorisation des échantillons passés du signal tandis que les couches cachées les transforment en vecteurs de caractéristiques. Chaque neurone d'une couche cachée modélise une caractéristique locale de la variation de la courbe étant donné que leur champ de vision est restreint à une fenêtre temporelle limitée : il n'y a donc plus de connectivité totale entre l'ensemble des neurones. De plus, pour que toutes les caractéristiques

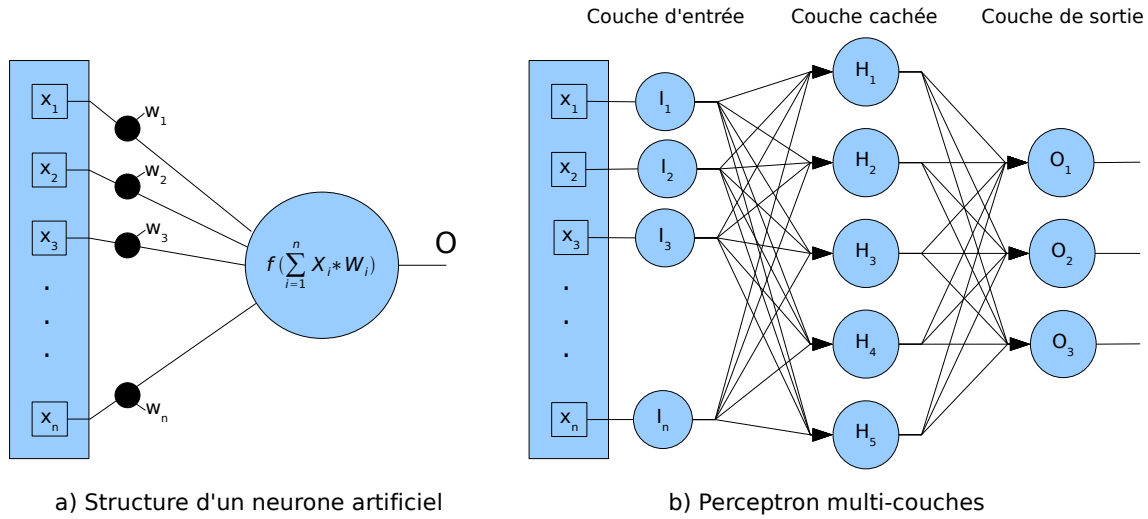


FIG. 4.3 – Neurone artificiel et perceptron multi-couches.

locales soient enregistrées dans les couches supérieures, la dimension caractéristique est augmentée alors que la dimension temporelle est réduite.

Les TDNN ont été utilisés en traitement de la parole, d'abord pour la reconnaissance de phonèmes [Waibel, 1989], puis pour la reconnaissance de mots à partir de sons ou d'images [Lavagetto, 1995], pour l'identification de locuteurs [Bennani et Gallinari, 1991], ou par exemple en biomédical pour la compression du signal ECG [Chatterjee et al., 2005]. De par leur structure, ils peuvent être comparés à des modèles auto-régressifs mais qui modélisent une fonction de prédiction non-linéaire : les neurones d'entrées sont des échantillons temporels retardés et le neurone de sortie peut éventuellement être un échantillon prédit. Les TDNN se révèlent tolérants au décalage temporel par contre leur apprentissage est relativement lourd.

Les TDNN montrent aussi des limites pour modéliser des signaux très longs. En effet, la taille de la couche d'entrée est relative à la longueur de la fenêtre temporelle à analyser. Les réseaux de neurones prédictif liés (LPNN) [Tebelskis et al., 1991], en reliant des réseaux de neurones modélisant chacun des phonèmes différents, permettent de synthétiser des séquences plus longues et d'apprendre une grande variété de mots. Cependant l'introduction d'une mémoire à long terme, interne au réseau, semble plus appropriée pour certaines applications [Pearlmutter, 1989, Robinson, 1994]. C'est cet aspect qui est introduit dans les réseaux récurrents.

Réseaux de neurones récurrents : De même que les TDNN réalisent des filtres auto-régressifs mais avec des non-linéarités, les réseaux de neurones récurrents (RNN) peuvent être considérés comme la correspondance en non-linéaire des filtres Auto Regressif Mouving Average (ARMA) [Connor et al., 1992]. Il s'agit d'un réseau de neurones où les sorties de certains neurones sont rebouclées sur les entrées de couches inférieures. Cependant l'expérience a montré que l'apprentissage n'est pas aussi simple qu'avec la version modifiée de l'algorithme de backpropagation qui est appliquée au TDNN et des instabilités apparaissent. Des approches alternatives associant des algorithmes génétiques ont donc été proposées [Blanco et al., 2001].

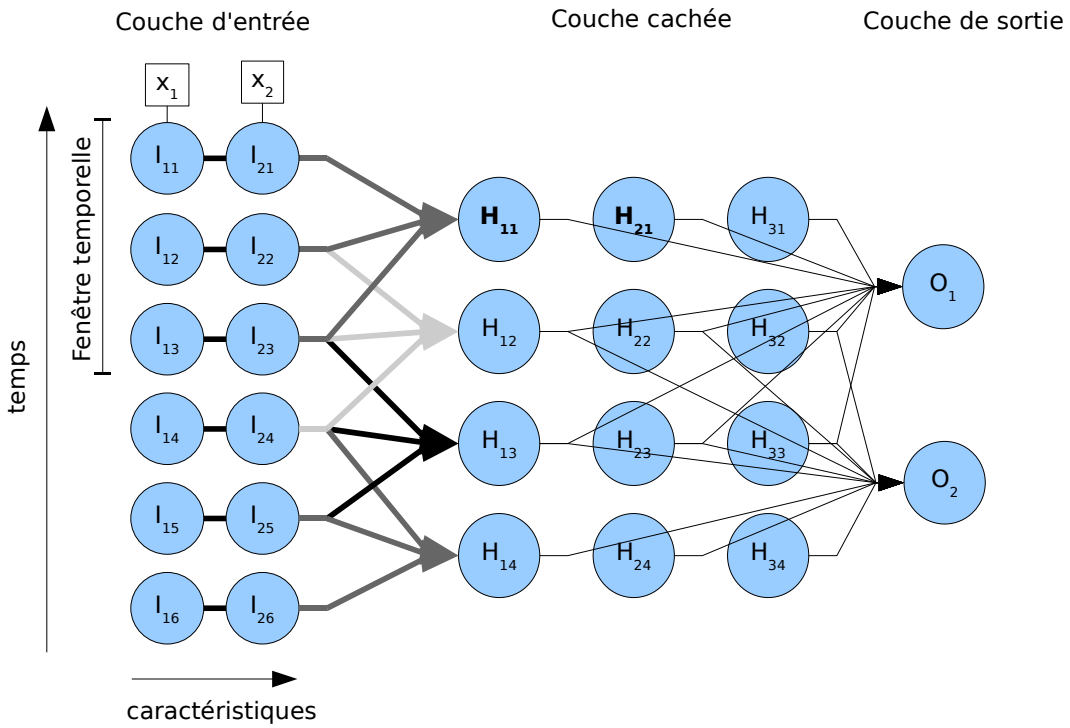


FIG. 4.4 – Architecture classique d'un Time-Delay Neural Network.

La figure 4.5 représente un réseau récurrent simple, où seule la sortie est rebouclée sur une couche inférieure. Une extension plus complète de ce type de réseau modélisant des process ARMA est le réseau de Jordan. Ce réseau contient des connexions allant de la couche de sortie vers la couche d'entrée mais aussi des récurrences entre la sortie et l'entrée d'un même neurone. Bien que n'étant pas exclusivement conçu pour la modélisation de séries temporelles, ces réseaux se sont révélés utiles à plusieurs reprises [DeCruyenaere et Hafez, 1992, Lee et Park, 1992].

Les réseaux de neurones ont aussi été adaptés à la modélisation de trajectoires dans les espaces d'états, notamment en considérant que les états correspondent à la sortie de la couche cachée. Le réseau de Elman [Elman, 1990] modélise la relation suivante, qui prend en compte, en plus d'un simple espace d'états, l'observation précédente :

$$\overrightarrow{s}(t) = \sigma(A\overrightarrow{s}(t-1) + D\overrightarrow{x}(t-1))$$

où $\overrightarrow{s}$ et $\overrightarrow{x}$ sont respectivement le vecteur d'état et le vecteur d'observations, tandis que $\sigma(\overrightarrow{a})$ est l'application d'une fonction sigmoïde sur chacun des éléments a_i de $\overrightarrow{a}$. Ces réseaux ont été mis en oeuvre dans plusieurs applications [Dorffner, 1996]. Il est cependant à noter que la transformation non-linéaire est très restreinte (seulement l'application d'une fonction sigmoïde et non pas une combinaison de fonctions non-linéaires) et ne représente pas du tout la forme générale d'une modélisation non-linéaire d'un espace d'état.

Les réseaux récurrents sont certainement plus adaptés que le TDNN simple pour modéliser des dynamiques temporelles. En effet, la récurrence permet la prise en compte d'un historique plus grand du signal. Cependant des problèmes de convergence et aussi une perte de l'information temporelle explicite dans les boucles de récurrences sont relevés dans [Dorffner, 1996].

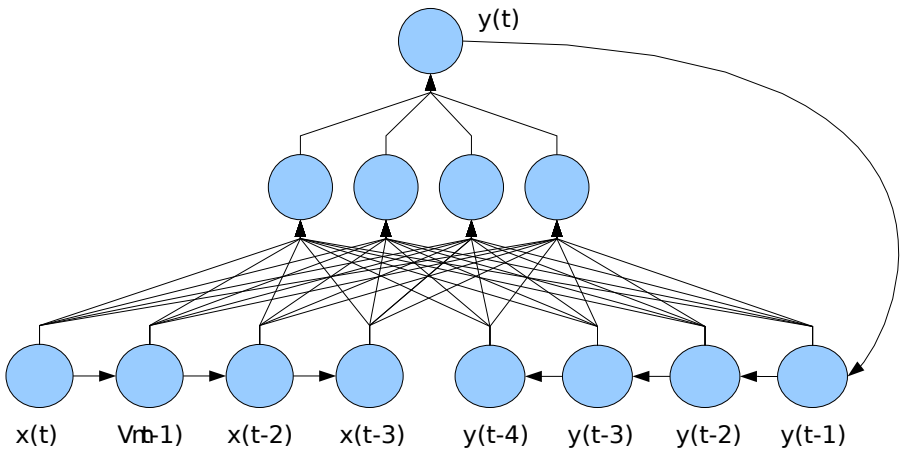


FIG. 4.5 – Architecture d'un Réseau de Neurones Récurrent.

Choix des délais Le choix des délais à introduire dans le réseau de neurones a aussi été examiné. Par exemple [Ramirez-Beltran et Montes, 2002] effectue une recherche exhaustive des délais à prendre en compte pour la modélisation d'un système multivariable. Une autre approche plus souple est proposée dans [Aussem et al., 1995] où les synapses du réseau de neurones sont remplacées par des modèles auto-régressifs. Il n'est donc plus nécessaire de présenter les différents instants choisis au réseau de neurones, celui-ci retient de lui-même l'information passée qui lui est utile.

Limitations des réseaux de neurones La limitation majeure des réseaux de neurones est qu'ils correspondent à une boîte noire : les nombreux paramètres et leurs interconnexions complexes rendent l'analyse des modèles extrêmement difficile. Ils sont donc principalement appliqués à des tâches de détection, de classification ou de prédiction mais ne permettent pas de réaliser d'autres tâches de fouilles de données telles que l'extraction de motifs ou l'extraction de connaissance en remontant à des segments particuliers des séries temporelles analysées. Une autre limitation, toujours liée à la complexité du modèle, concerne l'apprentissage. Il est reconnu que les réseaux de neurones requièrent de grandes quantités de données à cause de leur fort degré de liberté et l'adaptation des algorithmes d'apprentissage à des structures plus complexes (les TDNN, les RNN) se heurte à des problèmes de stabilité ou de temps de calcul. Un dernier problème réside dans la conception de la structure : le bon fonctionnement d'un réseau de neurones dépend non seulement d'un bon apprentissage mais aussi d'une structure (nombre de couches, nombre de neurones par couches, niveau de récurrence à intégrer) en adéquation avec la complexité des relations qu'il faut modéliser.

4.2.4 Modèles de Markov cachés

Les modèles de Markov cachées sont des automates à nombre d'états fini qui décrivent un système de manière stochastique non-déterministe. La théorie de base a été définie par Baum et Petrie à la fin des années 60 [Baum et Petrie, 1966], cependant les premières applications, notamment en traitement de la parole, ont seulement été proposées dans les années 70, par exemple par Jelinek [Jelinek et al., 1975, Jelinek, 1976]. La structure de base d'un modèle de Markov consiste en un ensemble d'états $S = (S_1, S_2, ..., S_N)$ connectés entre eux par des probabilités définies dans une table de transition. L'adjectif "caché" traduit le fait que l'émission des

observations, à partir d'un état, suit une loi aléatoire. C'est ce caractère aléatoire des mesures qui, ajouté aux propriétés des processus markoviens, fait la souplesse et la puissance de cette approche. Dans le cas des modèles du premier ordre, l'état du système à un instant t dépend uniquement de l'état du système à l'instant t-1, ce qui définit un processus Markovien :

$$P(q_t|q_{t-1}, q_{t-2}, ..., q_1) = P(q_t|q_{t-1})$$

Les modèles de Markov d'ordres supérieurs déterminent la prédiction de l'état suivant à l'aide des n états précédents : $P(S_q|q_{t-1}, q_{t-2}, ..., q_1) = P(q_t|q_{t-1}, ..., q_{t-n})$ et les procédures d'apprentissage et de décodage de la séquence d'états optimale ont été étendues à ces modèles [Kriouile, 1990]). Cependant celles-ci sont beaucoup plus complexes que les procédures standards. Dans le cadre d'un signal continu, une autre alternative, plus simple, est de traiter les dérivées du signal comme des variables supplémentaires.

Dans leur forme la plus simple, les MMC sont composés de 3 ensembles de probabilités :

- les probabilités des états initiaux $\pi = \{\pi_i\}$ où $\pi_i = P(q_1 = S_i), 1 \leq i \leq N$, qui définissent la probabilité que la série temporelle débute dans l'état S_i.
- les probabilités de transitions $A = \{a_{ij}\}$ où $a_{ij} = P(q_{t+1} = S_j|q_t = S_i)$, $1 \leq i, j \leq N$, qui définissent les probabilités de transiter d'un état S_i à un état S_j.
- les probabilités d'émission (aussi appelées probabilités d'observations) $b_j = P(O_t|q_t = S_i)$, qui définissent la probabilité d'obtenir un vecteur d'observation O_t, généré par l'état S_i. Ces dernières probabilités peuvent s'exprimer de différentes manières, suivant que les observations soient continues ou discrètes.

Un exemple de MMC est présenté figure 4.6, où les probabilités d'observations respectives à chaque état i sont définies par des lois normales de paramètres μ_i et σ_i.

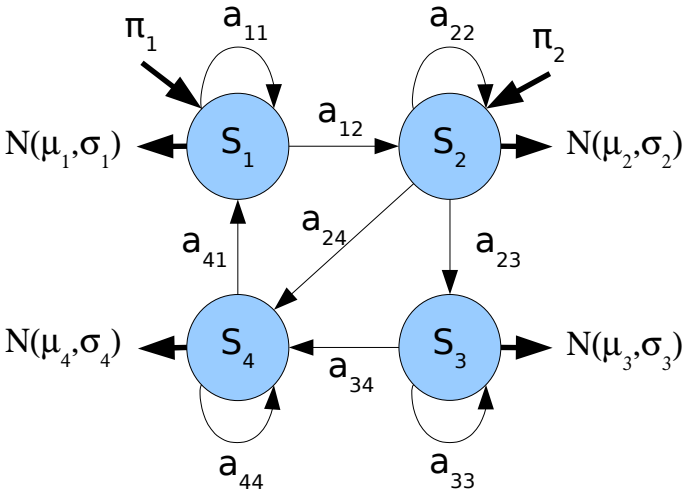


FIG. 4.6 – Exemple de MMC, avec des probabilités initiales π_i, des probabilités de transitions a_{ij}, des probilités d'observations régies par des lois gaussiennes de moyennes μ_i et d'écarts types σ_i.

Les MMC représentent un cadre théorique adapté à la résolution de 3 problèmes spécifiques largement décrits dans [Rabiner, 1989] :

Problème 1 : Étant donné une séquence d'observation $O = \{O_1, O_2, ..., O_T\}$ et un modèle M de paramètres θ, quelle est la probabilité, notée $P(O|M(\theta))$, que cette séquence soit

générée par $M(\theta)$? La résolution de ce problème permet la classification de séries temporelles en fonction d'un ensemble de modèles disponibles.

Problème 2 : Étant donné une séquence d'observation O et un modèle M de paramètres θ, comment peut-on retrouver la séquence d'états $Q = \{q_1, q_2, ..., q_T\}$ qui explique les observations O ? La résolution de ce problème permet la description de la dynamique d'une série temporelle multivariée en une séquence d'états.

Problème 3 : Étant donné une ou plusieurs séquences d'observations issues d'un corpus d'apprentissage λ, comment peut on ajuster les paramètres θ, d'un modèle M, pour maximiser $P(\lambda|M(\theta))$? La résolution de ce problème permet la description de dynamiques, apprises sur plusieurs séries temporelles, au sein d'un unique modèle.

C'est l'existence de solutions élégantes à ces trois problèmes qui a contribué à l'intérêt croissant des MMCs dans de nombreux domaines ainsi qu'au développement d'un grand nombre d'extensions. Les plus importantes sont présentées mais pour une bibliographie plus complète le lecteur pourra se référencer à [Murphy, 2002] :

- MMC à densité d'observation continue : les premiers MMC avaient des vecteurs d'observations discrets. Les CDMMC (continuous density MMC) représentent des probabilités d'observations de variables aléatoires continues, par exemple à l'aide de mixture de gaussiennes multivariées.
- Les MMC couplés [Saul et Jordan, 1995] : les MMC couplés mettent en évidence les relations pouvant exister entre plusieurs systèmes interconnectés. Il y a alors plusieurs séquences d'états, dont les probabilités de transiter d'états en états dépendent à la fois de leur propre état antérieur et de l'état antérieur des séquences voisines. Une utilisation typique est proposée dans [Chu et Huang, 2002] où la reconnaissance de la parole est effectuée en utilisant à la fois un flux audio et un flux vidéo capturant le mouvement des lèvres. Le lien entre ces deux flux est alors réalisé à l'aide d'un MMC couplé, avec deux chaînes internes.
- Les modèles semi-Markoviens [Russel et Moore, 1985, Levinson, 1986] : dans les modèles MMC standards les probabilités de rester dans un même état sont définis suivant une loi géométrique, de coefficient a_{ii}. Les modèles semi-Markoviens re-définissent de manière plus explicite le temps passé dans chaque état par des distributions paramétriques qui leur sont spécialement affectées et qui sont apprises comme les autres paramètres. La littérature récente présente de nouveaux développements autour des MSMC, notamment sur les méthodes d'apprentissage [Azimi et al., 2005, Yamagishi et Kobayashi, 2005], [Guédon, 2007].
- Les MMC auto-régressifs : ces modèles présentent des probabilités d'observations dépendant non seulement de l'état courant mais aussi de l'observation précédemment observée ; cette probabilité s'écrit donc : $P(O_t|S_t, O_{t-1})$.

Les modèles de Markov cachés se retrouvent dans de très nombreuses applications, notamment en acoustique, en bio-informatique, en climatologie, en télé-communications, en économétrie, en reconnaissance de texte et de la parole et pour de nombreux problèmes en traitement du signal et de l'image. La reconnaissance de la parole a toujours été l'application majoritaire, et ce notamment dans les années 90. Cependant, depuis les années 2000, de nouvelles applications émergent, par exemple en biologie pour l'analyse des séquences de protéines [Grundy et al., 2004, Edgar et Sjölander, 2004] et d'ADN [Fridlyand et al., 2004], [Majoros et al., 2005].

4.2.5 Approches hybrides ANN/MMC

En reconnaissance de la parole, les Modèles de Markov cachés (MMC) ont présenté des performances plus intéressantes que les réseaux de neurones : bien que les modèles de Markov soient généralement plus simplificateurs, la classification par méthode bayésienne assure un taux d'erreur minimal. Au cours des années 90, de nombreux travaux ont donc eu pour objectifs de combiner les avantages des deux approches. Ces travaux résultent pour la plupart en une modélisation à deux étages :

- soit le réseau neuronal est en amont du modèle de Markov et il sert à définir les probabilités d'observation x en fonction des états q_i ($P(x|q_i)$) ;
- soit le réseau neuronal est en aval du modèle de Markov : le modèle de Markov sert de pré-processeur pour le réseau de neuronal qui est plus discriminant mais aussi plus complexe.

4.2.5.1 Réseau neuronal en amont du modèle de Markov :

Il a été montré dans [Morgan et Bourlard, 1995] que les réseaux de neurones RBF ou le perceptron peuvent générer des sorties interprétables comme des probabilités *a priori* de choix de classes conditionnées par l'entrée observée, notées $P(q_k|x_n)$ (ou q_k est la k ième classe et x_n est le n ième vecteur d'observation). Ainsi, dans [Boulard et Morgan, 1993, Morgan et Bourlard, 1995, Schwenk, 1999], le réseau de neurones prend en compte les caractéristiques d'un signal sonore sur une fenêtre donnée et sert ainsi à la discrimination de phonèmes. Le modèle de Markov implémenté en aval, ou un simple décodeur Viterbi, utilise les probabilités*a posteriori* pour effectuer un alignement temporel.

Cependant des tests effectués par [Morgan et Bourlard, 1995] révèlent des erreurs d'approximation, avec notamment des sous-estimations des probabilités pour des sorties faibles du MLP tandis que des valeurs fortes surestiment ces probabilités.

4.2.5.2 Réseau neuronal en aval du modèle de Markov

Dans le cadre de la reconnaissance de phrases à partir d'un signal audio, les modèles de Markov peuvent être utilisés pour fournir les séquences de mots les plus probables, puis un système de traitement complémentaire est chargé de trouver la meilleure solution, en prenant en compte plus d'information. L'algorithme procurant l'ensemble de séquences les plus probables avec les modèles de Markov est appelé N-best search [Schwartz et al., 1992]. Il est par exemple utilisé dans [Mari et al., 1994] et [Zavaliagkos et al., 1994]. Les solutions conservées par cet algorithme sont ensuite présentées à un réseau de neurones de type MLP qui retourne lui aussi un score pour chaque séquence. Les scores respectifs du MMC et du réseau neuronal sont ensuite combinés pour trouver la meilleure solution.

4.2.5.3 Modèle Hybride Unificateur

Cette approche consiste à implémenter des MMC à partir de réseaux de neurones. En effet, Niles [Niles et Silverman, 1990] montre la relation existant entre les algorithmes d'apprentissages des deux modèles (le forward-backward et le backpropagation). Cette implémentation permet notamment la description d'états plus complexes, par exemple basés sur le principe de la quantification vectorielle qui permet une indexation optimale des vecteurs d'observations.

4.2.6 Bilan de l'état de l'art

Cette section bibliographique présente quatre types de modèles appliqués à la représentation de séries temporelles.

La représentation de séries temporelles dans les espaces des phases reconstruits présente une solution intéressante mais l'introduction d'observations mutli-dimensionnelles augmente très vite l'espace de représentation et complexifie l'apprentissage du modèle. Deux autres problèmes limitent leur utilisation aux données extraites de l'ECG : (i) le bruit d'observation n'est pas directement pris en compte dans le modèle or les données extraites peuvent présenter beaucoup de bruit, (ii) la modélisation de la trajectoire dans l'EPR nécessite plusieurs réalisations des observations, ce qui explique leur utilisation principalement pour des systèmes périodiques ou a-périodique.

Les algorithmes à base de réseaux de neurones ont l'avantage de pouvoir modéliser des dynamiques complexes et complètement non-linéaires. Cependant, bien que des techniques de guidage existent, le choix de l'architecture à employer demeure un problème complexe, souvent résolu de manière empirique. Pour le problème spécifique de fouille de données adressé dans ce travail, les réseaux de neurones présentent surtout l'inconvénient d'être vu comme une boite noire, et qu'interpréter les paramètres du modèle, par exemple pour faire un retour sur les caractéristiques mêmes des signaux, à partir des coefficients du modèle, est impossible.

Le filtrage de Kalman classique peut être utilisé avec l'algorithme EM dans le cadre de modèles Markovien linéaires et à états continus. Le filtre sert alors à éstimer les états cachés et l'algorithme EM détermine les paramètres du modèle en maximisant la vraisemblance de ce dernier. Les fondements théoriques solides du filtrage de Kalman expliquent son utilisation dans un grand nombre d'applications. Cependant, il est à noter que, dans le cadre de dynamiques non-linéaires, il faut connaître la forme de l'équation paramétrique et ensuite l'apprentissage permet l'identification des paramètres. Ceci ne peut donc correspondre à un cadre de fouille de données. Dans le cas du modèle linéaire il demeure toujours les problèmes du choix de la dimension des états cachés, de l'initialisation très importante de l'algorithme EM pour l'identification des paramètres ainsi que l'inconvénient des temps de calculs longs, comparés aux autres méthodes.

Les modèles de Markov cachés représentent les dynamiques de manière simple, avec des modèles qui peuvent être comparés entre eux. Ils sont bien adaptés au traitement multivariable et des versions rapides des algorithmes d'apprentissage des paramètres existent. Un avantage supplémentaire des MMC est de proposer une représentation graphique de la dynamique grâce aux états. Ces constats nous ont amené à privilégier les MMC dans notre application, où l'analyse de séries temporelles multivariées et à valeurs continues a guidé notre choix vers un modèle semi-Markovien à densité d'observation continu.

4.3 Modèles semi-Markovien pour l'analyse de séries temporelles multivariées

Le problème posé de caractérisation de dynamiques, nous incite à intégrer au mieux la notion de temps dans notre modèle. En effet, pour notre application, le temps passé dans un état peut être d'une importance capitale, ou du moins, tout aussi important que les transitions d'états en états : il est notamment possible que certains états se distinguent par des temps de

passage très court et d'autres par des temps de passage beaucoup plus élevés. Les MMC standards proposent, pour chaque état, une distribution géométrique pour l'évaluation du temps de passage, qui n'est pas toujours adaptée car elle suppose que les temps de passage courts sont plus probables que les temps de passage longs (de part la nature de la loi géométrique). Une distribution suivant une loi Gaussienne est donc proposée comme alternative. Comme la probabilité d'être à un instant t dans un état n'est plus uniquement dépendante de l'état dans lequel on est à l'instant $t-1$, la propriété de Markov n'est plus totalement respectée, d'où le terme de modèle semi-Markovien caché (MSMC). Cette différence entre un état de MMC et de MSMC est exposée figure 4.7.

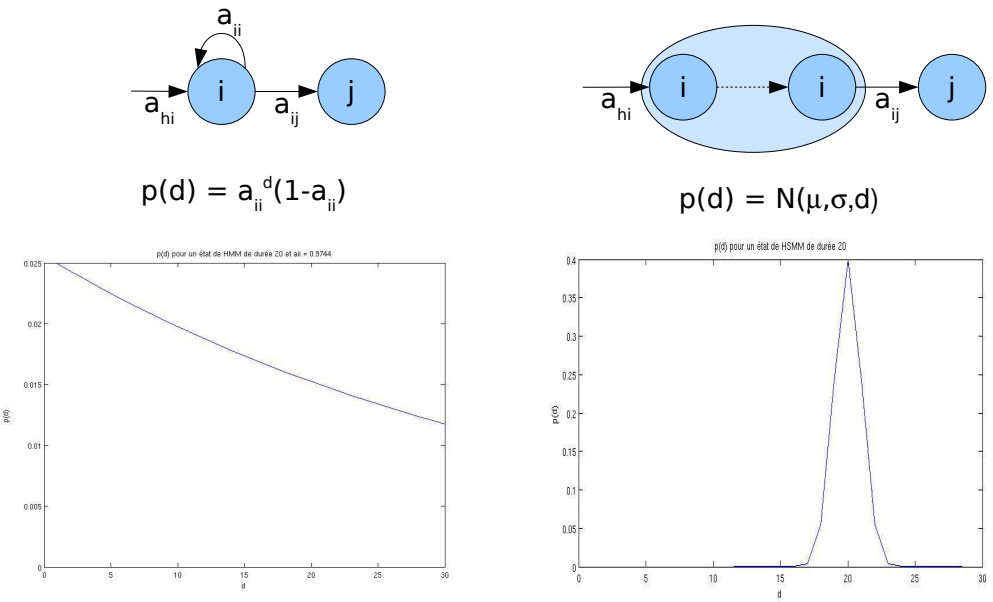


FIG. 4.7 – Différence MMC/MSMC sur la modélisation du temps passé dans un état : à gauche un état de MMC dont la probabilité de boucler sur lui-même suit une loi géométrique de paramètre a_{ii}, à droite un état de MSMC dont la probabilité de rester dans le même état suit une loie normale de paramètres μ et σ.

L'aspect continu des variables à modéliser est simplement pris en compte par des Gaussiennes multivariées. Des mixtures de Gaussiennes sont fréquemment employées mais ceci peut être équivalent à considérer plus d'états et la prise en compte d'une seule mixture simplifie l'apprentissage du modèle. D'ailleurs l'équivalence entre un modèle à N états et M mixtures et un modèle à N*M états et une seule mixture est montrée dans [Bicego et al., 2003].

Les paramètres du modèle, notés θ, à apprendre sont donc les suivants :
- les probabilités des états initiaux $\pi = \{\pi_i\}$,
- les probabilités de transitions $A = \{a_{ij}\}$ ou $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, $1 \leq i, j \leq N, i \neq j$, qui définissent les probabilités de transiter d'un état i à un état j (mais pas de rester dans le même état),
- les vecteurs μ_i et matrices Σ_i qui définissent respectivement les centres et les covariances des gaussiennes multivariées de chaque état,
- les scalaires μ_i^d et V_i^d qui définissent respectivement les moyennes et les variances des gaussiennes assignées à chaque état pour représenter les temps de passage.

La sous-section suivante présente la procédure d'apprentissage de ces paramètres.

4.3.1 Apprentissage des paramètres d'un MSMC

L'apprentissage des paramètres a pour objectif de trouver, pour un modèle M d'architecture donnée, l'ensemble des paramètres θ^* qui maximise le critère de vraisemblance :

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{r=1}^{R} P(O^r | M(\theta)) \tag{4.4}$$

où R et r sont respectivement le nombre de séries temporelles observées et l'index de ces séries temporelles.

Deux procédures d'apprentissage, la procédure de Baum-Welsch [Baum et al., 1970] et la procédure de Viterbi [Forney, 1973], issues toutes deux d'une approche itérative EM, sont majoritairement employées pour la résolution de ce problème.

Ces deux méthodes suivent le même principe :

- dans l'étape E, les paramètres courants sont employés pour déterminer les probabilités d'être dans un état S_i à l'instant t : $P(q_t = S_i | O, M(\theta))$, ainsi que les probabilités de transitions d'un état S_i à un état S_j : $P(q_t = S_i, q_{t+1} = S_j | O, M(\theta))$,
- dans l'étape M, les probabilités estimées dans l'étape E sont utilisées pour la ré-actualisation des paramètres.

Alors que la procédure Baum-Welch calcule les probabilités de l'étape E en considérant l'ensemble des observations, la méthode de Viterbi ne prend en compte que le chemin optimal. Par exemple, les probabilités $P(q_t = S_i | O, M(\theta))$ deviennent 1 si le chemin trouvé passe par l'état i à l'instant t et 0 sinon.

L'extension de l'algorithme Baum-Welch pour l'apprentissage des paramètres d'un modèle semi-Markovien a été proposée dans [Levinson, 1986] et a largement été réutilisée [Thoraval, 1995, Zen et al., 2004]. L'algorithme de Viterbi converge plus rapidement vers un maximum local que la procédure Baum-Welch mais elle peut rendre des estimées biaisées, notamment suite à une mauvaise initialisation [Rabiner, 1989]. L'algorithme de Viterbi a cependant été choisi car l'utilisation en fouille de données nécessite des apprentissages successifs et plusieurs modèles seront réalisés, nécessitant ainsi un temps de calcul réduit. La procédure de Viterbi est aussi moins sensible aux problèmes de stabilité numérique, par exemple lorsque les vraisemblances observées deviennent très petites, ce qui est d'autant plus fréquent avec les modèles semi-Markoviens. Enfin les modèles seront appris avec le plus grand nombre de séries temporelles possible, limitant ainsi le risque de tomber dans des minima locaux erronés. La figure 4.8 résume la procédure de Viterbi appliquée à un MMC. Les deux étapes E et M sont détaillées dans les deux paragraphes suivants, pour le cas spécifique des modèles semi-Markoviens.

4.3.1.1 Etape E : Estimation du chemin optimal par l'algorithme de Viterbi étendu aux MSMC

Comme présenté figure 4.8, l'algorithme de Viterbi se décompose en deux étapes. La première est l'étape de récursion qui parcourt le signal à partir de $t = 0$, tout en enregistrant les états et les temps de passage qui permettent de maximiser la vraisemblance observée à l'instant t. La seconde est l'étape de rétropropagation qui parcourt le signal en sens inverse,

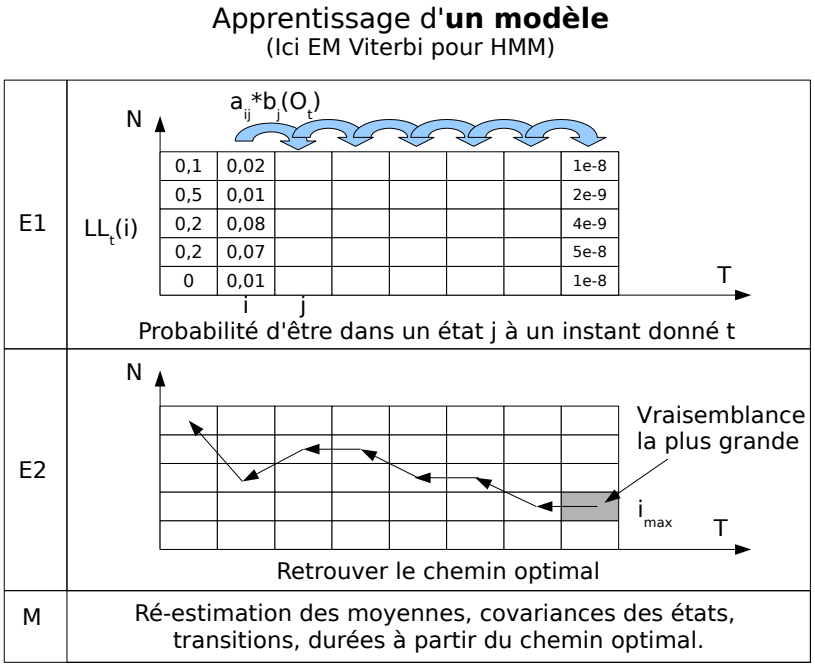


FIG. 4.8 – Apprentissage des paramètres par l'algorithme de Viterbi.

en partant de l'état pour laquelle la vraisemblance à T est maximum et en tenant compte des états précédents et des temps de passage enregistrés dans l'étape de récursion.

Etape récursion (E1) : La vraisemblance d'être dans un état i à l'instant t est déterminée par la relation suivante :

$$ll_t(i) = \max_{1 \leq d \leq Dmax} \left(p_i(d) \prod_{t-d+1}^{t} b_j(O_t) \max_{1 \leq j \leq N} (ll_{t-d}(j) a_{ji}) \right) \tag{4.5}$$

$$\text{où} \qquad t = \{D_{max}, D_{max}+1, ..., T-1, T\}$$

Cette vraisemblance est maximisée en fonction de l'état précédent j et de son temps de passage d. Pour chaque état et pour chaque instant, les valeurs de d_{max} et j_{max} maximisant $ll_t(i)$ sont enregistrées :

$$\psi_t(i) = j_{max}; \tag{4.6}$$

$$D_t(i) = d_{max}; \tag{4.7}$$

puis utilisées dans l'étape de rétropropagation.

Etape de rétropropagation (E2) : La séquence d'états $Q = \{q_1, q_2, ..., q_T\}$ est déterminée de manière à maximiser la vraisemblance $P(Q|O, M(\theta))$ et à l'aide de l'algorithme 1.

Au cours de cette étape, plusieurs variables, résumées table 4.1, sont enregistrées, pour permettre la mise à jour des paramètres dans l'étape de maximisation.

Algorithme 1 : Viterbi - rétropropagation

INITIALISATION

/* Recherche de l'état i_{max} maximisant la vraisemblance à T */

$i_{max} = \max_i\{ll_T(i)\}$;

/* Rester dans l'état i_{max} entre T - $D_T(i_max)$ et T */

pour $t = T - D_T(i_{max})$ *à* T **faire**

  $q(t) = i_{max}$;

/* t_{fin} : dernier instant avant une transition entre états */

$t_{fin} = T - D_T(i_{max}) - 1$;

$t = t_{fin}$;

BOUCLE DE RETROPROPAGATION

tant que $t > 0$ **faire**

  /* Récuperation de l'état */

  $etat = \psi_{t_{fin}+1}(q(t_{fin}+1))$

  tant que $t > t_{fin} - D_{t_{fin}}(etat)$ **faire**

    /* Rester dans état entre $t_{fin} - D_{t_{fin}}(etat)$ et t_{fin} */

    $q(t) = etat$;

    $t--$;

  /* Mise à jour de t_{fin} pour passer à l'état précédent */

  $t_{fin} = t$;

n_i	Nombre de passages dans l'état i
S_i	Somme des initialisations dans l'état i
S_{ij}	Somme des transitions de l'état i à l'état j
S_id	Somme des durées dans l'état i
S_id^2	Somme des durées au carré dans l'état i

TAB. 4.1 – Variables calculées et enregistrées au cours de l'étape de retro-propagation.

4.3.1.2 Etape M : Mise à jour des paramètres

Les paramètres du modèle sont ajustés à partir des résultats de l'étape E intégrés sur l'ensemble des individus. Les matrices de transitions et d'initialisation sont aisément recalculées à l'aide du vecteur S_i et de la matrice S_{ij} :

$$\pi_i = \frac{S_i}{\sum_j^N S_j} \tag{4.8}$$

$$a_{ij} = \frac{S_{ij}}{\sum_j^N S_{ij}} \tag{4.9}$$

Le vecteur des moyennes de la gaussienne caractérisant chaque état est redéfini par :

$$\mu_i = \frac{\sum_t \omega_i^t O_t}{\omega_i} \tag{4.10}$$

où ω_i^t est égal à 1 lorsque le chemin optimal se trouve dans l'état i à l'instant t, et à zéro sinon, et $\omega_i = \sum_t w_i^t$.

La covariance est calculée par la relation suivante :

$$\begin{aligned}\hat{\Sigma}_i &= \frac{1}{\omega_i}\sum_t \omega_i^t (O_t - \mu_i)(O_t - \mu_i)' \\ &= \frac{1}{\omega_i}\sum_t \omega_i^t (O_t O_t' - \mu_i O_t' - O_t \mu_i' + \mu_i \mu_i') \\ &= \frac{1}{\omega_i}(\sum_t \omega_i^t (O_t O_t') - \sum_t \omega_i^t (\mu_i O_t') - \sum_t \omega_i^t (O_t \mu_i') + \sum_t \omega_i^t (\mu_i \mu_i'))\end{aligned} \tag{4.11}$$

ce qui donne, en considérant la stationnarité des états et en sortant les termes en μ_i des sommes :

$$\begin{aligned}\hat{\Sigma}_i &= \frac{\sum_t \omega_i^t (O_t O_t')}{\omega_i} - \frac{\mu_i \sum_t \omega_i^t O_t'}{\omega_i} - \frac{(\sum_t \omega_i^t O_t)\mu_i'}{\omega_i} + \frac{\mu_i \mu_i' \sum_t \omega_i^t}{\omega_i} \\ &= \frac{\sum_t \omega_i^t (O_t O_t')}{\omega_i} - \mu_i \mu_i' \text{ suivant l'équation 4.10}\end{aligned} \tag{4.12}$$

Enfin la moyenne et variance des durées de chaque état sont mises à jour de la même manière :

$$\mu_i^d = \frac{S_i d}{n_i} \tag{4.13}$$

$$V_i^d = \frac{S_i d^2}{n_i} - (\mu_i^d)^2 \tag{4.14}$$

4.3.2 Apprentissage de l'hyper-paramètre, N, le nombre d'états du modèle

Comme expliqué précédemment, les états sont définis par une seule gaussienne multivariée. Seul le nombre d'état est donc à ajuster pour assurer un apprentissage correcte des dynamiques temporelles. Trouver le nombre d'états approprié est important : un modèle avec trop d'états va sur-apprendre les données, et a plus de chance de regrouper différentes dynamiques dans un unique modèle. D'un autre coté, un modèle avec trop peu d'états décrira mal les dynamiques des séries qui lui sont affectées. Trouver le nombre d'états optimum est généralement effectué en calculant en fonction de N la vraisemblance marginale des données $P(X|N)$. Cette vraisemblance s'écrit :

$$P(X|N) = \int_\theta P(X|\theta, N) P(\theta|N) d\theta \tag{4.15}$$

Cependant il est impossible d'intégrer sur l'ensemble de l'espace des paramètres et des approximations sont donc appliquées. Les plus populaires sont l'approximation de Laplace [Heckerman et al., 1995], l'approximation de Cheeseman-Stutz [Cheeseman et Stutz, 1996] ou le critère d'information bayesienne (BIC) [Schwarz, 1978]. Dans le contexte de la fouille de donnée, le critère BIC, qui est plus rapide et ne nécessite pas l'introduction de connaissances *a priori* a été retenu dans de nombreuses applications [Li et al., 2002, Katz, 1981, Berchtold et Raftery, 2002] :

$$N_{choisi} = \underset{N}{\text{argmax}}(logP(X|N, \hat{\theta}(N)) - f(N)/2 * logI) \tag{4.16}$$

où $f(N)$ est une fonction qui retourne le nombre de paramètres du modèle en fonction du nombre d'états et I est le nombre de séries temporelles.

4.3.3 Représentation d'un MSMC sur les séries temporelles

Après la détermination du nombre d'états et l'apprentissage des paramètres qui ont été abordés précédemment, il est possible, lorsque l'on se restreint à des séries bivariées, de représenter le modèle obtenu sur ces séries. La figure 4.9 réalise une superposition entre les séries générées par un système de Lorenz et le MSMC chargé de les modéliser. Les ellipses décrivent les matrices de covariances de chaque état alors que les chiffres dans ces ellipses correspondent au temps moyen passé dans ces états. Les flèches d'épaisseurs variables représentent les coefficients de la matrice de transitions.

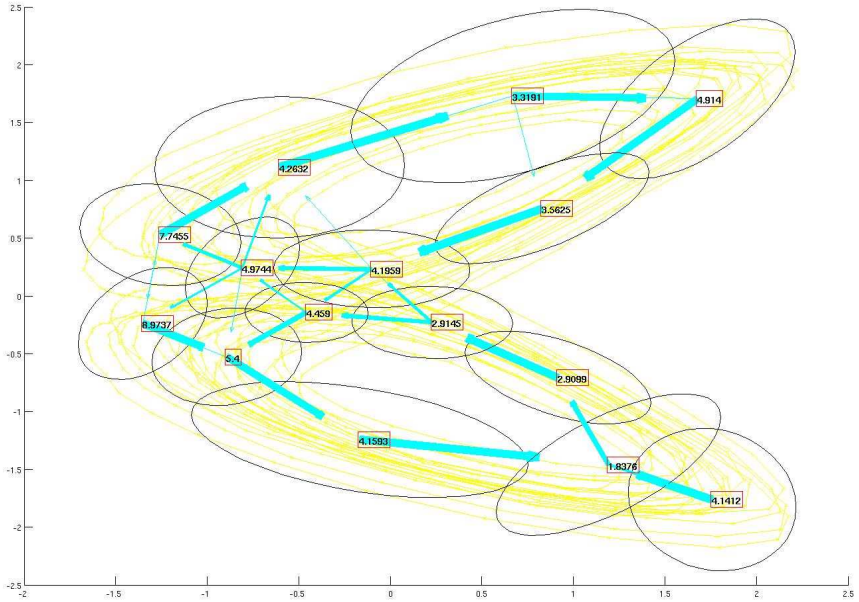


FIG. 4.9 – Modélisation du système de Lorenz avec un MSMC.

4.3.4 Exploitation des MSMC

La sous-section précédente montre donc la possibilité d'exploiter graphiquement les MSMC pour décrire les séries temporelles. Cependant, dans le cadre de ce travail, les modèles seront principalement employés aux tâches de classification et de clustering.

La classification assigne une série temporelle à un modèle d'index k_{win} pour lequel la vraisemblance obtenue est maximale parmi un ensemble de K modèles en compétitions :

$$k_{win} = \underset{1 \leq k \leq K}{\operatorname{argmax}} \{P(O_i, \theta_k)\}$$

Le clustering réalise plusieurs itérations d'apprentissages et de tests de vraisemblance afin de retrouver les groupes naturels d'individus. La section suivante présente une méthode de clustering basée sur les MSMC.

4.4 Clustering avec les MSMC

Le clustering est l'opération qui permet de réaliser des groupes naturels d'individus en se basant sur un espace de représentation. Une fois le clustering réalisé, il devient aisé d'analyser les distances entre individus et par la suite d'expliquer les regroupements ou le fait que certains individus soient isolés. Cette tâche de fouille de données est d'autant plus intéressante qu'elle requière très peu de connaissances *a priori* pour son initialisation et qu'elle permet aussi la recherche des variables utiles à la discrimination des individus. Dans le cadre de notre étude, cette tâche consiste à retrouver les groupes naturels de séries temporelles. Un algorithme de clustering basé sur l'utilisation des MSMC est donc proposé dans cette section.

Les méthodes de clustering typiques sont divisées en méthodes de partitionnement ou en méthodes hiérarchiques. Les méthodes de partitionnement partent d'un nombre fixé de clusters et échangent, au cours d'un certain nombre d'itérations, les individus entre ces clusters. Les méthodes hiérarchiques sont ascendantes ou descendantes : dans le premier cas les clusters sont regroupés, dans l'autre ils sont divisés. Ces regroupements ou divisions sont choisies suivant une fonction dite de score. Le clustering avec modèle représente une méthode hybride, où il est possible de regrouper ou de diviser des modèles, tout en réaffectant les individus aux nouveaux modèles et en ré-estimant ces derniers de manière itérative.

Dans notre problématique de recherche de dynamiques globales, il apparaît raisonnable de supposer que le nombre de classes est faible par rapport au nombre d'individus à analyser. Ceci est d'ailleurs le cas dans la plupart des applications de clustering. De plus, dans le cas d'une approche basée modèle, une méthode de clustering descendante permet de limiter les problèmes de généralisation puisque plusieurs séries sont généralement affectées à un seul modèle. Un algorithme de type descendant, avec augmentation progressive du nombre de modèles a donc été choisi.

4.4.1 Procédure de clustering descendante proposée

Des approches de clustering descendant avec les modèles de Markov ont déjà été proposées dans la littérature [Li et al., 2002]. Cependant, il n'existe pas, à notre connaissance, d'introduction des MSMC à la place des MMC standard. Notre approche est nouvelle aussi sur la manière dont le nombre de modèles est augmenté. Un clustering flou, donnant des niveaux d'appartenance entre 0 et 1, est effectué sur les individus afin de les assigner aux différents modèles. Cet aspect flou est aussi conservé lors de l'apprentissage des modèles par l'algorithme EM, et ainsi, à chaque étape du clustering, l'appartenance de chaque individu à chacun des clusters est évaluée.

La figure 4.10 présente cette procédure de clustering dont les détails sont les suivants :

Étape 1 : Les séries sont aléatoirement affectées à l'un des deux groupes initiaux. L'appartenance des individus est binaire pour cette première étape.

Étape 2.1 : L'estimation du nombre d'états N de chaque MSMC est effectuée avec le critère BIC, en utilisant l'équation 4.16. En théorie, le nombre d'états doit être ré-estimé à chaque itération de la boucle EM (comme les étapes 2.2 à 2.5), cependant le choix a été fait de le fixer à chaque ré-initialisation de l'algorithme EM, ce qui permet un gain de calcul notable.

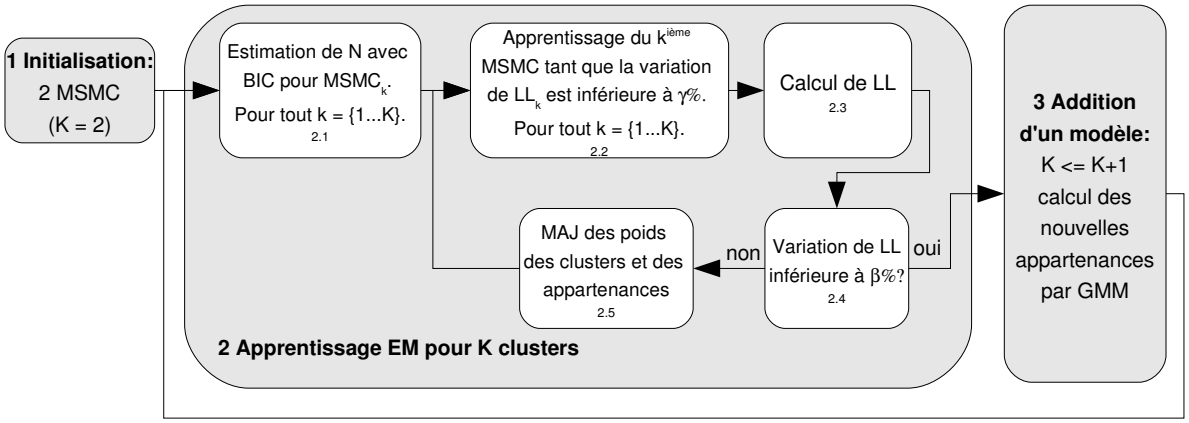


FIG. 4.10 – Schéma de la procédure de clustering proposée.

Étape 2.2 : L'apprentissage des K MSMC est effectué avec l'algorithme de Viterbi auquel est intégré l'appartenance des individus. L'équation 4.4, qui permet d'actualiser les paramètres pour maximiser la vraisemblance devient :

$$\hat{\theta_k} = \underset{\theta_k}{\operatorname{argmax}} \sum_i P(k|O_i, \Lambda) log P(O_i|\theta_k) \tag{4.17}$$

où $P(k|O_i, \Lambda)$ est l'appartenance *a priori* de l'individu i au groupe k, sachant la structure des modèles Λ.

La log-vraisemblance LL_k du modèle k est alors calculée par :

$$LL_k = P(\sum_i P(k|O_i, \Lambda) log P(O_i|\hat{\theta_k})) \tag{4.18}$$

Lorsque LL_k ne varie pas de plus de $\alpha\%$, il est considéré que le modèle a été suffisament ajusté aux données et l'apprentissage s'arrête.

Étape 2.3 : La log-vraisemblance du partitionnement courant (LL) est calculée comme étant la somme des log-vraisemblances sur le nombre d'individus et sur l'ensemble des modèles :

$$LL = log P(X|\Lambda) = \sum_i log(\sum_k \alpha_k P(O_i|\theta_k)) \tag{4.19}$$

où les α_k sont les probabilités *a priori* des modèles.

Étape 2.4 : Lorsque LL varie de moins de $\beta\%$ entre l'itération courante et l'itération précédente de l'algorithme EM, il est considéré que l'apprentissage du partitionnement avec le nombre K de clusters est achevé. Dans ce cas, un nouveau cluster est créé. Sinon une nouvelle itération de l'algorithme EM est relancée (Etape 2.5).

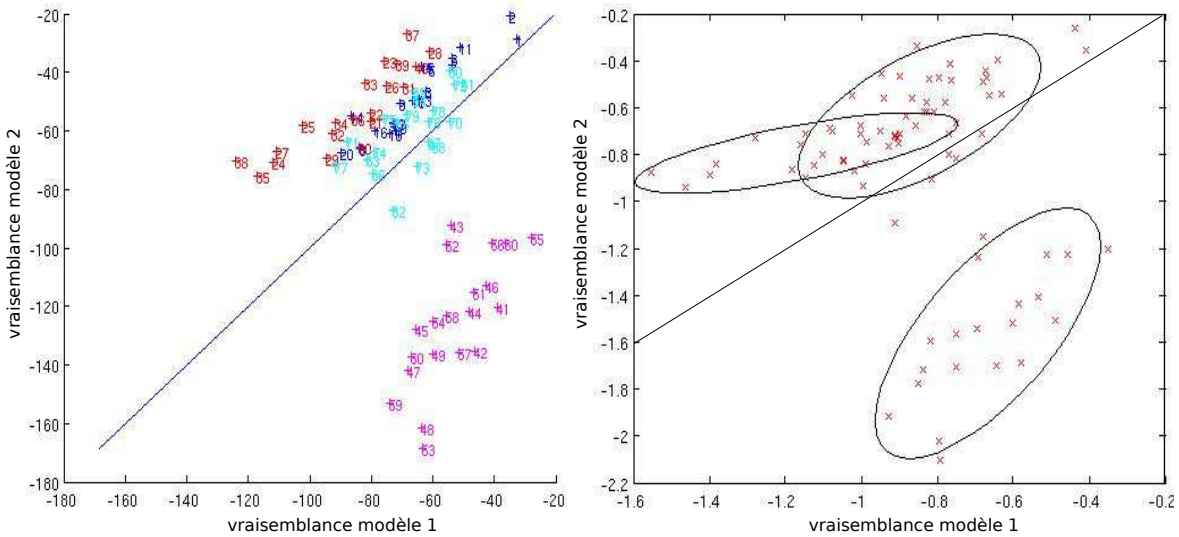
Étape 2.5 : Les mises à jour de l'appartenance de chaque individu et des poids de chaque cluster (α_k) sont calculées à chaque itération de la boucle EM, respectivement avec les équations 4.20 et 4.21 :

$$P(k|O_i, \Lambda) = \frac{\alpha_k P(O_i|\theta_k)}{\sum_j \alpha_j P(O_i|\theta_j)} \tag{4.20}$$

$$\alpha_k^{new} = \frac{1}{N} \sum_n P(k|O_i, \Lambda) \tag{4.21}$$

Étape 3 : Lorsqu'un nouveau cluster est créé, plusieurs méthodes pour affecter les individus à ce nouveau cluster sont proposées :

- Une méthode simple consiste à sélectionner les individus présentant une vraisemblance faible pour les modèles déjà générés et donc à les affecter au nouveau modèle. Par exemple Li [Li et Biswas, 2000] prend l'individu qui a la vraisemblance la plus faible et crée un nouveau cluster uniquement à partir de cet individu. Le partitionnement est ensuite initié avec les anciens et le nouveau modèle. Cependant l'individu n'est pas choisi sur des critères de représentativité d'un groupe, d'ailleurs il se peut même que ce soit un outsider, ce qui peut poser problème. Enfin, créer un cluster à partir d'un seul individu peut entraîner aussi un problème de généralisation.
- Au lieu d'utiliser la vraisemblance d'un seul individu, il est aussi possible de créer le nouveau partitionnement à partir de l'ensemble de l'espace des vraisemblances, où chaque axe représente la vraisemblance associée à un modèle. L'avantage va être de prendre en compte la position de tous les individus dans ce bon espace de représentation des dynamiques. La méthode proposée se base sur l'algorithme des mixtures de Gaussiennes (GMM) [McLachlan et Peel, 2000] : dans un espace de vraisemblance crée avec K modèles, K+1 gaussiennes sont ajustées et la probabilité de chaque point d'être issu de ces gaussiennes définit l'appartenance flou des séries temporelles aux nouveaux modèles qui seront générés. La figure 4.11 illustre la réalisation de 3 clusters (qui donneront 3 modèles) à partir d'un espace de vraisemblance planaire (2 modèles).



a) plan de vraisemblance, avec deux modèles b) clustering avec un 3ième modèle

FIG. 4.11 – Clustering dans l'espace des vraisemblances avec l'algorithme GMM : un exemple de passage de deux à trois modèles.

4.4.2 Évaluation du nombre de clusters et risque de sur-apprentissage

Le nombre correct de clusters peut être évalué par inspection de l'évolution de la log-vraisemblance de l'ensemble de la structure de modèles (notée LL), et tracée en fonction du nombre de cluster. Un palier peut être observé lorsque le nombre correct est atteint. Des critères de pénalités sur la complexité comme le critère BIC ou Cheeseman-Stutz sont aussi employés. Dans le cadre de notre fouille de données, le choix est laissé à l'utilisateur sur le nombre de clusters à conserver pour les analyses. Les critères de pénalité sont présentés à titre indicatif.

Un choix doit être fait, via l'ajustement du paramètre γ, sur le nombre d'itérations de l'algorithme de Viterbi pour l'apprentissage de chaque modèle (étape 2.2) sur une itération de la boucle EM (étapes 2.2 à 2.5). L'idéal serait de procéder à une seule itération et de réévaluer immédiatement l'appartenance des individus aux nouveaux clusters mais cette solution est difficile à mettre en oeuvre en pratique pour des raisons de temps de calcul. D'un autre coté, avec la méthode choisie pour déterminer le nombre d'états, il y a un risque de surestimer ce dernier, d'autant plus que l'appartenance des individus est initialement inconnue et que les clusters ne sont donc pas exacts au début de l'apprentissage. Apprendre peu au début et ré-estimer régulièrement l'appartenance des individus est donc obligatoire. En fait ce problème d'ajustement de la précision d'apprentissage est proche d'un problème de recuit simulé. Au début les modèles sont ajustés de manière grossière aux données, ce qui leur permet de changer de clusters, comme le recuit simulé avec coefficient de température élevé. A la fin, pour avoir des modèles fiables, le nombre d'itérations est augmenté, ce qui correspond à réduire les probabilités qu'un individu change de cluster, comme lorsque le coefficient de température est diminué. Une solution simple à ce problème est proposée : au début les modèles sont appris jusqu'à ce que leur vraisemblance varie de moins de $\gamma = 1\%$ entre deux itérations. Lorsque la vraisemblance globale LL varie de moins de $\beta = 1\%$, la précision d'apprentissage est réduite à $\gamma = 0.01\%$ et les modèles sont à nouveau appris jusqu'à ce que LL varie de moins de $\beta = 1\%$.

4.5 Validation sur données simulées

L'analyse de dynamiques des séries temporelles peut être décomposée en trois tâches distinctes, représentées figure 4.12. Ces trois tâches sont l'apprentissage des paramètres d'un modèle à partir de plusieurs séries temporelles, le test de vraisemblance qu'une série temporelle soit générée par les modèles proposés et la simulation des séries temporelles à partir d'un modèle.

Pour évaluer l'aptitude des MSMC à résoudre ces différentes tâches, trois tests sur des données simulées sont proposés dans cette section :

- Le premier test, très simple, consiste i) à apprendre la dynamique d'une série temporelle avec un MMC et un MSMC, ii) à simuler ces deux modèles et iii) à observer si les dynamiques sont bien reconstituées.
- Le second test présente un problème de classification où différentes méthodes sont utilisées pour apprendre, puis classer, un ensemble de séries temporelles. Dans ce test les MSMC sont comparés avec d'autres méthodes de classification de séries temporelles.
- Le troisième test est dédié à l'évaluation des capacités de l'algorithme de clustering basé sur les MSMC et présenté dans la section précédente.

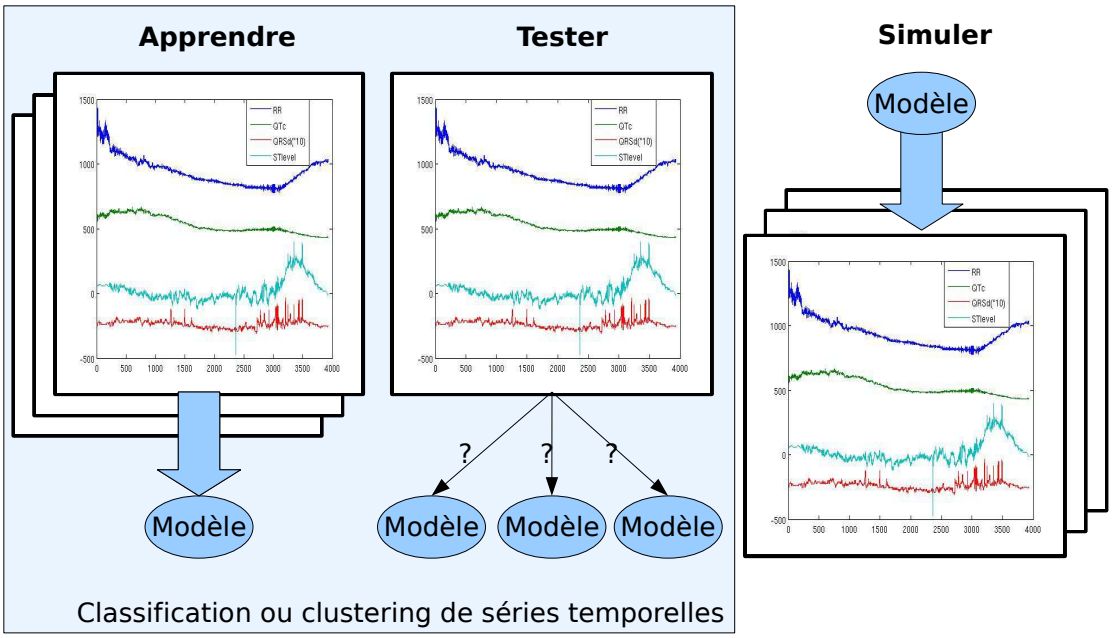


FIG. 4.12 – Les 3 tâches affectées aux MMC/MSMC.

4.5.1 Apprentissage et simulation avec MMC et MSMC

La figure 4.13 montre les différentes étapes de ce test :
- deux séries temporelles sont générées : elles se présentent sous la forme de paliers sur lesquels sont additionnés un bruit blanc gaussien ;
- ces deux séries sont apprises sur un MMC et un MSMC à quatre états ;
- trois séries temporelles sont simulées avec ces deux modèles.

Les résultats montrent que les séries simulées par le MSMC correspondent beaucoup plus aux séries initiales que les séries simulées par le MMC. Les temps passés sur chacun des paliers sont mieux restitués lorsque les états sont équipés d'une loi normale dédiée à représentation du temps que lorsque le simple paramètre de bouclage sur soi-même est conservé. Ceci illustre la performance accrue du MSMC pour des tâches de simulation ou de prédiction sur le MMC. Le second test l'évalue aussi sur un problème de classification.

4.5.2 Evaluation des performances en classification : étude comparative

Un problème simple de classification est proposé pour évaluer les performances de différents modèles de représentation de séries temporelles. Les modèles comparés sont : le modèle AR, un modèle basé sur une représentation dans un espace de phase reconstruit, un MMC et un MSMC tandis que les séries temporelles sont extraites d'un système de Rössler. Ce type de système dynamique a été choisi en raison de sa fréquente utilisation dans la littérature pour des tests sur données simulées. Les équations du système de Rössler sont les suivantes :

$$\begin{cases} \frac{dx}{dt} = -y - z \\ \frac{dy}{dt} = x + ay \\ \frac{dz}{dt} = b + zx - zc \end{cases} \tag{4.22}$$

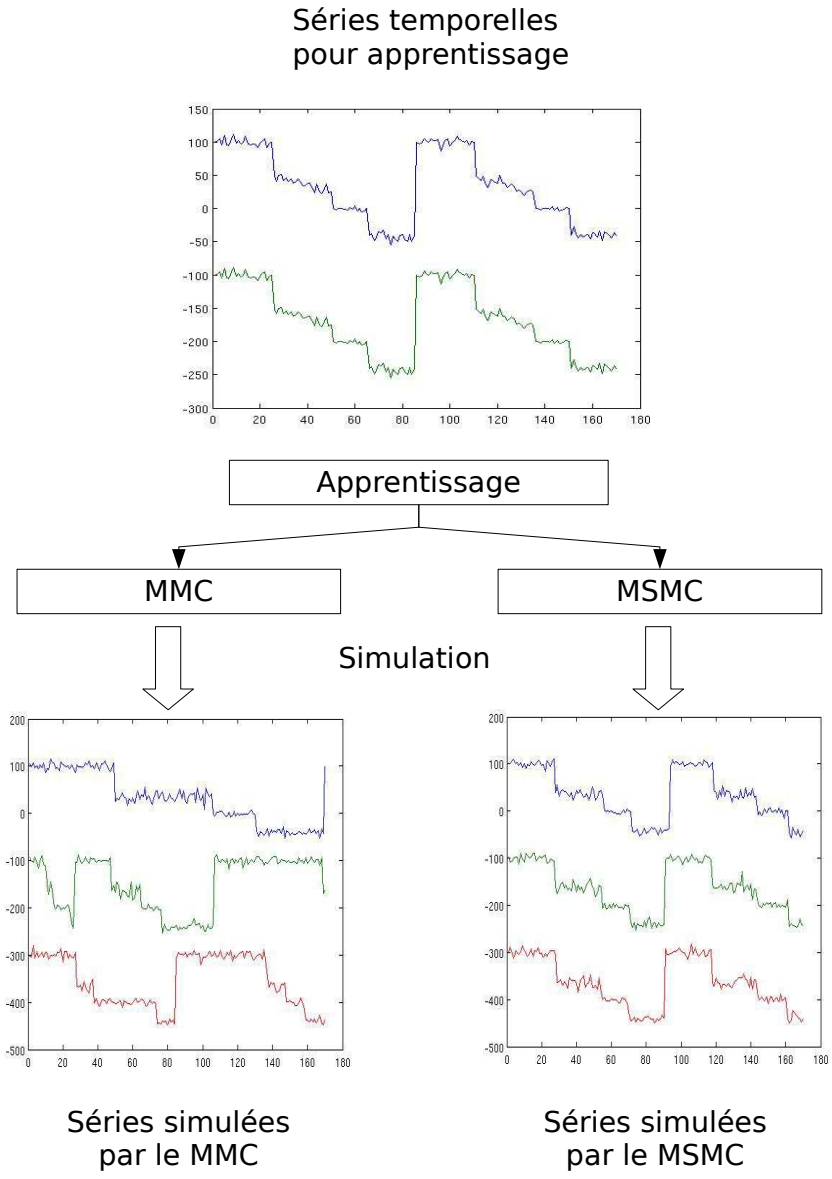


FIG. 4.13 – Simulation de séries temporelles à partir d'un MMC et d'un MSMC ayant appris les mêmes dynamiques.

A partir de ces équations, deux ensembles de séries temporelles sont produits, le premier est réalisé avec un paramètre a tiré suivant une loi uniforme entre [0.01 0.11] et le second toujours suivant une loi uniforme mais sur l'intervalle [0.13 0.23]. Pour ces deux ensembles les paramètres b et c sont fixes, avec comme valeurs 0.2 et 2.5 et les séries réalisées correspondent à la variable z du système de Rössler. Les deux ensembles sont composés de 230 séries chacun : 30 pour l'apprentissage et 200 pour le test. A ces séries temporelles est superposé un bruit blanc gaussien dont l'amplitude est ajustée pour correspondre à différents rapports signal à bruit (RSB) : 5, 10 et 15dB.

L'apprentissage et la classification se passent de la manière suivante, qui est différente pour les modèles AR, les modèles de trajectoires dans l'EPR et les modèles Markoviens :

4.5.2.1 Modèle AR

Apprentissage Le critère d'Akaike est utilisé pour déterminer l'ordre des modèles puis ceux-ci sont calculés pour chacune des séries temporelles d'apprentissage. Il est ensuite recherché de manière supervisée les coefficients les plus discriminants entre les deux groupes.

Classification Les modèles AR sont calculés pour l'ensemble des individus de test puis séparés en deux groupes en appliquant un algorithme des k-means (avec un nombre de groupe égal à deux) sur les coefficients discriminants trouvés pendant l'apprentissage.

4.5.2.2 Modélisation dans l'EPR

Apprentissage Les méthodes d'information mutuelle et des plus proches voisins sont utilisées pour déterminer le délai et la dimension de l'EPR. Les séries d'apprentissage sont intégrés dans cet attracteur puis leur trajectoire est modélisée par une mixture de gaussiennes multivariées [Povinelly, 2005], avec un modèle pour chacune des deux classes.

Classification Les séries de tests sont intégrées dans l'attracteur et leur appartenance aux deux modèles précédemment construits est évaluée. Les séries sont classées suivant le maximum de vraisemblance.

4.5.2.3 Modélisation par MMC et MSMC

Apprentissage Un modèle est créé pour chacune des deux classes, avec les algorithmes précédemment décrits.

Classification Les séries de test sont classées suivant le maximum de vraisemblance.

Les taux d'erreurs de classification suivant les différents modèles et les différents RSB sont présentés table 4.2.

Les modèles EPR et MSMC ont approximativement les mêmes taux d'erreur pour des bruits faibles mais l'EPR montre des performances très réduites avec le RSB de 5dB. Le modèle AR a toujours des performances plus faibles que les autres modèles, quel que soit le niveau de bruit. Il y a aussi une nette différence entre le MMC et le MSMC en faveur du MSMC. Ceci montre l'aptitude des MSMC à l'apprentissage de dynamiques avec notamment leur intérêts dans le cas de séries temporelles bruités ainsi que le gain apporté par les MSMC comparés aux MMC standards.

		AR	EPR	MMC	MSMC
RSB	5dB	46.5	32.5	21	17
		24.5	47.5	26	16
		26.5	46.5	14.5	5
		19.5	46.5	17.5	4
	10dB	17	4	14	11
		9.5	5.5	12.5	3.5
		16.5	5	12	4.5
		14	8	10.5	8.5
		9	5.5	5	3.5
	15dB	9.5	4.5	8	4.5
		9.5	5	13	4.5
		13	7	10	7.5
		9	3	9	4
		11	3	12	7

TAB. 4.2 – Résultats (taux d'erreurs exprimés en %) de la classification des séries temporelles du système de Rössler. Plusieurs ensembles d'apprentissage et de test ont été générés pour chacun des niveaux de bruit.

4.5.3 Evaluation de l'algorithme de clustering

L'algorithme de clustering décrit dans la section précédente est appliqué à différents ensembles de tests pour évaluer ses performances. Il est à noter que cet algorithme est générique et peut être appliqué à des ensembles de séries temporelles très variées, mais ayant comme caractéristique d'être différentiables suivant leurs dynamiques.

Dans le cadre de cette validation, l'algorithme est appliqué sur des ensembles de tests qui contiennent des séries issues de différents groupes. Ces séries sont générées à partir d'un ensemble de règles et de paramètres et les différentes valeurs de paramètres définissent les groupes. L'algorithme de clustering doit donc permettre de reclasser de manière non supervisée les séries qui lui sont présentées suivant leurs groupes initiaux. Les ensembles de tests utilisés sont les suivants :

- T_1 : Les deux groupes de ce test diffèrent par les pentes des séries temporelles. Un décalage aléatoire est aussi introduit entre chaque série et un bruit Gaussien est additionné (RSB de 63dB). Cet exemple est intéressant car toutes les séries ont la même distribution et il est possible d'évaluer la robustesse des modèles de Markov en fonction de décalages temporels.
- T_2 : Quatre MMC différents, trois de 4 états et un de 5 états sont réalisés et utilisés pour générer les séries temporelles des différents groupes. La taille des séries temporelles est fixée à 80 points.
- T_3 : Un système de Lorenz, dont les équations sont les suivantes, est utilisé pour générer 4 groupes de séries temporelles :

$$\begin{cases} \frac{dx}{dt} = a * (y - x) \\ \frac{dy}{dt} = b * x - y - x * z \\ \frac{dz}{dt} = x * y - c * z \end{cases} \tag{4.23}$$

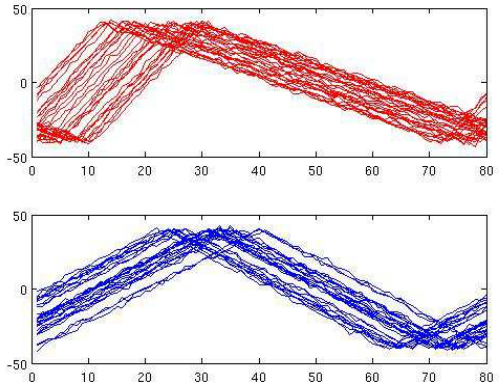
La configuration des paramètres régissant les différents groupes est listée table 4.3. Chaque groupe est constitué de 20 séries temporelles de 200 points. Un bruit blanc gaussien est aussi additionné (RSB de 16dB). T3 est lui même divisé en deux ensembles : dans T_{3-1} seul la variable d'état z est conservée pour le clustering alors que dans T_{3-2}

	a	b	c
G1	10	$U(28, 28.5)$	$U(4.2, 4.8)$
G2	10	$U(25, 25.5)$	$U(4.2, 4.8)$
G3	10	$U(28, 28.5)$	$U(3.3, 4)$
G4	10	$U(25, 25.5)$	$U(3.3, 4)$

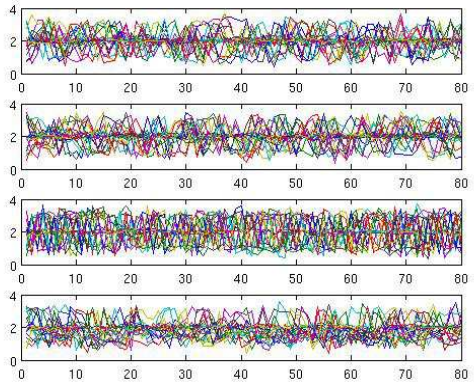
TAB. 4.3 – Paramètres du système de Lorenz pour 4 groupes de séries temporelles. U(a,b) désigne la loi de probabilité uniforme sur un intervalle [a,b].

un clustering de séries multivariées sera réalisé en analysant les variables y et z.

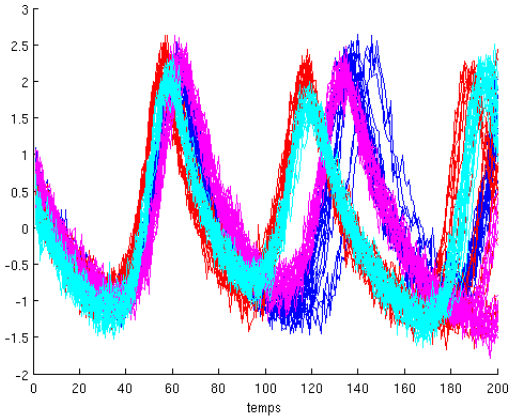
Les séries temporelles de ces ensembles de test sont présentées table 4.4. Pour T_{3-2} les séries sont intégrées dans l'espace d'état définit par les variables y et z.



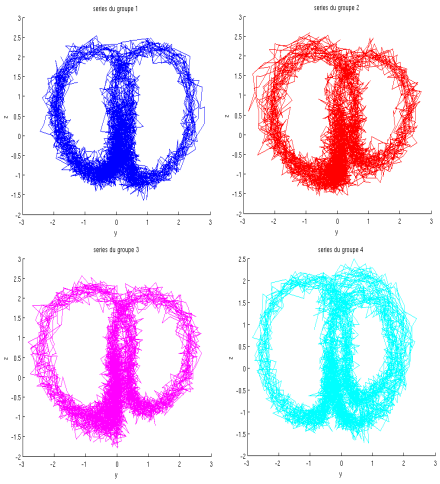
T_1



T_2



T_{3-1}



T_{3-2}

TAB. 4.4 – Les différentes séries temporelles présentées à l'algorithme de clustering.

L'évaluation des performances est effectuée à travers les tableaux de contingences 4.5, 4.6 et 4.7, obtenus à l'issu du clustering. Les lignes représentent les groupes d'où sont issues les séries temporelles et les colonnes les groupes trouvés par l'algorithme de clustering.

	G1	G2
I1	27	3
I2	11	19
sans dérivée Taux d'erreur = 23%		

	G1	G2	G3
I1	0	25	5
I2	27	3	0
avec 3 groupes Taux d'erreur = 13%			

	G1	G2
I1	0	30
I2	30	0
avec dérivée Taux d'erreur = 0%		

TAB. 4.5 – Tableaux de contingence pour le clustering de T_1, avec différents MSMC .

	G1	G2	G3	G4
I1	19	0	0	1
I2	0	0	18	2
I3	0	20	0	0
I4	4	0	0	16

Taux d'erreur = 8.75%

TAB. 4.6 – Tableaux de contingence pour le clustering de T_2 .

	G1	G2	G3	G4
I1	8	3	2	7
I2	13	6	0	1
I2	7	12	0	1
I2	0	6	0	14

T_{3-1} : Taux d'erreur = 48.75%

	G1	G2	G3	G4
I1	20	0	0	0
I2	0	12	8	0
I2	0	1	19	0
I2	0	0	0	20

T_{3-2} : Taux d'erreur = 11.25%

TAB. 4.7 – Tableaux de contingence pour le clustering de T_{3-1} et T_{3-2} .

Analyse des résultats : Les résultats montrent que la méthode EM associée aux modèles semi-markoviens permet de distinguer des groupes de séries à partir d'un ensemble de séries non labellisées. Le test T_1 prouve l'intérêt de la méthode pour analyser des séries qui peuvent présenter des décalages temporelles. Il faut cependant remarquer, tableau 4.5, qu'initialement l'algorithme sépare de manière approximative les deux groupes de séries (taux d'erreur de 23%). Raffiner le clustering avec un troisième modèle fournit une séparation plus marquée des deux groupes (taux d'erreur de 12%). Enfin, le fait d'ajouter la dérivée comme variable supplémentaire dans les MSMC entraine une séparation parfaite des deux groupes. Ce résultat est rassurant puisque les séries de cet ensemble se distinguent précisément par leurs pentes. Le test T_2 (tableau 4.6) montre que des séries Markoviennes standards peuvent aussi être caractérisées par les modèles semi-Markoviens, avec l'algorithme d'apprentissage spécialement conçu. Ce test est aussi utile pour analyser le comportement de l'algorithme en cas d'un nombre de groupes de séries plus élevé (4). Le dernier test (tableau 4.7) est sûrement le plus intéressant car il présente des séries temporelles multivariées avec des caractéristiques plus complexes. Ces séries sont aussi plus proches de celles que l'on peut retrouver dans le domaine du biomédical. Il est important de noter que le résultat du clustering dans le cas de T_3 est nettement amélioré lorsque deux variables sont analysées au lieu d'une seule, le taux d'erreur passant de 48.75% dans le premier cas à 11.25% dans le deuxième.

4.6 Conclusion

Ce chapitre aborde le cadre général de la fouille de données de séries temporelles. Il a été relevé en premier lieu qu'une approche de modélisation des séries temporelles permet d'appréhender efficacement les différentes problématiques de la fouille de données. En second lieu, la bibliographie sur la modélisation des séries temporelles montre que les modèles Markoviens sont particulièrement adaptés et requièrent peu de connaissances *à priori*. Les propriétés des modèles de Markov leur permettent effectivement de résoudre de nombreuses tâches de fouille de données. Dans le cadre de séries temporelles continues, la section 4.3 de ce chapitre montre l'intérêt des modèles semi-Markoviens et d'un algorithme d'apprentissage, basé sur un premier modèle Markovien puis sur l'algorithme de Viterbi qui permet de réaliser un apprentissage rapide des paramètres du modèle à partir d'un certain nombre de séries temporelles. Les problèmes de fouilles de données de séries temporelles telles que (i) la classification, (ii) la simulation, (iii) la représentation des données sous forme compacte, sont directement résolus par l'utilisation des MSMC. Le clustering demeure un problème plus complexe et est donc l'objet de la quatrième section de ce chapitre. Une approche de clustering innovante, basée sur une augmentation progressive du nombre de modèles utilisés pour représenter les séries a été réalisée. Cette approche allie un algorithme EM pour apprendre les paramètres de l'ensemble des modèles et un codage flou pour définir l'appartenance des séries aux modèles. De plus, l'affection des individus aux modèles, lorsque leur nombre est augmenté, est gérée à l'aide de la répartition obtenue à la fin de l'itération précédente. Cet algorithme de clustering est testé avec succès sur un ensemble varié de données simulées. Il reste cependant à noter qu'en pratique, les résultats d'un tel algorithme seront naturellement dépendants du nombre de séries temporelles disponibles et du choix des variables prises en compte pour l'apprentissage des modèles.

Bibliographie

[Abarbanel et al., 1994] Abarbanel, H. D. I., Carroll, T. A., Pecora, L. M., Sidorowich, J. J., et Tsimring, L. S. (1994). Predicting physical variables in time-delay embedding. *Phys. Rev. E*, 49(3) :1840–1853.

[Antunes et Oliviera, 2001] Antunes, C. et Oliviera, A. (2001). Temporal data mining : an overview. *Lecture Notes in Computer Science*.

[Aussem et al., 1995] Aussem, A., Murtagh, F., et Sarazin, M. (1995). Dynamical recurrent neural networks–towards environmental time series prediction. *Int J Neural Syst*, 6(2) :145–170.

[Azimi et al., 2005] Azimi, M., Nasiopoulos, P., et Ward, R. (2005). Offline and online identification of hidden semi-Markov models. *Signal Processing, IEEE Transactions on*, 53(8) :2658–2663.

[Baum et Petrie, 1966] Baum, L. et Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37 :1554–1563.

[Baum et al., 1970] Baum, L., Petrie, T., Soles, G., et Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilitic functions of Markov chains. *Annals Math Stat*, 41(1) :164–171.

[Bennani et Gallinari, 1991] Bennani, Y. et Gallinari, P. (1991). On the use of TDNN-extracted features information in talker identification. *Acoustics, Speech, and Signal Processing, International Conference on*, 1 :385–388.

[Berchtold et Raftery, 2002] Berchtold, A. et Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statist. Sci.*, 17(3) :328–356.

[Berndt et Clifford, 1994] Berndt, D. et Clifford, J. (1994). Using dynamic time warping to find patterns in times series. *KDD-94*, NA :359–370.

[Bezruchko et al., 2001] Bezruchko, B., Dikanev, T., et Smirnov, D. (2001). Role of transient processes for reconstruction of model equations from time series. *Physical Review E*, 64(3) :29–45.

[Bicego et al., 2003] Bicego, M., Murino, V., et Figueiredo, M. (2003). A sequential pruning strategy for the selection of the number of states in hidden markov models. *Pattern Recognition Letters*, 24 :1395–1407.

[Blanco et al., 2001] Blanco, A., Delgado, M., et Pegalar, M. (2001). A real coded genetic algorithm for training recurrent neural networks. *Neural Networks*, 14 :93–105.

[Boulard et Morgan, 1993] Boulard, H. et Morgan, N. (1993). *Connectionist Speech Recognition : A Hybrid Approach*. Kluwer Academic Publishers.

[Buzug et Pfister, 1992] Buzug, T. et Pfister, G. (1992). Comparison of algorithms calculating optimal embedding parameters for delay time coordinates. *Physica D Nonlinear Phenomena*, 58 :127–137.

[Casdagli, 1989] Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, 35 :335–356.

[Cellucci et al., 2003] Cellucci, C., Albano, A., et Rapp, P. (2003). Comparative study of embedding methods. *Physical Review*, 67.

[Chatterjee et al., 2005] Chatterjee, A., Nait-Ali, A., et Siarry, P. (2005). An input-delay neural-network-based approach for piecewise ECG signal compression. *Biomedical Engineering, IEEE Transactions on*, 52(5) :945–947.

[Cheeseman et Stutz, 1996] Cheeseman, P. et Stutz, J. (1996). *Advances in Knowledge discovery and data-mining*, chapter Bayesian classification (autoclass) : Theory and results, pages 153–180. Cambridge, MA : MIT press.

[Chu et Huang, 2002] Chu, S. et Huang, T. (2002). Audio-visual speech modeling using coupled hidden Markov models. *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, 2 :2009–2012.

[Connor et al., 1992] Connor, J., Atlas, L., et Martin, D. (1992). Recurrent networks and NARMA modeling. In et als, M. J., editor, *Neural Information Processing Systems 4*, volume 14, pages 301–308, San Mateo, CA.

[Das et al., 1997] Das, G., Dimitrios, D., et Mannila, H. (1997). Finding similar time series. *First European Symposium on Principles of Data mining and knowledge discovery*, NA :NA.

[DeCruyenaere et Hafez, 1992] DeCruyenaere, J. et Hafez, H. (1992). A comparison between Kalman filter and recurrent neural networks. In *IJCNN International Joint Conference on Neural Networks, IEEE*, pages 247–251, Baltimore.

[Dempster et al., 1977] Dempster, A., Laird, N., et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

[Dorffner, 1996] Dorffner, G. (1996). Neural network for time series processing. In COMPANY, T. I., editor, *Neural Network World*, volume 6-4, pages 447–468.

[Edgar et Sjölander, 2004] Edgar, R. et Sjölander, K. (2004). COACH : profile-profile alignment of protein families using hidden Markov models. In *Bioinformatics*, volume 20, pages 1309–1318. Oxford Univ Press.

[Elman, 1990] Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14-2 :179–212.

[Farmer et Sidorowich, 1987] Farmer, J. et Sidorowich, J. J. (1987). Predicting chaotic time series. *The American Physical society. Letters in Physical Review*, 59-8 :845–848.

[Forney, 1973] Forney, G. D. (1973). The Viterbi algorithm. *Proc. IEEE*, 61 :268–278.

[Fraser, 1989] Fraser, A. M. (1989). Reconstructing attractors from scalar time series : a comparison of singular system and redundancy criteria. *Physica*, 34 :391–404.

[Fridlyand et al., 2004] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., et Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, 90(1) :132–153.

[Garcia et Almeida, 2005] Garcia, S. et Almeida, J. (2005). Multivariate phase space reconstruction by nearest neighbor embedding with different time delays. *Phys. Rev.*, E 72.

[Ghahramani et Hinton, 1996] Ghahramani, Z. et Hinton, G. (1996). Parameter estimation for linear dynamical systems. Technical report, Dept. of Computer Science, University of Toronto.

[Grundy et al., 2004] Grundy, W., Bailey, T., Elkan, C., et Baker, M. (2004). meta-MEME : Motif-based hidden Markov models of protein families. In *Bioinformatics*, volume 13, pages 397–406. Oxford Univ Press.

[Guédon, 2007] Guédon, Y. (2007). Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics and Data Analysis*, 51(5) :2379–2409.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., et Chickering, D. M. (1995). A tutorial on learning with bayesian networks. *Machine Learning*, 20 :197–243.

[Hung et Ho, 1999] Hung, Y. et Ho, H. (1999). Real-time mobility tracking algorithms for cellular networks based on Kalman filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[Jebara et Pentland, 1997] Jebara, T. et Pentland, A. (1997). Parametrized structure from motion for 3d adaptive feedback tracking of faces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, page 144.

[Jelinek, 1976] Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proc. IEEE*, 64 :532–536.

[Jelinek et al., 1975] Jelinek, F., Bahl, L., et Mercer, R. (1975). Design of a linguistic statistical decoder for the recocgnition of continuous speech. *IEEE Trans. Information Theory*, 21 :250–256.

[Julier et Uhlmann, 1997] Julier, S. et Uhlmann, J. (1997). A new extension of the Kalman filter to nonlinear systems. *PROCEEDINGS- SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, 3068 :182–193.

[Kalman, 1960] Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82 :35–45.

[Katz, 1981] Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, 23(3) :243–249.

[Kennel et al., 1992] Kennel, M., Brown, R., et Abarbanel, H. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review*, A46-6 :3111–3118.

[Kim et Woods, 1998] Kim, J. et Woods, J. (1998). 3-D Kalman filter for image motion estimation. *Image Processing, IEEE Transactions on*, 7(1).

[Kraskov, 2004] Kraskov, A. ; Stögbauer, H. G. P. (2004). Estimating mutual information. *Physical Review -Serie E-*, 69(6).

[Kriouile, 1990] Kriouile, A. (1990). *La reconnaissance automatique de la parole et les modèles Markoviens cachés*. PhD thesis, Université de Nancy 1.

[Lavagetto, 1995] Lavagetto, F. (1995). Speech articulatory analysis through time delay neural networks. In *Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on*, pages 306–309.

[Laxman et Sastry, 2006] Laxman, S. et Sastry, P. (2006). A survey of temporal data mining. *Sadhana*, 31-2 :173 ?198.

[Lee et Park, 1992] Lee, C. et Park, K. (1992). Prediction of monthly transition of the composition stock price index using recurrent back-propagation. In I, A. et J, T., editors, *Artificial Neural Networks 2*, pages 1629–1632, North-Holland, Amsterdam.

[Levinson, 1986] Levinson, S. (1986). Continuously variable duration hidden Markov models for speech analysis. *Proceedings IEEE/ICASSP*, pages 1241–1244.

[Li et Biswas, 2000] Li, C. et Biswas, G. (2000). a bayesian approach to temporal data clustering using hidden Markov models. *ICML*.

[Li et al., 2002] Li, C., Biswas, G., Dale, M., et Dale, P. (2002). Matryoshka : A hmm based temporal data clustering methodology for modeling system dynamics. *Intelligent Data Analysis*, 6(3) :281–308.

[Liu et al., 1998] Liu, Q., Islam, S., Rodriguez-Iturbe, I., et Le, Y. (1998). Phase-space analysis of daily streamflow : characterisation and prediction. *Advances in Water Resources*, 221 :463–475.

[Majoros et al., 2005] Majoros, W. H., Pertea, M., Delcher, A. L., et Salzberg, S. L. (2005). Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinformatics*, 6 :16.

[Mari et al., 1994] Mari, J.-F., Fohr, D., Auglade, Y., et Junqua, J.-C. (1994). Hidden Markov models and selectively trained neural networks for connected confusable word recognition. In *Proc. Int. Conf. Spoken Language Processing*, volume 3, pages 1519–1522. Acoustical Soc. Japan.

[Matthies et al., 1989] Matthies, L., Kanade, T., et Szeliski, R. (1989). Kalman filter-based algorithms for estimating depth from image sequences. In *International Journal of Computer Vision*, volume 3, pages 209–238.

[McLachlan et Peel, 2000] McLachlan, G. et Peel, D. (2000). *Finite Mixture Models*.

[Morgan et Bourlard, 1995] Morgan, N. et Bourlard, H. (1995). Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE*, 83-5 :742–772.

[Murphy, 2002] Murphy, K. (2002). *Dynamic Bayesian Networks : Representation, Inference and Learning*. PhD thesis, University of California, Berkeley.

[Mörchen, 2006] Mörchen, F. (2006). *Time Series Knowledge Mining*. PhD thesis, Philipps-Universitat Marburg.

[Niles et Silverman, 1990] Niles, L. et Silverman, H. (1990). Combining hidden Markov model and neural network classifiers. *IEEE Int. Conf. Acoust. Speech, Signal Processing*, 1 :417–420.

[Pearlmutter, 1989] Pearlmutter, B. (1989). Learning state space trajectories in recurrent neural networks. *Neural Networks, IJCNN., International Joint Conference on*, 2 :365–372.

[Povinelly, 2005] Povinelly, R. J. (2005). Towards the prediction of transient st changes. *Computers in Cardiology*, 32 :663–666.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speeach recognition. *Proc. of the IEEE*, 77(2) :257–285.

[Radoji et al., 2002] Radoji, I., Gautama, T., et Mandi, D. (2002). A comparison of two novel methods for characterisation of heart rate variability series. *Biosignal*, 69.

[Ramirez-Beltran et Montes, 2002] Ramirez-Beltran, N. et Montes, J. (2002). Neural networks to model dynamics systems with time delays. *IIE Transactions*, 34 :313–327.

[Ramjattan et Cross, 1995] Ramjattan, A. N. et Cross, P. A. (1995). A Kalman filter model for an integrated land vehicle navigation system. In *Journal of Navigation*, pages 293–302.

[Robinson, 1994] Robinson, A. J. (1994). An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2) :298–305.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychol Rev.*, 65(6) :386–408.

[Roweis, 2000] Roweis, S.; Ghahramani, Z. (2000). An EM algorithm for identification of nonlinear dynamical systems.

[Rumelhart et al., 1986] Rumelhart, D., Hinton, G., et Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323 :533–536.

[Russel et Moore, 1985] Russel, M. et Moore, R. (1985). Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 10 :5–8.

[Saul et Jordan, 1995] Saul, L. et Jordan, M. (1995). Exploiting tractable substructures in intractable networks. *Advances in Neural Information Processing Systems*, ISSUE 8 :486–492.

[Schwartz et al., 1992] Schwartz, R., Austin, S., Kubala, F., Makhoul, J., Nguyen, L., Placeway, P., et Zavaliagkos, G. (1992). New uses for the N-Best sentence hypotheses within the BYBLOS speech recognition system. In *Proc. ICASSP92*, pages 1/1–1/4.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annuals of Statistics*, 6 :461–464.

[Schwenk, 1999] Schwenk, H. (1999). Using boosting to improve a hybrid HMM/neural network speechrecognizer. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 2 :1009–1012.

[Shumway et Stoffer, 1982] Shumway, R. H. et Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4) :253–264.

[Suzuki et al., 2003] Suzuki, T., Ikeguchi, T., et Suzki, M. (2003). Mutlivariable nonlinear analysis of foreign exchange rates. *Physica*, A323 :591–600.

[Takens, 1980] Takens, F. (1980). Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence. Lecture Notes in Mathmeatics. Springer-Verlag*, 898 :336–381.

[Tebelskis et al., 1991] Tebelskis, J., Waibel, A., Petek, B., et Schmidbauer, O. (1991). Continuous speech recognition using linked predictive neural networks. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 61–64vol.1.

[Thoraval, 1995] Thoraval, L. (1995). *Analyse statistique de signaux électrocardiographiques par modèles de Markov cachés*. PhD thesis, Université de Rennes.

[Waibel, 1989] Waibel, A. Hanazawa, T. H. G. S. K. L. K. (1989). Phoneme recognition using time-delay neural networks. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 37, pages 328–339.

[Wan et Han, 2007] Wan, Y. et Han, M. (2007). Prediction of multivariate chaotic time series based on optimized phase space reconstruction. *Control Conference, CCC 2007*, pages 169–173.

[Weigend et al., 1990] Weigend, A., Huberman, B., et Rumelhart, D. (1990). Predicting the future : A connectionist approach. *International Journal of Neural Systems*, 1-3 :193–209.

[Wells, 1995] Wells, C. (1995). *The Kalman Filter in Finance*. Springer ; 1 edition.

[Yamagishi et Kobayashi, 2005] Yamagishi, J. et Kobayashi, T. (2005). Adaptive training for hidden semi-Markov model. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1 :365–368.

[Zainab et Mark, 2005] Zainab, R. et Mark, B. (2005). Real-time mobility tracking algorithms for cellular networks based on Kalman filtering. *IEEE Transactions on Mobile Computing*, 4(2) :195–208.

[Zavaliagkos et al., 1994] Zavaliagkos, G., Zhao, Y., Schwartz, R., et Markoul, J. (1994). A hybrid segmental neural net/hidden Markov model system for continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2-1 :151–160.

[Zen et al., 2004] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., et Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. *Proc. of ICSLP*, pages 1180–1185.

Chapitre 5

Analyse de caractéristiques temporelles évolutives lors d'un épisode ischémique et d'un test d'effort

Le chapitre précédent a tenté de montrer comment les modèles de Markov cachés pouvaient être astucieusement utilisés pour classer des signaux ayant des dynamiques différentes. Une démonstration de faisabilité a été effectuée sur des données simulées. Cet ultime chapitre nous ramène au coeur du problème clinique à l'origine de ce travail. Il présente des résultats de classification et de clustering des séries temporelles extraites de l'ECG à l'aide des algorithmes qui ont été présentés dans le chapitre 3. Deux exemples cliniques distincts sont traités : tout d'abord, nous nous intéresserons à la détection d'épisodes ischémiques à partir d'enregistrements Holter et dans un second temps nous analyserons des ECG d'efforts acquis chez des patients atteints du syndrome de Brugada.

5.1 Classification et clustering des épisodes ischémiques

L'objectif est ici de montrer comment la prise en compte du temps peut améliorer la détection des épisodes ischémiques. La détection de ce type d'épisodes est toujours fondée sur la recherche de déviations du segment ST mais nous nous efforcerons d'augmenter la spécificité de cette détection par l'analyse des dynamiques d'autres indicateurs extraits de l'ECG.

Les épisodes ischémiques analysés sont extraits de la base de données LTST [Jager et al., 2003]. Cette base a été créée dans l'objectif de fournir un large ensemble d'épisodes ischémiques qui puissent être utilisés pour l'évaluation des détecteurs automatiques. Elle contient donc un grand nombre de décalages du segment ST prétraités automatiquement puis corrigés et annotés manuellement par des experts.

De manière résumée, les différentes déviations sont annotées suivant quatre groupes :
- les épisodes ST transitoires, associés à un événement ischémique : ST-IS (ischémie). Ils sont caractérisés par des changements de morphologie du segment ST, potentiellement accompagnés de changements dans le rythme cardiaque. Des informations cliniques supplémentaires suggérant l'ischémie sont prises en compte.
- les épisodes ST non-ischémiques dus à des changements du rythme cardiaque ST-RC (Rythme Cardiaque). Ils sont distingués par des changements de morphologies du segment ST, accompagnés de variations du rythme cardiaque et quand les informations

cliniques supplémentaires ne suggèrent pas l'ischémie.
- les décalages ST dus à des changements de position du patient ST-AS (Axis shift). Ils sont distingués par des changements de morphologies du segment ST brusques ou progressifs, accompagnés de changements d'amplitudes dans les ondes du QRS.
- les décalages ST dus à des changements de conduction ventriculaire ST-CC. Ils sont repérables par des complexes QRS particuliers, souvent élargis. Ces épisodes sont très peu représentés dans la base LTST (dans 3 enregistrements sur les 86). Pour des raisons de simplicité, de temps de calcul et de fiabilité statistique, ils ne seront pas pris en compte dans la suite de cette étude.

Les épisodes analysés respectent certaines contraintes, notamment au niveau de la déviation du segment ST observée [Jager et al., 2003]. Comme l'illustre la figure 5.1, pour qu'un épisode soit considéré, il doit présenter un décalage d'au moins V_{min} par rapport au niveau V_{ref} et pendant un temps T_{min}. Les instants notés T_{deb} et T_{fin} sont respectivement les débuts et fins des épisodes. Ils sont enregistrés dans les annotations de la base et servent de points de référence pour effectuer l'extraction des indicateurs.

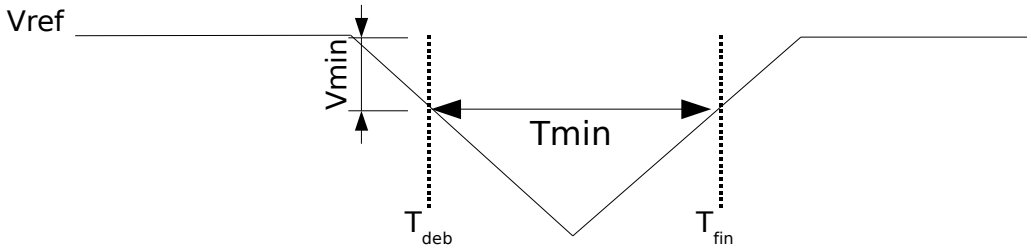


FIG. 5.1 – Définition d'un décalage du segment ST significatif en fonction de V_{min} et de T_{min}, tel que présenté dans la base LTST.

Les algorithmes de segmentation et d'optimisation des seuils présentés au chapitre 2, avec son algorithme d'optimisation des seuils, ont donc été utilisés. Un moyennage des battements sur une durée de 16 secondes a été appliqué avant la segmentation, en accord avec la méthodologie d'analyse du segment ST employée dans [Smrdel et Jager, 2004]. Au final nous récuperons un ensemble de séries temporelles dont un exemple pour chaque type d'épisodes (ST-IS, ST-RC et ST-AS) est illustré figure 5.2. La série labélisée "Annot." correspond à la série de l'amplitude du segment ST controlée par des experts et fournie dans la base de données. Il importe de souligner que l'extraction automatique effectuée par notre algorithme en est qualitativement très proche.

Pour nos analyses trois hypothèses ont été posées :
- **Sur l'amplitude et la durée des épisodes** : nous nous sommes intéressés aux épisodes enregistrés dans la base LTST sous le protocole B, c'est à dire que ces épisodes présentent une déviation d'au moins $V_{min} = 100\mu V$ pendant un temps $T_{min} = 60s$.
- **Sur la période extraite** : pour des raisons évidentes de temps de calcul, les indicateurs ne sont pas extraits sur la durée totale de chaque enregistrement Holter de la base LTST mais uniquement autour des épisodes de déviation du segment ST. Les annotations correspondant au protocole B ont donc été prises en compte pour retrouver le début et la fin de chaque épisode. Une fenêtre temporelle d'analyse de 15 minutes avant T_{deb} et de 5 minutes après T_{fin} est réalisée. Les indicateurs sont extraits dans cette fenêtre

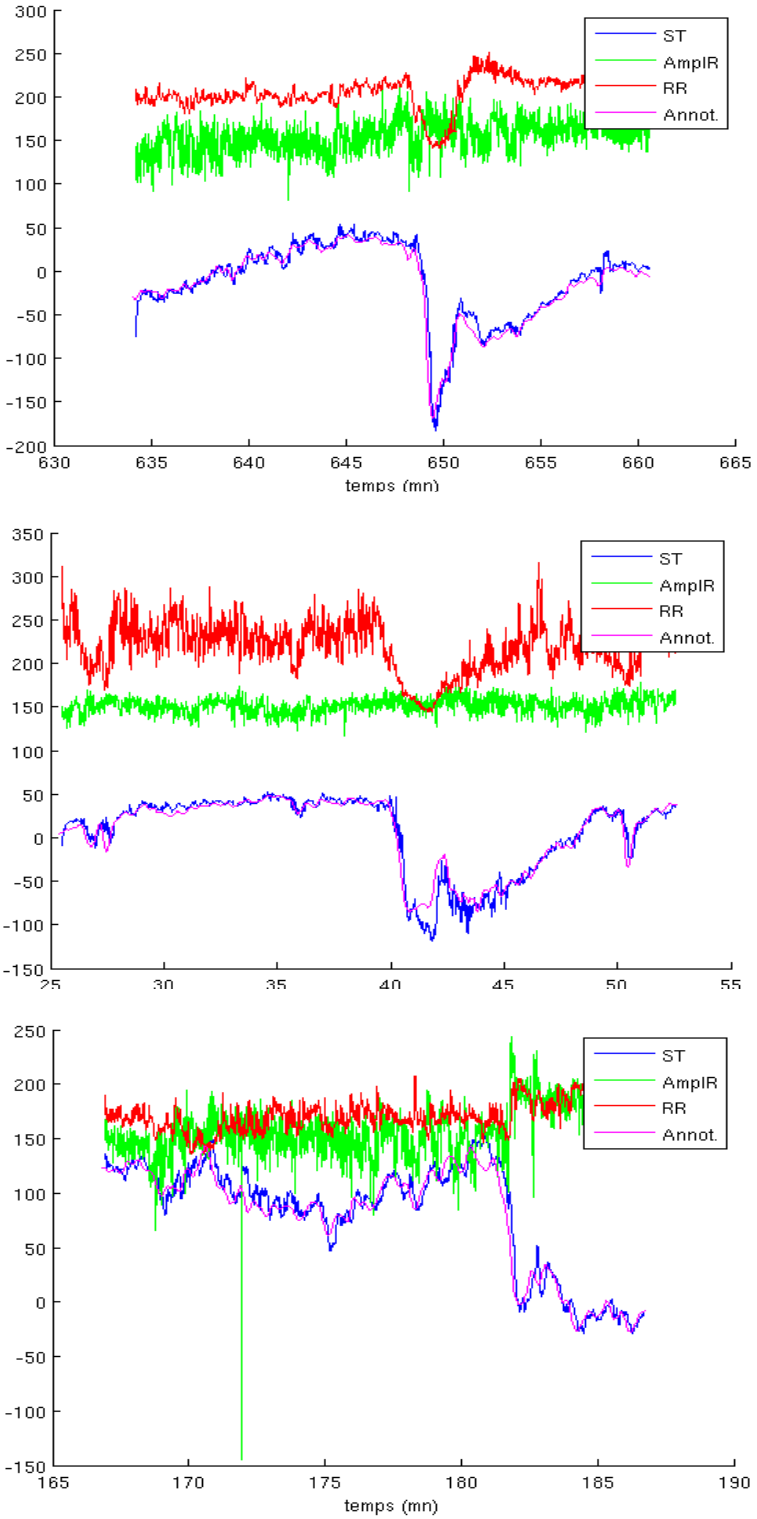


FIG. 5.2 – Exemples des trois types d'épisodes de déviations du segment ST extraits de la base LTST et représentés par différentes variables. ST : série ST extraite, en μV ; $AmplR$: amplitude de l'onde R, en dizaine de μV ; RR : intervalle RR, en nombre d'échantillons (à 250Hz) ; Annot : série temporelle de l'amplitude du segment ST controlée par des experts, en μV.

temporelle.
- **Sur le pré-traitement des indicateurs** : avant d'être présentées aux MSMC pour l'apprentissage, les séries temporelles de chaque indicateur sont normées et centrées.

5.1.1 Résultats en classification supervisée

La classification supervisée des épisodes ischémiques consiste à :

1. apprendre 3 modèles, M_{IS}, M_{RC}, M_{AS}, à partir des séries temporelles correspondantes. Plus précisément, il s'agit de trouver les paramètres $\theta_{opt}\{IS, RC, AS\}$ qui maximisent les probabilités $P(X_{\{IS,RC,AS\}}|\theta_{\{IS,RC,AS\}})$ où $X_{\{IS,RC,AS\}}$ désigne les ensembles d'apprentissage des trois groupes d'épisodes.
2. affecter les épisodes des ensembles de tests à un des trois modèles appris. Cette étape se base sur le maximum de vraisemblance :

$$k_{win} = \underset{k\in\{IS,RC,AS\}}{\operatorname{argmax}} \{P(O_i|M_k)\}$$

où O_i désigne l'épisode à tester.

Nous avons choisi de réaliser des ensembles d'apprentissage et de tests avec respectivement 2/3 et 1/3 des épisodes, tirés de manière aléatoire. Pour limiter la sensibilité des résultats aux individus tirés dans les ensembles de test et d'apprentissage, 15 tirages sont effectués. Les résultats présentés par la suite correspondent donc au taux d'erreur moyen (rapport entre le nombre d'épisodes mal classés et le nombre total d'épisodes dans l'ensemble de test) obtenus sur ces différents tirages.

5.1.1.1 Configuration de l'apprentissage

Afin d'optimiser la classification, il est important de configurer les séries temporelles sur lesquelles sont apprises les MSMC. Il faut notamment déterminer :
- les indicateurs dont les dynamiques conjointes sont les plus discriminantes entre les groupes.
- la fenêtre temporelle, définie autour du début de l'épisode T_{deb}, et dans laquelle des dynamiques significatives seront analysées. En effet, une fenêtre trop restreinte ne permettra pas d'acquérir les dynamiques qui discriminent les épisodes alors qu'une fenêtre trop grande introduira des observations qui ne sont plus liées aux épisodes ischémiques.

Une recherche empirique des indicateurs et de la fenêtre optimales a donc été effectuée, ceci en supposant que ces deux paramètres puissent être optimisés de manière indépendante. Les ensembles d'indicateurs sont tout d'abord testés avec une fenêtre arbitrairement fixée entre T_{deb} - 180 s et T_{deb} + 180 s puis cette fenêtre est ajustée avec les variables précédemment trouvées. Les tableaux 5.1 et 5.2 montrent les taux d'erreurs obtenus respectivement en fonction des variables et de la fenêtre temporelle.

C'est donc avec les variables {Niveau ST, intervalles RR et Amplitude des ondes R et T} et sur un intervalle de temps de 4mn30 avant et de 3mn30 après T_{deb} que la discrimination entre les trois groupes d'épisodes ST est optimale. Les variables prises en compte sont conformes à celles jugées intéressantes dans la littérature : les variations de l'onde R pour l'évaluation des changements d'axe, les amplitudes du segment ST et de l'onde T pour la distinction des

Indicateurs	Niv ST	Niv ST RR	Niv ST $AmplR$	Niv ST RR $AmplR$	Niv ST RR $AmplR$ $QRSd$	Niv ST RR $AmplR$ Niv ST_{V2}	Niv ST RR $AmplR$ $AmplT$
Tx. Err. (%)	56.0	43.5	41.2	36.1	38.5	33.2	31.4

TAB. 5.1 – Recherche de l'ensemble de variables produisant le taux d'erreur de classification minimum. Niv ST_{V2} est l'amplitude du segment ST mesurée sur la seconde voie (la première étant celle où le décalage d'au moins V_{min} est observé en premier).

Deb \ Fin	T_{deb}+120s	T_{deb}+150s	T_{deb}+180s	T_{deb}+210s	T_{deb}+240s
T_{deb}-150s	31,9	31,8	31,3	32,4	33,8
T_{deb}-180s	31,4	31,8	31,4	31,5	32,7
T_{deb}-210s	32,6	30,5	30,1	31,1	31,3
T_{deb}-240s	33,4	32,2	30,7	29,2	30,3
T_{deb}-270s	33	30,8	31,4	29	29,3
T_{deb}-300s	32,7	30,9	30,6	29,8	29,1
T_{deb}-330s	31,4	30,2	30	30,3	29,6

TAB. 5.2 – Taux d'erreurs de classification (en %) obtenus en fonction de la position de la fenêtre d'analyse des séries temporelles. Les bornes gauche (en colonne) et droite (en ligne) sont toutes deux réferencées par rapport à T_{deb}.

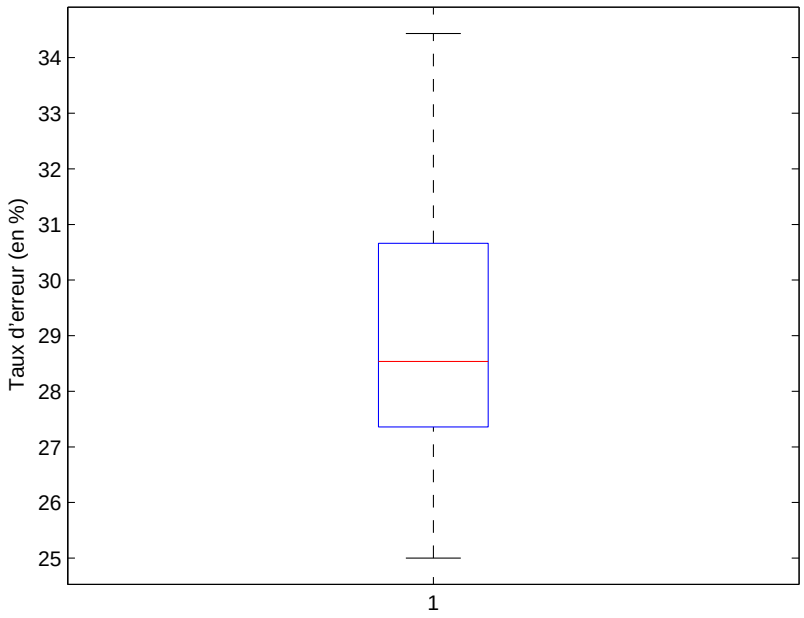
changements liés à l'ischémie et l'intervalle RR pour l'appréciation de l'influence du rythme cardiaque. Il est aussi à noter que ce sont les variables dont l'extraction est fiable : les pics des ondes, l'intervalle RR et l'amplitude du segment ST sont plus aisés à détecter que la durée du QRS ou l'intervalle QT.

L'impact des populations d'apprentissage et de test sur le taux d'erreur est aussi évalué en observant sa variation pour les 15 tirages. La boite à moustache obtenue est présentée figure 5.3. La moyenne et l'écart type sont donc respectivement de 29% et 2.9%, avec un minimum trouvé à 25% et un maximum à 34.4%. Ce biais, lié aux ensembles de tests et d'apprentissages, justifie la prise en compte de la moyenne sur les 15 tirages comme référence de résultats. Il demeure néanmoins raisonnable.

5.1.1.2 Comparaison des résultats de la classification

La table de contingence obtenue dans la configuration optimale est présentée table 5.4.

Ces résultats peuvent être comparés à [Zimmerman et al., 2003] et [Langley et al., 2003] qui ont utilisé la même base de données. [Langley et al., 2003] affichent le meilleur taux de classification (81.4%) mais appliquent leur algorithme sur des signaux ST qui ont été débarrassés manuellement des déviations liés aux changements d'axes et intègrent pourtant ce type d'épisodes dans leur classification. Il est à rappeler que les signaux que nous utilisons n'ont pas été traités manuellement et sont directement extraits de l'algorithme présenté chapitre 3. Les résultats de classification sont donc principalement comparables à ceux de [Zimmerman et al., 2003]. En effet ces derniers travaillent également directement sur les signaux ECG, en intégrant le segment ST et l'onde T dans un espace de phase reconstruit (EPR) puis en classant les signaux dans l'EPR à l'aide de modèles de mixtures de gaussiennes.



TAB. 5.3 – Boxplot des taux d'erreur réalisés sur 15 tirages différents de la population d'apprentissage et de test, pour l'ensemble des variables et la fenêtre temporelle trouvés à partir des tableaux 5.1 et 5.2.

	Classifié en tant que				
	ST-IS	ST-AS	ST-RC	Sensibilité	Sensibilité (ZI)
ST-IS	852	163	95	76.8%	80.6%
ST-AS	402	1166	82	70.7%	48.6%
ST-RC	150	42	228	54.3%	33.3%
Tx d'erreur (en %)				29.0	46.7

TAB. 5.4 – Résultats de la classification sur les épisodes de l'ensemble de test. La dernière colonne présente les résultats (sensibilité et taux d'erreur) publiés dans [Zimmerman et al., 2003].

Ils présentent aussi des résultats détaillés, adaptés pour réaliser une comparaison des performances. Les sensibilités de détection des différentes classes sont donc présentées dans la table 5.4, ainsi que la table de contingence obtenue par notre approche. Le taux d'erreur final obtenu sur l'ensemble de test est de 46.7% pour [Zimmerman et al., 2003] contre 29% dans notre cas.

5.1.2 Résultats en classification non supervisée

L'algorithme de clustering est ici appliqué sur les épisodes ischémiques, en conservant les variables et la fenêtre optimale trouvée en classification. Il convient de rappeler qu'il s'agit ici de trouver naturellement les classes sans connaissance *a priori*.

L'expérience montre que les résultats, obtenus à l'issue de la première itération de l'algorithme de clustering, ne sont pas identiques suivant les individus aléatoirement assignés aux deux modèles lors de l'initialisation. Ceci est dû au fait que l'algorithme EM converge vers un minimum local de nos données qui dépend de l'initialisation. Pour rendre notre algorithme plus robuste à l'initialisation, la solution proposée est de réaliser plusieurs tirages et de moyenner les vraisemblances obtenues à l'issu de l'apprentissage. Au préalable au moyennage, il est aussi important de vérifier la concordance des modèles M_1 et M_2 entre les itérations. En effet le modèle M_1 d'une itération ne correspond pas obligatoirement au modèle M_1 d'une autre itération : le groupe qu'il représente peut être aléatoirement indexé par M_1 ou M_2. Les indices des modèles sont donc parfois changés de manière à maximiser la concordance entre les groupes.

A l'issu du clustering avec deux modèles, le plan de vraisemblance, figure 5.3, a été obtenu et la table 5.5 représente la répartition des 3 types d'épisodes dans ces deux modèles.

	ST-IS	ST-AS	ST-RR
M_1	197	148	79
M_2	21	177	2

TAB. 5.5 – Répartition des épisodes dans les deux modèles créés par clustering.

	ST-IS	ST-AS	ST-RR
M_1	20	118	4
M_2	68	145	14
M_3	130	62	63

TAB. 5.6 – Répartition des épisodes dans 3 modèles créés par clustering.

L'expérience montre que les épisodes ST-AS ressortent du clustering, avec le modèle M2 qui leur est quasiment dédié. Le taux de séparation, calculé en considérant que les épisodes ST-IS et ST-RR sont affectés à M1 et que les épisodes ST-AS sont affectés à M2, est de 72%. Il n'est pas possible de distinguer les épisodes ST-IS des ST-RC, ceci semble indiquer que les différences entre ces deux groupes soient moins importantes que les différences entre le groupe ST-AS et les autres ou même que les différences intra-groupes. Ce résultat peut s'expliquer par la variété importante des pathologies (sténose mitrale, syndrome de Wolf-Parkinson-White, hypertension) des patients dont sont extraits les épisodes ST-RC, empêchant ainsi leur regroupement. Passer à 3 modèles (table 5.6) conduit toujours à une séparation des épisodes ST-AS des autres épisodes. Ceci est visible dans le plan de vraisemblance des modèles

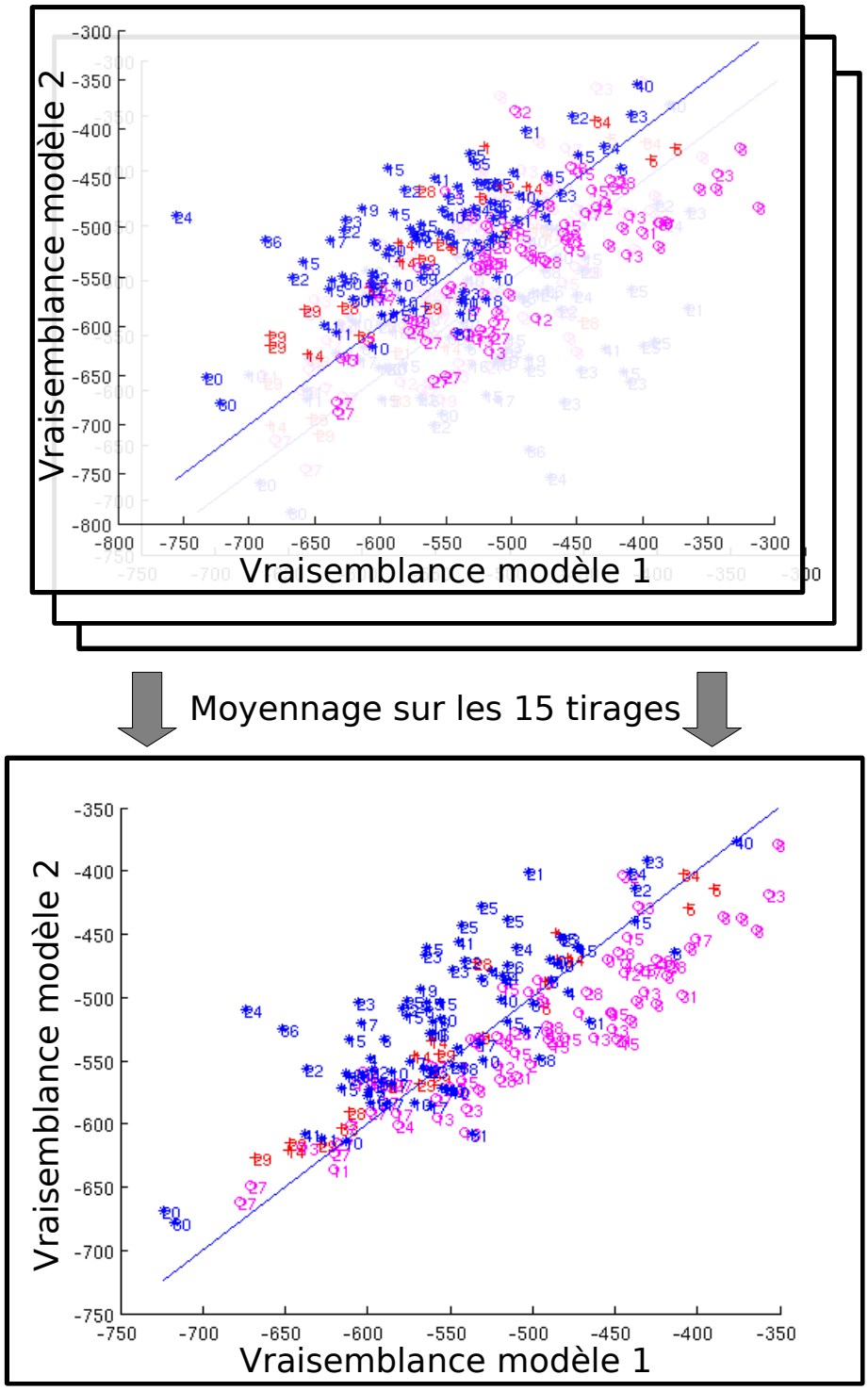


FIG. 5.3 – Plan de vraisemblance des épisodes de la base LTST, avec deux modèles et un moyennage sur 15 tirages.

1 et 3, figure 5.4. Augmenter le nombre de modèles (des tests ont été réalisés jusqu'à 6 modèles) n'a pas permis d'améliorer la séparation des différents groupes.

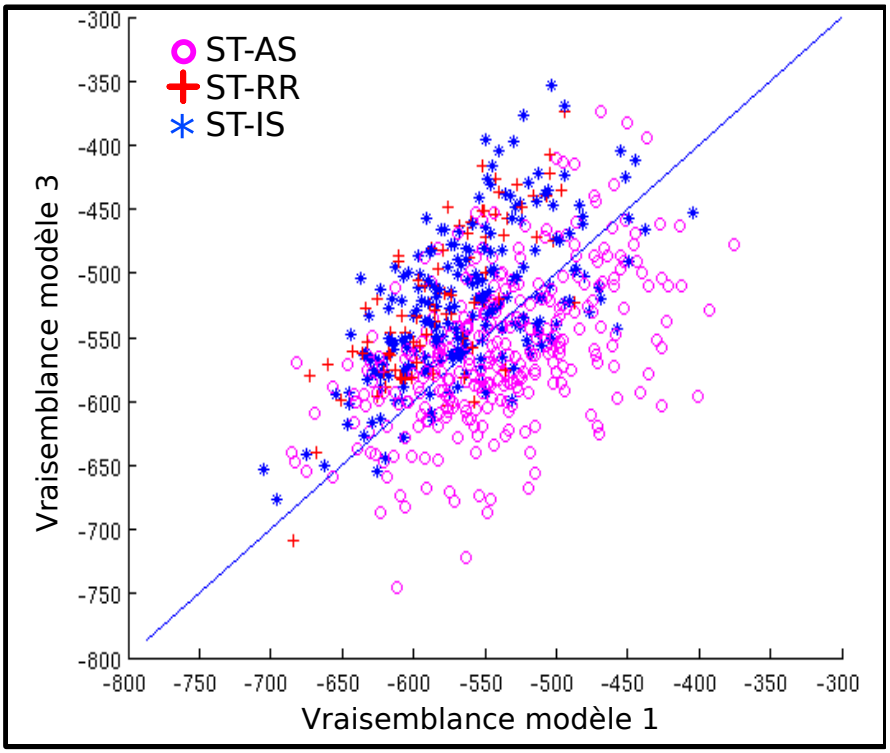


FIG. 5.4 – Episodes de la base LTST répresentés par leur vraisemblance d'être générés par les modèles M_1 et M_3 et créés lors du clustering.

5.1.3 Discussion

La méthode proposée permet la classification des déviations du segment ST en analysant les dynamiques des indicateurs extraits de l'ECG sur une fenêtre temporelle. Après sélection des indicateurs pertinents et de la fenêtre temporelle d'analyse, un taux d'erreur de classification de 29% est obtenu. Bien qu'inférieur à d'autres résultats publiés dans la littérature, ce taux d'erreur demeure relativement élevé. Toutefois, il est important de noter que nous travaillons uniquement sur des données électrocardiographiques alors qu'en pratique les cardiologues portent un diagnostic en intégrant d'autres sources d'informations (cliniques, historiques ...). La finalité du classifieur présenté ici pourrait donc être de dénombrer, à titre indicatif, les épisodes ischémiques survenant au cours d'enregistrements de type Holter et en limitant le nombre de faux positifs par rapport aux approches existantes.

Pour le clustering, l'espace des vraisemblances obtenu est intéressant pour représenter l'ensemble des épisodes, même si seulement les épisodes de types ST-AS se différencient aisément, avec un taux de séparation de 72%.

Il est aussi à noter que la détection de l'ischémie aiguë non induite, à partir d'enregistrements Holter par exemple, est particulièrement difficile par rapport à la détection d'ischémies induites par angioplastie. Dans ce second cas, et avec des critères de décision basés sur le vectocardiogramme, une sensibilité de 98% et une spécificité de 96% ont été reportés dans [Fayn et al., 2007].

5.2 Application sur le syndrome de Brugada

Le syndrome de Brugada, décrit par les frères Brugada en 1992 [Brugada et Brugada, 1992], est une maladie génétique rare, affectant les canaux sodiques et cause de mort subite par fibrillation ventriculaire. Il serait responsable de 4% à 12% de l'ensemble des morts subites et de 20% des morts subites sans cardiopathie structurelle sous-jacente. Sa prévalence est estimée à 5/10000 mais est très difficile à évaluer dans la population générale d'autant qu'il existe une forte disparité géographique [Antzélévitch et al., 2005].

Comme le syndrome de Wolf-Parkinson-White, le syndrome de Brugada repose sur des éléments cliniques et électrocardiographiques. En ce qui concerne l'électrocardiographie, trois types d'ECG ont été définis. L'aspect typique est le type 1 et est caractérisé par un sus-décalage du segment supérieur à 0.2mV dans plus d'une dérivation précordiale droite (V1 à V3), avec un aspect en dôme (figure 5.5). Cependant il est montré que cette manifestation de type 1 est intermittente, d'où des risques de faux négatifs et l'intérêt porté aux types 2 et 3 qui sont permanents et suggèrent le Brugada. Le type 2 se distingue par un sus-décalage concave du segment ST, d'au moins 0.2mV et accompagné d'une onde T positive ou biphasique, tandis que le type 3 présente seulement un sus-décalage, convexe ou concave, du segment ST d'au moins 0.1mV.

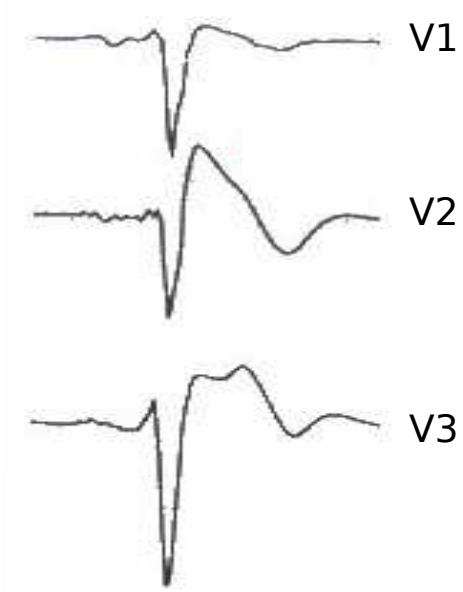


FIG. 5.5 – Syndrome de Brugada de type 1 observée dans les voies V1 à V3.

Au sein des patients atteints du syndrome de Brugada, deux groupes se distinguent :

le groupe "symptomatique" : il caractérise des patients ayant présenté des symptômes (arrêt cardio-circulatoire par fibrillation ventriculaire, syncopes par troubles du rythme ventriculaire ou syncopes inexpliquées, convulsions nocturnes, pertes d'urines nocturnes) associés à un aspect de Brugada de type 1 sur l'électrocardiogramme de surface. Cet aspect ECG de Brugada de type 1 peut être spontané ou induit par les tests pharmacologiques (ajmaline ou acétate de flécaïnide).

le groupe "asymptomatique" : il réunit des sujets indemnes de manifestations cliniques présentant sur l'ECG un aspect de Brugada de type 1. Chez ces patients, le phénotype ECG de Brugada est découvert soit lors de l'enquête familiale réalisée à partir d'un sujet symptomatique, soit fortuitement lors de la réalisation d'un ECG systématique.

5.2.1 Objectif

L'objectif de cette étude est l'analyse comparative des deux groupes symptomatique et asymptomatique suivant des critères électrocardiographiques relevés lors d'un test d'effort. En effet, l'influence du système nerveux autonome à la fois sur l'expression ECG du syndrome de Brugada et sur les troubles conductifs est fortement suspectée mais n'a été que très peu évaluée. Cette étude, intégrant 8 patients symptomatiques et 15 patients asymptomatiques, est basée sur un protocole d'effort en trois phases où l'ECG 12 dérivations standard est enregistré à une fréquence de 1kHz :

- **la phase de repos**, avec deux minutes en position allongée puis deux minutes en position assise sur le vélo ;
- **la phase d'exercice**, avec un départ d'une charge de 30 Watts (W) pour les femmes et de 50W pour les hommes. La charge est ensuite augmentée par paliers de deux minutes jusqu'au maximum d'effort. Pour les femmes, l'augmentation est de 20W par paliers et pour les hommes de 30W au début puis de 20W ;
- **la phase de récupération**, qui se décompose en une phase de récupération active, de trois minutes, avec un palier fixé à 20W pour les femmes et à 30W pour les hommes, suivi de la phase de récupération passive, de trois minutes sans aucun effort.

La section suivante expose les différents indicateurs extraits du signal ECG, en considérant les difficultés liées à l'ECG d'effort.

5.2.2 Indicateurs extraits

Les indicateurs extraits caractérisent à la fois les étages ventriculaire et atrial.

Pour l'étage atrial :
La durée de l'onde P et celle de l'intervalle PR sont mesurées sur la dérivation standard DII. Cependant, afin de nous affranchir de mesures inexactes, liées au rythme cardiaque élevé au maximum d'effort et au chevauchement fréquent des ondes P et T (figure 5.6), nous avons préféré étudier de façon arbitraire l'intervalle $PfinR$ mesuré entre la fin de l'onde P et le début du QRS. La durée de l'onde P, quant à elle, est exclue de l'analyse. Le nouvel intervalle $PfinR$ est défini par la formule : durée de PfR = durée de PR - durée de P.

Pour l'étage ventriculaire :
Les durées de l'intervalle QRS et de l'intervalle QT sont mesurées dans les dérivations précordiales V1 ou V2, en fonction de la qualité et de la reproductibilité de la mesure de ce paramètre dans ces dérivations. Cependant, toujours en raison des superpositions des ondes P et T, la fin de l'onde T est souvent indétectable à effort élevé. Nous avons donc préféré étudier, de façon arbitraire, l'intervalle $QTpic$, mesuré entre le début du QRS et le pic de l'onde T.

Afin de nous affranchir de l'influence de la fréquence cardiaque sur l'intervalle $QTpic$, cet intervalle $QTpic$ est corrigé selon deux méthodes :

- au repos : en appliquant la formule de Bazett ($QTcb = \frac{QT}{\sqrt{RR}}$), formule la plus usitée et reconnue valable pour des fréquences cardiaques comprises entre 60 et 80 battements par

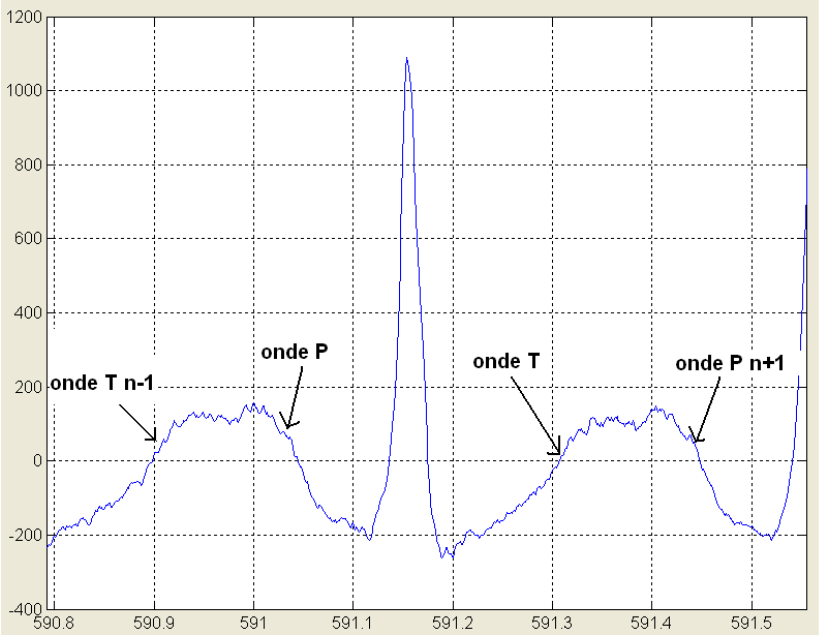


FIG. 5.6 – Exemple de problème de chevauchement des ondes P et T apparaissant aux alentours du maximum d'effort et entraînant l'impossibilité de détecter correctement les fins d'ondes T et les débuts d'ondes P.

minute ; en revanche, elle n'est pas appliquée pour corriger le $QTpic$ à l'effort en raison du risque connu de surestimation du QT corrigé par cette formule pour des fréquences cardiaques élevées.
- lors de l'épreuve d'effort : en appliquant la formule de Fridericia ($QTcf = QT * RR^{-1/3}$).

Le point J (jonction entre la fin du QRS et le début du segment ST) est également étudié dans ces mêmes dérivations V1 et V2. Il s'agit du seul paramètre soumis uniquement à un positionnement manuel par les opérateurs, grâce à l'interface homme-machine de la station d'analyse développée et décrite au chapitre 2. L'élévation, par rapport à la ligne isoélectrique, du point ainsi positionné a ensuite été déterminée de façon automatique.

Les variables RR, $QRSd$, $QTpic$, $PfinR$ et $AmplJ$ ont donc été enregistrées au cours de l'épreuve d'effort pour être ensuite exploitées dans l'analyse des différents patients. Les variables RR, $QRSd$ et l'amplitude au niveau du point J sont typiques pour analyser le rythme cardiaque (le RR), la durée de la dépolarisation ventriculaire (le $QRSd$) et vérifier la présence d'un palier entre la dépolarisation et la repolarisation (l'amplitude au point J, $AmplJ$). A l'effort, les ondes P et T présentent des chevauchements, rendant difficile la détection de la fin de l'onde T et le début voire même le pic de l'onde P. Les intervalles QT et PR habituellement analysés ont donc été remplacés par les intervalles $QTpic$ et $PoffR$.

L'enregistrement de ces variables a été effectué en trois étapes :
- **E1**, la segmentation de l'ensemble des battements de manière automatique avec l'algorithme présenté en 3.2.
- **E2**, la vérification par le cardiologue d'un battement à la fin de chaque palier. Ces battements sont choisis pour leur aspect (faible bruit et morphologie typique) et la segmentation automatique est éventuellement corrigée. Ils sont ensuite enregistrés.

– **E3**, la segmentation semi-automatique, de l'ensemble des battements, suivant la méthode présentée en 3.5.1 et en ré-utilisant les battements de E2.

Des analyses quantitatives de ces données, orientés vers la discrimination des groupes "Symptomatiques" et "Asymptomatiques", ont été menées en appliquant une analyse factorielle dans un premier temps et une modélisation par les MSMC dans un second temps. L'analyse factorielle basée sur l'Analyse en Composantes Principales (ACP) permettra notamment d'exprimer les corrélations entres variables et de retrouver celles qui semblent être les plus discriminantes entre les deux groupes. Elle sera effectuée sur des battements enregistrés dans l'étape E2 et ne prendra pas en compte les dynamiques. Ce dernier point sera justement abordé avec une adaptation de l'Analyse Factorielle des Correspondances Multiples proposée au laboratoire [Gueguin et al., 2008] puis par les MSMC, en réalisant un modèle pour chacun des deux groupes. Dans ces deux derniers cas ce seront les segmentations de l'étape E3 qui seront utilisées.

5.2.3 Investigation par analyses factorielles

Devant le peu de connaissances sur la distinction des groupes symptomatiques et asymptomatiques, il nous a semblé opportun d'explorer notre base de données sans utiliser d'*a priori* sur les patients et sur les indicateurs. L'objectif de cette analyse factorielle est donc de pouvoir représenter tous les patients (symptomatiques et asymptomatiques) dans un repère commun pour mettre à jour leurs différences/ressemblances suivant les indicateurs extraits. Les analyses en composantes principales sont tout particulièrement adaptées à ce problème. Elle permettent notamment de transformer un tableau de données, usuellement avec les individus en ligne et les variables en colonne, en un autre tableau, avec toujours les individus en ligne mais avec des nouvelles variables (les composantes principales) en colonne. Les composantes principales sont calculées de façon à maximiser la variance entre les individus et en réalisant une base orthonormée. Elles peuvent aussi être rangées suivant l'importance de la variance qu'elles expliquent. La base orthonormée construite à partir des composantes expliquant le plus de variance présente donc un repère tout à fait adapté pour représenter les individus. Cependant la principale difficulté inhérente à ce type d'analyse est la prise en compte du temps, lorsque les individus dans le tableau de données initiales correspondent à des séries temporelles. La solution généralement proposée consiste à sélectionner les instants les plus importants (par introduction d'*a priori* sur les données) et à considérer les variables à chacun de ces instants. Cette démarche a été suivie dans un premier temps pour analyser les séries temporelles des patients atteints du Brugada.

5.2.3.1 Analyses en Composantes Principales

Comme expliqué précédemment, une première analyse est effectuée à l'aide d'une ACP sur nos données en sélectionnant plusieurs instants spécifiques. Quatre instants particuliers (figure 5.7), choisis par un cardiologue pour leur correspondance aux différentes étapes de l'épreuve d'effort, sont ici considérés :

– le repos en position assise sur le vélo (I1), au début de l'épreuve, juste avant de pédaler,
– au maximum de l'effort (I2),
– la fin de la récupération active (I3), juste avant d'arrêter de pédaler et donc 3 minutes après I2,
– la fin de la récupération passive ou fin de l'épreuve d'effort (I4), 3 minutes après I3.

Toujours dans l'optique de mettre à jour des différences entre les deux groupes de patients, une analyse directe sur les variables et aux 4 instants est effectuée, suivie par une analyse des variations entre chaque instants.

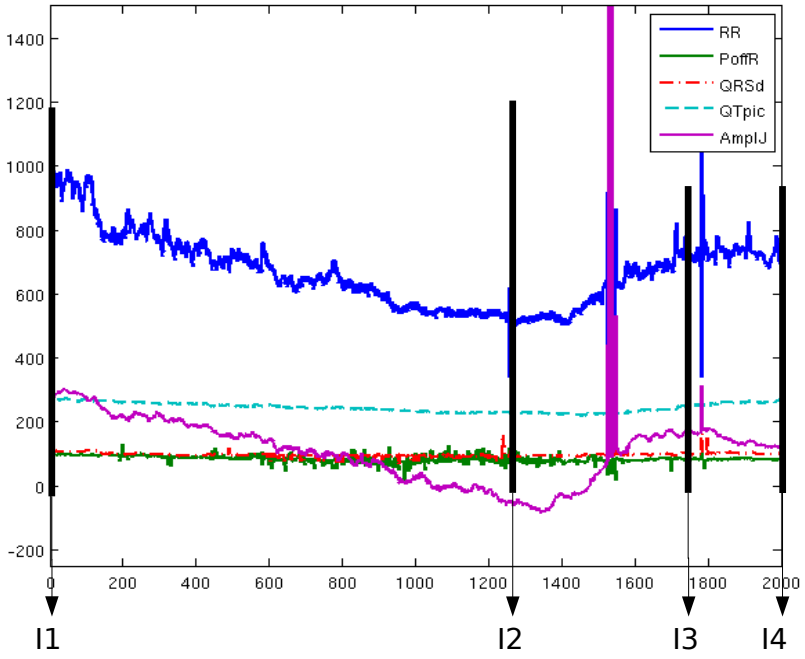


FIG. 5.7 – Séries temporelles disponibles et instants analysés.

Analyse des données initiales brutes

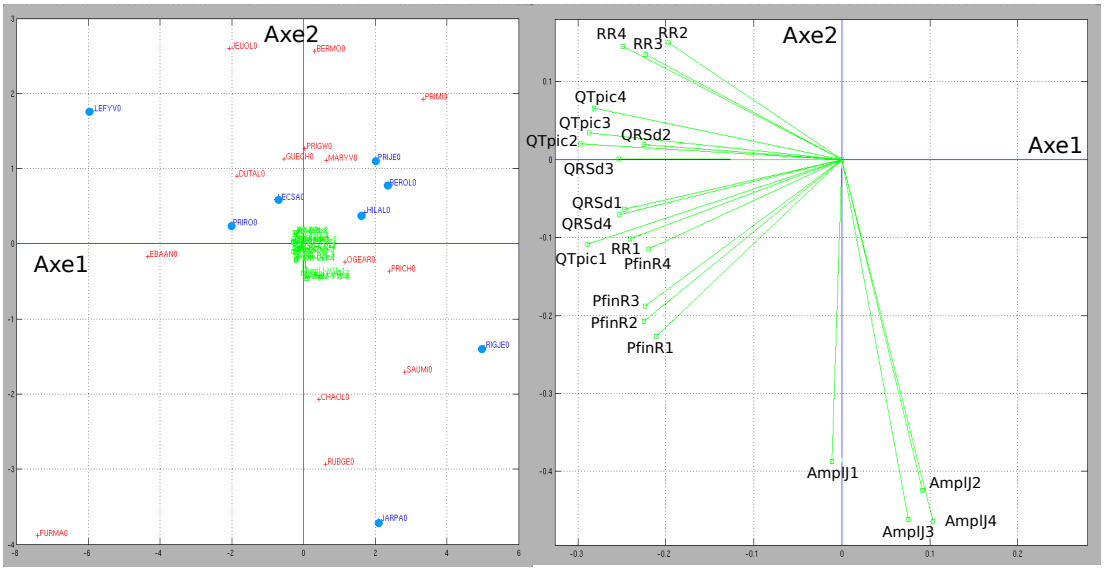
Le premier tableau de données pris en compte dans l'ACP se présente avec en ligne les 8 patients symptomatiques et les 15 patients asymptomatiques, et en colonne 20 variables : chacune des 5 variables initiales(RR, $QTpic$, $PfinR$, $QRSd$, $AmplJ$), indexées suivant les 4 instants choisis. Ces 20 variables sont normées et centrées. L'application de l'ACP sur le tableau ainsi obtenu fourni les valeurs propres de la table 5.7, pour les axes 1 à 7.

n	λ	%	% cumulé
1	9.1	45.46	45.46
2	3.53	17.67	63.14
3	2.78	13.89	77.03
4	1.93	9.66	86.69
5	0.88	2.17	91.09
6	0.43	1.78	93.26
7	0.36	1.60	95.03

TAB. 5.7 – Valeurs propres des composantes principales

Les quatres premiers axes expliquent une portion très importante (87%) de la variabilité présente dans le tableau de données et représentent tous plus que la variabilité moyenne contenue dans une variable (car leurs valeurs propres sont supérieures à 1). Pour des raisons pratiques, nous n'afficherons ici que les projections des individus et des variables sur les deux premiers axes (figure 5.8) qui représentent tout de même 63% de l'inertie totale, soit une perte

inférieure à 40% lors du passage de $\mathbf{R}^{20}$ à $\mathbf{R}^2$.



a) représentation des individus sur les axes 1 et 2

b) représentation des variables

FIG. 5.8 – Analyse en composantes principales sur les données initiales. a) la projection des individus sur les deux premiers axes (individus symptomatiques en points bleus et individus asymptomatiques en croix rouges), en b) la projection des variables (zoom), pour évaluer leur importance dans la constitution de ces axes.

La représentation des variables (figure 5.8-b) permet d'évaluer leurs corrélations et leurs contributions aux axes 1 et 2 : les variables finales, prises aux différents instants, sont principalement regroupées suivant les variables initiales, ce qui montre que les variations au cours de l'épreuve d'effort expliquent une variabilité moins grande que celles observées entre deux variables distinctes. On retrouve aussi des corrélations logiques, avec le *QTpic* fortement corrélé avec le *RR* alors que l'amplitude du point *J* est beaucoup moins corrélée avec les autres variables et participe principalement à la construction de l'axe 2.

La projection des individus sur ces deux axes ne fait pas ressortir de distinction particulière entre les deux groupes d'individus. D'autres combinaisons d'axes ont été explorées (projections sur Axe1/Axe3, Axe2/Axe3, ...) mais n'ont pas montrées de différences en terme de séparation des groupes.

Analyse des variations entre les instants I1 à I4

Pour évaluer principalement l'évolution au cours de l'effort, il a été choisi de représenter les individus par leur variation entre chacun des 4 instants conservés. Nous obtenons donc, comme variables finales, 3 valeurs pour chacune des variables initiales, telles que présentées table 5.8. Ces données sont ensuite normées et centrées.

L'analyse des valeurs propres des composantes trouvées par l'ACP (table 5.9) montre que les 3 premières composantes suffisent à expliquer 57.9% de la variance entre les individus. Des valeurs propres supérieurs à 1 sont relevées sur les six premiers axes, et une inspection graphique de nos individus sur ces six axes (pris deux à deux) a donc été effectuée. Cependant il apparaît que dans le cadre de la séparation des patients symptomatiques et asymptomatiques, les deux premiers axes sont les plus intéressants.

	$\Delta(I2-I1)RR$ RR_1	$\Delta(I3-I2)RR$ RR_2	$\Delta(I4-I3)RR$ RR_3	$\Delta(I2-I1)QTpic$ $QTpic_1$	$\Delta(I3-I2)QTpic$ $QTpic_2$	...
8 individus symptomatiques						
15 individus asymptomatiques						

TAB. 5.8 – Configuration du tableau de données de la deuxième ACP.

n	λ	%	% cumulé
1	4.2	28.05	28.05
2	2.43	16.18	44.24
3	2.05	13.69	57.9
4	1.69	11.28	69.20
5	1.42	9.44	78.64
6	1.05	7.01	85.655
7	0.56	3.73	89.39

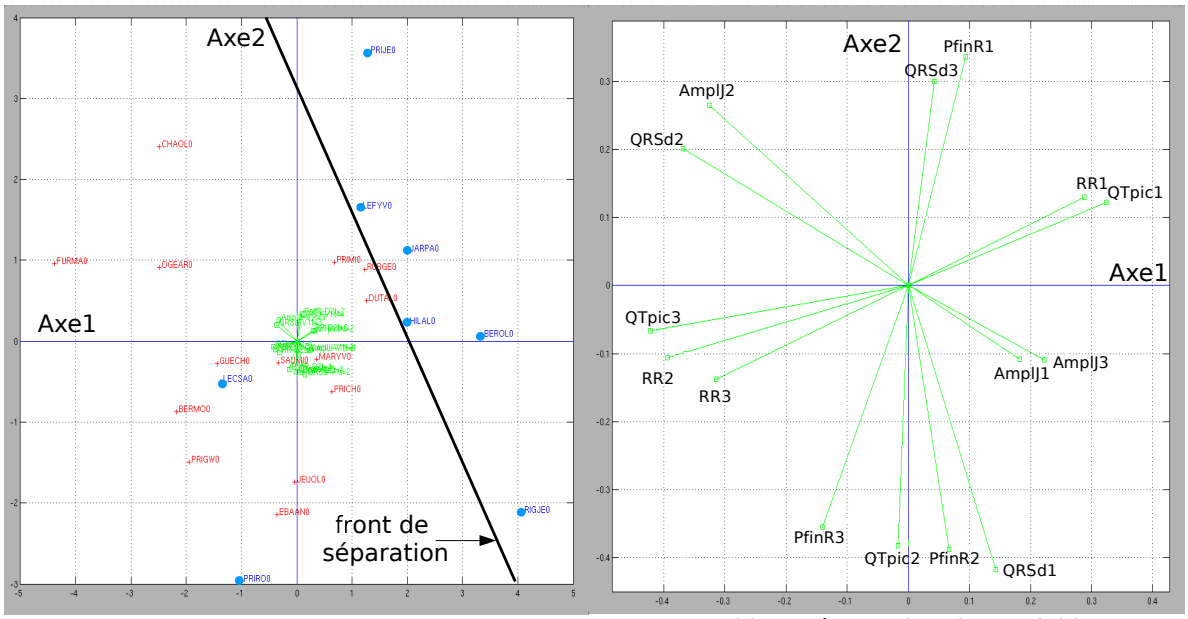
TAB. 5.9 – Valeurs propres des composantes principales.

La projection des individus sur les deux premiers axes est reproduite figure 5.9 ainsi que la contribution de chaque variable à la construction de ces deux axes. Il est notamment important de relever que les variables concernant le RR et le QT sont très corrélées (sauf le $QTpic2$) et participent majoritairement à la construction de l'axe 1. Les variables d'indices 2 et 3, qui correspondent à l'évolution pendant la période de récupération, sont aussi généralement opposées aux variables d'indices 1, qui correspondent à l'évolution jusqu'au maximum d'effort. Pour l'axe 2 ce sont principalement les variables $PoffR$ ($PoffR1$, $PoffR2$ et $PoffR3$) et $QRSd$ ($QRSd1$ et $QRSd3$) qui interviennent.

La position des individus sur ces deux axes reflètent qu'une frontière entre les deux groupes peut être tracée et la direction discriminante correspond principalement aux variables RR et $QTpic$ ($QTpic$ 1 et 3). Il est à noter que deux patients symptomatiques (LECSA et PRIRO) se retrouvent positionnés au milieu des patients asymptomatiques. Ces patients ont été examinés plus en détail par un cardiologue et il ressort, notamment dans le cas de LECSA, que son ECG peut être assimilé à celui des patients asymptomatiques. Ceci peut expliquer sa position dans notre analyse mais semble confirmer que les symptômes de ce patient ne s'accompagnent pas d'un comportement anormal sur l'ECG d'effort.

Cette analyse permet donc d'observer que les sujets symptomatiques présentent, à l'effort, un moindre raccourcissement de l'intervalle RR par rapport aux sujets asymptomatiques. En récupération active et passive, ils présentent aussi un moindre allongement. Ceci est visible lorsque les dérivées du signal RR sont tracées pour les deux groupes, figure 5.10.

Pour vérifier le pouvoir discriminant de cette variation de l'intervalle RR, un test d'hypothèse de Man-Witney a été appliqué sur la variable qui semble être la plus significative : le ΔRR entre le maximum d'effort et le repos ($\Delta(I2-I1)RR$). Ce test porte sur l'identité des distributions (non-paramétriques) des données qui lui sont fournies. Pour la variable citée et pour les deux groupes de patients, une p value de 0.020 est estimée, ce qui est significatif ($<$ 0.05).



a) représentation des individus sur les axes 1 et 2

b) représentation des variables

FIG. 5.9 – Analyse en composantes principales pour l'analyse des variations entre les instants I1 à I4. a) la projection des individus sur les deux premiers axes (individus symptomatiques en points bleus et individus asymptomatiques en croix rouges), en b) la projection des variables (zoom), pour évaluer leur importance dans la constitution de ces axes.

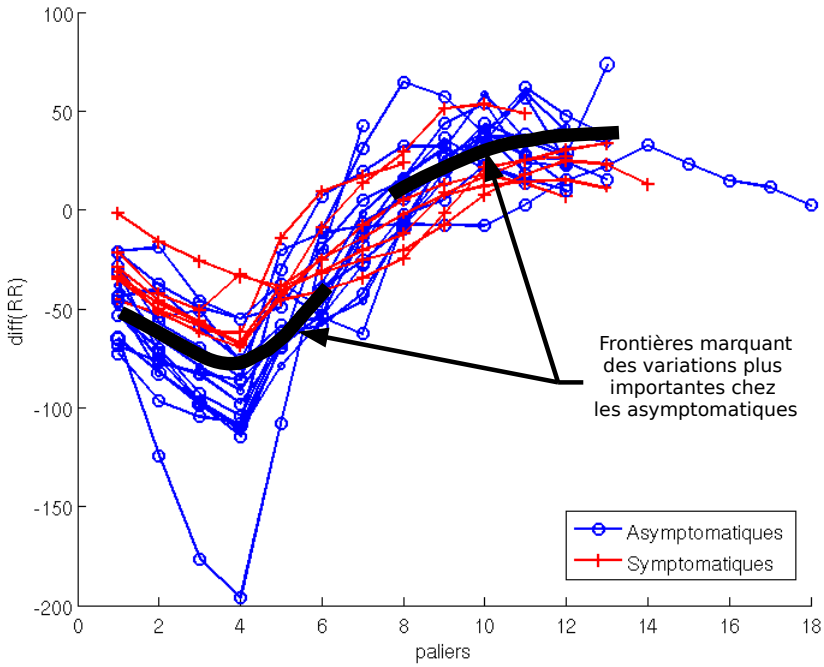


FIG. 5.10 – Variations de la variable RR entre chaque paliers pour les deux groupes de patients.

Enfin, en classification supervisée, la construction d'un arbre de régression (figure 5.11), à partir de l'algorithme décrit dans [Breiman, 1993], montre aussi que l'intervalle RR intervient de manière prépondérante pour la discrimination des deux groupes d'individus. En séparant les deux groupes suivant le premier seuil lié au RR, seulement trois erreurs sont commises : un patient symptomatique (S) dans le groupe des asymptomatiques (AS) et deux patients AS dans le groupe S. Ensuite, il apparait que c'est la durée du complexe QRS et l'intervalle $PfinR$ qui interviennent dans la séparation des deux groupes. Cependant, le nombre de patients étudiés ici est trop faible pour conclure sur l'influence de ces variables.

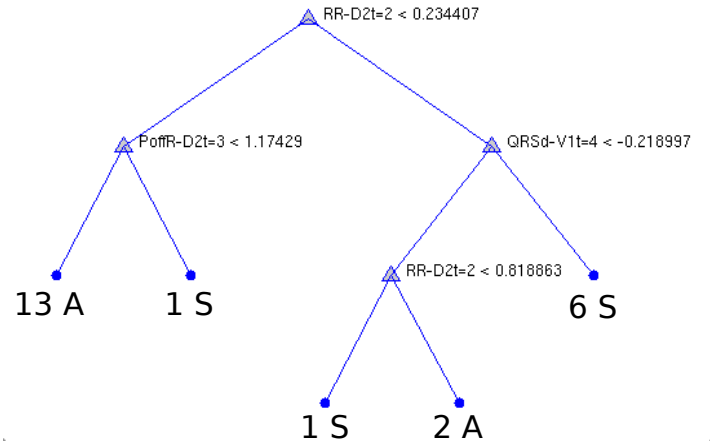


FIG. 5.11 – Arbre de décision pour la discrimination des deux groupes symptomatique et asymptomatique. Le nombre d'individus à l'extrémité de chaque feuille est indiqué, avec la lettre A pour les asymptomatiques et S pour les symptomatiques.

5.2.3.2 Analyses Factorielle des Correspondances Multiples (AFCM)

L'analyse précédente a permis de montrer que la prise en compte de la dynamique semble importante pour discriminer les patients symptomatiques des patients asymptomatiques. En parallèle à cette étude, le laboratoire a mis au point une méthode d'analyse de séries temporelles, fondée sur l'AFCM ([Gueguin et al., 2008]). Il était donc intéressant de juger cette méthode sur les données de Brugada.

Succinctement, l'AFCM est destinée aux données qualitatives et exploite des tableaux disjonctifs. Les séries temporelles multivariées extraites des épreuves d'efforts doivent donc être transformées. L'application de l'AFCM se décompose alors en deux étapes : (1) la création du tableau de données par codage spatio-temporel flou [Loslever et Bouilland, 1999], (2) l'analyse factorielle en elle-même, effectuée sur ce tableau de données. Comme dans l'ACP, des variables factorielles sont calculées, sur lesquelles il est possible de projeter les variables initiales et les individus. Le codage spatio-temporel flou, illustré figure 5.12, considère le domaine temporel de chaque variable à travers un ensemble de fenêtres temporelles floues $T = \{T_1, ..., T_j, ...T_{N_t}\}$ où $\mu_{T_j}(t_q)$ désigne la valeur d'appartenance, de l'instant q_t à la fenêtre temporelle T_j. Elle est comprise dans l'intervalle $[0, 1]$ et respecte la condition $\sum_j \mu_{T_j}(t_q) = 1$. De même, les valeurs $V_n(t_q)$, (la variable n et à l'instant t_q) sont affectées aux fenêtres spatiales floues d'un ensemble $S_n = \{S_{1,n}, ..., S_{i,n}, ..., S_{N_s,n}\}$, avec une appartenance $\mu_{S_{i,n}}$.

La valeur d'appartenance à la fenêtre spatio-temporelle $W_{i,j}^n$, pour la variable V_n est définie par [Loslever et Bouilland, 1999] :

$$\mu_{W_{i,j}^n} = \frac{1}{\sum_{q=1}^{Q} \mu_{T_j}(t_q)} \sum_{q=1}^{Q} \mu_{T_j}(t_q).\mu_{S_{j,n}}(V_n()t_q) \tag{5.1}$$

avec Q le nombre d'instants dans la série temporelle.

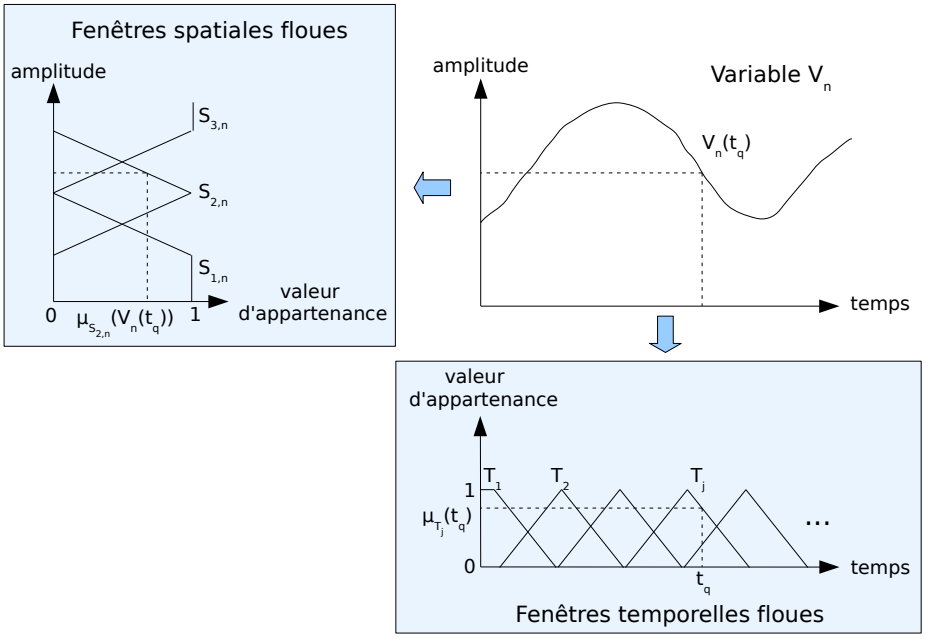


FIG. 5.12 – Codage temporel et spatial flou d'un signal continu.

Par expérimentation, nous avons choisi une configuration avec des fenêtres temporelles de taille fixe (5 secondes et des signaux échantillonnés à 2Hz) et avec une description de chaque variables en 9 modalités. Le tableau d'analyse prend alors la forme suivante :

	$S_{1,1}$	$S_{2,1}$	...	$S_{N_s,1}$	$S_{1,2}$	...	$S_{N_s,2}$	...	$S_{i,n}$	...	$S_{1,N}$	...	$S_{Ns,N}$
T_1													
⋮													
T_J									$\mu_{W_{i,j}^n}$				
⋮													
T_{N_t}													

TAB. 5.10 – Tableau de données employés dans l'AFCM. N_s est fixé à 9 et N_t varie en fonction de la taille de la série temporelle.

Les différentes modalités sont projetées sur les axes 1 et 2 (figure 5.13). On observe que ce sont majoritairement les modalités des variables RR et $QTpic$, respectivement notées 1 et 4, qui contribuent aux 2 axes étudiés et qui sont par ailleurs très corrélées. Ceci rejoint les résultats obtenus avec l'ACP effectuée sur les variations entre les instants. Les variables $PfinR$ et $AmplJ$ (notées 2 et 5) ont une position plus centrée et contribuent moins à l'axe 1.

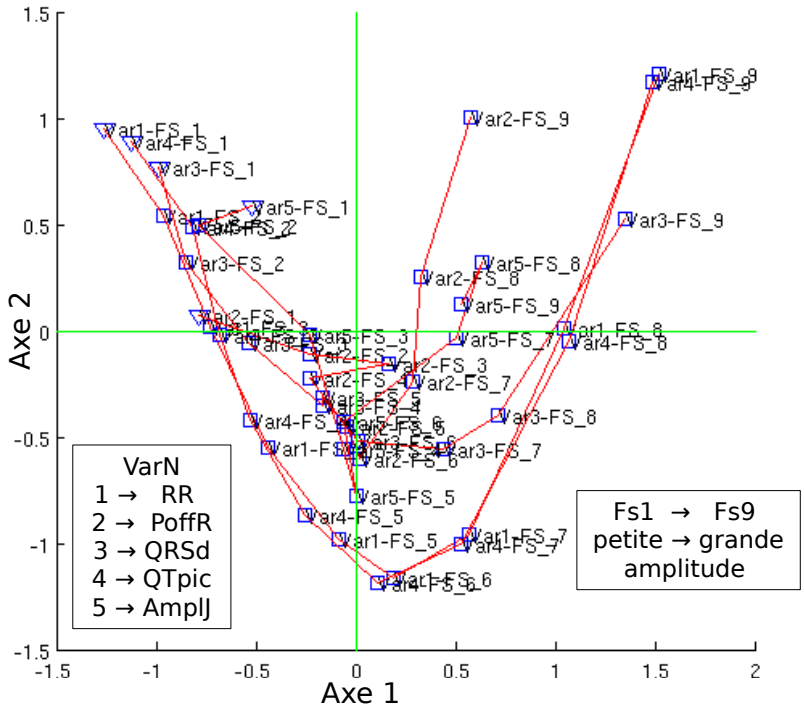


FIG. 5.13 – Les variables projetées sur les axes 1 et 2.

La trajectoire des individus est représentée figure 5.14. On retrouve bien une évolution correspondant aux différentes phases de l'épreuve d'effort : les individus partent de la position où le *RR* et le *QTpic* sont élevés (fenêtre spatiale 9) puis le rythme s'accélère jusqu'au maximum d'effort (fenêtres spatiales 1 ou 2, en fonction du maximum d'effort atteint) et finalement un retour vers les modalités plus élevées s'opère pendant la phase de récupération.

L'analyse des trajectoires souligne également des dynamiques d'amplitudes plus réduites pour les patients asymptomatiques. Ceci confirme les résultats fournis par l'ACP qui suggèrent que les patients asymptomatiques ont une amplitude plus importante que les symptomatiques. Il nous a paru important d'évaluer l'amplitude de ces dynamiques lorsque les séries temporelles sont projetées sur les axes 1 et 2. Une mesure d'étendue (variation maximale) a donc été effectuée sur ces deux axes et est représentée figure 5.15.

Il apparaît que les patients asymptomatiques ont en effet une étendue plus importante que les autres, avec quelques exceptions : le patient LECSA, symptomatique, présente des étendues élevées et 3 patients asymptomatiques ont des étendues faibles.

5.2.4 Application d'une classification par les MSMC

Comme déjà souligné, l'ACP revient à étudier les patients aux 4 instants choisis sur la durée totale de l'épreuve d'effort (environ une dizaine de minutes et dépendant du nombre de paliers avant d'atteindre le maximum d'effort). Il était donc important d'apprécier l'impact de notre méthodologie à base de MSMC sur l'épreuve d'effort complète et de le resituer par rapport à l'AFCM. Les MSMC ont donc été appliqués sur les séries temporelles extraites et centrées (figure 5.7). La normalisation effectuée dans les analyses factorielles n'est pas ici nécessaire car les états des MSMC, basés sur les gaussiennes multivariées, s'accommodent bien des différences d'ordre de grandeur des variables qu'ils modélisent. La mise en oeuvre, comme

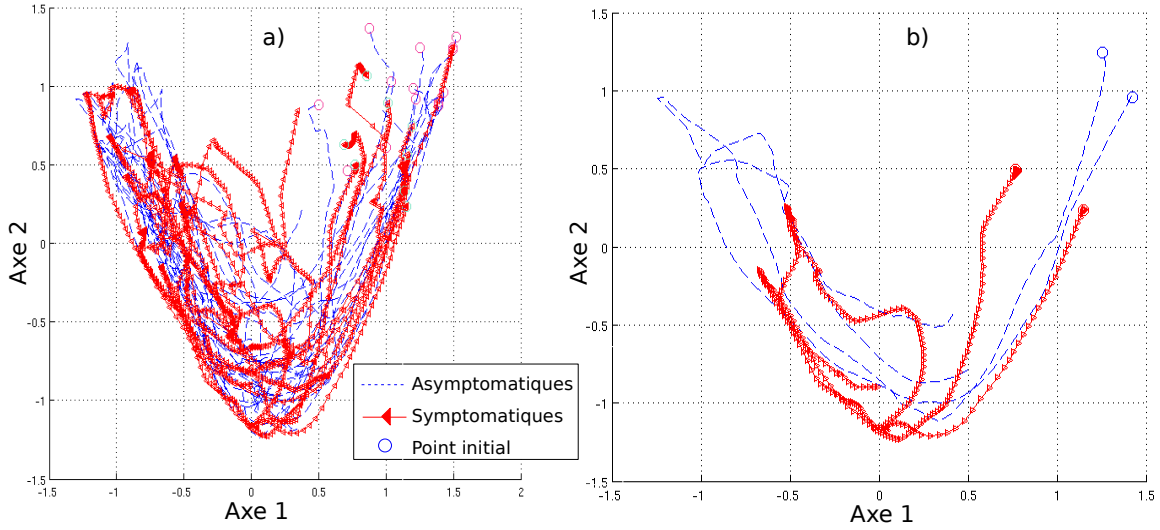


FIG. 5.14 – Trajectoires des individus obtenues par projection sur les axes factoriels 1 et 2. En a) tous les individus sont représentés, en b) seuls deux individus de chaque groupe sont tracés. Les cercles désignent le début de l'épreuve de d'effort.

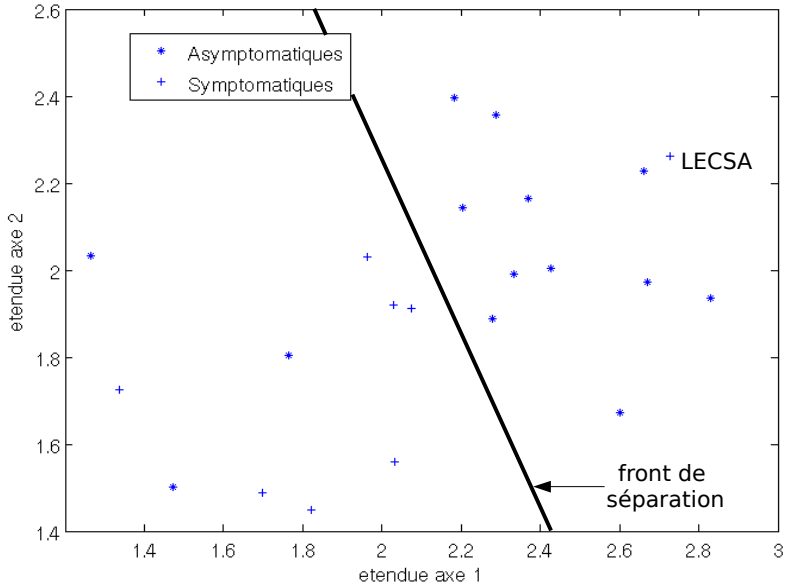


FIG. 5.15 – Etendues observées sur les axes 1 et 2 pour chaucun des patients. Le patient LECSA présente des étendues élevées.

pour l'exemple de l'ischémie, nécessite la configuration de l'apprentissage, avant de réaliser la classification des deux groupes de patients.

Apprentissage des MSMC :

Comme pour l'ischémie, il est nécessaire de déterminer quelles sont les variables à introduire dans les modèles. La table 5.11 récapitule donc les taux d'erreurs obtenus avec plusieurs ensembles de variables. Pour chacun de ces ensembles, le nombre d'états est fixé au préalable à l'aide du critère BIC. Par exemple, pour le cas du RR seul, la recherche d'un maximum global sur la courbe de vraisemblance marginale (figure 5.16) suggère un nombre d'états de 38. Le même nombre d'états est obtenu pour les autres ensembles testés.

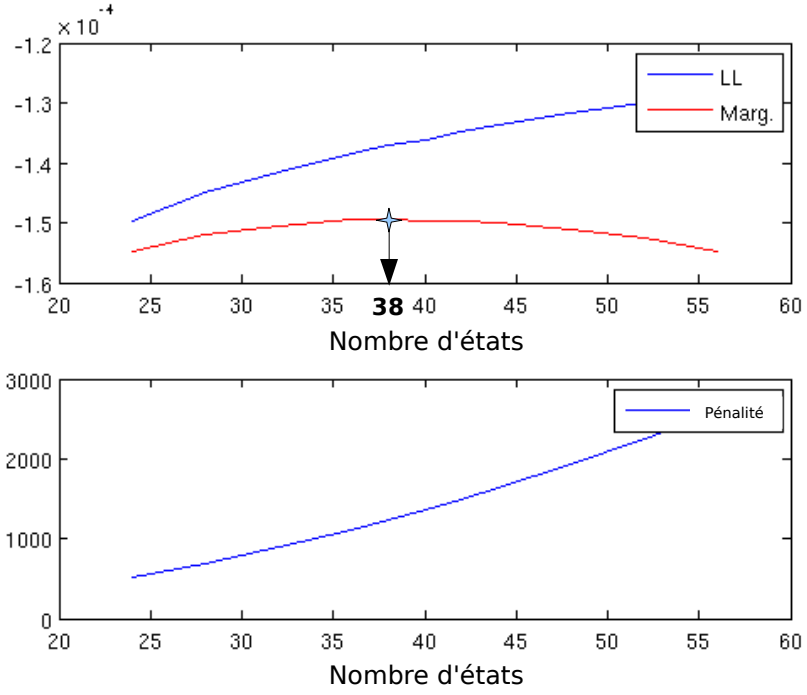


FIG. 5.16 – Application du critère BIC pour la détermination du nombre d'états sur les séries RR des patients Brugada.

	RR	QT	$AmplJ$	$PfinR$	$QRSd$	$RR+QT$
Taux d'erreurs (en %)	24.5	34,5	55.8	46	42	32

TAB. 5.11 – Taux d'erreurs obtenus avec les MSMC et avec différents ensembles de séries temporelles.

Il apparaît table 5.11 que c'est avec la variable RR seule que la distinction entre les deux groupes de patients est la mieux réalisée, ce qui correspond aux résultats trouvés avec les analyses factorielles.

5.2.5 Comparaison des performances de classification

A partir des résultats précedement évoqués, quatre méthodes de classification des patients symptomatiques/asymptomatiques peuvent être employées. Ces méthodes sont basées sur :

- un seuil de décision sur le ΔRR entre le repos et le maximum d'effort (**M1**),
- la projection des individus sur les axes 1 et 2 de l'ACP réalisée à partir des variations entre les instants analysés (**M2**),
- les étendues de chaque individu après projection de leurs séries temporelles sur les axes 1 et 2 (**M3**),
- les MSMC avec une modélisation de la dynamique de la variable RR (**M4**).

Pour M_4, la classification se déroule comme pour l'ischémie, suivant le maximum de vraisemblance. Les méthodes **M1**, **M2** et **M3** sont configurées comme suit :

M1, détermination du seuil de décision du ΔRR :
En raison de la différence entre le nombre de patients symptomatiques et le nombre de patients asymptomatiques, le seuil de décision S est ajusté de manière à minimiser la mesure D :

$$D = \sqrt{(100 - SE)^2 + (100 - SP)^2} \tag{5.2}$$

où SE et SP sont respectivement la sensibilité et la spécificité, exprimées en pourcent, de la détection des patients symptomatiques.

Les individus dépassant le seuil trouvé en apprentissage sont classés comme asymptomatiques, les autres comme symptomatiques.

M2 et M3, détermination d'une séparatrice linéaire dans $\mathbf{R}^2$:
Un classifieur par hyperplan est employé : une séparation linéaire nous a semblé adaptée car le faible nombre d'individus et leur répartition ne semble pas nécessiter un grand degré de liberté dans la détermination de la séparatrice. Cette dernière est déterminée suivant le même critère que dans les machines à vecteurs de support (SVM), en maximisant la marge entre les individus "vecteurs supports" et elle-même [Loosli et al., 2005].

Pour évaluer les performances de classification des patients symptomatiques/asymptomatiques une procédure basée sur plusieurs itérations d'apprentissage et de tests est proposée. En effet, étant donné le faible nombre d'individus, un seul tirage d'un groupe d'apprentissage et d'un groupe de test ne peut être représentatif de la performance du classifieur. Il a donc été choisi, par expérimentation, de réaliser 300 tirages (parmi $C_8^3 * C_{15}^5 = 168168$ combinaisons d'ensembles de test et d'apprentissage possibles), avec toujours 2/3 des patients en apprentissage et 1/3 en test.

A chaque itération de cette procédure,

1. les ensembles de test et d'apprentissage sont tirés pour les deux classes,
2. l'apprentissage du seuil optimal, des séparatrices ou des modèles est effectué,
3. la classification des données de tests est opérée,
4. la matrice de contingence est mise à jour avec les résultats de la classification.

Résultats :

La figure 5.17 représente l'évolution des taux d'erreurs des quatre méthodes, calculée à partir des différentes mises à jour des matrices de contingences, au cours des itérations. Il apparaît

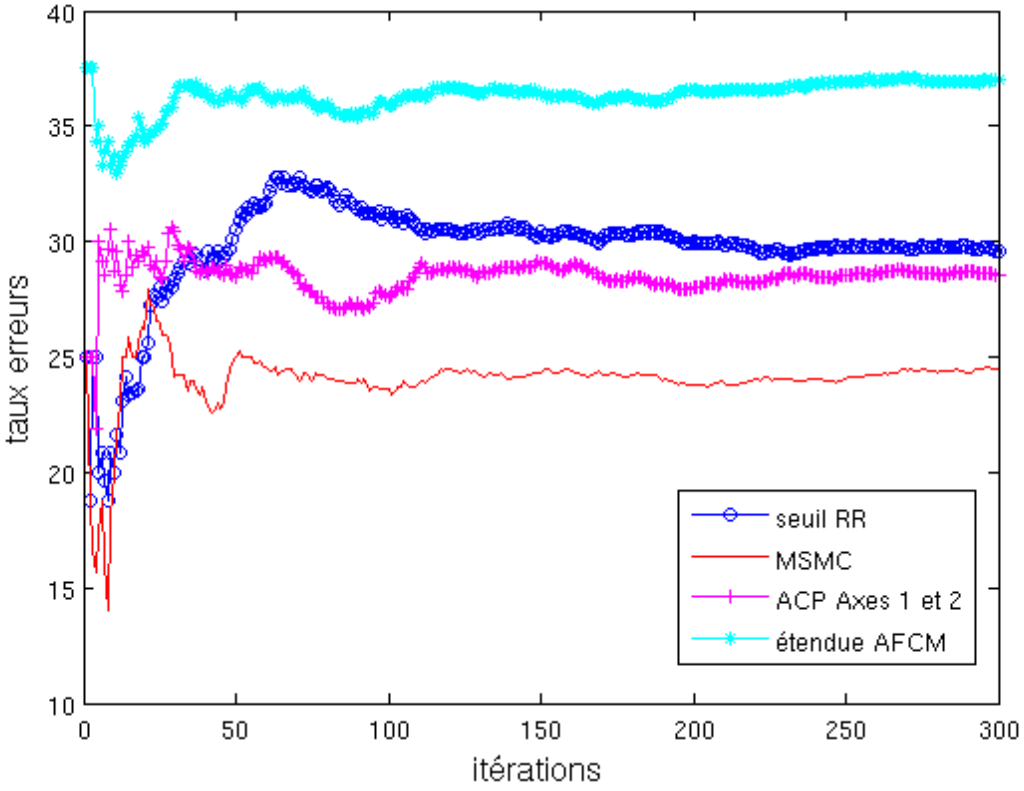


FIG. 5.17 – Evolution et stabilisation du taux d'erreur (pourcentage de mal classés) au cours des itérations pour les 4 classifieurs proposés.

que le choix de 300 tirages assure un taux d'erreur stabilisé et indépendant des patients tirés pour constituer les populations d'apprentissages et de tests.

Le tableau 5.12 compare les deux méthodes en termes de taux d'erreurs, de spécificité et de sensibilité.

	Tx d'erreurs (%)	SE (%)	SP (%)	Dist (%), par rapport à SE et SP = 1
Seuil RR (M1)	29.6	70.5	70.2	42.4
ACP axes 1 et 2 (M2)	28.58	57.9	79.3	46.8
étendue AFCM (M3)	37.0	39.6	76.9	64.7
MSMC (M4)	24.5	64.8	81.9	42.0

TAB. 5.12 – Comparaison entre le seuil sur la variation du RR et les MSMC

Le classifieur basé sur l'ACP est plus performant que celui basé sur l'étendue, ce qui parait normal puisque la séparation des deux groupes symptomatiques et asymptomatiques était plus visible dans le premier cas (figure 5.9) que dans le second (figure 5.15). Le meilleur taux d'erreur (24.5%) et la plus petite distance par rapport à la sensibilité et la spécificité égales à 100% sont obtenus avec les MSMC, démontrant ainsi l'importance de la dynamique sur la durée totale de l'épreuve d'effort. Cependant, en terme de sensibilité et de spécificité l'utilisation du seuil RR est intéressante aussi, car la sensibilité est la plus élevée des différentes méthodes et

est équivalente à la spécificité. Ce résultat est notamment du au fait que le seuil est optimisé suivant un critère basé sur la sensibilité et la spécificité.

5.2.6 Discussion

Les analyses factorielles sont des outils d'investigations qui ont largement été utilisées dans le cadre d'étude sur des pathologies cardiaques. Ces méthodes sont notamment intéressantes de part leur aspect multivariable et de part les possibilités de représentations graphiques des données. Cependant, l'analyse factorielle basée sur l'ACP est dans l'incapacité de traiter tous les points d'une série temporelle et nécessite une sélection des instants à analyser. Le codage flou des données et l'AFCM permettent de palier ce problème. Ce second type d'analyse permet la representation conjointe des variables et des trajectoires sur les axes factorielles, avec pour seul inconvénient la determination de paramètres pour la réalisation des fenêtres spatiales et temporelles. Une fois la projection réalisée, la classification (supervisée ou non-supervisée) nécessite la définition d'une mesure de similarité entre séries temporelles. Des mesures comme le DTW ont été proposées [Gueguin et al., 2008] mais nous avons choisi une mesure plus simple, basée sur l'étendue des trajectoires et fondée sur les observations précédentes de l'ACP. Enfin, les MSMC permettent une modélisation complète des séries temporelles, la détermination automatique du nombre d'états et la recherche des ensembles de variables à intégrer qui optimisent les taux de classification. Il est également possible de représenter les individus de manière simple, dans les espaces de vraisemblance. Cependant, les MSMC ne permettent pas d'évaluer les corrélations entre variables de manière aussi explicite que les analyses factorielles. D'un point de vue clinique, cette étude prospective sur l'électrocardiographie à l'effort des patients atteints du syndrome de Brugada, montre l'importance de la dynamique du rythme cardiaque dans la distinction des patients "symptomatiques" des "asymptomatiques". Ce résultat est apparu tout d'abord dans les ACP et l'influence de la dynamique s'est retrouvée dans les analyses avec les MSMC, avec un taux de classification, basé sur la série RR extraite de l'épreuve d'effort, supérieur à 75%.

Bibliographie

[Antzélévitch et al., 2005] Antzélévitch, C., P, P. B., et Brugada, J. (2005). The brugada syndrome, from bench to bedside. *Blackwell Futura*.

[Breiman, 1993] Breiman, L. (1993). *Classification and Regression Trees*. Boca Raton.

[Brugada et Brugada, 1992] Brugada, P. et Brugada, J. (1992). Right bundle branch block, persistent st segment elevation and sudden cardiac death : a distinct clinical and electrocardiographical syndrome. a multicenter report. *J. Am. Coll. Cardiol.*, 20 :1391–6.

[Fayn et al., 2007] Fayn, J., Rubel, P., Pahlm, O., et Wagner, G. S. (2007). Improvement of the detection of myocardial ischemia thanks to information technologies. *International Journal of Cardiology*, 120 :172–180.

[Gueguin et al., 2008] Gueguin, M., Roux, E., Hernandez, A. I., Poree, F., Mabo, P., Graindorge, L., et Carrault, G. (2008). Exploring time-series retrieved from cardiac implantable devices for optimizing patient follow-up. *IEEE Trans on Biomed Eng.*

[Jager et al., 2003] Jager, F., Taddei, A., Moody, G. B., Emdin, M., Antolic, G., Dorn, R., Smrdel, A., Marchesi, C., et Mark, R. G. (2003). Long-term ST database : a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. *Medical Biological Engineering Computing*, 41(2) :172–183.

[Langley et al., 2003] Langley, P., Bowers, E. J., Wild, J., Drinnan, M. J., Allen, J., Sims, A. J., Brown, N., et Murray, A. (2003). An algorithm to distinguish ischaemic and non-ischaemic ST changes in the Holter ECG. *Computers in Cardiology 2003*, 30 :239–242.

[Loosli et al., 2005] Loosli, G., Canu, S., Vishwanathan, S. V. N., et Smola, A. J. (2005). Invariances in classification : an efficient svm implementation. *ASMDA 2005 - Applied Stochastic Models and Data Analysis*.

[Loslever et Bouilland, 1999] Loslever, P. et Bouilland, S. (1999). Mariage of fuzzy sets and multiple correspondence analysis : examples with subjective interval data and biomedical signals. *Fuzzy Sets and Systems*, 107 :255–275.

[Smrdel et Jager, 2004] Smrdel, A. et Jager, F. (2004). Automated detection of transient ST-segment episodes in 24h electrocardiograms. *Med Biol Eng Comput*, 42(3) :303–311.

[Zimmerman et al., 2003] Zimmerman, M. W., Povinelli, R. J., Johnson, M. T., et Ropealla, K. (2003). A reconstructed phase space approach for distinguishing ischemic from non-ischemic ST changes using Holter ECG data. *Computers in Cardiology, 2003*, 30 :707–710.

Conclusion

En cardiologie, comme dans un grand nombre de domaines, l'information temporelle d'une grandeur et son évolution sont des données prépondérantes. Typiquement, dans le cadre de ce travail sur le signal ECG, elle se doit d'être exploitée pour analyser l'apparition de pathologies telles que l'ischémie ou bien les modifications d'électrophysiologie engendrées par un test d'effort. La prise en compte de cette information temporelle peut améliorer la spécificité en évitant les fausses alarmes liées aux bruits ou à des observations transitoires peu pertinentes. En effet, en opposition à une mesure unique à un instant t, prendre en compte l'évolution temporelle consiste à réaliser une mesure plus complète des phénomènes électrophysiologiques. En outre, la caractérisation des dynamiques observées peut aussi être d'une importance fondamentale pour détecter des phénomènes particuliers ou pour différencier des sujets dans une analyse de population.

Dans cette optique, le travail exposé dans ce mémoire s'articule autour de deux axes : l'extraction de caractéristiques du signal ECG et l'exploitation de leurs dynamiques.

Pour réduire la dimension des données et ainsi analyser directement l'information pertinente contenue dans l'ECG, nous avons été amenés à segmenter chaque battement pour en extraire les durées et amplitudes des ondes. Reproduire de manière automatisée une segmentation manuelle fiable effectuée par un cardiologue est un problème difficile, surtout à l'effort. Une chaîne de traitement du signal intégrant le débruitage de l'ECG, la détection des battements puis leur segmentation a été réalisée. La littérature expose un grand nombre de méthodes pour la segmentation et, parmi elles, les méthodes basées sur les transformées en ondelettes sont apparues particulièrement adaptées pour la prise en compte des non-stationnarités. Notre méthode s'inspire donc des travaux de [Martinez et al., 2004] et de [Li et al., 1995] mais un soin tout particulier a été apporté à l'ajustement des paramètres d'analyses (seuils et fenêtres temporelles). En effet, ces derniers sont difficiles à ajuster en pratique et ont pourtant une influence non-négligeable sur la qualité de la segmentation. Une procédure d'optimisation a donc été proposée et constitue la première contribution de ce travail : elle s'appuie sur la définition d'une fonction de coût visant à accroître les performances par rapport aux résultats déjà publiés et emploie un couple d'algorithmes évolutionnaires pour minimiser cette fonction de coût. L'évaluation de l'algorithme de segmentation ([Dumont et al., 2005, Dumont et al., 2008a]) avec ses paramètres optimisés a ensuite été réalisée sur une base de données d'ECG annotés et a montré que la qualité de la segmentation était améliorée par rapport aux résultats de la littérature.

Les indicateurs extraits de l'ECG ne sont pas uniquement les durées et amplitudes des ondes. D'autres sont mesurés tels que l'amplitude et la pente du segment ST mais aussi les paramètres spectraux de la variabilité cardiaque. Une seconde contribution a donc été de développer une station d'analyse recueillant l'ensemble de ces indicateurs, sous forme de séries temporelles. L'originalité de la plateforme est d'inclure une méthode alternative de segmentation semi-automatique du signal ECG. Cette méthode offre au clinicien un droit de regard et

de correction aux segmentations proposées. L'intérêt est alors d'exploiter l'information apportée par le clinicien pour améliorer la segmentation automatique en proposant un recalage par programmation dynamique.

La troisième contribution majeure de ce travail a été de proposer des méthodes d'exploitation des séries temporelles extraites. Nous nous sommes orientés vers une approche modèle pour plus de souplesse d'utilisation. En effet, les modèles permettent une représentation compacte des dynamiques, et peuvent ensuite être employés aisément pour réaliser des simulations, de la classification (supervisée ou non), de la détection etc... Les modèles semi-Markovien Cachés sont apparus ici bien adaptés à l'analyse d'observations multivariées et continues. Leur apprentissage ne nécessite pas de connaissance à priori sur les données et le seul hyper-paramètre à ajuster est le nombre d'états du modèles. Par contre, leur structure les limite à une représentation stationnaire par morceaux des séries à modéliser.

Au cours de ce travail, les MSMC ont été appliqués en classification supervisée et non supervisée (ou clustering) de séries temporelles. Cette seconde application est traitée ici à l'aide d'un algorithme de type clustering descendant, basé sur une augmentation progressive du nombre de clusters. La méthode proposée combine à la fois les MSMC pour modéliser les séries, l'exploitation de l'espace de vraisemblance pour créer les nouveaux clusters et un algorithme de type Expectation-Maximization (EM) flou pour effectuer le partitionnement individus/modèles. Les performances de cet algorithme ont été vérifiées, avec succès, sur plusieurs ensembles de séries simulées.

Sous-jacent à ce travail méthodologique, se trouvent deux applications cliniques fondamentales et particulièrement difficiles : la caractérisation d'épisodes ischémiques et l'analyse d'épreuves d'efforts de patients atteints du syndrome de Brugada. Pour la première, l'enjeu consiste à améliorer la classification des déviations du segment ST suivant que cette déviation soit réellement due à un épisode ischémique ou bien qu'elle soit liée à une accélération du rythme cardiaque ou à un changement de direction de l'axe électrique du coeur. Le taux d'erreur de classification obtenu est de 29%. Ce résultat, publié dans [Dumont et al., 2008b] et fondé uniquement sur des critères électrocardiographiques, est intéressant par rapport aux autres publiés dans la littérature. Pour obtenir des meilleurs taux de classification, l'introduction d'information clinique, en plus des résultats donnés par la classification avec les MSCM, devrait être envisagée. L'intérêt de notre approche est qu'elle peut être utilisée en clustering. Dans ce cas, il est apparu que les épisodes liés à des changements d'axes sont rapidement différenciés des autres. Notre analyse souligne aussi l'importance d'utiliser comme variables le niveau ST, l'intervalle RR et l'amplitude des ondes R et T pour effectuer la discrimination. La seconde application a permis la séparation des groupes "symptomatiques" et "asymptomatiques" des patients atteints du syndrome de Brugada. Dans ce dernier cas, on a constaté l'importance du rythme cardiaque à travers des analyses factorielles, ainsi que l'importance de son évolution, au cours de l'épreuve d'effort, avec les MSMC. Cependant, les résultats obtenus dans cette étude doivent être lus avec précaution dans la mesure où le nombre de patients est restreint. L'intérêt de la modélisation avec les MSMC se situe principalement au niveau de l'apport méthodologique. Les MSMC permettent de caractériser différents groupes d'individus suivant les dynamiques qu'ils présentent au cours de l'épreuve d'effort, et offrent la possibilité d'une analyse complète de la série temporelle, nous libérant de la contrainte de devoir sélectionner des instants précis.

Les perspectives concernent à la fois la segmentation du signal ECG, la modélisation des dynamiques à partir des MSMC, la méthode de clustering et les applications cliniques.

La segmentation est pour l'instant réalisée en traitant indépendamment les voies de l'ECG

or deux types d'approches multi-voies peuvent être envisagées. La première, simple, est dite "décentralisée". Elle est basée sur une segmentation individuelle de chacune des voies puis sur la fusion des différentes détections, par exemple en sélectionnant la meilleure voie ou en réalisant une moyenne des détections. La deuxième, plus complexe, est dite "centralisée". Elle exploite directement l'information de chacune des voies pour produire une détection unique. Ce type de méthode peut, par exemple passer, par la construction du vectocardiogramme (VCG) à partir d'un ECG 12 dérivations ou exploiter directement le contenu des différentes voies ECG. Il est par ailleurs évident que réaliser une détection multi-voie plus robuste ne peut être possible qu'au détriment d'une augmentation de la complexité de l'algorithme de segmentation alors que les résultats ne sont pas toujours assurément meilleurs (notamment en fonction du niveau de bruit). Au final, le choix multivoie/monovoie doit donc être fait en fonction de l'application et des signaux ECG disponibles. Toujours à propos de l'algorithme de segmentation, un soin particulier a été porté sur l'ajustement des seuils de décision mais ceux-ci demeurent fixes au cours du temps. Il pourrait être intéressant de les rendre adaptatifs, par exemple en fonction du bruit ou de la fréquence cardiaque (ceci est déjà fait de manière simplifiée pour le choix des fenêtres temporelles).

Les MSMC se sont révélés très utiles pour la modélisation des différentes séries temporelles multivariées extraites de l'ECG mais nous sommes restés dans le cadre de données extraites d'une seule modalité (l'ECG), avec des fréquences d'échantillonnage identiques et avec des dynamiques assez proches. Intégrer d'autres modalités (la respiration, la pression artérielle,...) avec des fréquences d'échantillonnage et des dynamiques différentes pourrait être intéressant, d'autant plus que les modèles de Markov couplés semblent adaptés pour cette problématique. La représentation des séries temporelles à partir de modèles linéaires à états continus et l'utilisation du filtrage de Kalman pourraient être aussi considérés pour réaliser ensuite une comparaison avec les modèles à états discrets que sont les MSMC.

Concernant la classification non-supervisée (clustering), l'estimation du nombre d'états pour chaque modèle est une difficulté majeure, puisque les modèles varient au cours des itérations de l'algorithme proposé. Il a été choisi, pour des raisons de simplicité et de temps de calcul, d'employer le critère BIC au début de l'algorithme EM et à l'initialisation de chaque nouveau modèle. Il serait cependant intéressant d'évaluer, sur des données simulées, l'impact de ce choix sur le résultat du clustering. Des comparaisons avec d'autres méthodes d'estimation du nombre d'états ou en réalisant les estimations à chaque itération de l'algorithme EM pourraient être effectuées dans ce sens. D'un autre coté, il pourrait être envisagé des améliorations sur l'aspect fouille de données. L'espace de vraisemblance créé avec les MSMC permet une représentation graphique simple pour évaluer la proximité des séries temporelles en termes de dynamiques, mais la compréhension de la nature des regroupements nécessite aussi l'affichage d'informations supplémentaires sur la provenance de ces séries temporelles. Par exemple, dans les applications en cardiologie, un accès aisé à des informations telles que l'âge des patients, le sexe ou l'historique du traitement serait d'une grande utilité car elles pourraient expliquer les regroupements observés au cours du clustering. L'inspection des dynamiques, par exemple, pour la recherche de différences entre groupes de séries temporelles, pourrait aussi être approfondie en exploitant plus en détail le contenu des modèles, notamment les matrices de probabilités de transition entre états obtenues après apprentissage. En effet, avec des états communs entre les modèles, les probabilités de transitions peuvent être comparées pour aisément retrouver des différences.

Finalement, la méthodologie proposée pourrait être étendue à d'autres applications. A titre d'exemple, en néonatalogie, la détection des épisodes d'apnée bradycardique chez les prématurés peut être abordée de manière identique au problème de la détection des épisodes ischémiques chez l'adulte. En effet, la détection de la période d'apnée se base actuellement

sur la seule variable RR et peut vraisemblablement être améliorée en intégrant de nouvelles modalités ou de nouveaux signaux. La détection précoce de l'apnée est d'importance majeure et est étudiée au sein du laboratoire dans le cadre du PHRC Intem. Il est cependant à noter que les ECG de nouveaux nés présentent de nombreuses différences par rapport aux ECG d'adultes. Leur segmentation nécessitera donc probablement un ajustement spécifique des paramètres de l'algorithme, ce qui pourra être effectué à partir des algorithmes évolutionnaires déjà proposés.

Enfin, concernant le syndrome de Brugada, un nouveau protocole doit bientôt commencer et porter sur une investigation approfondie des liens existant avec le système nerveux autonome.

Bibliographie

[Dumont et al., 2005] Dumont, J., Hernandez, A., et Carrault, G. (2005). Parameter optimization of a wavelet-based electrocardiogram delineator with an evolutionary algorithm. *Computers in Cardiology, 2005*, pages 707– 710.

[Dumont et al., 2008a] Dumont, J., Hernandez, A., et Carrault, G. (2008a). Improving ECG Beats Delineation with an Evolutionary Optimization Process. *IEEE Trans. Biomed. Eng. (Accepté, en cours de publication).*

[Dumont et al., 2008b] Dumont, J., Hernandez, A., Fleureau, J., et Carrault, G. (2008b). Modelling temporal evolution of cardiac electrophysiological features using Hidden Semi-Markov Models. *IEEE Annual Conference of the Engineering in Medicine and Biology Society (EMBC 2008) (Accepté, en cours de publication).*

[Li et al., 1995] Li, C., Zheng, C., et Tai, C. (1995). Detection of ECG characteristic points using wavelet transforms. *IEEE Trans Biomed Eng*, 42(1) :21–28.

[Martinez et al., 2004] Martinez, J. P., Almeida, R., Olmos, S., Rocha, A. P., et Laguna, P. (2004). A wavelet-based ECG delineator : evaluation on standard databases. *IEEE Trans Biomed Eng*, 51(4) :570–581.

Table des figures

Résumé

Ce mémoire s'intéresse à l'analyse de dynamiques de séries temporelles observées en cardiologie. La solution proposée se décompose en deux étapes. La première consiste à extraire l'information utile en segmentant chaque battement cardiaque à l'aide d'une décomposition en ondelettes, adaptée de la littérature. Le problème difficile de l'optimisation des seuils et des fenêtres temporelles est résolu à l'aide d'algorithmes évolutionnaires. La deuxième étape s'appuie sur les modèles Semi-Markovien Cachés pour représenter les séries temporelles composées de l'ensemble des variables extraites. Un algorithme de classification non-supervisée est proposé pour retrouver les groupements naturels. Appliquée à la détection des épisodes ischémiques et à l'analyse d'ECG d'efforts de patients atteints du syndrome de Brugada (pour la distinction des patients symptomatiques et asymptomatiques), la solution proposée montre des performances supérieures aux approches plus traditionnelles.

Mots clés : *Electrocardiogrammes, Traitement du Signal, Algorithmes Génétiques, Modèles de Markov, Algorithmes d'Expectation-Maximisation, Classification.*

Abstract

This manuscript focuses on the problem of analysing dynamics of time series observed in cardiology. The proposed solution is divided into two steps. The first one consists in the extraction of useful information from the ECG by segmenting each beat with a wavelet decomposition algorithmn, adapted from the litterature. The difficult problem of optimising both thresholds and time windows is solved with evolutionary algorithms. The second step relies on Hidden Semi-Markovian models to represent the time series made up of the extracted variables. An algorithm of unsupervised classification is proposed to retrieve the natural groups. The application of this method to the detection of ischemic episodes and to the analysis of stress ECG from patients suffering from Brugada syndrome presents a higher performance than more tradionnal approaches.

Keywords : *Electrocardiogram, Signal Processing, Genetic Algorithms, Markov Models, Expectation-Maximisation Algorithms, Classification.*