Rodolfo Telo Martins de Abreu

Licenciatura em Ciências de Engenharia Biomédica

Algorithms for Information Extraction and Signal Annotation on long-term Biosignals using Clustering Techniques

Dissertação para obtenção do Grau de Mestre em Engenharia Biomédica

Orientador: Prof. Doutor Hugo Filipe Silveira Gamboa

	Júri:
Presidente:	Prof. Doutor Mário António Basto Forjaz Secca
Vogais:	Prof^{a}. Doutora Valentina Borissovna Vassilenko Prof. Doutor Hugo Filipe Silveira Gamboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro, 2012

Algorithms for Information Extraction and Signal Annotation on long-term Biosignals using Clustering Techniques

Acknowledgements

This dissertation, along with the conclusion of this course of studies, represents a crucial step in my life in which many people deserve my sincere gratitude.

First, I am sincerely grateful to my advisor, Prof. Hugo Gamboa for all the support and shared expertise that greatly contributed in achieving the purposed objectives for this dissertation. I also thank the given opportunity to work in such a healthy business environment, allowing me to grow and evolve personally and professionally.

I would like to thank *PLUX - Wireless Biosignals, S.A.* and all of its team members for creating a healthy environment where moments of hard work and pure entertainment were reconciled harmoniously. Then, I thank Neuza Nunes, Nídia Batista, Lídia Fortunato, Nuno Cardoso and Tiago Araújo. I also owe a special thanks to Joana Sousa for all the support under various circumstances and for always having the right word when I most needed. Finally, also a special thanks goes to my thesis' colleagues Ricardo Chorão, Diliana Santos, Angela Pimentel and Nuno Costa for all the support, laughs and lunches that we shared over these last months.

To my friends Cátia, Mafalda, Margarida, Teresa Gabriel, Sofia, Ricardo, Teresa Neves, Ana and Filipa for making these last five years unforgettable. A very special thanks to my long-time friends Catarina, Ana, Marcos José, Marcos André and Patrícia for supporting me from the beginning (almost literally).

Dirijo o meu último agradecimento aos meus pais, Carlos e Otília, e ao meu irmão Nuno, por todo o apoio recebido e orgulho demonstrado à medida que fui alcançando objectivos na minha vida.

Abstract

One of the biggest challenges when analysing data is to extract information from it, especially if we dealing with very large sized data, which brings a new set of barriers to be overcome. The extracted information can be used to aid physicians in their diagnosis since biosignals often carry vital information on the subjects.

In this research work, we present a signal-independent algorithm with two main goals: perform events detection in biosignals and, with those events, extract information using a set of distance measures which will be used as input to a parallel version of the *k*-means clustering algorithm. The first goal is achieved by using two different approaches. Events can be found based on peaks detection through an adaptive threshold defined as the signal's root mean square (RMS) or by morphological analysis through the computation of the signal's *meanwave*. The final goal is achieved by dividing the distance measures into n parts and by performing *k*-means individually. In order to improve speed performance, parallel computing techniques were applied.

For this study, a set of different types of signals was acquired and annotated by our algorithm. By visual inspection, the L_1 and L_2 Minkowski distances returned an output that allowed clustering signals' cycles with an efficiency of 97.5% and 97.3%, respectively. Using the *meanwave* distance, our algorithm achieved an accuracy of 97.4%. For the downloaded ECGs from the Physionet databases, the developed algorithm detected 638 out of 644 manually annotated events provided by physicians.

The fact that this algorithm can be applied to long-term raw biosignals and without requiring any prior information about them makes it an important contribution in biosignals' information extraction and annotation.

Keywords: Biosignals, Waves, Events detection, Features extraction, Pattern recognition, *k*-means, Parallel computing, Signal processing.

Resumo

Um dos maiores desafios quando se estão a analisar dados é a capacidade de extrair informação dos mesmos, principalmente se estivermos a lidar com dados de grandes dimensões, o que traz um novo conjunto de barreiras a serem ultrapassadas.

Neste trabalho de investigação, é apresentado um algoritmo independente do tipo de biosinal que estiver a ser analisado e que apresenta dois objectivos principais: o primeiro prende-se com a detecção de eventos. Posteriormente, são retiradas medidas de distância que irão ser colocadas como *input* numa nova versão paralela do algoritmo de *clustering* *k*-means. A aplicação de técnicas de *clustering* irá permitir a extração de informação relevante dos sinais.

O primeiro objectivo é concretizado recorrendo a duas abordagens distintas. Numa vertente mais simples e computacionalmente mais leve, os eventos podem ser detectados através da computação dos picos do sinal, onde o limiar é adaptável e definido como o valor quadrático médio do sinal; por outro lado, a detecção de eventos pode resultar de uma análise morfológica e da computação da onda média representativa do sinal. O último objectivo é concretizado dividindo as medidas de distância previamente obtidas em n partes e aplicando o algoritmo *k*-means em cada uma delas individualmente. De modo a diminuir o tempo de processamento foram utilizadas técnicas de programação em paralelo.

Para este estudo foram adquiridos e anotados pelo nosso algoritmo diversos tipos de sinais. De modo a realizar a validação do algoritmo, recorreu-se a uma inspecção visual dos sinais processados, obtendo-se uma eficiência de 97.5% e de 97.3% quando utilizadas as distâncias de Minkowski L_1 e L_2, respectivamente. Utilizando a distância à onda média, o nosso algoritmo atingiu uma precisão de 97.4%. Relativamente aos ECGs que foram obtidos nas bases de dados da *Physionet*, o algoritmo desenvolvido conseguiu detectar 638 das 644 anotações clinicamente relevantes fornecidas por médicos.

O facto do algoritmo desenvolvido poder ser aplicado em sinais *raw*, de longa duração e sem necessitar de qualquer informação prévia sobre os mesmos, faz com que represente uma importante contribuição na área do processamento de sinal e, mais especificamente, na anotação e extracção de informação de biosinais.

Palavras-chave: Biosinais, Ondas, Detecção de eventos, Extracção de características, Reconhecimento de padrões, *k*-means, Programação em paralelo, Processamento de sinal

Contents

List of Figures

List of Tables

Acronyms

AAL Ambient Assisted Living

ACC Accelerometry

ADC Analog-to-Digital Converter

ADL Activities of Daily Livings

AHD Atherosclerotic Heart Disease

AI Artificial Intelligence

AmI Ambient Intelligence

BVP Blood Volume Pressure

CAD Coronary Heart Disease

CNS Central Nervous System

ECG Electrocardiography

EEG Electroencephalography

EMG Electromyography

FFT Fast Fourier Transform

HR Heart Rate

HRV Heart Rate Variability

IC Inspiratory Capacity

LED Light-Emitting Diode

PVC Premature Ventricular Contraction

RMS Root Mean Square

RV Residual Volume

SNR Signal-to-Noise Ratio

TLC Total Lung Capacity

TV Tidal Volume

VC Vital Capacity

1

Introduction

1.1 Motivation

One of the biggest challenges nowadays is to increase people's life expectancy while reaching those late ages with a high quality of life, living independently for a longer period of time. Although technology is constantly evolving, these goals could not be reached if a closer and regular monitoring of patients was not taken. Thus, the inevitable increase of data to be analyse becomes a concern for physicians and technicians due to the exhaustive and time-consuming nature of this type of tasks.

The search and development of computer-aid techniques to automatically analyse data is continuously growing. In the signal processing field, these tools are usually developed to analyse only one type of signals, such as electrocardiography (ECG), electromyography (EMG), respiration, accelerometry (ACC), among others. The main goal of these tools is to monitor activities and vital signs in order to detect emergency situations or deviations from a normal medical pattern [8]. Besides, the barriers that arise when dealing with long records (provided by a continuous monitoring of ill patients at home, for example) are yet to be overcome.

Given these facts and since the concept of Ambient Assisted Living (where patients are comfortably monitored at home, using wearable sensors that acquire signals from their bodies and processing those signals in an unobtrusive way) is spreading, the development of a single processing tool able to detect emergency situations by automatically annotating the given input signals and which is also suitable for long-term records represents an important asset in the signal processing field and a major help for physicians in analysing electrophysiological data.

This dissertation was developed at *PLUX - Wireless Biosignals, S.A.*, under the Research and Development (R&D) department. One of the major goals of this department is to create innovative solutions for comfortably monitoring people under a variety of scenarios (at healthcare facilities, home, training facilities, among others). Besides, it also aims at developing signal processing tools for extracting information from the biosignals acquired during the people's monitoring. The possibility to contribute with the development of signal processing tools and the given opportunity of working in a business environment strongly encouraged this research work.

1.2 Objectives

This thesis aims at the development of a signal-independent processing algorithm able to perform clustering techniques in long-term biosignals and extracting information from them. With that information, the output of the algorithm is an annotated signal. Due to the high level of abstraction present in our algorithm, a set of different types of biosignals was acquired to test its performance, including ECG, EMG, blood volume pressure (BVP), respiratory and ACC signals. An events detection tool was first implemented, distance measures were taken using different distance functions and a parallel version of the *k*-means clustering algorithm was also designed and characterized.

Due to the suitability of our algorithm in long-term biosignals, parallel computing techniques were applied in order to improve performance.

1.3 Thesis Overview

In Figure 1.1 it is exposed the structure of the present thesis.

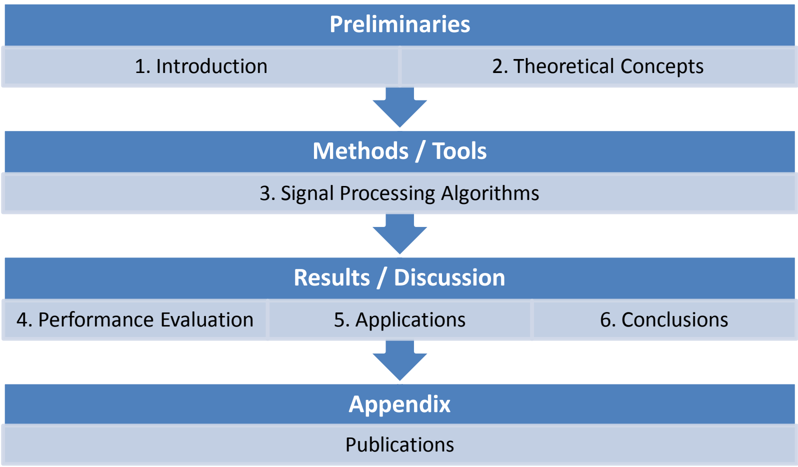


Figure 1.1: Thesis overview

In the present chapter, the motivation that lead to the development of this thesis is exposed and the main objectives are also briefly explained. In Chapter 2, a theoretical contextualization of the concepts used in this thesis is provided and a state of the art is also characterized. These two chapters comprise the thesis's Preliminaries.

In Chapter 3, the signal processing algorithms that were developed to fulfil the thesis's objectives are depicted. The two developed approaches to detect events on biosignals and the parallel version of the *k*-means clustering algorithm are exposed. This chapter accounts for the thesis's Methods/Tools.

The last three chapters address the results and discussion of this research work. In Chapter 4, the procedures to test our algorithm performance are exposed, including the visual inspections and comparison with the annotated signals from the Physionet databases. In Chapter 5 some specific applications of our algorithm are presented, showing its applicability in this research topic. To conclude, an overview of the developed work, results and contributions are presented and some future work suggestions are discussed in Chapter 6.

The thesis writing was done using the LaTeX environment [9]. The signal processing algorithms were developed in *Python* using the scipy [10] package.

2

Concepts

In this chapter the main concepts that were used in this dissertation are presented. Fundamentals on biosignals, biosignal acquisition, biosignal processing and machine learning will be stated.

2.1 Biosignals

Biosignals can be described as space-time records of biological events that generate physiological activities (e.g. chemical, electrical or mechanical) — *biological signals* — likely to be measured [2, 11].

2.1.1 Biosignals Classification

Biosignals are usually divided according to the physiological phenomenon that was behind their generation. Thus, the most current and important classifications are [2, 12]:

- **Bioeletric (Electrophysiologic) Signals**: When a stimulus is enough to depolarize a cell's membrane, an action potential is generated which leads to electrical changes that can be measured by electrodes. Electrocardiogram, electromyogram or electroencephalogram are examples of bioelectric signals.

- **Biomechanical Signals**: These signals are associated with biological motion that generates force, such as blood pressure.

- **Biochemical Signals**: These signals are generated due to changes in the concentration of certain chemicals, providing functional and physiological information.

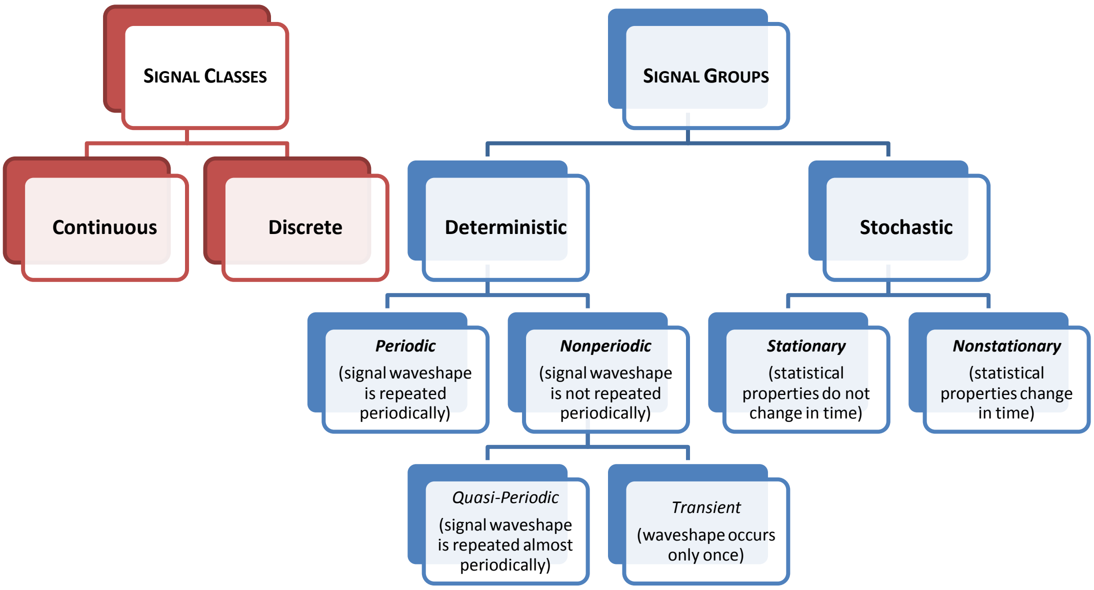

SIGNAL CLASSES
Continuous
Discrete
SIGNAL GROUPS
Deterministic
Stochastic
Periodic
(signal waveshape is repeated periodically)
Nonperiodic
(signal waveshape is not repeated periodically)
Stationary
(statistical properties do not change in time)
Nonstationary
(statistical properties change in time)
Quasi-Periodic
(signal waveshape is repeated almost periodically)
Transient
(waveshape occurs only once)

Figure 2.1: Biosignal's types and classification. Adapted from [1].

- **Biomagnetic Signals**: Magnetic fields can be generated by the human body due to the electrical changes that occur within it. Magnetoencephalogram is an example of biomagnetic signals.

Biosignals can also be classified as *deterministic* or *stochastic*. Deterministic signals can be described by mathematical functions or even by plots. Although most of the real-world signals are nondeterministic, it is quite usual to define a model (or function) that approximately describes the signal to be analysed.

One of the most important type of deterministic signals is the *periodic* signals, that are characterized by a time interval T that separates two successive copies of the signal. Thus, being $x(t)$ a periodic signal, it can be expressed as

$$x(t) = x(t + kT) \tag{2.1}$$

where k is a integer. Although most of the real-world signals are nonperiodic, there is a very important class — *quasi-periodic signals* — which includes signals that have some slight changes along their cycles, so they can be considered as almost periodic. It is the example of the ECG signal; in fact, the time between R peaks is always different but approximately equal, and so ECG's PQRST complex interval of one heart beat is almost the same as the next.

Unlike deterministic signals, stochastic signals cannot be described by mathematical functions due to the uncertainty that characterizes their parameters. In fact, these parameters have to be determined (estimated) using statistical analysis through probability functions or statistical measures. A good example is the EMG signal [1, 2].

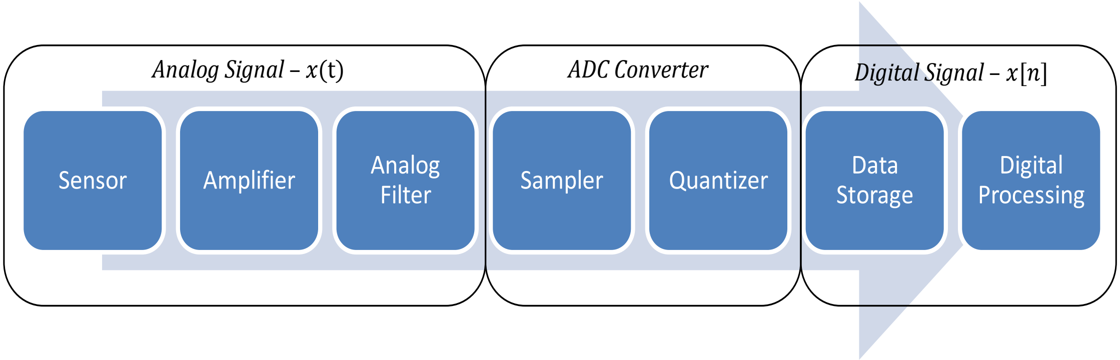


Figure 2.2: General block diagram for the aquisition of a digital signal. Adapted from [1, 2].

It is also common to divide biosignals in *continuous* and *discrete*. Continuous biosignals are described by functions that use continuous variables or by differential equations and can be represented mathematically by $x(t)$, where t is a continuous variable – time. Thus, continuous biosignals provide information at any instant of time. Most of the biological signals are continuous in time, such as ECG, EMG, BVP, among others.

On the other hand, discrete biosignals (or time-series) are described by functions that use discrete variables, i.e. digital sampled data. This type of biological signals can be represented mathematically by $x[n]$, where $n = 0, \ldots, L$ represents a subset of points of t, with L the length of the signal. Therefore, discrete biosignals only provide information at a given discrete point along the time axis [1, 2, 13].

2.1.2 Biosignals Acquisition

Biosignals are collected using sensors that are able to convert certain biological activities (e.g. heart beat, respiratory movements) into an electrical output. Since computers only have the capacity to store and process discrete amounts of information [14], it is necessary to convert the continuous data into discrete units in order to allow the processing. Besides, the most important developments in signal processing are related to discrete signals. Thus, it is common to convert a continuous signal into a discrete one using an Analog-to-Digital Converter (ADC). ADC is a computer controlled voltmeter that transforms continuous biological signals into digital sequences [1, 2].

The transformation of the ADC comprises two steps: *sampling* and *quantization*. Mathematically, the sampling process can be described as [1]:

$$x(t) = x(n)|_{n=tT_s} \tag{2.2}$$

where $x(t)$ and $x(n)$ are the continuous and sampled (discrete) signal, respectively, and

T_s the sampling interval. Thus, the sampling frequency, f_s, can be defined as:

$$f_s = \frac{2\pi}{T_s} \tag{2.3}$$

However, f_s must verify an important condition in order to guarantee that the discrete signal is no different (no information added or removed) from the original signal. This is very important in any area but since the result of biosignals processing might be used to assist physicians on diagnosis, this feature becomes even more essential [2]. In fact, being F the higher frequency present in the original signal, if $f_s = 2F$ it is said that the signal is Nyquist-sampled and f_s is called the Nyquist frequency [14]. Therefore, the Nyquist theorem defines a minimal sampling frequency given by [15]:

$$f_s > 2F \tag{2.4}$$

In the quantization process, each value of the signal amplitude can only assume a finite number of values. This process is related to the ADC resolution which is the number of bits that are going to be used to generate a digital approximation. Thus, this process brings inherently loss of information, which can be attenuated by increasing the available number of bits [2, 14].

2.1.3 Biosignals Processing

Biosignal processing is necessary in order to extract relevant information present in raw data. Although most of the processing techniques are applied in digitized signals, some analogue signal processing is usually necessary [16]. However, for the purpose of this dissertation, we will only focus on digital signal processing.

After the biosignal acquisition phase, the next step is to interpret the meaning of the acquired signal. To accomplish that, it is often necessary to apply different types of processing: *pre-processing* and *specialized processing* [17].

Signal pre-processing can also be called as signal "enhancement" because it allows to separate the acquired information from the inherent noise that appears due to most of the measurement systems, which is a limiting factor in the performance of instruments. Besides, external factors such as movement during the acquisitions might also induce noise appearance. However, the acquisition conditions also influence the quality of the acquired signal.

Thus, the first signal processing techniques emerge due to the necessity of removing noise artefacts, which were limiting the extraction of useful results disguised by them. These techniques usually consist in the application of some specific filters that allow noise removal without eliminating signal's (useful) information and improving, therefore, the *signal-to-noise ratio*, SNR, which is given by:

$$SNR = 20 \log \left(\frac{Signal}{Noise} \right) \tag{2.5}$$

where *Signal* represents the signal's useful information and *Noise* the signal's noise [1, 16, 17].

Once the pre-processing phase is finished, a specialized processing is applied to signals, along with classification and recognition algorithms [17].

In general, it is only after the pre-processing and specialized processing steps that the result of the acquisition process reveals the true meaning of the physical phenomenon that produced the signal under analysis.

2.1.4 Biosignals Types

In this sub-section a brief description of electrocardiography, electromyography, accelerometry, blood volume pressure and respiratory signals (the main biosignals addressed in this research work) will be presented.

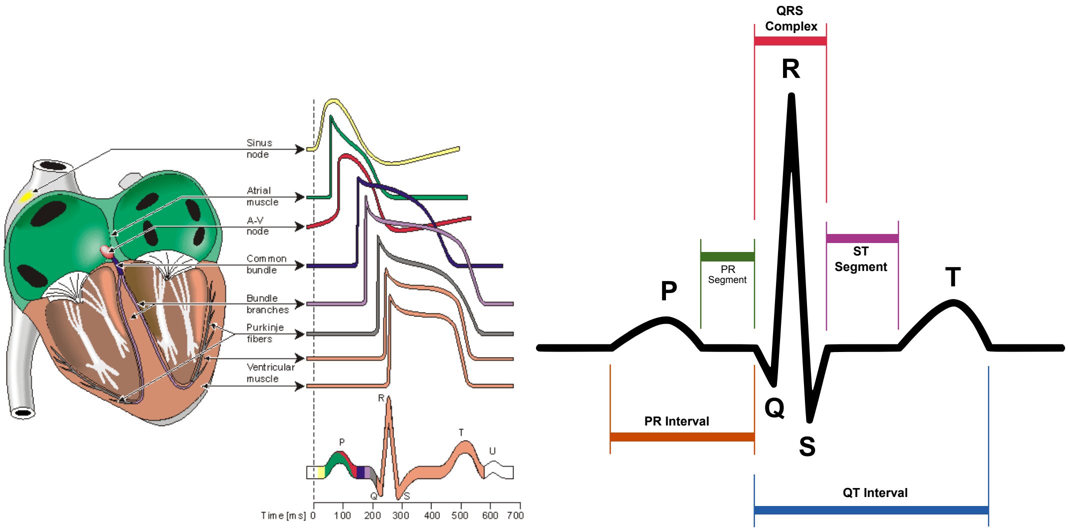
Electrocardiography

The electrocardiography (ECG) signal is one of the most well studied biosignals, because changes due to pathologies are easily detected and its acquisition process is quite simple. These factors contribute to its wide use in medicine [14].

Since every single muscle on the human body needs an electrical stimulus to contract or relax, the heart's muscle — myocardium — is no exception. In fact, as we can see in Figure 2.3 (a), there is a group of cells located on the right atrium that are responsible to generate a depolarization wave (due to a certain stimulus). These cells are designated as the pacemaker cells and together form the sinoatrial (SA) node. Then, the bundle of His assures that the depolarization wave propagates along the heart, which results on the contraction of myocardium. All this process generates currents and, consequently, a measurable electrical signal [14, 18].

As a result of the electrical conductivity of the human tissues, this electrical signal disperses all around the body, making it possible to be detected on the body surface. Therefore, the ECG wave form (Figure 2.3 (b)) results from the sum of the electrical changes within the heart. The first portion is designated as P wave and is related to the atria depolarization and the QRS complex is due to the ventricle depolarization. Then appear two more waves, T e U, assigned to the repolarization of ventricles and atria, respectively. Most of the times the U wave is masked by the ventricular depolarization of the following cardiac cycle [2, 14, 18].

By analysing ECG signals it is possible to detect a large variety of cardiac pathologies simply looking at the heart rate, ECG amplitude and heart rate variability (changes along the time interval between two consecutive beats), making these important diagnostic features.



(a) The origin of the electrical activity

(b) Pattern of a normal ECG signal

Figure 2.3: Electrophysiology of the heart. From [3].

Electromyography

The electromyography (EMG) signals relate to the electrical signals that are behind muscle contraction. These signals can be recorded by electrodes inserted in the muscle (intramuscular recordings) or located over the skin (surface recordings). Surface recordings are the most used method since they are non-invasive. In this research work, only the surface recording method was used in order to obtain this type of biosignals [14].

Just like it was mentioned before, every muscle needs an electrical stimulus to begin its activity. These signals are generated on nerve cells called motor neurons. A motor unit is defined as the joint of a motor neuron and all the muscles fibers that it innervates. When a motor unit becomes active, an electrical signal stimulates muscles to contract and this electrical signal corresponds to the resulting electromyographic signal [19, 20]. Therefore, the time when a motor unit becomes active represents the EMG's onset; likewise, the time when a motor unit becomes inactive represents the EMG's offset.

In Figure 2.4 the procedure to acquire EMG surface recordings is shown. When the EMG is acquired from electrodes mounted directly on the skin, the signal is a composite of all the muscle fiber action potentials occurring in the muscles underlying the skin. Therefore, in order to obtain individual motor unit action potentials, a signal decomposition is required [1, 4].

Since the EMG detects the electrical activity of muscles, it represents an important diagnostic tool in detecting muscle dysfunctions.

Accelerometry

Accelerometry, as the name suggests, is a technique that measures the acceleration of an object. Acceleration can be defined as the rate in change of direction or magnitude in the

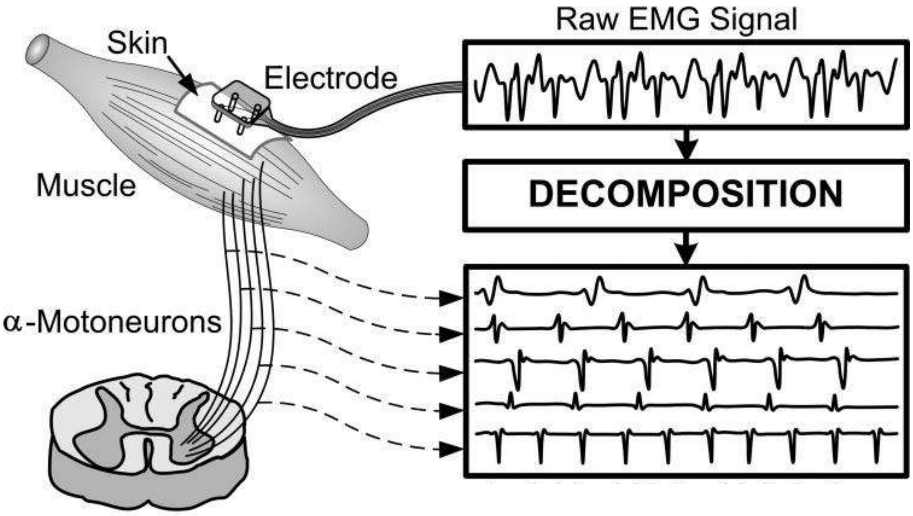


Figure 2.4: Surface EMG recording. Adapted from [4].

velocity of an object. Therefore, its units are m/s^2 or g units, where $1\ g = 9.81\ m/s^2$.

In order to measure the acceleration, accelerometers are used. Initially, accelerometers were designed only to be sensitive in one direction but nowadays, accelerometers are capable of measuring acceleration in each orthogonal axis [14]. However, the calibration process is quite important. In fact, the output of a stationary accelerometer pointing toward global vertical must be $1\ g$ (or $-1\ g$, depending the accelerometer's orientation) [21].

Since an accelerometer is usually small and its utilization is inexpensive, it has been widely used in monitoring human motion and in classifying human movement patterns. Besides, it is quite important that movement measures are unobtrusive to obtain a more accurate monitoring of human motion [22, 23].

Blood Volume Pressure

The peripheral pulse wave is a mechanical event closely following the ECG complex. It reflects the interplay between left ventricular output and the capacitance of the vascular tree [24].

In order to measure this peripherical pulse wave, a noninvasive device — the photoplethysmograph — was built, allowing the detection of blood volume changes in living tissues by optical means. The device was first described by Hertzman [25] and consists of a light source — light-emitting diode (LED) —, photodetector and AC amplifier. The electrical signal from the photodetector is related to blood volume changes in tissue which are caused by variations in the blood volume pressure (BVP), also allowing the detection of flow variations in the periphery during the cardiac cycle [26]. In fact, observing Figure 2.6, a maximum is noted in the pulse wave associated to the systolic pressure (when the blood pressure is maximum) and a minimum corresponding to the diastolic pressure.

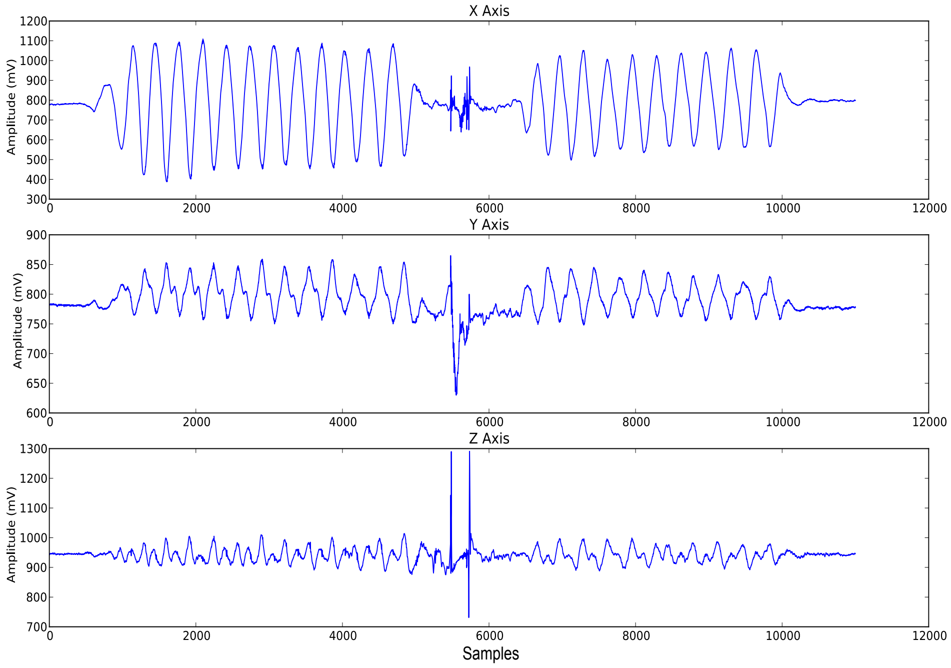


Figure 2.5: Acceleration signals acquired with a three axis sensor.

By measuring the time between each cycle of the BVP signal, the heart rate (HR) can be computed and a heart rate variability (HRV) analysis can be undertaken.

This signal provides a method for determining properties of the vessels and changes with ageing and disease.

Despite the complex source of the signal and problems of measurement, this technique has been widely used for the monitoring of arterial pressure, detecting anxiety, among other applications [27].

Respiration

Respiratory signals are directly or indirectly related to the lungs volume along each breath. The tidal volume (TV) is the amount of air flowing into and out of the lungs in each breath, which is typically 500 mL for an adult. However, the respiratory system has the ability to move much more air than the tidal volume. In fact, the inspiratory capacity (IC) represents the maximum volume of air that someone can inhale. On the other hand, it is impossible to exhale all the lungs' air and this volume represents the residual volume (RV). Finally, the total lung capacity (TLC) equals the vital capacity (VC) — the volume of air exhaled from a maximum inspiration to a maximum expiration — plus the residual volume [1, 18].

For the purpose of this research work, indirect measures of the respiration were taken.

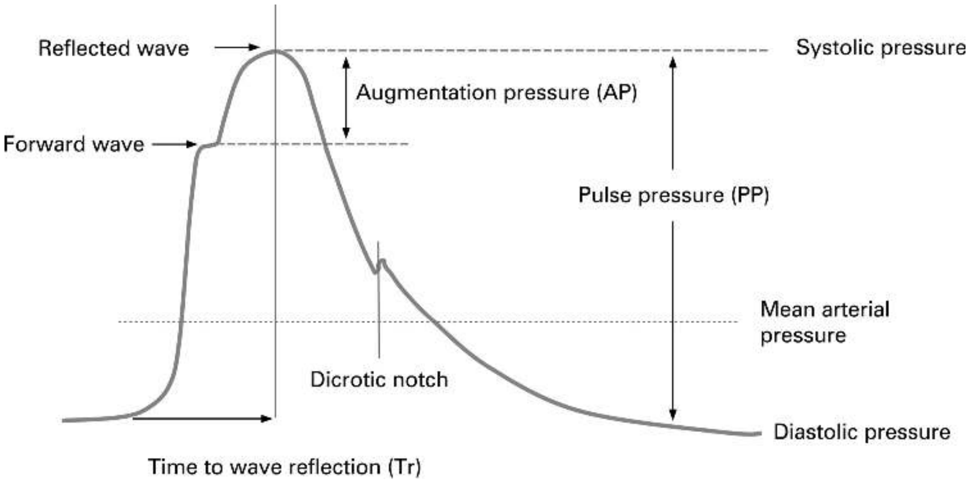


Figure 2.6: Example of a BVP signal with annotations of the main points in one cycle.

In fact, using a respiratory sensor which integrates a Piezo Film Technology (PVDF) sensor, changes in length related to the abdominal and thoracic movements can be measured, obtaining a respiratory signal where the respiratory cycles can be observed [28].

In Figure 2.7 an ECG and a respiratory signal are presented. Both signals are illustrated in order to show the relationship between them. In fact, with the R-peaks detection of an ECG signal, it is possible to estimate the correspondent respiratory signal, by computing the ECG envelope.

2.2 Clustering

2.2.1 Machine Learning and Learning Methods

Machine learning is a mechanism where due to the input of new data, computers (or machines) change their structure or program in a way that is expected an improvement on future performance [29]. Therefore, machine learning is related to *Artificial Intelligence* (AI), although most of the AI research does not concern with learning which make computers develop a task always the same way over time [30]. Google is an example where machine learning is a key feature [31]. In fact, when someone does a search, the input is classified and a set of pages that have the same classification are returned. Amazon [32] and any gaming company that have AI in their games also use machine learning.

Thus, it is usual to question the reason why machines have to learn or, on other words, why machine learning is important. In fact, there are cases where a task can be only defined by giving an example and sometimes there are hidden relations among data which can be extracted using *Data Mining* techniques (pattern recognition). Besides, environments change over time and humans are continuously discovering new information, turning machines that are not capable of learn obsolete [29]. It is worth noticing that Machine Learning and Data Mining are heavily related and rely on the induction process [33].

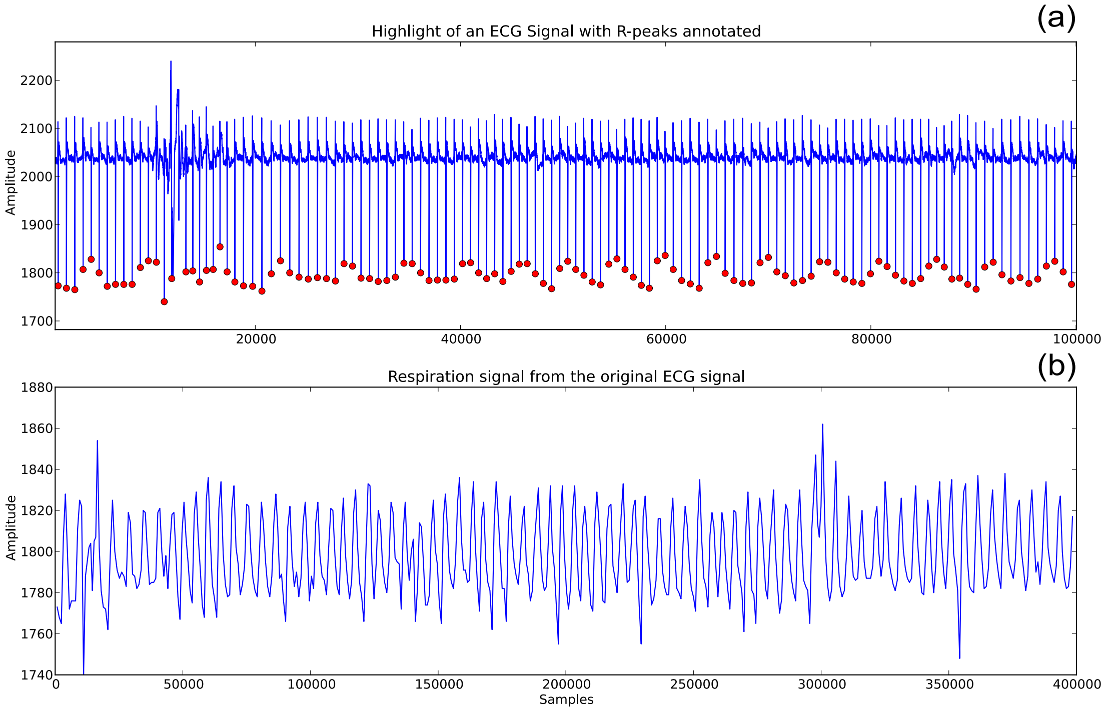


Figure 2.7: (a) Highlight of an ECG signal with the R-peaks annotated and (b) the computed ECG envelope allowing the estimation of the respiratory signal.

The performance improvement over time requires that humans establish a set of rules for computers to use them to learn along with the new incoming data. Therefore, one of the main goals of machine learning is to generate classifications on input data simple enough to be understood by humans [34]. However, computers are not always capable of making the right decision (classification) when applying those rules; in fact, there is always an error and the main goal of the learning process is to minimize that error which is usually expressed by a function – *error function*. This leads to a new world called *optimization* [29].

As it was previously mentioned, one important aspect of machine learning is data classification. In fact, one of the most important tasks when analysing data is trying to find patterns or relationships in data in order to group it into a set of categories [5].

To accomplish that, two types of classification are used: the *supervised learning* or the *unsupervised learning*, which is also known as *clustering*.

In the supervised learning approach, the learner (usually, the computer) receives input data, **D**, that will be divided into two sets: the *training set*, $\sum$, and the *testing set*, Ω. The training set, as the name suggests, will be used to teach the learner, establishing a

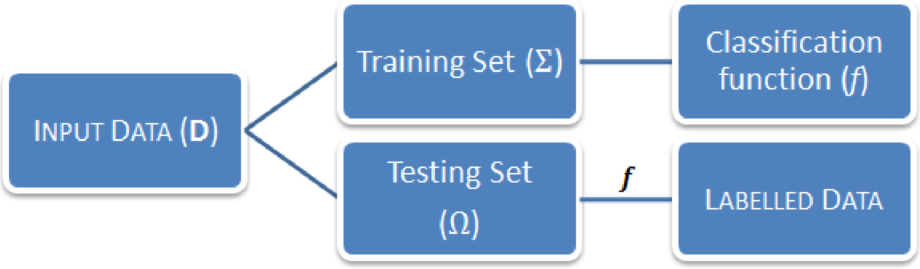


Figure 2.8: General procedure of supervised learning.

function f between the labelled data and its output. Analogously, the testing set contains data that will be classified due to the previous learning process (represented by the function f); if the obtained classifier is ideal, there will be no errors, which means that f perfectly describes the entire data (training set plus testing set) [29, 35]. Thus, the main objective in supervised learning is to label new data according to the previous labelled data received as input [36] and, as a result, data will be grouped into specific classes. In other words, it allows pattern recognition and this is why supervised learning is broadly used, despite its human and computational costs and limitations [37].

In opposition with the supervised learning, the unsupervised learning approach does not have a training set, which means that the learner only receives unlabelled data with no information about the class of each sample [37]. Therefore, the main goal of clustering is to separate data into a finite number of data structures using only the data itself [6, 36]. These structures have to be organized in such a way that the intra-group variability is minimized and the inter-group dissimilarity is maximized [38]. It is worth noticing that clustering algorithms can be applied on raw data, although it is common to pre-process it [39].

2.2.2 Clustering Methods

In order to implement a clustering algorithm, it is necessary to define two major issues: in which way data objects are grouped and what's the criteria to be used in the grouping process [40]. Therefore, according to the chosen grouping process, clustering algorithms can be broadly divided into five categories, which will be briefly described [5, 6, 36, 38, 39, 41]:

1. **Partitioning Methods.** These methods (or algorithms) separate data into k different clusters, where each cluster as at least one object. There is the possibility of each object be a part of only one cluster — *hard clustering* — or, on the other hand, one single object can be on two or more different clusters simultaneously — *fuzzy clustering*. Fuzzy clustering can be easily converted to hard clustering by allocating each object in the cluster where there is a maximum similarity. Partition-based

algorithms are always designed according to an objective function and work well on spherical-shaped clusters and on small to medium data set. The well-known partition-based algorithms are the *k*-means and *fuzzy c*-means [42].

2. **Hierarchical Methods.** These methods group data objects into a tree of clusters (which can be also called a dendrogram) where the root node represents the whole data set, the leaf node a single object and the intermediate nodes represent how similar objects are from each other. Hierarchical algorithms can be broadly divided in *agglomerative* and *divisive* algorithms. In agglomerative algorithms, each data object represents a cluster (singleton) and then the clusters are merged (according to their similarity) until a stop condition is achieved (if not, all the data objects will belong to the same group), which usually is related to the number of desired clusters – bottom up approach. On the other hand, in divisive algorithms, all the data objects represent one cluster and they are sliced into a specified number of clusters (if not, each data object will represent a single cluster) – top down approach. Most of the hierarchical clustering algorithms are variants of the single-link, complete-link or minimum-variance algorithms. CURE [43], BIRCH [44] and Chameleon [45] are examples of hierarchical algorithms.

3. **Density-Based Methods.** These algorithms are based on density in the neighbourhood of a certain cluster: until that density exceeds a certain threshold, the cluster grows continuously. DBSCAN [46] and OPTICS [47] are examples of density-based algorithms.

4. **Grid-Based Methods.** These methods quantize the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed, allowing a small processing time. STING [48] is an example.

5. **Model-Based Methods.** These methods divide data objects into clusters according to a specific model established for each cluster and it can be obtained using a statistical approach or a neural network approach.

2.2.3 Clustering Steps

Regardless of the various methods of clustering presented in the previous sub-section, cluster analysis can be divided into four major stages [5, 6, 36]:

1. **Feature selection or extraction.** Feature selection chooses the most distinct features from the original data features (called candidates) while feature extraction uses one or more transformations on the original features to produce useful and novel ones. Either one or both can be used to obtain an appropriate set of features and, ideally, these features should belong to different clusters, be immune to noise, easy to extract and interpret. If the feature selection is done adequately, the algorithm design process will be simplified.

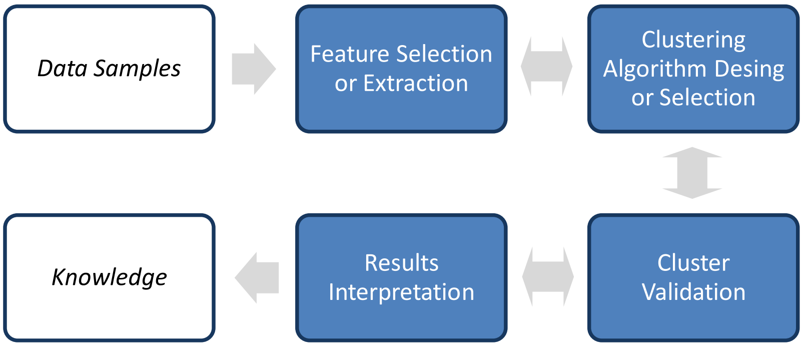

Data Samples
Feature Selection or Extraction
Clustering Algorithm Desing or Selection
Cluster Validation
Results Interpretation
Knowledge

Figure 2.9: Basic process for cluster analysis. From [5, 6].

2. **Clustering algorithm design or selection.** Almost all clustering algorithms are connected to some particular definition of distance measure. Therefore, clustering algorithm design usually consists of determining an adequate proximity measure along with the construction of a criterion function, affecting the way that data objects are grouped within clusters.

3. **Cluster Validation.** When comparing different approaches, it is common to obtain different results (clusters); besides, when comparing the same algorithm, a change in pattern identification or presentation order of input patterns may result in different clusters also. Thus, in order to accurately use clustering algorithms results, a cluster validation is necessary. This validation process must be objective and have no preferences to any algorithm but there is not an optimal (and general) procedure for clusters validation, except in well-prescribed subdomains.

4. **Results interpretation.** The main objective of clustering is to provide users information about the original data in order to solve the encountered problems. Since cluster results only represent a possible output, further analysis (e.g. using supervised learning techniques) are necessary to guarantee the reliability of results.

2.2.4 The *k*-means clustering algorithm

The *k*-means clustering algorithm [49, 50] is the best-known squared error-based clustering algorithm. In fact, this algorithm seeks an optimal partition of the data by minimizing the sum-of-squared-error criterion. If we have a set of objects $\mathbf{x}_j$, $j = 1, \ldots, N$, and we want to organize them into k partitions $C = \{C_1, \ldots, C_k\}$, then the squared error criterion is defined by [6, 51]:

$$J(\mathbf{M}) = \sum_{i=1}^{k} \sum_{j=1}^{N} ||\mathbf{x}_j - \mathbf{m}_i||^2 \tag{2.6}$$

where $\mathbf{m}_i$ is an element of the cluster prototype or centroid matrix $\mathbf{M}$.

In order to minimize this criterion, an iterative procedure is taken. Thus, the *k*-means belongs to the category of the hill-climbing algorithms. The basic steps of this clustering algorithm are summarized as follows [5, 38].

1. Initialize a *k*-partition randomly or based on some prior knowledge. Calculate the cluster prototype matrix $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_k]$.

2. Assign each object, $\mathbf{x}_j$, of the data set to the nearest cluster partition $C = \{C_1, \ldots, C_k\}$ based on the criterion,

$$\mathbf{x}_j \in C_w, \quad \text{if } ||\mathbf{x}_j - \mathbf{m}_w|| < ||\mathbf{x}_j - \mathbf{m}_i|| \tag{2.7}$$

   for $j = 1, \ldots, N$, $i \neq w$ and $i = 1, \ldots, k$.

3. Recalculate the cluster prototype matrix where each element will now be given by

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \tag{2.8}$$

4. Repeat steps 2)-3) until there is no change in each cluster or, in other words, in the cluster prototype matrix's elements.

In spite of its simplicity, easy implementation and relatively low time complexity, there are some major and well-known drawbacks, resulting in a large number of variants of the original *k*-means algorithm in order to overcome these obstacles. The major disadvantages are presented bellow [5, 6].

- There is no automatic method to identify the optimal number of partitions, which must be given as an input for the *k*-means algorithm.

- The random selection of the initial partition to initialize the optimization procedure affects the centroids convergence, returning different results for the same data.

- The iteratively optimal procedure of *k*-means cannot guarantee convergence to a global optimum.

- The sensitivity to outliers and noise distorts the cluster shapes. In fact, even if an observation is quite far away from a cluster centroid, the *k*-means forces that observation into a cluster.

- The mathematical definitions adjacent to the *k*-means algorithm limits its applications only to numerical data.

The parallel *k*-means algorithm that was implemented has n iterations in order to minimize the effect of the initial partition and since parallel computing techniques are applied, the complexity is (approximately) divided by the number of CPUs present in the computer.

2.3 Ambient Assisted Living

One of the main objectives of this research work is to apply the developed algorithm to Ambient Assisted Living. Therefore, in this section a brief description of what is Ambient Assisted Living, along with its objectives and relevance will be presented.

Ambient Assisted Living (AAL) belongs to a larger definition of assisting users in their activities called Ambient Intelligence (AmI). AmI is basically a digital environment designed to assist people in their daily lives without interfering with them [52, 53].

Through the use of wearable sensors, AAL aims at monitoring elderly and chronically ill patients at their homes. Therefore, one of the main goals of AAL is to develop technologies which enable users to live independently for a longer period of time, increasing their autonomy and confidence in accomplishing some daily tasks (known as ADL, Activities of Daily Livings) [54, 55].

Thus, AAL systems are used to classify a large variety of situations such as falls, physical immobility, study of human behaviour and others. These systems are developed using an Ubiquitous Computing approach [56] (where sensors and signals processing are executed without interfering on ADL) and must monitor activities and vital signs in order to detect emergency situations or deviations from a normal medical pattern [8]. Ultimately, AAL solutions automate this monitoring by using software capable of detecting those deviations.

2.4 State of the Art

Due to the constant evolution in sensing systems and computational power, biosignals acquisition and processing are always adapting to new technologies.

The main goal of clustering algorithms is to find information in data objects that allows to find subsets of interest — *clusters* — where objects in the same cluster have a maximum homogeneity. Therefore, the clustering base problem appears in various domains and is old, being traced back to Aristotle [51].

In fact, clustering algorithms can be typically applied to computer sciences, life and medical sciences, astronomy, social sciences, economics and engineering [5]. Due to these applications and area of research work, state-of-the-art developments in biosignals clustering and in Ambient Assisted Living (AAL) will be presented.

Applying clustering techniques to biosignals is an approach that has been used recently. Due to the large amount of data that is analysed nowadays, clustering techniques are used for feature extraction and pattern recognition on biosignals.

Clustering on ECG signals has been used in order to group the QRS complexes (or beats) into clusters that represent central features of the data. Lagerholm et al. (2000) [57] used a self-organizing network to perform beat clustering and detect different heart beats types. However, it was not made a heart rate variability analysis which could lead to a better detection of heartbeat types. On the other hand, Cuesta-Frau et al. (2002) [58] follow the previous approach but, after that, selected one single beat to represent all the beats in a cluster, allowing an automatic feature extraction of a long-term ECG. Since this approach requires a dissimilarity measure (measuring distances between arrays with different lengths) to obtain an input to the clustering algorithm, it has an extremely high computational cost. Similarly, Ceylan et al. (2009) [59] used a Type-2 *fuzzy c*-means (T2FCM) to improve the performance of a neural network, where T2FCM pre-classifies heart beats into clusters and a neural network is trained with the output of T2FCM, reducing also the training period. However, Chao et al. (2011) [60] used *c-medoids* to obtain optimal ECG templates that would be further used to train a classifier able to separate ECG signals from other types of biosignals. Thus, this work can only identify ECG signals and since there are more than 19 categories for these biosignals, in order to obtain a high accuracy it would require the construction of the templates of all categories.

Concerning clustering on EMG signals, Chan et. al (2000) [61] presented a classifier that was used to control prosthetics. However, in order to obtain high training speed, data features were clustered using the Basica Isodata algorithm and then fed to a backpropagation algorithm which input was used to determine which function would be executed by the prosthesis. Therefore, the delay between the onset of the EMG and the prosthesis control was reduced (estimated to be 300 ms) but this control resulted in a limited number (four) of movements. Also, Ajiboye et al. (2005) [62] presented an heuristic fuzzy logic approach to classify multiple EMG signals for multifunctional prosthesis control. In this study, the *fuzzy c*-means clustering algorithm was used to automate the construction of a simple amplitude-driven inference rule base, which is common when using a heuristic approach to solve a certain problem. The usage of simple inference rules allows a short delay (45.7 ms) but this algorithm only allows the control of one degree-of-freedom at a time, which limits the prosthesis in performing combined movements.

Due to the increasing amounts of data coming from all types of measurements and observations, some parallel computing techniques have been applied to clustering algorithms. These parallel techniques usually consist in performing *data parallel* and/or *task parallel* strategies. In the first strategy, the idea is to divide and distribute data into different processors and each one will compute the allocated data. The latter consists in dividing a main task into sub-tasks and dispatching them into different processors [63]. The master/slave strategy was used by [63, 64, 65], where the main program is run by the host, being in charge of data distribution and cluster results gathering. On the other hand, [66, 67] achieved running time improvements using a wider bus system and not more processors. Our modified *k*-means algorithm only uses the own computer's processors and therefore, does not require a system network implemented.

Since one of this research work applications is the Ambient Assisted Living, a set of interesting projects and studies will be presented next.

AAL aims at developing technologies that allow elderly people and chronically ill patients living in their home environment for a long period of time by assisting them in accomplishing their activities independently [68].

To accomplish AAL goals, a large number of projects were developed. In fact, the Aware Home [69], I-Living [70] and Amigo [71] projects are based on building intelligent environments (also called smart houses) where a software infrastructure allows electronic equipments to work together, providing a set of centred services that assist elderly people and chronically ill patients. However, the social component of this environments is usually undervalued and the COPLINTHO project [72] tries to solve this failure [54]. Unlike the aforementioned projects, AAL4ALL also develops services and technologies to assist AAL users and it has the goal to enter in the business market and commercialize their products [56].

Along with AAL projects, there is some recent studies that contributed to AAL objectives. In fact, Steinhage et al. (2008) [73] created a sensor network embedded on the floor where feet pressure generates events that allowed fall detection and elderly people's activity monitoring. Thus, users do not need to wear sensors embedded on cloths or be monitored by cameras, compromising their privacy, but the collected data has a high complexity and the installation costs of the system are quite high. On the other hand, Goshorn et al. (2008) [55] implemented a classifier capable of recognizing hand gestures, which generated commands for AAL communications, improving the recognition rates and increasing the available vocabulary since the set of hand gestures is based on the alphabet of anatomic hand postures. Since this approach is a supervised one, a limited number of hand gestures will be recognized. Besides, elderly people or chronically ill patients performing correct and stable hand gestures can represent a difficult task for them.

Clustering techniques are not commonly used on Ambient Assisted Living applications. Nevertheless, Hein et al. [74] used clustering techniques in order to identify high-level activities through wearable sensors, saving a quite large amount of time by by making unnecessary the annotation of the activities; besides, human annotation is error-prone and inaccurate because it does not occur *in situ* but afterwards. However, most clustering results can not be interpreted directly by a human and, consequently, it is necessary to include a classifier. Since k-Means needs a preset value for k, this number was determined by searching for a local maxima in the resulting accuracy. Therefore, it is impossible to state that the correct number of higher-level activities to be recognized are k, which can also vary along with the subjects. Similarly, Rashidi et al. [37, 75] propose an unsupervised method for finding discontinuous and varied-order activity patterns in a real world setting as part of a project. In fact, using a supervised approach requires a pre-definition of activities and the same activity can be performed by various means. However, this algorithm is unable to detect abnormalities present in activities, which could contribute to detect diseases or emergency situations on elderly people.

In fall detection domain, Luštrek et al. [76] developed a system where users wear accelerometers along with location sensors, allowing fall detection with contextual information. In this study, a variety of approaches were used, including unsupervised ones. However, in this study, it is necessary that subjects use many sensors in order to gather the required information, which can fail in reaching an ubiquitous approach, essential when performing human activity monitoring.

As could be verified, most of the presented approaches rely on developing tools that are specific to a certain type of biosignals, limiting its application in different types of biosignals that can be acquired from the human body. In this research work, we present a signal-independent algorithm able to detect events and perform clustering techniques in long-term biosignals. This algorithm will return an annotated signal that can be interpreted next, allowing the extraction of relevant information. Given these features, people monitoring becomes a clear application and since it does not require any prior information about the input signal (including its type), this algorithm represents a powerful tool in the signal processing field and in AAL.

3

Signal Processing Algorithms

In this chapter the implemented signal processing algorithms are thoroughly explained. The events detection algorithm based on the concept of *meanwave* [77, 78] is briefly explained and its improvements depicted. The concept of adaptive threshold using the signals' RMS to detect its events is characterized. Finally, a parallel version of the *k*-means clustering algorithm is also presented.

3.1 Events detection algorithms

3.1.1 Peaks detection approach

An event can be broadly defined as a change in state of the system under study [79]. When analysing cyclic biosignals, it is usual to use as reference points for events the signal peaks. In order to find those peaks, a threshold must be defined. In this study, the chosen threshold was the signal's root mean square (RMS). The mathematical definition of this feature is given by Equation 3.1.

$$RMS = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}|x[n]|^2} \tag{3.1}$$

with n ranging from 1 to N, where N represents the length of the signal.

In order to obtain a higher accuracy in detecting signal events, our algorithm updates the threshold every ten seconds. Figure 3.1 presents an example of a respiratory signal with the peaks detected using this approach; one can also observe that the horizontal lines are representing the evolution of the threshold (RMS) over time.

Although this approach brings interesting results, using the concept of waves and

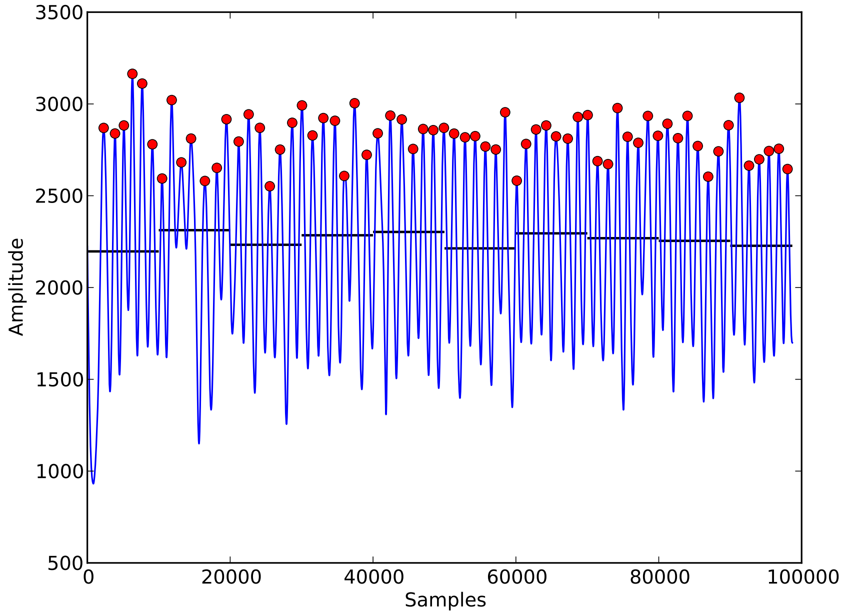


Figure 3.1: Peaks detections using the RMS of the signal as an adaptive threshold. The horizontal lines represent the evolution of the threshold over time.

meanwave in signals with noise, significant morphological changes or baseline deviations showed better results. Nevertheless, due to its simplicity and low computational cost which is fundamental since our algorithm is also designed to be applied in long duration records, using the signal's RMS as an adaptive threshold for peaks detections is also an interesting method for accomplishing this step of our algorithm. Once the peaks are detected, a *meanwave* is also constructed, obtaining one more source of morphological comparison of waves.

3.1.2 *Meanwave* approach

In this sub-section, the implemented events detection algorithm based on the concept of *meanwave* is presented. First, an algorithm overview is shown (see sub-section 3.1.2.1). In sub-section 3.1.2.2, the basic concepts used on our algorithm are characterized, providing the necessary contextual information and in sub-section 3.1.2.3, all the algorithm improvements are depicted.

3.1.2.1 Algorithm overview

The signal processing algorithms developed for this study can be divided into two distinct steps: the first one consists on detecting signal events; the latter performs a parallel version of the *k*-means clustering algorithm to the extracted information from the events detection step. Both phases are adapted to be run in long-term biosignals using parallel computing techniques to improve the execution speed of the algorithm.

An overview of our events detection algorithm is produced in Figure 3.2, where the main building blocks are shown. Due to its applicability to long-term signals, first the signal is divided into N parts and each one will be processed individually, determining the events of each part. Finally, the results are assembled and the signal events detected.

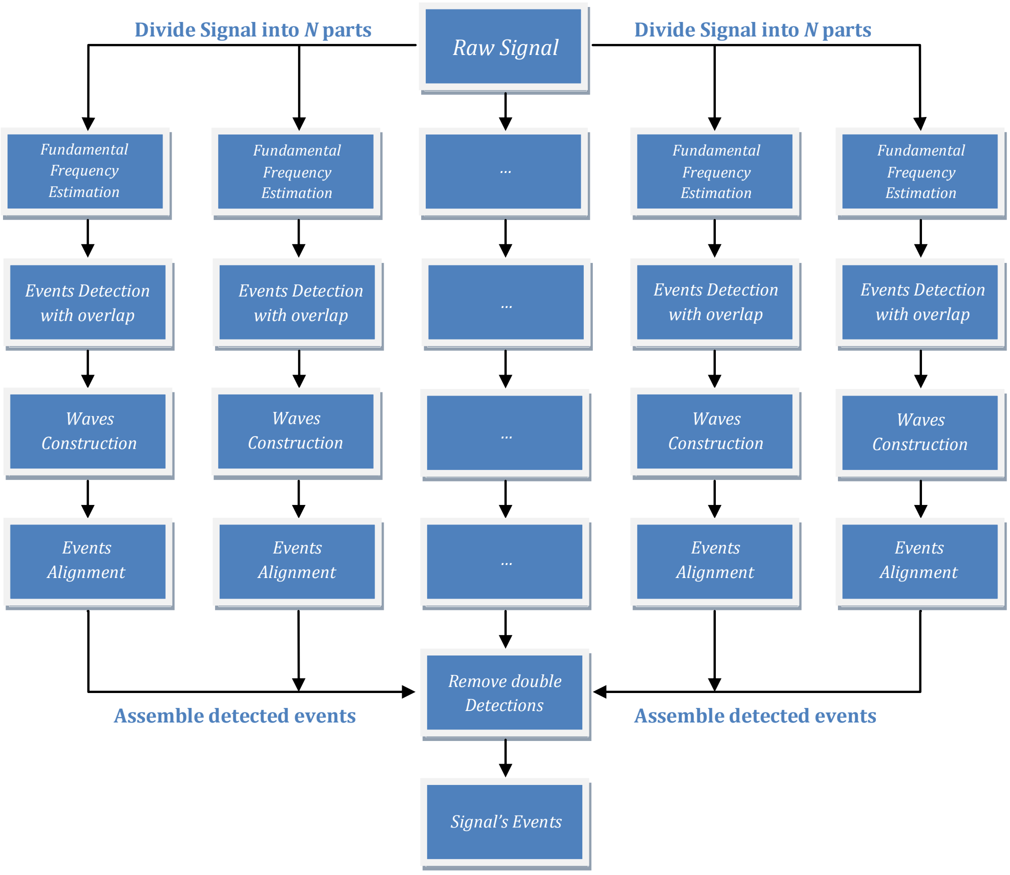


Figure 3.2: Signal Processing Algorithm Steps.

3.1.2.2 *autoMeanWave* algorithm

The *autoMeanWave* algorithm has the main goal of detecting events on biosignals and for that, a *meanwave* is automatically computed, capturing the signal's behaviour. In order to construct the *meanwave*, the signal must be cyclic and those cycles must be separated, making the fundamental frequency (f_0) estimation an essential part of the process [77, 78].

In the *autoMeanWave* algorithm, the first step consists of estimating the cycles size in samples. If the signal is periodic (or quasi-periodic), then the cycles size will be the smallest repeating unit of a signal – the repeating period pattern length. By calculating the signal first harmonic, it is possible to estimate the cycles size. For this purpose, in this algorithm the Fast Fourier Transform (FFT) of the signal is computed and the first peak found (after applying a smoothing filter). This peak is assumed to correspond to the signal first harmonic and, consequently, the signal f_0. Hence, the cycles size — *winsize* — is estimated and given by:

$$winsize = \frac{f_s}{f_0} \times 1.2 \tag{3.2}$$

Then, a random part of the original signal (window) with a length of *winsize* (N) is selected and a correlation function is applied to calculate a distance signal showing the difference between each overlapped cycle (signal$[i : i + N]$) and the window selected at the first place. In Equation 3.3 it is defined the correlation function used in the *autoMeanWave* algorithm where d_i represents an element of the distance signal, with i ranging from 1 to $M - N$, being M the length of the signal.

$$d_i = \frac{\sum_{j=1}^{N} |\text{signal}[i : i + N]_j - \text{window}_j|}{N} \tag{3.3}$$

Then, the local minima of the distance signal are found and assumed to be the signal events. Finally, the *meanwave* is computed and the signal events are aligned using a reference point which can be chosen among a set of options.

3.1.2.3 Algorithm improvements

Fundamental frequency estimation Although the basic concept of the events detection algorithm was shown in the previous sub-section, some improvements were made and the possibility of applying this algorithm in long-term biosignals was added. In fact, the fundamental frequency estimation is one of the most important parts of our algorithm and an accurate method is extremely necessary.

Being f_e and $winsize_e$ the estimated fundamental frequency by the algorithm and the cycles sizes, respectively, two scenarios might arise:

- If $f_e \ll f_0$, then $winsize_e \gg winsize$. Thus, the length of the distance signal computed by the correlation function will be much smaller and, therefore, the number of local minima will also be smaller. Since the local minima represent the signal events, a great number of events will be dismissed, resulting in a poor estimation of signal's number of cycles.

- If $f_e \gg f_0$, then $winsize_e \ll winsize$. In opposition to the previous scenario, the length of the distance signal computed by the correlation function will be much greater and, therefore, the number of local minima will also be greater. Thus, a great number of cycles that do not exist will be considered.

Therefore, instead of determining the first peak of the signal FFT, we used a time-domain method for f_0 estimation based on the autocorrelation of time series. Since the correlation can be defined as a measure of the similarity between two waves, the autocorrelation function is the correlation of a wave with itself. Given that we deal with periodic (or quasi-periodic) signals, the autocorrelation function is also periodic. In fact, if the time lag is none, then the waves are in phase and the autocorrelation function reaches a maximum; as the time lag increases to half of the period, the autocorrelation function reaches a minimum since the wave and its time-delayed copy are out of phase. Once the time lag reaches the length of one period, both waves are again in phase and the autocorrelation increases back to a maximum [80].

If we are dealing with an infinite time series, $y[n]$, the mathematical definition of the autocorrelation function is given by the Equation 3.4. Since biosignals are finite time series that could be expressed by $x[n]$, the autocorrelation function is given by Equation 3.5.

$$R_y(v) = \sum_{n=-\infty}^{+\infty} y[n]y[n+v] \tag{3.4}$$

$$R_x(v) = \sum_{n=0}^{N-1-v} x[n]x[n+v] \tag{3.5}$$

In order to compute the autocorrelation function, we used the convolution between the signal FFT and the FFT of the signal reversed in time. A representation of the autocorrelation function and the extracted information from it is presented in Figure 3.3.

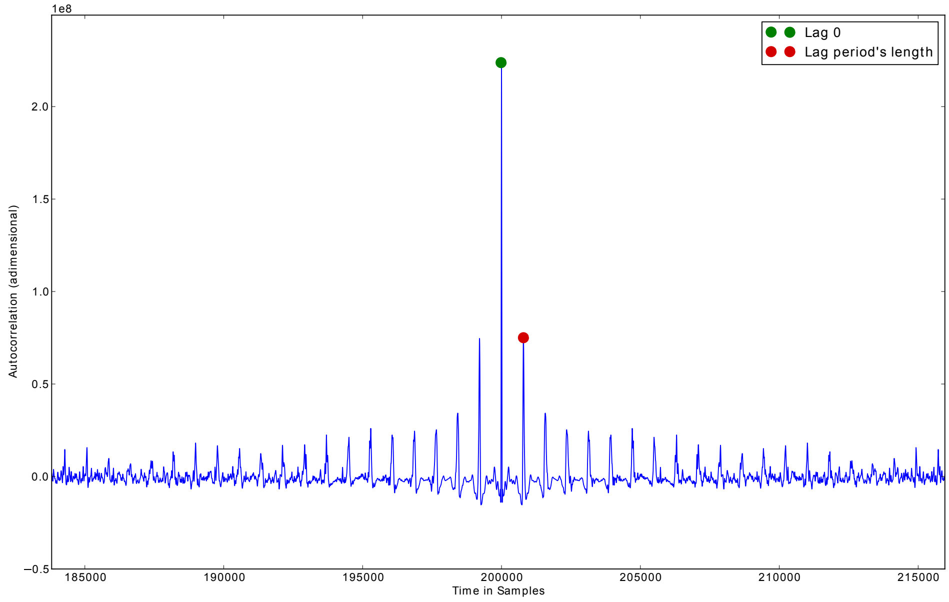


Figure 3.3: Autocorrelation function of a cyclic signal.

After throwing away the negative lags, the first peak is found, corresponding to the first instant where both signals (the original and the reversed in time) are in phase with lag 0. Since the next peak will represent the next instant when the lag reaches the length of one period, finding this peak will allow to estimate the period length and, consequently, the fundamental frequency.

In order to obtain a more accurate estimation of the fundamental frequency, a quadratic interpolation for estimating the true position of an inter-sample maximum when nearby samples are known was used. Given a function f and an index i of that function, the coordinates x e y of the vertex of a parabola that goes through point i and its two neighbours are given by:

$$x = \frac{1}{2} \times \frac{f(i-1) - f(i+1)}{f(i-1) - 2f(i) + f(i+1)} + i$$

$$y = f(i) - \frac{1}{4} \times [f(i-1) - f(i+1)](x-i)$$

Hence, giving as input the autocorrelation function and the peak sample where the lag reaches the length of one period, we obtain the window size using Equation 3.2, where $f_0 = x$. However, we widened the window by 30% due to the higher accuracy returned by this method for estimating the fundamental frequency.

It is well known that there are many ways of computing the fundamental frequency since this is a current and active research topic. In fact, an ultimate method for f_0 estimation is yet to be discovered [80]. However, the autocorrelation approach proved to return more accurate results than the previously presented method.

Events alignment When a more accurate estimation for f_0 was implemented, we also improved the signal events alignment step. In fact, in order to obtain an accurate morphological comparison between waves, an almost perfect alignment of the signal events is required. In [77, 78] the alignment is achieved by selecting a notable point from the computed *meanwave*. For certain types of signals, this led to an inaccurate events alignment and, therefore, an incorrect distance measure between waves and between the *meanwave*.

In order to solve this issue, our algorithm performs two phases of events alignment. First, the events are aligned using a notable point (minimum or maximum value) from the computed *meanwave*. This notable point is defined as an input of the alignment algorithm and if none is given, the maximum value is the default one. Subsequently, our algorithm builds all the waves based on the previously aligned events, where each wave is centred in the correspondent event. The length of a wave, l_{wave}, is given by computing the differences between two consecutive events and averaging those values; thus, the events will be located at the sample $\frac{l_{wave}}{2}$ of each wave. Finally, our algorithm runs through all the computed waves and relocates the events to the minimum or maximum value of each wave. Being $event_i$ (with i ranging from 1 to l_{wave}) the final wave-sample of one event, the $shift_i$ that will be applied to relocate it is given by:

$$shift_i = event_i - \frac{l_{wave}}{2} \tag{3.6}$$

Computing this for all waves, an array is obtained containing the shift values to be applied to all signal-samples of the events. Therefore, being $event_j$ (with j ranging from 1 to the signal length) the signal-sample of one event, its final position will be given by:

$$event_j = event_j + shift_i \tag{3.7}$$

An example of this further events alignment is shown in Figure 3.4.

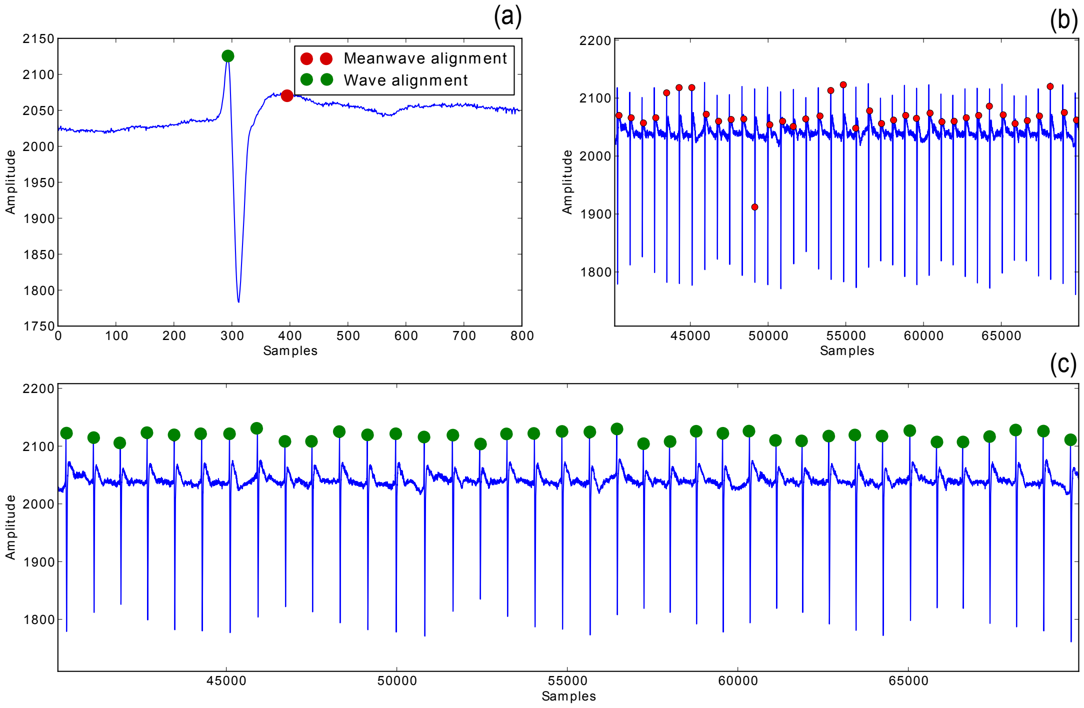


Figure 3.4: (a) Highlight of event alignment using a *meanwave* and a wave reference point; Events alignment in an ECG signal using (b) the *meanwave* and (c) the wave reference points.

After this final alignment, new waves based on these events are constructed and the distance between each wave and its *meanwave* is computed, returning accurate wave alignments.

The differences between the two alignment approaches can be observed in Figure 3.5. It is important to notice that not only an inaccurate waves alignment leads to incorrect distance measures between the waves and the *meanwave* but it also affects its construction.

Long-term biosignals applicability The last improvement made on the *autoMeanWave* algorithm is the ability to run over long-term biosignals. In order to accomplish this goal, our algorithm divides the signals into N parts and each part is processed individually. Hence, being L the length of the original signal, each part will have a length of $L_p = L/N$; it is important to notice that the last part might be smaller than the remaining ones. The length of each part is an input for our algorithm and affects its sensitivity to the fundamental frequency evolution. In fact, if there are significant changes in the

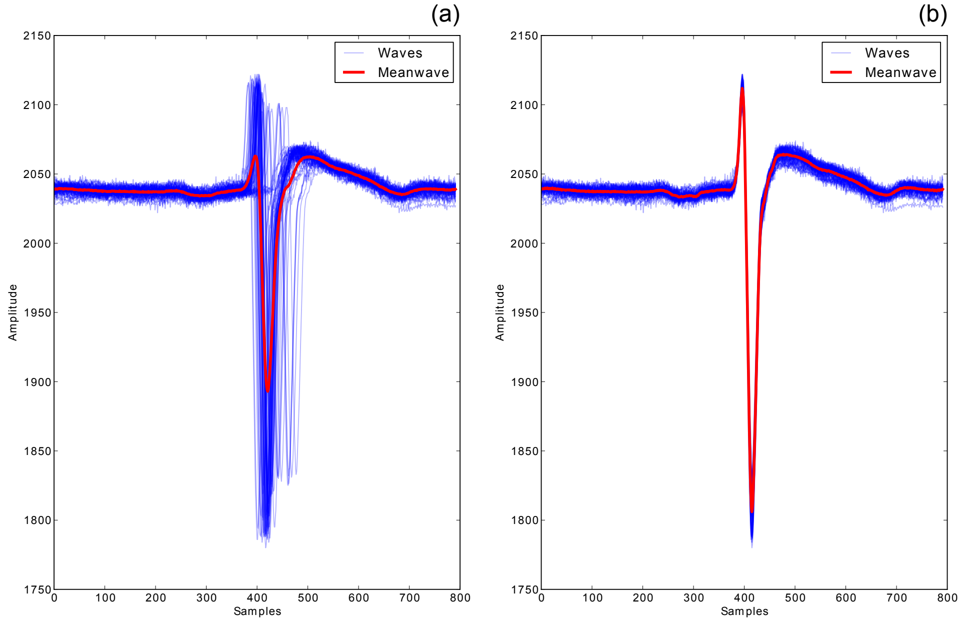


Figure 3.5: Waves and *meanwave* alignment and construction using (a) *meanwave* reference point and (b) waves reference points.

signal's fundamental frequency, using larger parts might cause loss of sensitivity to those changes.

To guarantee that no information is lost among transition zones, we introduce a f_0-dependent overlap with length L_o, resulting in a total length for each part of $L_{pf} = L_p + L_o$. The overlap is defined as follows:

1. Select a random part of the signal with length L_p. In our algorithm, the selected part of the signal is

$$signal_{part} = signal\left[halflen - \frac{L_p}{2}; halflen + \frac{L_p}{2}\right]$$

   with $halflen$ being half of the signal length.

2. Compute the fundamental frequency using the approach described in 3.1.2.3.

3. Estimate the cycles size, $winsize_{part}$, according to the given information about the original signal.

4. Define the overlap as

$$overlap = winsize_{part} \times 6$$

With this overlap it is expected that at most six events are detected twice during the transition from one part to the next. This will result in double detections and to remove them, the concept of neighbourhood of a number is applied. We define the neighbourhood (with a radius of ϵ) of a number n as the set:

$$V_\epsilon(n) =]n - \epsilon; n + \epsilon[\tag{3.8}$$

Using Equation 3.8 and defining $\epsilon = 0.3 \times winsize$ and $n = e_{i-1}$ (with i ranging from 2 to N), where e_{i-1} is the last detected event of the part $i-1$, we define the neighbourhood of e_{i-1} as the set:

$$V(e_{i-1}) =]e_{i-1} - 0.3 \times winsize; e_{i-1} + 0.3 \times winsize[$$

In order to remove the double detections, if the event e_i from part i belongs to the set $V(e_{i-1})$, then all the events that precede e_i (including it) are eliminated.

To overcome the obstacle of dealing with large amounts of data, the HDF5 format was used. Hence, the storage of large sized data and its fast access is possible, being these features the main advantages of HDF5 files [81].

Once the events are correctly detected and aligned, distance measures are taken and clustering techniques are applied to obtain signal annotations.

3.2 Distance measures

The previously extracted events allow a duly indication of the signal's waves. Therefore, distance measures can be taken between waves and between each wave and the signal's *meanwave*.

There are several distance measures that can be applied to one-dimensional arrays and more specifically, to time-series. In order to obtain inputs to our parallel *k*-means algorithm, we use a set of different distance functions. First of all, the Minkowski-form Distance defined as [61]

$$L_p(P,Q) = \left(\sum_i |P_i - Q_i|^p\right)^{1/p} \quad , \quad 1 \leq p \leq \infty \tag{3.9}$$

In this study, we will use the L_1, L_2 and L_∞ distance functions, which are defined as

$$L_1(P,Q) = \sum_i |P_i - Q_i|$$

$$L_2(P,Q) = \sqrt{\sum_i (P_i - Q_i)^2}$$

$$L_\infty(P,Q) = \max_i |P_i - Q_i|$$

The squared version of L_2, L_2^2, will also be used. Besides, the χ^2 histogram distance given by [82]

$$\chi^2(P,Q) = \frac{1}{2}\sum_i \frac{(P_i - Q_i)^2}{P_i + Q_i} \tag{3.10}$$

will also be utilized to obtain distance measures.

Due to the suitability of our algorithm in long-term biosignals, these distance measures are not represented by a distance matrix. In fact, biosignals can be seen as time-series which have an important feature that allows distance measures without building an extremely high computational cost distance matrix when dealing with long records: the order of relationship between two consecutive samples.

Figure 3.6 represents a distance matrix for an ECG signal and the highlight of the first 30×30 elements.

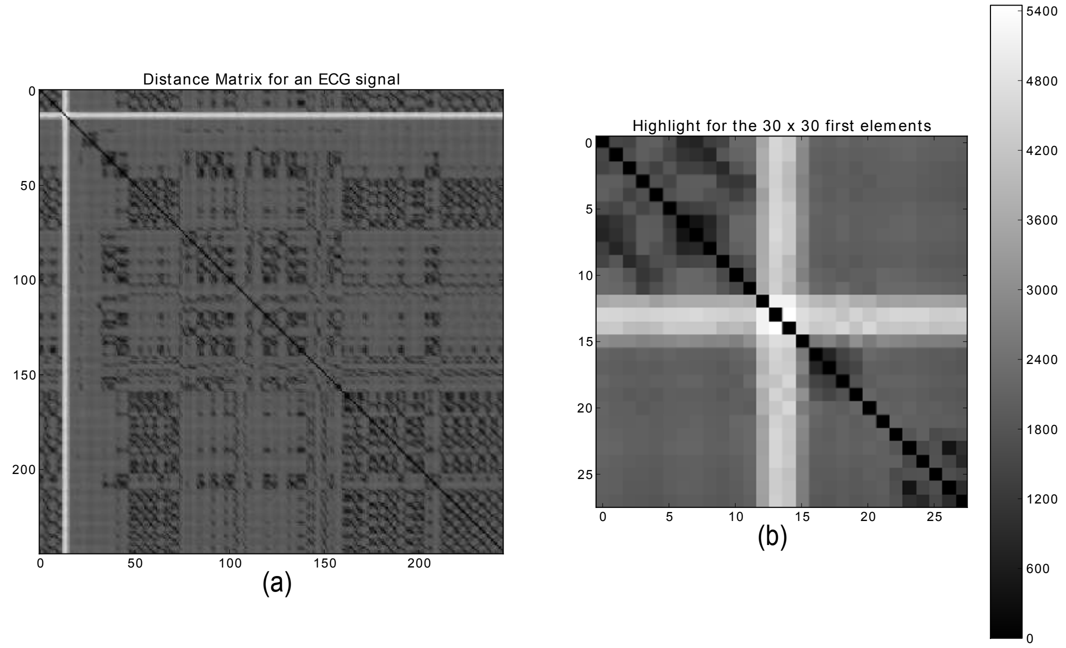


Figure 3.6: (a) Distance matrix for an ECG signal and (b) the highlight of the first 30×30 elements of the distance matrix.

Observing the 30×30 distance matrix, it is clear that the observations number 11, 12, 13 and 14 are significantly different from the other ones, resulting in two transition zones: the transition from the 10^{th} observation to the 11^{th} and from the 14^{th} observation to the 15^{th}. Since we are dealing with time-series, the concept of transition zone is valid and therefore, instead of searching for the resemblance between each observation and the other ones to build a distance matrix, it is only necessary to find the transition zones.

In order to find the transition zones, the distances between each observation and the consecutive one must be computed. In a distance matrix, these distances correspond to those elements which are adjacent to the diagonal of the matrix. Hence, morphological comparisons between waves, w_i, result in a *distance array* where each element, a_i, is given by:

$$a_i = f(w_i, w_{i+1}), i = 1, \ldots, n - 1 \tag{3.11}$$

being f the distance function and n the number of waves, representing also the number of events detected by the previous step of our algorithm.

In the particular case where $w_i = mw$, where mw denotes the signal's *meanwave*, then each element of the distance array will be:

$$a_i = f(mw, w_i), i = 1, \ldots, n \tag{3.12}$$

Although the distance matrix carries richer information about waves resemblance than the distance array, its high computational cost makes it impracticable in long records.

Using this set of distance functions, a comparison between the efficiency of each one as an input for our clustering algorithm will be made, allowing to state which is the most adequate distance function for morphological analysis in biosignals.

3.3 Parallel *k*-means algorithm

As the final step of our algorithm, a parallel *k*-means was implemented, being able to perform unsupervised learning on long-term biosignals. The main concept of the *k*-means algorithm was kept [49, 50], which is a partitioning method for clustering where data is divided into *k* partitions [38]. The optimal partition of the data is obtained by minimizing the sum-of-squared error criterion with an interactive optimization procedure. Clustering algorithms can perform hard-clustering when each cluster can only be assigned to one partition; otherwise, they perform fuzzy-clustering. Our algorithm was designed to perform hard-clustering since it was based in the *k*-means hard clustering algorithm. As an additional modification to the original *k*-means, we also introduce n iterations to the algorithm in order to minimize one of the biggest drawbacks related to the initial partition. In fact, different initial partitions usually converge to different cluster groups [5, 6].

As Figure 3.7 suggests, the parallel k-means algorithm receives as input a set of observations. These observations come from the distance measures previously presented. In order to run this parallel version of the k-means algorithm over large sized data, the observations (with the initial length of M) are divided into N parts. Due to the fixed size of each part, the last part might be smaller than the remaining ones.

After dividing the observations into parts, our algorithm performs the *k*-means clustering algorithm in each part individually. For each part a different set of k centroids will be obtained. The i^{th} part, for example, will have the centroids $[\mathbf{a}_i, \mathbf{b}_i, \ldots, \mathbf{k}_i]$, with i ranging from 1 to N. It is important to notice that if each observation is n-dimensional, each centroid will also be n-dimensional.

Since the *k*-means algorithm randomly assigns clusters to the computed *k* partitions, different clusters assignment when performing *k*-means repeatedly over the same set of

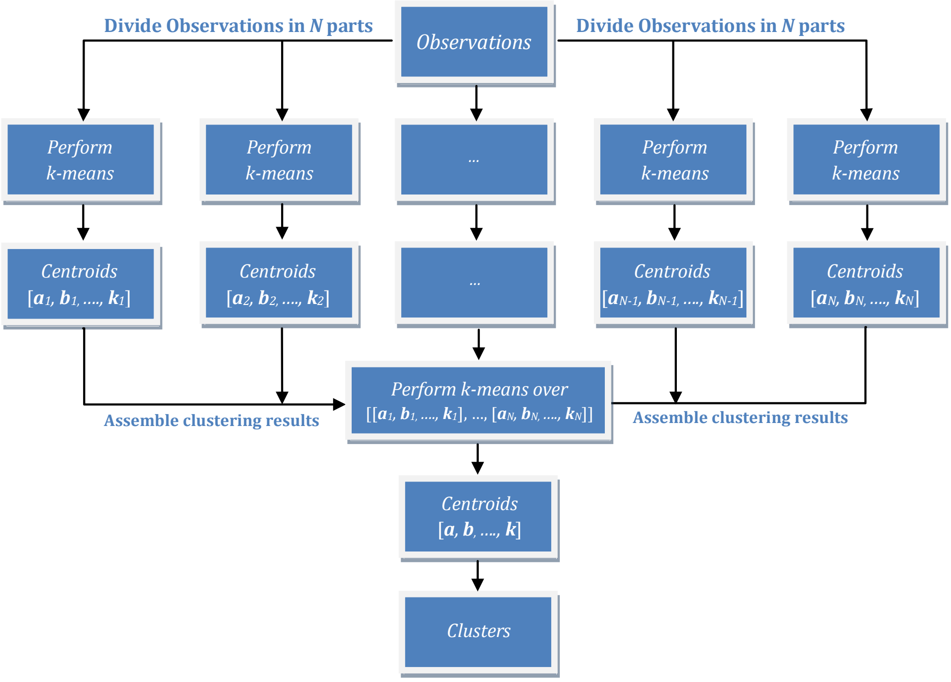

Divide Observations in N parts
Observations
Divide Observations in N parts
Perform k-means
Perform k-means
...
Perform k-means
Perform k-means
Centroids [a1, b1, ..., k1]
Centroids [a2, b2, ..., k2]
...
Centroids [aN-1, bN-1, ..., kN-1]
Centroids [aN, bN, ..., kN]
Perform k-means over [[a1, b1, ..., k1], ..., [aN, bN, ..., kN]]
Assemble clustering results
Assemble clustering results
Centroids [a, b, ..., k]
Clusters

Figure 3.7: Parallel k-means algorithm schematics.

observations will be obtained. Hence, the probability to find observations belonging to different parts but being similar between them and having different clusters assigned is elevated. In order to solve this problem, a way to assemble the computed set of centroids from each part must be found in order to obtain a single set of centroids representing the observations as a whole. In fact, assembling all the N set of centroids, we obtain:

$$[[\mathbf{a}_1, \mathbf{b}_1, \ldots, \mathbf{k}_1], \ldots, [\mathbf{a}_N, \mathbf{b}_N, \ldots, \mathbf{k}_N]]$$

Considering this as new set of k-dimensional observations, our algorithm runs one more time the k-means algorithm, obtaining the *global centroids*, $[\mathbf{a}, \mathbf{b}, \ldots, \mathbf{k}]$, of the k partitions of the original observations (with length M). With these centroids, the Euclidean Distance (defined as the L_2 distance) is computed between each centroid and each observation, resulting in a $k \times M$ matrix. Searching for the line where the minimum element of each row is located gives us the cluster which that observation will be assigned to. Due to this approach, all of the k-means issues will be amplified but still, the obtained results are quite satisfactory.

3.4 Signal processing algorithms overview

After presenting a detailed description of the signal processing algorithms implemented for this study in the previous sections, an overview is provided in Figure 3.8.

First, an events detection algorithm based on peaks detection through an adaptive

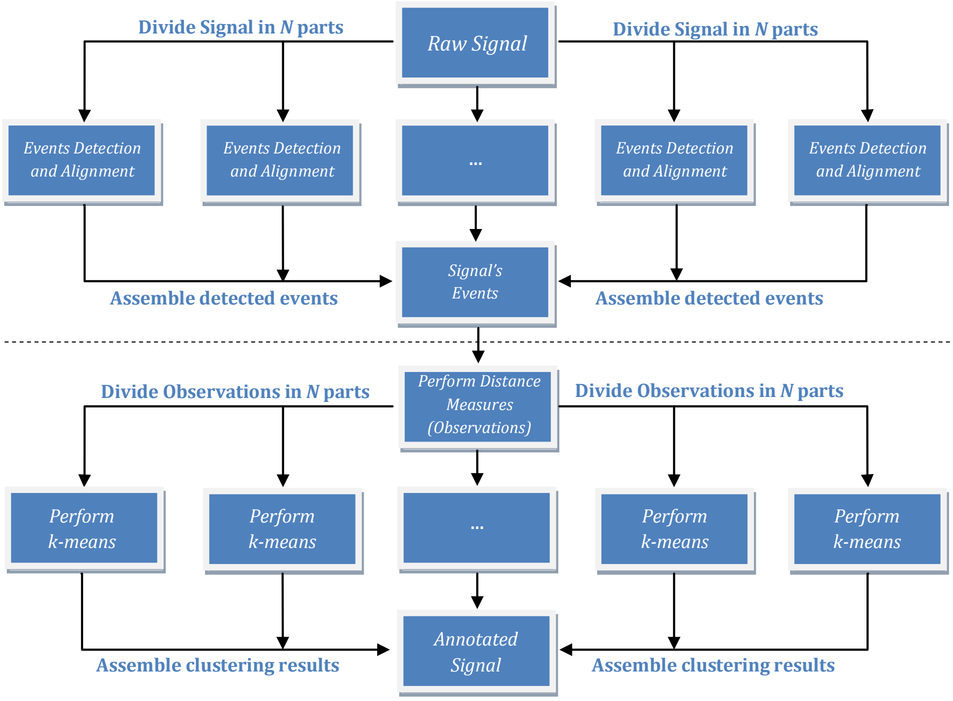

Raw Signal
Divide Signal in N parts
Divide Signal in N parts
Events Detection and Alignment
Events Detection and Alignment
...
Events Detection and Alignment
Events Detection and Alignment
Signal's Events
Assemble detected events
Assemble detected events
Perform Distance Measures (Observations)
Divide Observations in N parts
Divide Observations in N parts
Perform k-means
Perform k-means
...
Perform k-means
Perform k-means
Annotated Signal
Assemble clustering results
Assemble clustering results

Figure 3.8: Global schematics for the signal processing algorithms.

threshold defined as the signal's RMS or based on the concept of *meanwave* is applied. Then, distance measures are taken in order to obtain inputs for the parallel *k*-means clustering algorithm, which will perform unsupervised classification and return an annotated signal.

The validation of our algorithm will be presented in the next chapter.

4

Algorithms Performance Evaluation

In this chapter the performance of the events detection and the parallel *k*-means algorithms is assessed. Both evaluations will be exposed together due to the relationship between them.

4.1 Evaluation overview

The performance evaluation of our algorithms will be taken in parallel. Since the implemented parallel *k*-means has as input distance measures based on a previous events detection, the assessment of this first step is always issued before discussing the clustering results.

In order to test our algorithm, a set of different types of biosignals was acquired and some of them representing long-term signals. This set-up enabled to show the suitability of the implemented algorithms in processing different types of biosignals and without requiring any prior information about them; besides, the ability to be run in long-term signals is also shown.

A set of signals from a public database (PhysioBank) [83, 84] at PhysioNet was also used in this performance evaluation, since these signals were already annotated, bringing a good comparison to our algorithm annotations.

Finally, a set of synthetic signals was also generated, representing a controlled environment since the events and clustering results were previously known.

4.2 Evalution using Synthetic Signals

For the first part of our algorithm evaluation, a set of synthetic signals was generated based on a random walk with a varying number of samples. These signals have a different number of modes but no major variations on the fundamental frequency.

The obtained results for the events detection and clustering procedure are presented in Tables 4.1 and 4.2, respectively. For our evaluation, we defined an *error* as a cycle that is wrongly identified or clustered and a *miss* as a cycle that is not identified or clustered. These results were based on five synthetic signals that are represented by Synthetic$_i$ (with i ranging from 1 to 5) in the following tables.

Signal	**Cycles**	**Detected cycles**	**Errors**	**Misses**
Synthetic$_1$	52	51	0	1
Synthetic$_2$	62	61	0	1
Synthetic$_3$	540	539	0	1
Synthetic$_4$	619	618	0	1
Synthetic$_5$	948	947	0	1

Table 4.1: Results concerning the developed events detection algorithm on synthetic signals.

Signal	**Cycles**	**Correctly clustered cycles**	**Errors**	**Misses**
Synthetic$_1(k = 3)$	52	51	0	1
Synthetic$_2(k = 4)$	52	51	0	1
Synthetic$_3(k = 3)$	540	539	0	1
Synthetic$_4(k = 4)$	619	618	0	1
Synthetic$_5(k = 6)$	948	947	0	1

Table 4.2: Clustering results on synthetic signals. The value of k represents the number of modes present in the signal and also the number of clusters.

It is important to notice that only the clustering results from the computed distances between each signal's cycles and the signal's *meanwave* are presented due to the poor results obtained when using the other distance functions. This might be related to the fact that the distance measures lead to a distance array instead of a distance matrix, which has a quite lower computational cost but also carries poorer information about the cycles resemblance. This issue was already addressed in section 3.2.

Since our algorithm is running through synthetic signals, a high performance was expected. An example of an annotated synthetic signal is produced in Figure 4.1.

In fact, analysing the events detection results, our algorithm detected 2216 out of 2221 events, representing an efficiency of 99.8%. All the detected events were perfectly aligned with the chosen reference point (the local maxima as default). Besides, it is important to state that the single miss that was found in all synthetic signals corresponded

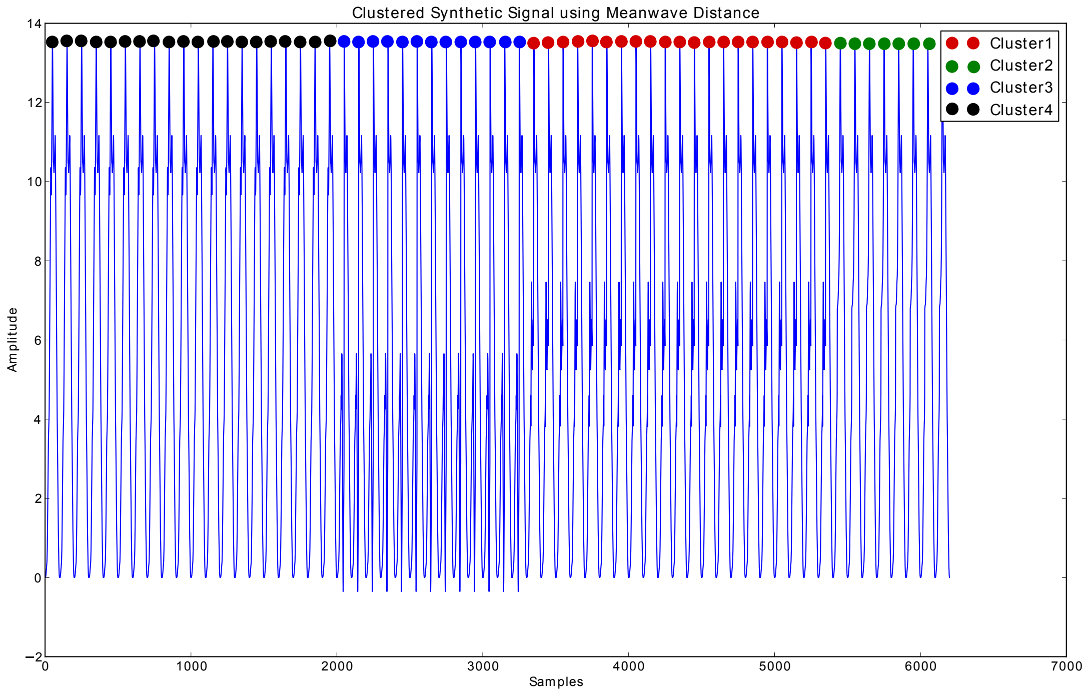


Figure 4.1: Clustered synthetic signal with four different modes and with no changes in the fundamental frequency.

to the last cycle of the signal, which can be smaller than the remaining ones. Hence, our algorithm doesn't run through that cycle since a complete morphological comparison between waves is impossible.

Relatively to the clustering results, our algorithm was able to separate all the modes that were present in the synthetic data. However, the synthetic signals that were generated using more than six modes started to return poor results, which is an obvious limitation of our algorithm.

4.3 Evaluation using Acquired Signals

4.3.1 Acquisition System

In order to accomplish this phase of our algorithm performance evaluation, a set of different types of biosignals was acquired. For that, four types of sensors were used: an ECG sensor (*ecgPlux*), a triaxial accelerometer sensor (*xyzPlux*), a BVP sensor (*bvpPlux*), a respiratory sensor (*respPlux*) and an EMG sensor (*emgPlux*).

All the sensors were connected to the channels of a device — *bioPlux* research unit — and the signals were acquired continuously and in real time. The *bioPlux* is responsible for the signal analogue to digital conversion since it has an integrated 12 bit ADC. The maximum sampling frequency is 1000 Hz, which is the one used for the acquired signals for this research work.

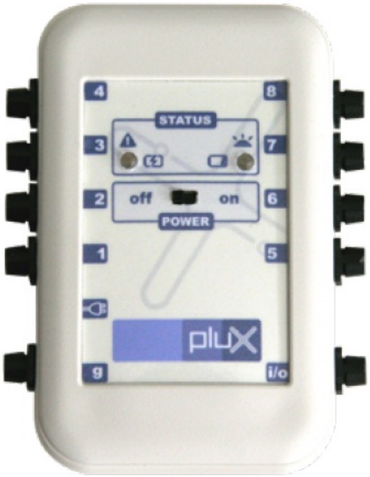


Figure 4.2: *bioPlux* Research system.

The acquisition system is also portable, small sized and light-weighted and is also responsible for the wireless transmission of the digital signals to a computer using a communication protocol based on the Bluetooth technology.

Biosignals based on different scenarios were acquired and some of them were long-termed in order to test the suitability of our algorithm in large sized data.

4.3.2 Acquired Signals

The ECG signals were obtained in different contexts. A 7 hour signal was acquired during a night of sleep of a person diagnosed with amyotrophic lateral sclerosis, a progressive neurodegenerative disease characterized by the loss of neurons at all level of the motor system. It is also considered one of the most puzzling diseases concerning its pathogenesis [85, 86]. Several patients are being monitored during the night through the previously presented *bioPlux* under the project *wiCardioResp* [7].

A set of ECG signals was also acquired under the project *ICT4Depression*. The project goal is to monitor home patients with depression and collect their extracted biosignals, in order to provide a more efficient way of treatment (in other words, a personalized treatment). For research and evaluation purposes, two ECG signals were acquired right after a subject performed some exercise and then, at rest. During the acquisition of these ECG signals, BVP signals were simultaneously obtained. Two ECG signals were also acquired from pregnant women in an hospital environment in order to monitor their vital signs but also their growing fetus. For these acquisitions, the electrodes were placed in the abdominal region in order to detect fetal ECG.

Respiratory signals (Resp) were acquired under both *wiCardioResp* and *ICT4Depression* projects and also for research and evaluation purposes.

Finally, a group of accelerometry (ACC) signals was acquired representing different tasks. For the acquired ACC signals under the *ICT4Depression* project, subjects were walking at average speed. Besides these signals, we created some scenarios that would enable the acquisition of ACC signals with different modes. These signals are described as follows.

- **Activity 1: Walk, Run, Walk, Jump**. In this task, the accelerometer was located on the right hip along with the *bioPlux*, so that the y axis of the accelerometer was pointing downward. It was asked to the subjects to walk (for about 1 minute and half), run, walk again and jump on the same place (each for about 1 minute). These four modes were executed non-stop.
- **Activity 2: Crouching, leg flexion and leg elevation**. In this task, the subject was standing straight with both feet completely on the ground and was asked to perform 10 squats followed by 10 vertical leg flexions —- moving the heel towards the gluteus — and 10 leg elevations, moving the knee towards the chest. The subjects used an accelerometer located at the right hip and oriented so the y axis was pointing downward.
- **Activity 3: Jumping, leg flexion and single leg vertical jumping**. In this task, the following procedure was executed: normal vertical jumping, leg flexion and single leg vertical jumping. Each mode was repeated 10 times. The subjects used an accelerometer located at the right hip and oriented so the y axis of the accelerometer was pointing downward.

An EMG signal was also acquired from subjects performing Activity 2. To conclude the presentation of the acquired signals for this study, it is important to state that every signal was acquired with a sampling frequency of 1000 Hz.

4.3.3 Pre-processing phase

Although it is not mandatory for our algorithm, a signal-specific pre-processing phase was applied to some signals. For the ECG signals, a 25 Hz lowpass filter was applied in order to remove most of the noise. A lowpass filter was also applied to respiratory signals, but in this case with a threshold of 1.6 Hz. The ACC signals were low-passed with a smoothing (lowpass) filter with a moving average of 200 samples. The EMG signal was also smoothed with a moving average of 200 samples after computing the signal envelope. Finally, no pre-processing was applied to the BVP signals.

4.3.4 Results

Since our algorithm aims at extracting information in a broad perspective, two different approaches were taken, resulting in two independent types of clustering results. First, our algorithm verifies the time-samples difference; finally, it performs a morphological comparison between waves (see Equation 3.11 and 3.12).

A visual inspection for performance evaluation was taken and different criteria were used for the different types of clustering results. However, the concepts of *error* (when a cycle is wrongly classified or identified) and *miss* (when a cycle is not classified) are used in both type of results.

The validation process was taken using only the *meanwave* approach to obtain signals events (see sub-section 3.1.2) since its higher accuracy is necessary to obtain more reliable clustering results.

Clustering using time-samples difference information

Despite its conceptual simplicity, an almost perfect events detection and alignment can lead to a time-sample variability analysis between those events. In fact, if the events $\mathbf{e} = \{e_1, \ldots, e_n\}$ are perfectly aligned in a cyclic signal, a time-sample difference array, Δ, can be computed where each element, Δ_i is given by:

$$\Delta_i = e_{i+1} - e_i \quad , \quad i = 1, \ldots, n-1 \tag{4.1}$$

Since we are dealing with cyclic signals, analysing this information will allow us to study the variability over the duration of each cycle. An example is shown in Figure 4.3 where an ECG signal is represented as well as the clustered events based of the time-sample variability.

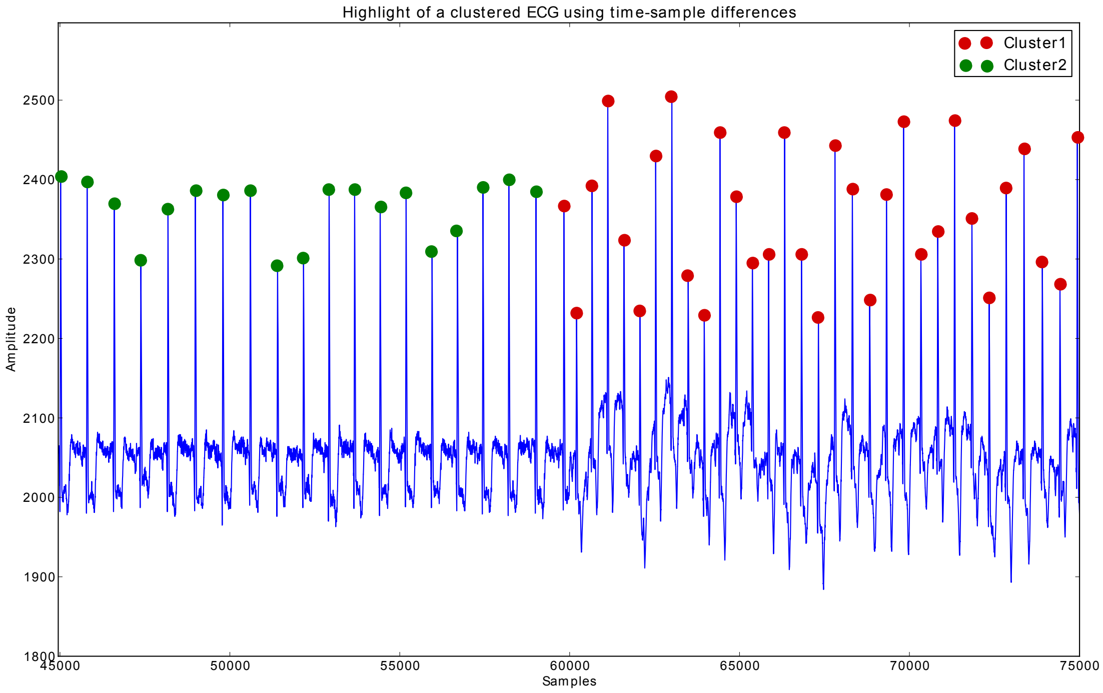


Figure 4.3: Clustering using events time-samples variability. A clear transition between the rest and the exercise state of the subject is annotated using this information.

With this information only, we are able to conclude that during the first minute (recall that the x axis is in samples and that the sampling frequency for all the acquired signals was 1000 Hz) the subject was at rest due to the signal's lower frequency represented by a longer Δ_i intervals and afterwards, the subject started some activity or exercise verified by Δ_i intervals reduction.

After running this clustering method in order to divide the data into k partitions, the obtained results are presented in Table 4.3.

Signal	Cycles	Correctly clustered cycles	Errors	Misses
ECG$_1$ (wiCardioResp)	24551	24279	0	272
Resp$_1$ (wiCardioResp)	5210	5163	1	46
ECG$_2$ (Exercise/Rest)	224	222	0	2
BVP$_1$ (Exercise/Rest)	224	223	0	1
ECG$_3$ (Rest/Exercise)	165	162	2	1
BVP$_2$ (Rest/Exercise)	165	162	2	1
ECG$_4$ (Walking - ICT)	199	198	0	1
ECG$_5$ (Resting - ICT)	170	165	2	3
ACC$_1$ (Walking - ICT)	132	131	0	1
ACC$_2$ (Walking - ICT)	184	183	0	1
Resp$_2$	67	65	1	1

Table 4.3: Clustering results using Δ_i with $k = 2$ clusters

The ACC signals extracted from the previously described activities were not used is this part of the algorithms evaluation since their major difference between modes resides in their morphology, having minimum changes in the fundamental frequency.

Since the ECG$_2$, BVP$_1$, ECG$_3$ (represented in Figure 4.3) and BVP$_2$ are signals whose major difference is their change in frequency, using Δ_i information returns better results then the morphological comparison if the goal is to separate the resting state from the exercising state. However, morphological analysis can achieve other type of information which might also be relevant in aiding physicians' manual analysis, for example. For the rest of the ECG signals, a heart rate variability (HRV) analysis should be taken in order to assess the clustering results.

Analysing the presented results, our events detection algorithm identified 30960 out of 31291 cycles, achieving an efficiency of 98.9%. Concerning the clustering results, our algorithm correctly clustered 30953 out of 31291, achieving an efficiency of 98.9%. It is important to state that, as for the synthetic signals, most of the failures occurred in the last cycle, which is usually smaller then the remaining ones, hindering our algorithm to run along those smaller cycles.

Clustering using morphological comparison

Next, a morphological analysis was taken in order to obtain signals annotations. As it was mentioned before in Section 3.2, our algorithm uses a set of distance functions in order to study which one returns better results. An example of an annotated ECG signal is illustrated in Figure 4.4.

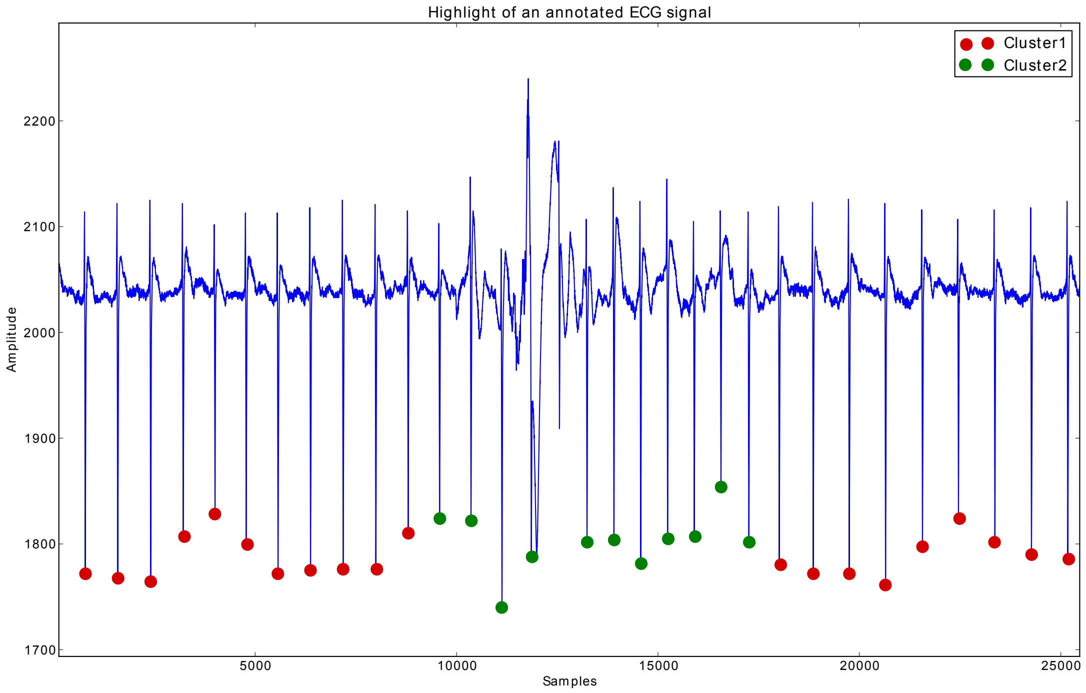


Figure 4.4: Highlight of an annotated ECG signal where the noise area (green dots) is indicated as different from the remaining areas of the represented signal (red dots).

The obtained results are presented in Tables 4.4 and 4.5. The global results are presented in Table 4.6 along with the respective accuracies for each distance function used for the clustering procedure.

The accuracy results present in Table 4.6 were obtained using the formula

$$accuracy(\%) = \frac{\text{number of correctly clustered cycles}}{\text{total number of cycles}} \times 100\% \tag{4.2}$$

Thus, the total number of cycles must also equal the number of correctly clustered cycles plus the number of missed ones.

In ECG$_1$, an approximately 7 hour signal, we were able to test our algorithm to perform events detection and clustering on a long record. The highest accuracy was obtained using the L_1 distance, although the L_2 and *meanwave* distances also led to a high algorithm performance. It is also important to analyse the number of missed cycles; in fact, the signal's visual inspection was taken by dividing it into 26 parts and, therefore, we recorded the number of missed cycles for each part. The mean value was 10.5 but the standard deviation was 14.4. This shows that in some parts of the signal, there were significant changes in fundamental frequency, resulting in a high number of missed cycles, while in other parts, the number of misses was extremely low. In order to minimize the number of missed cycles, smaller parts could be analysed allowing a more sensitive perception of the fundamental frequency's temporal evolution. However, sensitivity for noise presence is also augmented, producing poor results when determining cycles sizes.

In the analysed ACC signals, only the *meanwave* distance resulted in a high algorithm performance. These results are possibly related to the higher sensitivity of *meanwave* distance measures. In fact, when a signal is divided into n parts, it will be constructed n *meanwaves* that will be used to calculate distances between them and each cycle present among the n parts. Besides, as it was discussed before, when using the other distance functions, our algorithm builds a distance array as input for the parallel version of the *k*-means algorithm, which has a lower computational cost but also carries poorer information about the cycles resemblance.

It is also interesting to analyse the obtained results using the *meanwave* distance for the BVP_2 and ECG_7 signals. In fact, an unexpected high number of errors were encountered when comparing the results of the remaining signals. This is due to the fact that these signals have almost no variations on their cycles morphology and since the *meanwave* brings highly sensitive distance measures, this sensitivity "excess" returned poor results.

Analysing the results globally, the L_1 and L_2 distances returned a total of 787 and 841 errors out of 31310 cycles, achieving 97.5% and 97.3% of accuracy, respectively. Besides, the *meanwave* distance returned a total of 850 errors out of 32439 cycles, achieving 97.4% of accuracy. The other distance functions performed poorly and should not be considered as good distance functions for clustering biosignals.

4.4 Evaluation using Physionet Library signals

To finalize our algorithm's performance assessment using visual inspection, a set of ECG signals was chosen from different Physionet databases. These signals were manually annotated by physicians. Thus, the ambiguous interpretations provided by visual inspection are overcame since a reliable source previously annotated the ECGs events and also their medically relevant annotations for each beat. However, it is not expected that our algorithm only detects the medically relevant episodes on the ECGs signals since it is also sensitive to morphological changes in each cycle (e.g. due to the presence of noise). All the ECG signals were filtered with a low-pass filter with a cut-off frequency of 50 Hz and a high-pass filter with a cut-off frequency of 2 Hz to remove some possible baseline fluctuations and other respiratory artefacts.

The obtained results with the ECGs from the Physionet databases are shown in Table 4.7. ECG_8 was downloaded from the *Post-Ictal Heart Rate Oscillations in Partial Epilepsy: Data and Analysis* database [84, 87], representing an ECG acquired from a patient with partial epilepsy and sampled with a frequency of 200 Hz. ECG_9, ECG_{10} and ECG_{11} were downloaded from the *St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database* database [84], representing ECGs from patients diagnosed with coronary heart disease (and arterial hypertension), acute myocardial infarction and Wolff-Parkinson-White syndrome, respectively. Both signals were acquired with a sampling frequency of 257 Hz.

For these signals, only the L_1, L_2 and *meanwave* distances were used since it was previously concluded that the remaining distance functions should not be used for biosignals clustering.

All the ECGs were approximately half-hour-signals and in all of them, our algorithm returned good results. In fact, using the L_1 distance function, 10335 out of 10382 cycles were correctly clustered, achieving an efficiency of 99.5%. Using the L_2 distance function, 10341 cycles were correctly clustered, achieving an efficiency of 99.6%. Finally, the *meanwave* distance returned 10350 cycles correctly clustered along with an accuracy of 99.7%, representing the highest efficiency among the most suitable distance functions for clustering in biosignals.

In Chapter 5, a more thorough analysis of the previous results will be made since the ECG signals downloaded from the Physionet databases have also medically relevant annotations. Therefore, the ability of our algorithm for detecting medically relevant episodes in an ECG record will be assessed.

Signal	Cycles	Misses	D. Functions	Correct cycles	Errors
ECG$_1$	24551	272	L_1	24028	251
			L_2^2	23579	700
			L_2	23998	281
			L_∞	23447	832
			χ^2	23565	714
			Mw	24021	258
Resp$_1$	5210	46	L_1	5068	98
			L_2^2	4928	250
			L_2	5056	113
			L_∞	4983	192
			χ^2	4879	295
			Mw	5078	93
ECG$_2$	224	2	L_1	217	5
			L_2^2	198	24
			L_2	215	7
			L_∞	186	36
			χ^2	208	14
			Mw	216	6
BVP$_1$	224	2	L_1	183	39
			L_2	191	31
			$\sqrt{L_2}$	197	25
			L_∞	185	37
			χ^2	188	34
			Mw	183	39
ECG$_3$	165	1	L_1	159	5
			L_2^2	142	22
			L_2	156	8
			L_∞	151	13
			χ^2	134	30
			Mw	160	4
BVP$_2$	165	1	L_1	123	41
			L_2^2	98	66
			L_2	124	42
			L_∞	111	53
			χ^2	94	70
			Mw	114	50
ECG$_4$	199	1	L_1	191	7
			L_2^2	178	20
			L_2	193	5
			L_∞	172	26
			χ^2	182	16
			Mw	193	5

Table 4.4: Clustering results using morphological comparison with $k = 2$ clusters.

Signal	Cycles	Misses	D. Functions	Correct cycles	Errors
ECG$_5$	170	3	L_1	158	9
			L_2^2	127	40
			L_2	155	12
			L_∞	148	19
			χ^2	132	35
			Mw	162	5
ECG$_6$ (Pregnant)	231	1	L_1	228	2
			L_2^2	216	14
			L_2	226	4
			L_∞	221	9
			χ^2	175	55
			Mw	228	2
ECG$_7$ (Pregnant)	104	2	L_1	103	1
			L_2^2	102	2
			L_2	103	1
			L_∞	93	11
			χ^2	77	25
			Mw	82	20
Resp$_2$	67	1	L_1	65	3
			L_2^2	64	4
			L_2	46	19
			L_∞	44	21
			χ^2	60	5
			Mw	52	13
ACC$_1$	185	1	Mw	179	5
ACC$_2$	133	0	Mw	130	3
ACC$_3$ (Act 1)	672	1	Mw	666	5
ACC$_4$ (Act 2)	56	1	Mw	53	2
EMG (Act 2)	56	1	Mw	47	8
ACC$_5$ ($k = 3$, Act 3)	27	1	Mw	24	2

Table 4.5: Clustering results using morphological comparison with k clusters (continuation).

Distance Function	All Cycles	Correctly clustered cycles	Accuracy
L_1	31310	30523	97.5%
L_2^2	31310	29823	95.2%
L_2	31310	30469	97.3%
L_∞	31310	29741	94.9%
χ^2	31310	29654	94.7%
Mw	32439	31588	97.4%

Table 4.6: Accuracy obtained using different distance functions to obtain inputs for the clustering algorithm.

Signal	Cycles	Misses	D. Functions	Correct cycles	Errors
ECG$_8$	2545	7	L_1	2531	5
			L_2	2537	3
			Mw	2532	8
ECG$_9$	2753	2	L_1	2741	10
			L_2	2743	8
			Mw	2750	1
ECG$_{10}$	2450	4	L_1	2441	5
			L_2	2437	9
			Mw	2443	3
ECG$_{11}$	2634	8	L_1	2622	4
			L_2	2624	2
			Mw	2625	1

Table 4.7: Clustering results for the ECG signals downloaded from the Physionet databases.

5

Applications

In this chapter, a set of practical applications of the developed algorithm is presented. Although it aims at extracting information from biosignals in a broad perspective and uses non-supervised learning, objective interpretations can be taken from the algorithm clustering results.

5.1 Home Monitoring

As it was mentioned in Chapter 2, one of the practical applications of the developed algorithm was to assist home monitoring by detecting events on signals and finally, extract information from them in a broad perspective. For that, a set of signals acquired under different projects were used. Next, a brief explanation of each project will be presented, along with the importance of applying our algorithm and using its outputs.

wiCardioResp Project

A schematic representing the basic steps of this project is illustrated in Figure 5.1.

The *wiCardioResp* project aims at monitoring patients with cardio-respiratory disorders while they are comfortably at home or in other places if it is required (e.g. health care facilities). To accomplish these goals, a set of sensors are provided in order to continuously monitor patients through the acquisition of biosignals. Then, these biosignals are transmitted via Bluetooth to a device and, afterwards, the collected data is sent to a remote central that can be conveniently located at a hospital or clinical center. The final step consists in processing the biosignals that reach the remote central. This step will be accomplish by using previously developed signal processing tools [7].

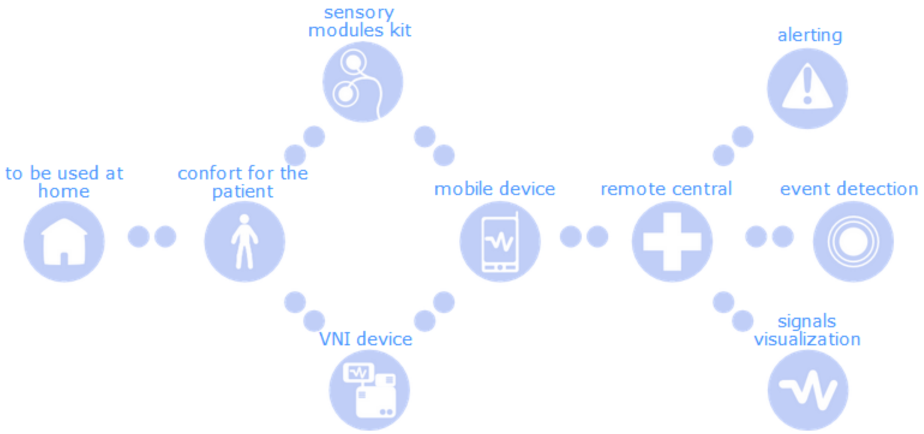


Figure 5.1: *wiCardioResp* Project schematics. From [7].

The signal processing step must return the following outputs:

- **Alerts.** If a signal processing tool is able to detect abnormalities, an alert can be sent to the remote central. This will allow a considerable reduction in the number of hospital visits from the patients that are being monitored. Besides, a more patient-specific diagnostic and treatment can be administrated.

- **Events Detection.** By detecting events on signals, significant transitions in the state of the signal under study can be determined. Besides, if we are dealing with cyclic signals (which represent most of the biosignals), the detection of each cycle might be useful to analyse morphological changes within each one or to evaluate the time-variability.

- **Signal Visualization.** If a visualization tool is developed for real-time biosignals visualization, a closer patient monitoring would be provided.

It is important to notice that these signal processing tools must be able to analyse long-term records since a continuous monitoring usually results in signals with several hours which gives rise to two types of problems: the hardware limitations that hinder the processing of large sized data and once these limitations are bypassed, the time required to process those signals.

Once the main objectives of this project are depicted the applications of the developed algorithm in the signal processing step will become evident.

In fact, the input that reaches the remote central comprises several types of biosignals, such as ECG, EMG, respiratory signals, among others. Since our algorithm is signal-independent, only one tool is necessary to process all these signals. Besides, the developed events detections algorithm returns one of the necessary outputs that were previously enumerated. The implemented clustering procedure allows biosignals annotation and information extraction. Once this information is automatically (using supervised learning, for example) or manually (through the physicians expertise) interpreted, alerts can be sent and a medical response can be taken. To minimize the signal processing time-consuming issue, parallel computing techniques are applied.

Under the light of this project, two long-term biosignals were processed and the results were presented and discussed in Chapter 4.

ICT4Depression Project

Similary to the *wiCardioResp* project, the *ICT4Depression* project aims at continuously monitoring people that are comfortably at home but who were diagnosed with depression. This monitoring consists of acquiring biosignals from patients through the use of sensory systems and to process those signals. The output provided from the developed signal processing tools must be able to automatically assess the patient's state and help to find patient-specific treatments. Besides, a patient interface to the system is developed in order to perform a psychological evaluation, behavioural activation, cognitive reconstruction, relapse prevention, among others [88].

The developed tools for biosignals processing must be suitable for ECG, ACC and respiratory signals. However, it would be more user-friendly if only one tool was able to process different types of signals and, therefore, our algorithm proves to be once more a valuable asset in the signal processing field. In order to test the suitability of our algorithm for the scenario provided by this project, four signals were used and processed, obtaining the results previously discussed in Chapter 4.

Pregnant Women Monitoring

For this research work, two ECGs from pregnant women were processed. The main objective that was behind the signals acquisition was to detect fetal ECGs. These signals were acquired under hospital environment and therefore, do not enter in the Ambient Assisted Living (AAL) field. However, it is expected to monitor pregnant women in their homes using a network similar to the already presented in *wiCardioResp* and *ICT4Depression* projects.

Since ECG and respiratory signals are simultaneously acquired, our algorithm proves once more to be an asset in hospital and home monitoring and biosignal processing fields. In fact, its independence regarding the signals types makes it a powerful tool in detecting events and performing signal annotations. However, our algorithm was not able to distinguish the maternal ECG from the fetal ECG, because of the predominance of the maternal ECG when finding the signal events.

5.2 Different Modes Identification

Since the distance measures that are taken to serve as input for the developed parallel k-means algorithm aim at quantifying the morphological changes among each signals cycles, building a set up in which signals are acquired when subjects are performing distinct tasks might return annotated signals where those tasks are clearly identified.

These facts were exposed in the previous chapter and they represent a major application of our algorithm. Actually, this feature can be very useful in sports. For example, if an athlete is practising a sport in which different techniques are applied, if a signal processing tool is able to automatically identify those techniques, an easier assessment of the athlete performance can be made by manually analysing the acquired signals or by automatically extracting other types of information using a specific tool.

In the Ambient Assisted Living environment (AAL), the study of human behaviour is one the most important goals. In fact, the AAL systems (that includes sensing systems and signal processing tools) must be able to classify a large variety of situations such as falls, physical immobility, among others. Besides, they also should be suitable for automatically monitor activities and vital signs in order to detect emergency situations or deviations from a normal medical pattern. These situations and variations are usually connected with morphological changes within the acquired signals which will possibly be detected by our algorithm due to its sensitivity to those types of changes in signals.

5.3 Medically Relevant Annotations

5.3.1 ECG signals Overview

In the previous chapter, four ECG signals were downloaded from the Physionet databases and a brief description of them was presented. In this section, a more thorough analysis of the obtained clustering results will be made since we have access to medically relevant annotations of episodes that occurred during the signal acquisitions. All the ECG signals were filtered using a low-pass filter with a cut-off frequency of 50 Hz and a high-pass filter with a cut-off frequency of 2 Hz to remove some possible baseline fluctuations and other respiratory artifacts.

First, a more complete description of the ECG signals is presented as follows, along with a brief characterization of the diagnosed heart disease. The same nomenclature that was presented in Chapter 4 will be used.

Diagnosis: Partial Epilepsy

The ECG_8 represents an ECG signal downloaded from the *Post-Ictal Heart Rate Oscillations in Partial Epilepsy: Data and Analysis* database [84, 87]. It has approximately 30 minutes and it was sampled at a frequency of 200 Hz.

The patient was diagnosed with partial epilepsy. Epilepsies comprise a quite diverse collection of disorders that are associated with neuronal excitability. The effects of these disorders — seizures — are usually reduced with symptomatic therapies. In fact, drug administration might minimize seizure frequency and, in the best case scenario, suppress them. However, a cure is only available through surgery in which the resection of epileptic tissue is performed. From a neurophysiological point of view, a seizure can be defined as a transient change of behaviour due to the disordered, synchronous and rhythmic firing of populations of Central Nervous System (CNS) neurons. Hence, epilepsy refers to a disorder of brain function characterized by the periodic and unpredictable occurrence of seizures. These seizures have been classified into partial seizures in which the epileptic foci is located in the cortical site and generalized seizures in which an involvement of the cortex of both hemispheres takes place. The terms "ictal" and "interictal" are commonly used to define spikes in electroencephalography (EEG) signals from patients with epilepsy. The term "ictal" refers to seizure-like spike and the term "interictal" refers to a between seizures-like spike [89, 90, 91].

In this ECG signal, heart rate oscillations (0.01 - 0.1 Hz) followed by a seizure were recorded. Manual annotations of this oscillations were taken by physicians.

Diagnosis: Coronary Artery Disease

The ECG_9 represents an ECG signal downloaded from the *St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database* database [84]. It has approximately 30 minutes and it was sampled at a frequency of 257 Hz.

The patient was diagnosed with Coronary Artery Disease (CAD), also known as Atherosclerotic Heart Disease (AHD). This disease is the leading cause of death and therefore, its timely diagnosis is desired. The disease may result from a disorder of fat metabolism that affects people from all ages. This disorder can lead to pathological changes that affect the inner of arteries. These changes are characterized by focal thickening of the inner of arteries where stable lipids can settle. A severe lipid accumulation can lead to stenosis or even occlusion of coronary arteries, with or without thrombus (or blood clot) formation. The clinical picture of the atherosclerosis (disease associated with the inner of arteries) is that angina pectoris, coronary insufficiency or myocardial infarction. If the damage on the myocardium in enough to reduce the cardiac output, congestive failure might follow [92]. It was also detected Artery Hypertension, which is a chronic medical condition in which blood pressure in arteries is higher than normal. This condition is a risk factor for premature coronary heart disease, ventricular dysfunction, and rupture of aortic or cerebral aneurysms [93].

In the patient's ECG it was detected premature ventricular contractions (PVCs), which represent a common feature in patients with CAD. These episodes were manually annotated by physicians.

Diagnosis: Acute Myocardial Infarction

The ECG_{10} represents an ECG signal downloaded from the *St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database* database [84]. It has approximately 30 minutes and it was sampled at a frequency of 257 Hz.

The patient was diagnosed with Acute Myocardial Infarction (AMI), which is a common disease with serious consequences in mortality, morbidity and cost to the society. The term myocardial infarction is associated with the death of cardiac myocytes due to prolonged ischaemia. Therefore, myocardial infarction is an acute coronary syndrome that can occur during the natural course of coronary atherosclerosis. In advanced stages of the disease, atherosclerotic plaques develop. Initially, normal lumen cross-sectional area will be preserved, since coronary arteries undergo compensatory outward remodelling in relation to plaque area. These plaques might be subjected to erosion, arising an angiographically non-significant stenosis. However, if plaques are ruptured, a total occlusion of the epicardial coronary artery can be experienced, interrupting the coronary blood flow and blocking the delivery of nutrients to the myocardium. Long-term coronary occlusion results in a progressive increase of the infarct size. In myocardial infarction with ST-segment elevation, occlusive and persistent thrombosis prevails [94, 95, 96].

In the patient's ECG it was detected PCVs and ST-segment elevations. These episodes were manually annotated by physicians.

Diagnosis: Wolff-Parkinson-White Syndrome

The ECG_{11} represents an ECG signal downloaded from the *St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database* database [84]. It has approximately 30 minutes and it was sampled at a frequency of 257 Hz.

This patient was diagnosed with the Wolff-Parkinson-White (WPW) Syndrome. This syndrome was first described by Wolff, Parkinson and White [97] through the study of a group of young patients with bundle branch block with short PR intervals associated with paroxysmal (or acute) tachycardia. The physiological principle related to this syndrome is the activation or "pre-excitation" of the ventricles at a site other than the normal atrioventricular conduction system. Although many hypothesis have arisen to explain the symptoms, a surgical cure for the WPW syndrome has proved that outside the atrioventricular node (AV) there are atrioventricular node-like structures that can conduct electrical activation from the atria to the ventricles, allowing the latter to be excited aberrantly. Thus, an anatomical substrate for re-entry tachycardia is provided and the occurrence of paroxysmal tachycardia is observed [98, 99, 100].

In the patient's ECG it was detected PCVs which were manually annotated by physicians.

5.3.2 Clustering Results vs. Medical Annotations

To compare the clustering results returned by our algorithm with the medical annotations from physicians, only the *meanwave* distance will be used, since it demonstrated to return the best results for the Physionet database.

An illustration of the highlighted ECG_9 from a patient diagnosed with Coronary Artery Disease is presented in Figure 5.2. The coloured vertical lines represent the manual annotations from the physicians for the normal beats (black lines) and for the abnormal beats (brown lines). The abnormal beats can have different classifications which will be exposed when the results for each signal are analysed. The red and green circles are the automatic unsupervised classification provided by our algorithm. The red ones represent the normal and morphologically similar beats and the green ones, in this case, represent only the abnormal beats.

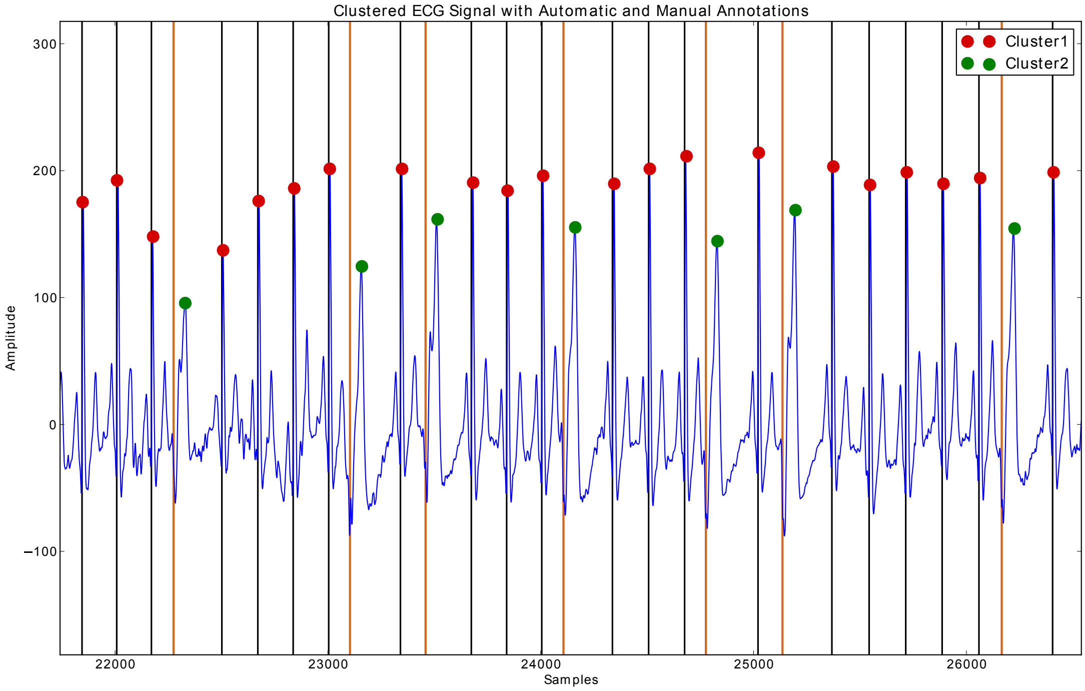


Figure 5.2: Highlight of an ECG signal recorded from a patient diagnosed with CAD. The vertical lines represent the manual annotations from the physicians and the circles represent the automatic annotations from our algorithm.

In order to demonstrate the suitability of our algorithm in discovering medically relevant events, the results obtained for this four ECGs signals are exposed in Table 5.1.

First, the ECG_8 signal had been annotated with four R-on-T premature ventricular contractions and six supraventricular premature or ectopic beats (atrial or nodal). Our

Signal	Diagnosis	Manual Annotations	Manual Annotations Detected
ECG_8	Partial Epilepsy	10	10
ECG_9	CAD	344	343
ECG_{10}	AMI	127	123
ECG_{11}	WPW Syndrome	164	163

Table 5.1: Assessment of our algorithm's automatic annotations compared with the manual annotations provided by physicians.

algorithm detected all the physicians annotations. For the ECG_9 signal, only premature ventricular contractions were detected by physicians and our algorithm annotated 99.7% of them. In the ECG_{10} record, physicians detected two atrial premature beats and the remaining ones were premature ventricular contractions. Our algorithm detected these annotations with an efficiency of 96.9%. Finally, the ECG_{11} record had also annotated two atrial premature contractions, being the remaining annotations premature ventricular contractions. For this signal, our algorithm achieved an efficiency of 99.4%.

As it was stated in the previous chapter, it is not expected that our algorithm only annotates the medically relevant episodes that occurred in the ECGs recordings since it is also sensitive to noise, baseline variations or morphological changes in cycles due to the various types of artefacts. Therefore, the number of clusters associated with abnormal beats, noise or artefacts will be greater than the number of manual annotations. Nevertheless, our algorithm's automatic annotations proved to detect 638 out of 644 manual annotations.

These facts demonstrate that our algorithm can be used to aid physicians in their diagnosis when analysing electrophysiological data. Besides, perform manual annotations in long-term records can be very time consuming and since our algorithm is also suitable to process large sized data, medical resources can be allocated to other tasks where their presence is required.

6

Conclusions

To conclude our work, an overview of the general contributions that this research provided for the signal processing area are exposed in this final chapter. A summary of the obtained results is also presented.

6.1 General Contributions and Results

The main objective of this dissertation was to develop an algorithm that performed clustering on time-series. In fact, giving as input a biosignal, the output should be an annotated biosignal based on an unsupervised classification. Since one of the major applications of the algorithm was to cluster biosignals from patients monitoring, the barriers associated with large sized data should also be overcame.

In order to accomplish these goals, a signal-independent algorithm for long-term signal processing and time series clustering was developed. First, an events detection step was taken where peaks were detected using an adaptive threshold defined as the signal's RMS or based on signals morphological analysis. Both approaches presented the ability to be run in long-term biosignals. Then, clustering techniques were applied using a parallel *k*-means capable of classifying large sized data. Our algorithm did not need any prior information on the signal and had a high performance speed-wise due to the order relationship between two consecutive samples, a key feature in time series that allowed the computation of a distance array instead of a distance matrix. Besides, the employment of parallel computing techniques in designing the parallel *k*-means also contributed to its speed increase.

To assess the algorithm's performance, three types of evaluation were taken. First, an evaluation based on synthetic signals was performed, achieving an efficiency of 99.8%. In contrast with the remaining evaluation procedures, signals with different number of modes were generated in order to assess the robustness of our algorithm with more than two clusters.

After testing with synthetic data, a set of different types of signals was acquired to test the algorithm with real data. First, the time-sample difference information was used to cluster biosignals. Despite its simplicity, interesting results were obtained with an efficiency of 98.9%. For ECG signals, for example, a HRV analysis should be taken. Finally, a morphological analysis was performed using a set of different distance functions. The L_1 and L_2 Minkowski distances returned an output that allowed to cluster signals cycles with an efficiency of 97.5% and 97.3%, respectively. Using the *meanwave* distance, our algorithm achieved an accuracy of 97.4%.

To conclude the algorithm's performance evaluation, four ECGs were downloaded from the Physionet databases. Since manual annotations were provided, a more thorough assessment could be made. For these signals, only L_1 and L_2 Minkowski distances and *meanwave* distance were used. An accuracy of 99.5%, 99.6% and 99.7% was obtained, respectively. Besides, the developed algorithm was able to detect the majority (638 out of 644) of the physicians annotations.

This research work also led to the publication of one chapter for a book and one paper. The chapter of the book is related to human behaviour recognition where clustering techniques play an important role. The algorithm's introduction and performance evaluation was exposed in the published paper. Both publications are presented in the Appendix A.

Summarizing, in this dissertation we presented a signal-independent algorithm with two main goals: perform events detection in biosignals and, with those events, extract information using a set of distance measures which will be used as input to a parallel *k*-means algorithm. The fact that a pre-processing phase is not mandatory, no prior information on the signals is needed and the suitability in clustering long-term biosignals makes the developed algorithm an important asset in the signal processing field. Besides, it also proved to have the potential to aid physicians in their analysis of electrophysiological data due to its ability in detecting and annotating medically relevant events in signals. Thus, the required time-consuming visual inspection is minimized, which is an important achievement since the time spent in analysing data manually is a very important issue.

6.2 Future Work

This research work covered several topics in the signal processing field and in some of them we intend to perform further research and development in the future.

- **Further validation:** Although a relatively large set of biosignals was used to assess the algorithm's performance, different experiment scenarios should be created to test the algorithm under other circumstances. Also, a more exhaustive validation of the signals downloaded from the Physionet databases should be taken in order to ascertain the potential of the developed algorithm as a diagnostic contribution.

- **Noise immunity tests:** One of the future goals is to run the developed algorithm over signals with different signal-to-noise (SNR) ratios. This is an important research topic since pre-processing algorithms are not mandatory for our algorithm's procedures.

- f_0 **estimation:** Since the computation of the fundamental frequency is still a current and active research topic, a constant monitoring of this area must be made, as well as a search for new approaches to overcome the drawbacks of each technique to estimate the fundamental frequency.

- **Optimal length of each part:** The presented algorithm receives as input the length of each part that will comprise the original signal. An automatic procedure to find the optimal length is a future goal, allowing a better monitoring of the temporal evolution of the fundamental frequency. This would lead to a significant reduction in the number of missed cycles.

- **Parallel techniques in the events detection algorithm:** Being events detection the most time consuming step of our algorithm, we aim to improve this point by using parallel computing techniques.

- **New distance functions:** In the present research work, a set of distance functions was used to analyse which one returned better inputs for the parallel k-means algorithm. The test of more distance functions is also a future goal, since better distance functions for clustering time-series might be found.

Bibliography

[1] J.D. Bronzino. *The biomedical engineering handbook*, volume 2. CRC Pr I Llc, 2000.

[2] J.D. Enderle and J.D. Bronzino. *Introduction to biomedical engineering*. Academic Pr, 2011.

[3] J.L. Rodríguez Sotelo et al. *Biosignal analysis for cardiac arrhythmia detection using non-supervised techniques*. PhD thesis, Universidad Nacional de Colombia-Sede Manizales, 2010.

[4] C.J. De Luca, A. Adam, R. Wotiz, L.D. Gilmore, and S.H. Nawab. Decomposition of surface EMG signals. *Journal of neurophysiology*, 96(3):1646–1657, 2006.

[5] R. Xu and D. Wunsch. *Clustering*. Wiley-IEEE Press, 2009.

[6] R. Xu, D. Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[7] wiCardioResp Project. http://wicardioresp.plux.info/pt-pt. [Accessed on August 18, 2012].

[8] G.N. Rodrigues, V. Alves, R. Silveira, and L.A. Laranjeira. Dependability analysis in the Ambient Assisted Living Domain: An exploratory case study. *Journal of Systems and Software*, 2011.

[9] LaTeX. http://www.latex-project.org/. [Accessed on August, 2012].

[10] Scipy. http://www.scipy.org/. [Accessed on August, 2012].

[11] H.H. Chang and J.M.F. Moura. *Biomedical signal processing*. McGraw Hill, 2009.

[12] EJ Ciaccio, SM Dunn, and M. Akay. Biosignal pattern recognition and interpretation systems. *Engineering in Medicine and Biology Magazine, IEEE*, 12(3):89–95, 1993.

[13] Charles S. Lessard. Signal Processing of Random Physiological Signals. *Synthesis Lectures on Biomedical Engineering*, 1(1):1–232, Jan 2006.

[14] M. Akay. *Wiley encyclopedia of biomedical engineering*. Wiley-Interscience, 2006.

[15] R. Silva H. Duarte-Ramos, F. Coito and M.D. Ortigueira. *Análise de Sinais em Engenharia Biomédica*. FCT-UNL, 2006.

[16] J.L. Semmlow. *Biosignal and biomedical image processing: MATLAB-based applications*, volume 1. CRC, 2004.

[17] I.G. Maglogiannis, K. Karpouzis, and M. Wallace. *Image and signal processing for networked e-health applications*, volume 2. Morgan & Claypool, 2006.

[18] Arthur C. Guyton and John E. Hall. *Textbook of Medical Physiology*. Elsevier, 2006.

[19] R. Merletti and P. Parker. *Electromyography: Physiology, Engineering, and Noninvasive Applications*. IEEE Press Series in Biomedical Engineering, 2004.

[20] M.B.I. Reaz, MS Hussain, and F. Mohd-Yasin. Techniques of EMG signal analysis: detection, processing, classification and applications. *Biological procedures online*, 8(1):11–35, 2006.

[21] J.J. Kavanagh and H.B. Menz. Accelerometry: A technique for quantifying movement patterns during walking. *Gait & posture*, 28(1):1–15, 2008.

[22] J.B.J. Bussmann, W.L.J. Martens, J.H.M. Tulen, FC Schasfoort, H.J.G. van den Berg-Emons, and HJ Stam. Measuring daily behavior using ambulatory accelerometry: the Activity Monitor. *Behavior Research Methods*, 33(3):349–356, 2001.

[23] M.J. Mathie, A.C.F. Coster, N.H. Lovell, and B.G. Celler. Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. *Physiological measurement*, 25:R1, 2004.

[24] W.B. Murray and P.A. Foster. The peripheral pulse wave: information overlooked. *Journal of Clinical Monitoring and Computing*, 12(5):365–377, 1996.

[25] A.B. Hertzman. The blood supply of various skin areas as estimated by the photoelectric plethysmograph. *American Journal of Physiology–Legacy Content*, 124(2):328–340, 1938.

[26] R. Martins and J. Medeiros. Development of a Blood Volume Pulse Sensor to measure Heart Rate Variability. In *Procedings of IBERSENSOR 2010*. Lisbon, Portugal, November 2010.

[27] MH Sherebrin and RZ Sherebrin. Frequency analysis of the peripheral pulse wave detected in the finger with a photoplethysmograph. *Biomedical Engineering, IEEE Transactions on*, 37(3):313–317, 1990.

[28] PLUX - Wireless Biosignals, S.A. `http://www.plux.info/`. [Accessed on August, 2012].

[29] N.J. Nilsson. Introduction to machine learning. 1996.

[30] D. Stoutamire. *Machine learning, game play, and Go*. 1991.

[31] Google Portugal. `https://www.google.pt/`. [Accessed on May, 2012].

[32] Amazon. `https://www.amazon.com/`. [Accessed on May, 2012].

[33] G. Paliouras. *Machine Learning and Its Applications*. 2001.

[34] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine learning, neural and statistical classification*. Citeseer, 1994.

[35] M. Alder. *An Introduction to Pattern Recognition*. 1997.

[36] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[37] P. Rashidi and D.J. Cook. Mining and monitoring patterns of daily routines for assisted living in real world settings. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 336–345. ACM, 2010.

[38] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[39] X. Zhang, J. Liu, Y. Du, and T. Lv. A novel clustering method on time series data. *Expert Systems with Applications*, 2011.

[40] X. Huang, X. Zheng, W. Yuan, F. Wang, and S. Zhu. Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. *Information Sciences*, 2011.

[41] S.A. Elavarasi, J. Akilandeswari, and B. Sathiyabhama. A survey on partition clustering algorithms. *learning*, 1(1), 2011.

[42] L. Kaufman, P.J. Rousseeuw, et al. *Finding groups in data: an introduction to cluster analysis*, volume 39. Wiley Online Library, 1990.

[43] S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, volume 27, pages 73–84. ACM, 1998.

[44] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM, 1996.

[45] G. Karypis, E.H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[46] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, volume 1996, pages 226–231. AAAI Press, 1996.

[47] M. Ankerst, M.M. Breunig, H.P. Kriegel, and J. Sander. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.

[48] W. Wang, J. Yang, and R. Muntz. STING: A statistical information grid approach to spatial data mining. In *Proceedings of the International Conference on Very Large Data Bases*, pages 186–195. INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS (IEEE), 1997.

[49] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[50] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[51] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical programming*, 79(1):191–215, 1997.

[52] A. Aztiria, J.C. Augusto, and A. Izaguirre. Autonomous learning of user's preferences improved through user feedback. *BMI*, 396:72–86, 2008.

[53] J.C. Augusto. Ambient intelligence: the confluence of ubiquitous/pervasive computing and artificial intelligence. *Intelligent Computing Everywhere*, pages 213–234, 2007.

[54] H. Sun, V. De Florio, N. Gui, and C. Blondia. Promises and challenges of ambient assisted living systems. In *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, pages 1201–1207. Ieee, 2009.

[55] R. Goshorn, D. Goshorn, and M. Kölsch. The Enhancement of Low-Level Classifications for Ambient Assisted Living. In *Proc. 2nd Workshop on Behavior Monitoring and Interpretation, BMI'08*, pages 87–101, 2008.

[56] AAL4ALL Project. `http://www.aal4all.org`. [Accessed on March, 2012].

[57] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo. Clustering ECG complexes using Hermite functions and self-organizing maps. *Biomedical Engineering, IEEE Transactions on*, 47(7):838–848, 2000.

[58] D. Cuesta-Frau, J.C. Pérez-Cortés, G. Andreu-García, and D. Novák. Feature extraction methods applied to the clustering of electrocardiographic signals. a comparative study. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 961–964. IEEE, 2002.

[59] R. Ceylan, Y. Özbay, and B. Karlik. A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network. *Expert Systems with Applications*, 36(3):6721–6726, 2009.

[60] S. Chao, F. Wong, H.L. Lam, and M.I. Vai. Blind biosignal classification framework based on DTW algorithm. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 4, pages 1684–1689. IEEE, 2011.

[61] F.H.Y. Chan, Y.S. Yang, FK Lam, Y.T. Zhang, and P.A. Parker. Fuzzy EMG classification for prosthesis control. *Rehabilitation Engineering, IEEE Transactions on*, 8(3):305–311, 2000.

[62] A.B. Ajiboye and R.F. Weir. A heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 13(3):280–291, 2005.

[63] Y. Zhang, Z. Xiong, J. Mao, and L. Ou. The study of parallel k-means algorithm. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, volume 2, pages 5868–5871. IEEE, 2006.

[64] H.R. Tsai, S.J. Horng, S.S. Tsai, S.S. Lee, T.W. Kao, and C.H. Chen. Parallel clustering algorithms on a reconfigurable array of processors with wider bus networks. In *Parallel and Distributed Systems, 1997. Proceedings., 1997 International Conference on*, pages 630–637. IEEE, 1997.

[65] S. Kantabutra and A.L. Couch. Parallel K-means clustering algorithm on NOWs. *NECTEC Technical journal*, 1(6):243–247, 2000.

[66] C.H. Wu, S.J. Horng, Y.W. Chen, and W.Y. Lee. Designing scalable and efficient parallel clustering algorithms on arrays with reconfigurable optical buses. *Image and Vision Computing*, 18(13):1033–1043, 2000.

[67] I. Dhillon and D. Modha. A data-clustering algorithm on distributed memory multiprocessors. *Large-Scale Parallel Data Mining*, pages 802–802, 2000.

[68] AN Belbachir, M. Drobics, and W. Marschitz. Ambient Assisted Living for ageing well–an overview. *e & i Elektrotechnik und Informationstechnik*, 127(7):200–205, 2010.

[69] Aware Home Project. `http://www.cc.gatech.edu/fce/ahri/`. [Accessed on April, 2012].

[70] I-Living Project. `http://lion.cs.uiuc.edu/assistedliving`. [Accessed on April, 2012].

[71] Amigo – Ambient Intelligence for the Networked Home Environment. `http://www.hitechprojects.com/euprojects/amigo`. [Accessed on April, 2012].

[72] COPLINTHO, IBBT. https://projects.ibbt.be/coplintho/. [Accessed on April, 2012].

[73] A. Steinhage and C. Lauterbach. Monitoring movement behavior by means of a large area proximity sensor array in the floor. In *2nd Workshop on Behaviour Monitoring and Interpretation (BMI'08)*, volume 396, pages 15–27, 2008.

[74] A. Hein and T. Kirste. Towards recognizing abstract activities: An unsupervised approach. In *BMI'08: Proc. of the 2nd Workshop on Behaviour Monitoring and Interpretation*, pages 102–114. Citeseer, 2008.

[75] P. Rashidi, D. Cook, L. Holder, and M. Schmitter-Edgecombe. Discovering activities to recognize and track in a smart environment. *Knowledge and Data Engineering, IEEE Transactions on*, (99):1–1, 2011.

[76] M. Luštrek, H. Gjoreski, S. Kozina, B. Cvetkoviü, V. Mirchevska, and M. Gams. Detecting Falls with Location Sensors and Accelerometers. In *Twenty-Third IAAI Conference*, 2011.

[77] Neuza Nunes, 2011. Algorithms for Time Series Clustering Applied to Biomedical Signals.

[78] N. Nunes, T. Araújo, and H. Gamboa. Two-modes cyclic biosignal clustering based on time series analysis. 2011.

[79] EJ Ciaccio, SM Dunn, and M. Akay. Biosignal pattern recognition and interpretation systems. *Engineering in Medicine and Biology Magazine, IEEE*, 12(3):89–95, 1993.

[80] University of Regina. Dept. of Computer Science and D. Gerhard. *Pitch extraction and fundamental frequency: History and current techniques*. Dept. of Computer Science, University of Regina, 2003.

[81] The HDF Group. http://www.hdfgroup.org/why_hdf/. [Accessed on August, 2012].

[82] O. Pele and M. Werman. The quadratic-chi histogram distance family. *Computer Vision–ECCV 2010*, pages 749–762, 2010.

[83] PhysioBank - Physiologic signal archives for biomedical research. http://physionet.ph.biu.ac.il/physiobank/. [Accessed on August 20, 2012].

[84] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[85] L.P. Rowland and N.A. Shneider. Amyotrophic lateral sclerosis. *New England Journal of Medicine*, 344(22):1688–1700, 2001.

[86] JD Mitchell and GD Borasio. Amyotrophic lateral sclerosis. *The lancet*, 369(9578):2031–2041, 2007.

[87] IC Al-Aweel, KB Krishnamurthy, JM Hausdorff, JE Mietus, JR Ives, AS Blum, DL Schomer, and AL Goldberger. Postictal heart rate oscillations in partial epilepsy. *Neurology*, 53(7):1590–1590, 1999.

[88] ICT4Depression Project. `http://www.ict4depression.eu/`. [Accessed on August 23, 2012].

[89] W. Penfield and T.C. Erickson. Epilepsy and cerebral localization. 1941.

[90] W. Penfield and H. Jasper. Epilepsy and the functional anatomy of the human brain. 1954.

[91] J.O. McNamara. Cellular and molecular basis of epilepsy. *The Journal of Neuroscience*, 14(6):3413–3425, 1994.

[92] B.A. Sachs. Arteriosclerotic Heart Disease. *The American Journal of Nursing*, 55(7):838–841, 1955.

[93] M.A. Gatzoulis, G.D. Webb, and P.E.F. Daubeney. *Diagnosis and management of adult congenital heart disease*. Churchill Livingstone, 2003.

[94] H.D. White and D.P. Chew. Acute myocardial infarction. *The Lancet*, 372(9638):570–584, 2008.

[95] E. Boersma, N. Mercado, D. Poldermans, M. Gardien, J. Vos, and M.L. Simoons. Acute myocardial infarction. *The Lancet*, 361(9360):847–858, 2003.

[96] F. Van de Werf, D. Ardissino, A. Betriu, D.V. Cokkinos, E. Falk, K.A.A. Fox, D. Julian, M. Lengyel, F.J. Neumann, W. Ruzyllo, et al. Management of acute myocardial infarction in patients presenting with ST-segment elevation. *European heart journal*, 24(1):28–66, 2003.

[97] L. Wolff, J. Parkinson, and P.D. White. Bundle-branch block with short PR interval in healthy young people prone to paroxysmal tachycardia. *American Heart Journal*, 5(6):685–704, 1930.

[98] L. Jacobson, K. Turnquist, and S. Masley. Wolff-Parkinson-White syndrome. *Anaesthesia*, 40(7):657–660, 1985.

[99] JJ Gallagher, M. Gilbert, RH Svenson, WC Sealy, J. Kasell, and AG Wallace. Wolff-Parkinson-White syndrome. The problem, evaluation, and surgical correction. *Circulation*, 51(5):767–785, 1975.

[100] A.D. Sharma and P.G. O'Neill. Wolff-Parkinson-White syndrome. *Current treatment options in cardiovascular medicine*, 1(2):117–125, 1999.

A

Publications

In this appendix it is presented the two publications that arose during the present research work. The first publication is entitled *'Clustering Algorithm for Human Behavior Recognition based on Biosignals Analysis'* and represents a chapter in the book *'Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security'*. Although the book is yet to be published, the chapter has already been accepted. The second and last publication is entitled *'A signal-independent algorithm for information extraction and signal annotation of long-term records'* and will be presented in BIOSIGNALS 2013 which is a co-located conference of the *'6th International Joint Conference on Biomedical Engineering Systems and Technologies'* (BIOSTEC 2013), held in Barcelona in February 2013.

A.1 Chapter for Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security

Clustering Algorithm for Human Behavior Recognition based on Biosignals Analysis.

Clustering Algorithm for Human Behavior Recognition based on Biosignals Analysis

NeuzaNunes[1], Diliana Santos [2], Rodolfo Abreu [2], Hugo Gamboa [2], Ana Fred [3,4]

[1] PLUX - Wireless Biosignals S.A., 1050-059 Lisboa, Portugal
[2] CEFITEC, Physics Department, FCT-UNL, 2829-516 Caparica, Portugal
[3] Electrical and Computer Engineering Department, IST – UTL, 1049 - 001 Lisboa, Portugal
[4] Pattern and Image Analysis, Instituto de Telecomunicações, 1049 – 001, Lisboa, Portugal

ABSTRACT

Time series unsupervised clustering has shown to be accurate in various domains and there's an increased interest in time series clustering algorithms for human behavior recognition.

We've developed an algorithm for biosignals clustering, which captures the general morphology of signal's cycles in one mean wave.

In this chapter we validate and consolidate the algorithm and compare our mean wave algorithm with a state-of-the-art algorithm with uses distances between data's cepstral coefficients to cluster the same biosignals. We were able to successfully replicate the cepstral coefficients algorithm and the comparison showed that our mean wave approach is more accurate for the type of signals analyzed, having a 19% higher accuracy value. We also tested the mean wave algorithm with biosignals with three different activities in it and achieved an accuracy of 96.9%. Finally, we performed a noise immunity test with a synthetic signal and noticed that the algorithm remained stable for signal-to-noise ratios higher than 2, only decreasing its accuracy with noise of amplitude equal to the signal.

The necessary validation tests that we performed in this study confirmed the high accuracy level of the developed clustering algorithm for biosignals which express human behavior.

INTRODUCTION

The constant chase for human well-being has led researchers to increasingly design new systems and applications for a continuous monitoring of patients through their biological signals. In the past, human activity tracking techniques focused mostly on observations of people and their behavior through a great amount of cameras. However, the use of wearable sensors has been increasingly sought because it allows continuous acquisitions in different locations, being independent from the infrastructures. The recognition of human behavior through wearable sensors has a vast applicability. In the sports field, for example, there is a need for wearable sensors to assess physiological signals and body kinematics during free exercise. Wearable sensors have also major utility in healthcare, particularly for monitoring elderly and chronically ill patients in their homes, through Ambient Assisted Living (AAL).

Human body has always been considered a complex machine in which all parts work harmoniously. Nevertheless, the endless pursuit for the optimal human performance has become an important work area of digital signal processing. Therefore, monitoring athletes is the logical way to achieve the best patterns that can be compared to pathological signals in order to contribute for patient's rehabilitation. Thereby, the continuous monitoring and evaluation of athletic performance allow the coaches to establish an optimal training program. In addition, it is useful for non-professional athletes to establish and achieve their personal goals (R. Santos et al., 2012).

The main goal of AAL is to develop technologies which enable users to live independently for a longer period of time, increasing their autonomy and confidence in accomplishing some daily tasks (known as ADL, Activities of Daily Livings). However, AAL was also designed to reduce the escalating costs associated with health-care services in elderly people.

Thus, AAL's systems are used to classify a large variety of situations such as falls, physical immobility, study of human behavior and others. These systems are developed using a Ubiquitous Computing approach (AAL4ALL Project, 2012) (where sensors and signals' processing are executed without interfering on ADL) and must monitor activities and vital signs in order to detect emergency situations or

deviations from a normal medical pattern (G.N. Rodrigues et al., 2010). Ultimately, AAL solutions automate this monitoring by software capable of detecting those deviations.
Signal-processing techniques have been developed to extract relevant information from biosignals which aren't easily detected in the raw data. However, most of these techniques are integrated in tools for specific biosignals, such as electrocardiography, respiration, accelerometry, among others. Thus, a single tool to recognize the morphology of the signal without prior information, analyzing and processing it accordingly is a recurrent necessity.
The smallest change in the signal's morphology over time may contain information of the utmost importance; hence, the detection of those changes has received much attention in this field. The recognition of different patterns in the signal's morphology is usually based on clustering or classification approaches. The ultimate goal is a generic and automatic classification system that doesn't require prior information and produces an efficient analysis whichever the type of the signal used.

In the following sections we summarize the scope and results of the developed algorithm and further evaluate it by testing it in a variety of contexts and validating it with a state-of-the art approach for time-series clustering.

RELATED WORK

N. Nunes (2012) presented an advanced signal processing algorithm for pattern recognition and clustering purposes applied to time varying signals collected from the human body. The recognition of differences in the signal's morphology produced by physiological abnormalities (arrhythmia, for example) or different conditions of the subject's state (walking or running, for example) was tested by collecting a set of cyclic biosignals with two distinctive modes. The acquired signals were the input of the generic algorithm. This algorithm knows beforehand the number of modes the signal has and comprises the computation of a mean wave, which is an averaging of all signal cycles aligned in a notable point. The algorithm automatically separates each signal's cycle using a k-means approach and traces the mean wave for any biosignal, capturing its morphology.

The algorithm then has a k-means clustering phase which uses the information gathered from the mean wave approach to separate the several modes of the original signal. As the implemented mean wave approach accurately identifies the morphology of a signal, it can be a powerful tool in several areas – as a clustering basis or for signal analysis. The algorithm produced is signal-independent with high level of abstraction, and therefore can be applied to any cyclic signal with no major changes in the fundamental frequency. This type of generic signal interpretation overcomes the problems of exhaustive, lengthy signal analysis and expert intervention highly used in the biosignal's classification field.

Several approaches for time series comparison have been proposed in literature. The most straightforward approach relies on similarity measures which directly compare observations or features extracted from raw data. Besides the measurements made directly between time series, distances can also be computed with models built from the raw data. By modeling the raw data with a stochastic model, similarities are detected in the dynamics of different time series.

Linear Predictive Coding (LPC) is one of the methods of model compression and is widely used in speech analysis. Linear prediction filters attempt to predict future values of the input signal based on past signals. The process of clustering time series models is usually a three-step procedure. Firstly, each time series is represented by a dynamical model, which is estimated using the given data. Secondly, a distance between the dynamical models is computed over all the models estimated in the first stage - this distance measure can be the same used to cluster data or features extracted from the data. And finally, a clustering or a classification mechanism is performed based on the distance metric defined (J. Boets et al., 2005). This general methodology has been applied previously in different application areas, by estimating similarity measures between the LPC coefficients (G. Antoniol, 2005; P. Souza, 1997). However, other method which estimates the cepstral coefficients from the LPC model and computes the distance between those coefficients has been widely used achieving state of the art results in this field (K. Kalpakis, 2001; M. Corduas, 2008; A. Savvides, 2008).

In this chapter we intend to further validate our algorithm, comparing its accuracy with the state-of-the-art cepstral coefficients algorithm, test the results with more than two different modes and perform a noise immunity test.

CEPSTRAL COEFFICIENTS ALGORITHM

Some of the publications that use the cepstral coefficients algorithm as a clustering mechanism used the Euclidean Distance between the LPC Cepstrum of two time series as their dissimilarity measure. The time series used in those publications were retrieved from a public database of ECG signals (Physionet, 2012).

The cepstral coefficients algorithm was replicated in our research and applied to the same public database, to achieve the same results as the ones documented. Our implementation was compared with the results exposed by Anthony Bagnall (2004), which uses an Euclidian distance between the cepstral coefficients to cluster the signals with a k-means clustering procedure.

We tested our implementation on a public ECG dataset which will be described below. The implementation and testing of the algorithm with the public ECG dataset will also be detailed next.

<u>ECG Dataset</u>

The public dataset of ECG signals used is divided into three groups.

Group 1: 22 recordings of people with malignant ventricular arrhythmia;
Group 2: 13 recordings of healthy people;
Group 3: 35 recordings of people with supraventricular arrhythmia.

Each recording comprises 2 seconds of acquisition.

Figure 1. Examples of a signal from each group from the public ECG dataset: a) malignant ventricular arrhythmia; b) normal ECG; c) supraventricular arrhythmia.

Figure 1 shows one example of a time series from each group. Two collections were defined in these researches: Collection 1 comprises the first two groups (35 signals), and Collection 2 gathers group 2 and 3 (48 signals). The cepstral coefficients algorithm received as input both collections to find two different clusters in each – representing the signals belonging to two different groups in each collection.

<u>Implementation</u>

Figure 2. Schematics for the cepstral coefficients algorithm implementation.

The first step of this algorithm is to fit a LPC model to the raw data, with a defined order. Among the direct transformations of LPC parameters, one is a filtering process to get the cepstral coefficients. We performed these steps using Python with the numpy and scikitstalkbox packages.
Using the LPC coefficients estimation we computed the five cepstral coefficients (order - 1) of each time series. After that, the Euclidean distance between the signals' coefficients was estimated. Using equation 1 for all signals, a distances matrix is computed.

$$distance = \sqrt{\sum_{i=1}^{order\ -1} (sig_1 cc_i - sig_2 cc_i)^2} \qquad (1)$$

Being $sig_1cc_i$and $sig_2cc_i$the $i$cepstral coefficient from the first and the second signal, respectively. Finally, by retrieving the distance values and introducing that matrix into a K-Means algorithm, the time series are separated into different clusters.

With this implementation, the cepstral coefficients algorithm was successfully replicated, achieving the same results as described by Anthony Bagnall with the ECG dataset.

RESULTS

Comparison with cepstral coefficients algorithm

For the actual comparison between the performance of our mean wave algorithm and the cepstral coefficients algorithm, the dataset used was the one described in our previous publication – for which the activities performed and the final clustering results are exposed in Table 1.

Table 1. Clustering results of the mean wave algorithm.From N. Nunes et al. (2012).

Task	*Number of Cycles*	*Cycles correctly clustered*	*Errors*	*Misses*
Synthetic	50	49	0	1
Walk and run	343	342	1	0
Run and jump	296	295	1	0
Jumps	85	84	1	0
Skiing	42	41	0	1
Elevation and squat	23	23	0	0
BVP rest and after exercise	165	159	4	2
All	1004	992	7	5

In the context tasks and signal types acquired, the accuracy was 99.3% for the separation of two different modes.

Table 2 gathers the clustering accuracy results obtained for each task and algorithms used (our mean wave algorithm and the cepstral coefficients algorithm). The accuracy percentage was computed using the equation 2.

$$accuracy = \frac{Cycles_{cc}}{Cycles_{N}} \times 100\% \qquad (2)$$

Being $Cycles_{CC}$ the number of correctly clustered cycles and $Cycles_N$ the total number of cycles.

Table 2. Comparison of the results obtained with the cepstral coefficients and the mean wave algorithm.

Task	*Accuracy of cepstral coefficient algorithm*	*Accuracy of mean wave algorithm*
Synthetic	100.0%	100.0%
Walk and run	92.4%	99.7%
Run and jump	68.2%	99.7%
Jumps	82.1%	98.8%
Skiing	90.2%	100.0%
Elevation and squat	56.5%	100.0%
BVP rest and after exercise	68.7%	96.4%
All	80.0%	99.3%

Our mean wave procedure presents a higher accuracy level for every signal but the synthetic waves, for which the accuracy is the same. Looking at the overall results, our algorithm achieved 99.3% of accuracy, and the cepstral coefficients algorithm only 80.0% for the same signals - which from the tests with this database makes our approach a better option for clustering cyclic signals. To note also that our algorithm can be applied to a continuous signal with different modes in it, automatically separating the signal's cycles and computing a distance metric for each. The cepstral coefficients algorithm, however, has to be applied in separated signals - in this study we had to isolate the cycles before applying the cepstral coefficients procedure.

In conclusion, the comparison between the two algorithms confirmed that the mean wave algorithm has a high accuracy level, reaching better results and is more suitable for the type of data analyzed than a state of the art algorithm in this area.

Validation

In this study we've also collected a new set of signals, with three different activities in each, composing over 2000 cycles.
To acquire the biosignals necessary for this study, we used a surface EMG sensor (*emgPLUX*) and a triaxial accelerometer (*xyzPLUX*). A wireless signal acquisition system, bioPLUX research (PLUX, 2012), was used for the signal's analogue to digital conversion and bluetooth transmission to the computer. This system has 12 bit ADC and a sampling frequency of 1000 Hz. In the acquisitions with the triaxial accelerometers, only the axis with inferior-superior direction was connected to the bioPLUX.

Several tasks were designed and executed in order to acquire signals with three distinct modes from 4 different subjects.

Before describing the activities executed in order to acquire the signals used for testing our algorithm, it is worthy to note that in all activities we used only the accelerometer's signal, except for activity 2 (walking, jumping, crouching), in which we tested also the EMG's signal.

Synthetic Signal:

To test our algorithm, a synthetic cycle was created using a low-pass filtered random walk (of 100 samples), with a moving average smoothing window of 10% of signal's length, and multiplying it by a *hanning*window. That cycle was repeated 296 times for the first mode, so all the cycles were identical. After a small break on the signal, the cycle was repeated 104 more times, but with an identical small change of 20 samples in all waves, creating a second mode. A third mode was created by changing the same 20 samples and repeating the new wave created 524 times. These three modes construct the synthetic wave represented in Figure 3.

Figure 3. Synthetic signal with three different modes (within each mode all waves are identical).

Activity 1 – Walking, Running, Walking, Jumping:

In this task, the accelerometer was located on the right hip along with the bioPLUX, so that the y axis's accelerometer was pointing downward. It was asked to the subjects to walk (for about 1 minute and half), run, walk again and jumpon the same place (each for about 1 minute). These four modes were executed non-stop. The signal acquired is demonstrated in Figure 4.

Figure 4. Activity 1: Walking, Running, Walking and Jumping.

Activity 2 – Walking, Jumping, Crouching:

With the accelerometer located on the right hip and oriented so that the y axis was pointing downward, the subjects performed a task of walking, vertical jumping and crouching continuously. Each mode was executed 10 times and it is worthy referring that in walking mode, each step was considered a cycle. The signal acquired is demonstrated in Figure5.

Figure 5. Activity 2 (ACC): Walking, Jumping and Crouching.

For the EMG's signal, electrodes were located on the ischiotibial of the right leg so that they were able to collect the muscle's activation signal during the activity. This signal was collected simultaneously with the accelerometer's signal. The signal acquired is demonstrated in Figure 6.

Figure 6. Activity 2 (EMG): Walking, Jumping and Crouching.

Activity 3 - Jumping, leg flexion and single leg vertical jumping

In this task, the following procedure was executed: normal vertical jumping, leg flexion and single leg vertical jumping. Each mode was repeated 10 times.
The subjects used an accelerometer located at the right hip and oriented so the y axis of the accelerometer was pointing downward. The signal acquired is represented in Figure 7.

Figure 7. Activity 3: Jumping, Leg Flexion and One Foot Jumping.

Activity 4 – Crouching, leg flexion and leg elevation

In this task, the subject was standing straight with both feet completely on the ground and was asked to perform 10 squats followed by 10 vertical leg flexions – moving the heel towards the gluteus- and 10 leg elevations – moving the knee towards the chest (Figure 8).
The subjects used an accelerometer located at the right hip and oriented so the y axis was pointing downward.

Figure 8. Activity 4: Crouching, Leg flexion and Leg Elevation.

In the new set of acquired signals we achieved the results reported on Table 3.

Table 3. Clustering results of the mean wave algorithm for signals with three modes.

Task	*Number of Cycles*	*Cycles correctly clustered*	*Errors*	*Misses*	*Accuracy (%)*
Synthetic	924	921	0	3	99,68
Act 1	693	692	0	1	99,86
Act 2 (Accelerometer)	175	149	7	20	85,14
Act 2 (Electromyography)	120	112	8	0	93,33
Act 3	180	154	7	19	85,56
Act 4	180	173	3	4	96,11
Total	2272	2201	25	47	96,88

We considered errors as misclassifications and misses as cycles which weren't classified at all.The misses encountered were mostly present in the borders of the signals. As can be seen from the last column, the algorithm implemented also shows a high accuracy when applied on activities with more than two modes.

That way we were able to check the performance of our algorithm when more than two clusters are involved. In the new set of acquired signals we achieved 96.88% of accuracy separating the three activities into different clusters, lowering our previous clustering result with two modes by only 2.4%.

Noise Immunity Test

With the intention of performing a noise immunity test in our algorithm we added Gaussian noise of mean zero and deviation error variable to the synthetic signal that we described previously. We compared the accuracy of the algorithm with the signal-to-noise ratio (SNR) for each situation. The results for this test are detailed in Table 4.

Table 4. Noise Immunity Test. Number of cycles = 924.

Signal-to-Noise Ratio	*Cycles correctly clustered*	*Errors*	*Misses*	*Accuracy (%)*
No noise	921	0	3	99.7%
SNR = 8.00	921	0	3	99.7%
SNR = 4.00	921	0	3	99.7%
SNR = 2.67	918	3	3	99.4%
SNR = 2.00	907	14	3	98.2%
SNR = 1.33	726	195	3	78.6%
SNR = 1.00	561	360	3	60.7%

From no noise to a SNR = 2, the results of the algorithm remained stable (99.7-98.2% of accuracy). With SNR values of 1.33 and 1.00 the accuracy lowered to 78.6 and 60.7%, respectively, which are acceptable considering an amount of noise with amplitude equal or superior than the signal to analyze.

CONCLUSION

In this work we further evaluated an algorithm previously presented, which is based on a generic mean wave approach to cluster the cycles of biosignals.

Our algorithm proves to be an accurate method in the detection of changes in the signal's morphology, achieving 99.3% of clustering accuracy in signals with only two different modes or activities. The algorithm accuracy for signals with three different modes was tested in this chapter, achieving an overall result of 96.9%.

We compared our clustering procedure with another method referenced in literature, the cepstral coefficients algorithm, which presented the best results to the date for time series data. We obtained better results using the same dataset of acquired data – the mean wave algorithm presents an accuracy 19% superior for the same data. The mean wave procedure is also much more appropriate for the analysis of continuous signals, as it automatically separates the signals' cycles and doesn't need different inputs for different signal's modes, unlike the cepstral coefficients algorithm.

Finally we performed a noise immunity test with a synthetic signal, adding Gaussian noise until the clustering procedure decreases in accurateness. Only with a noise of amplitude equal to the synthetic signal and therefore a signal-to-noise ratio of 1, the accuracy of the algorithm drops to 60.7%; however, the algorithm proved to be relatively stable for a SNR higher than 2.

The necessary validation tests that we performed in this study confirmed the high accuracy level of the developed algorithm for biosignals which express human behavior.

The continuously need to obtain more information, with more accuracy, more quickly and with less intervention from an expert has led to a growing application of signal processing techniques applied to biomedical data. The biosignal analysis and processing is a promising area with huge potential in medicine, sports and research.

In fact, pattern recognition and automatic classification of morphological and physiological deviations on biosignals using clustering techniques are essential on monitoring elderly people on their homes (AAL). Thus, the mean wave algorithm is a great asset in this context and can contribute to the main goal of AAL: increase the period of time in which elderly people are autonomous by being able to detect behavior and changes in biosignals (e.g. arrhythmia, fall detection, epilepsy episodes). Concerning sports, we can identify different actions in the same data resulting in correct group detachment. This outcome provides feature extraction in order to recognize and categorize patterns. Furthermore, it allows us to create a mathematical model which is capable to classify new movements that can be directly used on talent recruitment and/or sport's training optimizations.

Our algorithm can be applied to continuous cyclic time series, capturing the signal's behavior. The fact that this approach doesn't require any prior information and its good performance in different situations makes it a powerful tool for biosignals analysis and classification.

REFERENCES

AAL4ALL – Ambient Assisted Living for All. Last accessed in February 17, 2012 at http://www.aal4all.org/?lang=en.

Antoniol, G., Rollo, V., and Venturi, G., Linear predictive coding and cepstrum coefficients for mining time variant information from software repositories. *MSR2005: International Workshop on Mining Software Repositories*, 2005.

Bagnal, A. and Janacek, G., Clustering time series from arma models with clipped data. In *Procedings of KDD '04, the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, USA, Aug. 2004.

Boets, J., Cock, K., Espinoza, M., and Moor, B., Clustering time series, subspace identification and cepstral distances. *Communications in Information and Systems,Vol. 5, No. 1*, pages 69–96, 2005.

Kalpakis, K., Gada, D., and Puttagunta, V., Distance measures for accurate clustering of arima time-series. In *Procedings of the 2001 IEEE International Conference on Data Mining*, pages 273–280, 2001.

M. Corduas and D. Piccolo.Time series clustering and classification by the autoregressive metric.*Computational Statistics & Data Analysis, Vol 52*, pages 1860–1872, 2008.

Nunes, N., Araújo, T. and Gamboa, H. (2012) Time Series Clustering Algorithm for Two-Modes Cyclic Biosignals. In A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 233--245. Springer, Heidelberg.

PLUX – Wireless Biosignals, S.A. Last accessed in February 15, 2012 at www.plux.info.

Physionet – PhysioBank Archive Index. Last accessed in February 15, 2012 at http://www.physionet.org/physiobank/database.

Rodrigues, G., Alves, V., Silveira, R., Laranjeira, L., "Dependability analysis in the Ambient Assisted Living Domain: An exploratory case study," The Journal of Systems and Software 85 (2012) 112–131

Santos, R., Sousa, J., Sañudo, B., Marques, C., Gamboa, H., "Biosignals Events Detection A Morphological Signal-Independent Approach," Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2012), Vilamoura, Portugal, 2012.

Savvides, A., Promponas, V. and Fokianos, K.., Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition, Vol 41*, pages 2398– 2412, 2008.

Souza., P. Statistical tests and distance measures for LPC coefficients. *IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 25, Issue 6*, pages 554–559, 1977.

Key Terms & Definitions

Mean wave: Having a set of signal's cycles, the mean wave is a wave constructed by the mean value for each time-sample of those cycles.

Clustering: A method for assigning an observation to a specific group of observations which share some characteristics that differentiates them from others.

Biosignals: The human body produces physiological signals which can be measured. To those signals we call "biosignals".

A.2 Biosignals 2013

A signal-independent algorithm for information extraction and signal annotation of long-term records.

A signal-independent algorithm for information extraction and signal annotation of long-term records

Rodolfo Abreu[1], Joana Sousa[2] and Hugo Gamboa[1,2]
[1]*CEFITEC, Physics Department, FCT-UNL, Lisbon, Portugal*
[2]*PLUX - Wireless Biosignals, S.A., Lisbon, Portugal*
rodolfo.t.m.abreu@gmail.com, jsousa@plux.info, hgamboa@fct.unl.pt

Keywords: Biosignals, Waves, Events detection, Features extraction, Pattern recognition, *k*-means, Parallel computing, Signal processing.

Abstract: One of the biggest challenges when analysing data is to extract information from it, especially if we are dealing with huge amount of data, which brings a new set of barriers to be overcome. In this study, we present a signal-independent algorithm with two main goals: perform events detection in biosignals and, with those events, to extract information using a set of distance measures, which will be used as input to a modified *k*-means algorithm. The first goal is achieved by using two different approaches. Events can be found based on peaks detection through an adaptive threshold defined as the signal root mean square (RMS) or by morphological analysis through the computation of the signal *meanwave*. The final goal is achieved by dividing the distance measures into *n* parts and by performing *k*-means individually. For this study, a set of different types of signals was acquired and annotated by the presented algorithm. By visual inspection, we obtained an accuracy of 97.6% using L_1 Minkowski distance as distance function and 97.5% using L_2 Minkowski and *meanwave* distances. The fact that this algorithm can be applied to long-term raw biosignals and without requiring any prior information about them makes it an important contribution in biosignals' information extraction and annotation.

1 INTRODUCTION

Due to the constant evolution in sensing systems and computational power, biosignals acquisition and processing are always adapting to new technologies.

The main goal of clustering algorithms is to find information in data objects that allows to find subsets of interest – *clusters* – where objects in the same cluster have a maximum homogeneity. Therefore, the clustering base problem appears in various domains and is old, being traced back to Aristotle (Hansen and Jaumard, 1997).

In fact, clustering algorithms can be typically applied to computer sciences, life and medical sciences, astronomy, social sciences, economics and engineering (Xu and Wunsch, 2009).

Applying clustering techniques to biosignals is an approach that has been used recently. Due to the large amount of data that is analysed nowadays, clustering techniques are used for feature extraction and pattern recognition. Clustering on electrocardiography (ECG) signals has been used to group the QRS complexes (or beats) into clusters that represent central features of the data (Lagerholm et al., 2000; Cuesta-Frau et al., 2002). Also in electromiography (EMG), clustering algorithms have been used to cluster data features which will be used as input of a classifier, allowing a high training speed (Chan et al., 2000).

Due to the increasing amounts of data coming from all types of measurements and observations, some parallel computing techniques have been applied to clustering algorithms. These parallel techniques usually consist in performing *data parallel* and/or *task parallel* strategies. In the first strategy, the idea is to divide and distribute data into different processors and each one will compute the allocated data. The latter consists in dividing a main task into sub-tasks and dispatching them into different processors (Zhang et al., 2006). The Master/slave strategy was used by (Tsai et al., 1997; Kantabutra and Couch, 2000; Zhang et al., 2006), where the main program is run by the host, being in charge of data distribution and cluster results gathering. On the other hand, (Wu et al., 2000; Dhillon and Modha, 2000) achieved running time improvement using a wider bus system and not more processors. Our modified *k*-means algorithm only uses the own computer's processors and therefore, does not require a system network implemented.

In opposition to the previously presented studies, we present an algorithm that is signal-independent and which can be used in long-term biosignals, overcoming two main challenges while analysing biosignals. Although it is not mandatory, a signal-specific pre-processing can be applied, obtaining even better performance. Besides, it is required that biosignals are cyclic and without major variations in the fundamental frequency.

The developed algorithm aims at extracting information of biosignals and, with that information, apply clustering techniques to perform biosignals annotation. For that, detect signal events is required, which can be broadly defined as changes in state of the system under study (Ciaccio et al., 1993). In order to accomplish this, two different approaches were taken: one based on peaks detection through an adaptive threshold defined as the signal root mean square (RMS) and the other based on the computation of the signal *meanwave* by calculating the mean value for each time-sample of the signal's cycles (Nunes et al., 2011). Then, a morphological comparison between signal's cycles and the *meanwave* allows events detection which will be aligned using as reference point the cycle's minimum or maximum value. This precise alignment is essential to perform accurate distance measures between signal's cycles and between those and the respective *meanwave*. Both approaches consist of dividing the signal into parts and executing events detection in each part separately. Although the latter approach has a quite higher computational cost than the peaks detection approach, one should not discard the former due to its simple implementation and quickness which is an important feature when dealing with long-term biosignals.

Once the events detection step is complete, our algorithm performs a set of distance measures using different distance functions and methods in order to study which one returns the best results. These distance measures will be used as input for a modified *k*-means algorithm that divides the observations to be clustered in different parts, performs *k*-means for each part and finally assembles the results provided by each one. Thus, in this paper we present an approach that allows long-term signal classification without any prior information and with fast speed performance due to the employment of parallel computing techniques.

To test the performance of our algorithm, a set of long-term cyclic biosignals were acquired, such as accelerometry (ACC), blood volume pressure (BVP), electrocardiography (ECG) and respiration.

In the following section we present the detailed description of our algorithm steps and in section 3 it will be provided the results and discussion of our algorithm performance, concluding the paper in section 4 with an improvements' discussion.

2 SIGNAL PROCESSING ALGORITHMS

In this section it will be presented the acquisition system used to obtain the biosignals and also the detailed description of our algorithm.

2.1 Data Acquisition

In order to obtain the biosignals necessary to test our algorithm, we used a triaxial accelerometer sensor (*xyzPLUX*), an ECG sensor (*ecgPlux*), a BVP sensor (*bvpPlux*) and a respiration sensor (*respPlux*). These sensors were connected to a device – bioPlux reserach unit – responsible for the signal's analog to digital conversion and bluetooth transmission to the computer. Signals were sampled at a 1000 Hz frequency and converted using a 12 bit ADC (PLUX, 2012).

2.2 Algorithm implementation

As it was stated in the previous section, our algorithm can be divided in two distinct phases, which are depicted in Figure 1. This algorithm was implemented using Python with the scipy (Scipy, 2012) package.

As one can see, the first step consists of detecting signal events after dividing it into *N* parts. The last step consists of performing distance measures using a set of different distance functions and methodologies to be used as input for a modified *k*-means algorithm.

Both of these steps will be thoroughly discussed next.

2.2.1 Events Detection

As it was stated in the previous section, the first step in our algorithm is to detect events in cyclic biosignals. Since the main goal of this study is the extraction of information and to perform signals annotation, it is convenient that more than one approach is presented. In order to accomplish this, we propose two different methods for events detections which will be described next.

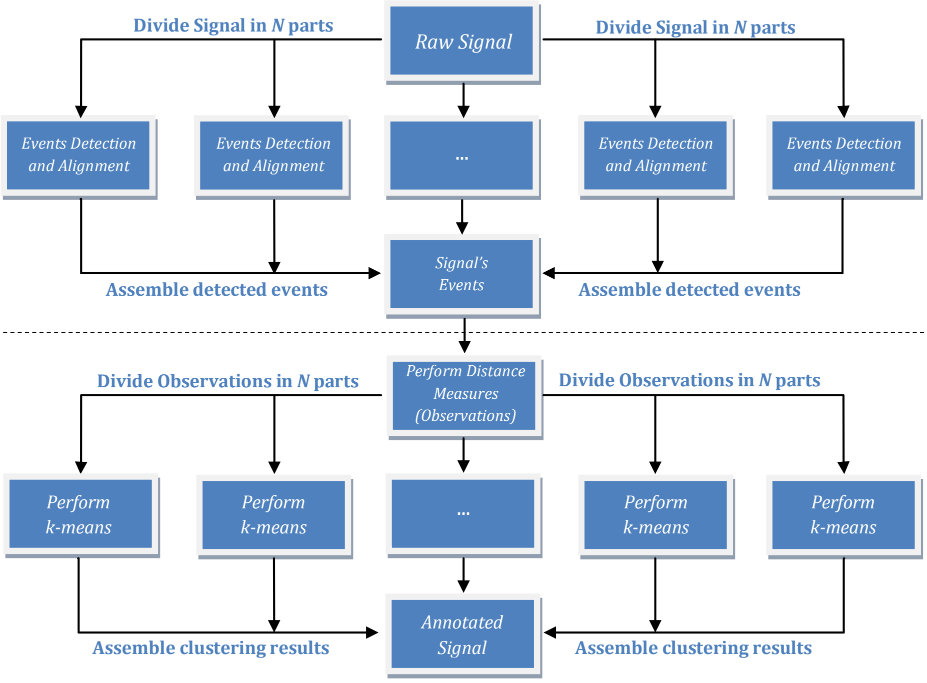

Divide Signal in N parts
Raw Signal
Divide Signal in N parts
Events Detection and Alignment
Events Detection and Alignment
...
Events Detection and Alignment
Events Detection and Alignment
Assemble detected events
Signal's Events
Assemble detected events
Divide Observations in N parts
Perform Distance Measures (Observations)
Divide Observations in N parts
Perform k-means
Perform k-means
...
Perform k-means
Perform k-means
Assemble clustering results
Annotated Signal
Assemble clustering results

Figure 1: Signal Processing Algorithm Steps.

Peaks detection approach. One of the biggest challenges when searching for peaks in biosignals is to find a suitable threshold. In our approach, we define this threshold as being the root mean square (RMS) of the signal, with the mathematical definition being presented in Equation 1 (H. Duarte-Ramos and Ortigueira, 2006).

$$RMS = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |x[n]|^2} \tag{1}$$

In order to obtain a higher accuracy in detecting signals events, our algorithm updates the threshold every ten seconds. In Figure 2 presents an example of a respiration signal with the peaks detected using this approach; one can also observe the horizontal lines representing the evolution of the threshold (RMS) over time.

Although this approach brings interesting results, using the concept of waves and *meanwave* provides more solid ones in signals with noise, significant morphological changes or baseline deviations. Nevertheless, due to its simplicity and low computational cost which is fundamental since our algorithm is also designed to be applied in long duration records, using the signal's RMS as an adaptive threshold for peaks detections is also an interesting method for accomplishing the first step of our algorithm. Once the peaks are detected, a *meanwave* is also constructed, obtaining one more source of morphological comparison of waves.

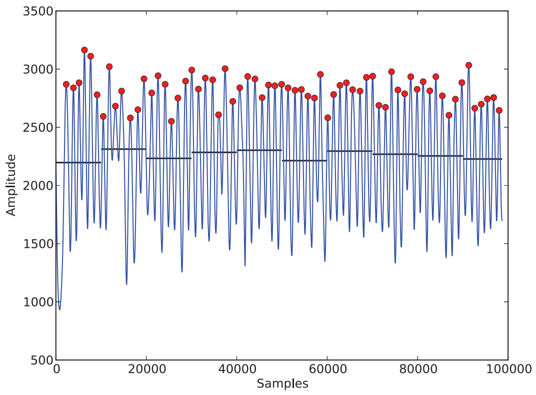


Figure 2: Peaks detections using signal's RMS as an adaptive threshold.

***Meanwave* approach.** In this approach, the basic concept was previously implemented by (Nunes et al., 2011) and, therefore, only a brief description will be presented.

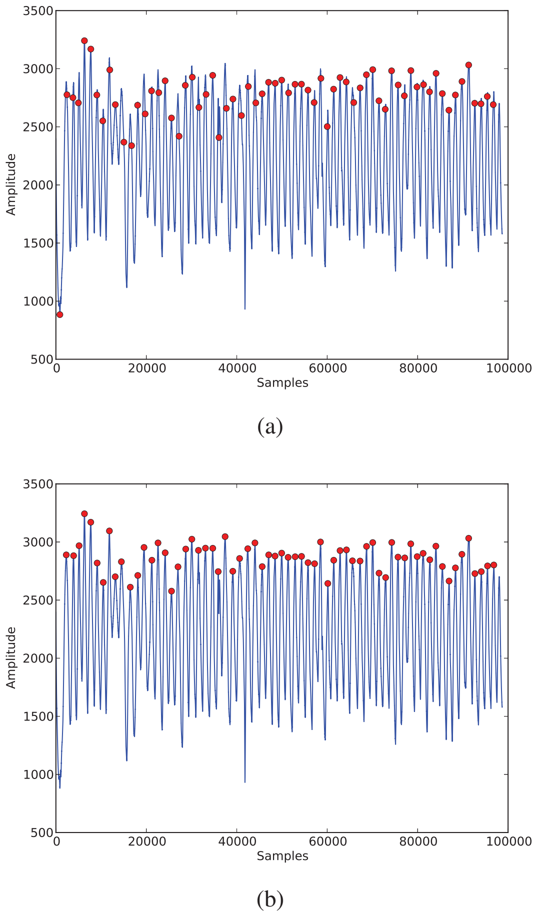
The *autoMeanWave* algorithm has the main goal the events detection on biosignals and for that, a *meanwave* is automatically computed, capturing the signal's behaviour. In order to construct the *meanwave*, the signal must be cyclic and those cycles must be separated, making the fundamental frequency (f_0) estimation an essential part of the process. In the *autoMeanWave* algorithm, the FFT of the signal is computed and the first peak found (after applying a smoothing filter), assuming that this peak corresponds to the signal's first harmonic and, consequently, the signal's f_0. Therefore, one can also estimate the cycles' size – *winsize* –, given by $^{f_s}/_{f_0}$, with a margin of 20%. Then, a random part of the original signal with a length of *winsize* is selected and a correlation function is applied to calculate a distance signal showing the difference between each overlapped cycle and the window selected at the first place; the local minima of the distance signal are found, assuming to be the signal's events. Finally, the *meanwave* is computed and the signal's events are aligned to a reference point which one can choose among a set of options.

Although the basic concept of the first step of our algorithm is previously shown, some improvements were made and the possibility of applying this algorithm in long-term biosignals was added. In fact, fundamental frequency estimation is one of the most important parts of our algorithm. Without an accurate value for f_0, and being f_{0_e} the estimated fundamental frequency, if $f_{0_e} \ll f_0$ then a high number of cycles will be despised; on the other hand, if $f_{0_e} \gg f_0$, many cycles that do not exist will be considered. Therefore, instead of determining the signal's FFT first peak, we use a time-domain method for f_0 estimation based on the autocorrelation of finite time series function of size N, $x[n]$ (representing our signal), where its mathematical definition is shown in Equation 2.

$$R_x(v) = \sum_{n=0}^{N-1-v} x[n]x[n+v] \tag{2}$$

It is well known that are many ways of computing the fundamental frequency since this is a current and active research topic. In fact, an ultimate method for f_0 estimation is yet to be discovered (Gerhard and of Regina. Dept. of Computer Science, 2003). However, the autocorrelation approach proved to return better results than the previously presented method.

Once a more accurate estimation for f_0 is implemented, we also improved the signal's events alignment step. In order to obtain an accurate morphological comparison between waves, an almost perfect alignment of the signal's events is required. In (Nunes et al., 2011) the alignment is achieved by selecting a notable point from the computed *meanwave*. For certain types of signals, this led to an inaccurate events alignment and, therefore, an incorrect distance measure between waves and between the *meanwave*. In order to solve this issue, after performing the alignment through the maximum value of the *meanwave* (one of the notable points presented as a trigger option), our algorithm runs through all the computed waves and relocates the events to the maximum value of each wave. An example of this further events alignment is shown in Figure 3. It is important to state that the alignment using the minimum value of each wave is also possible.



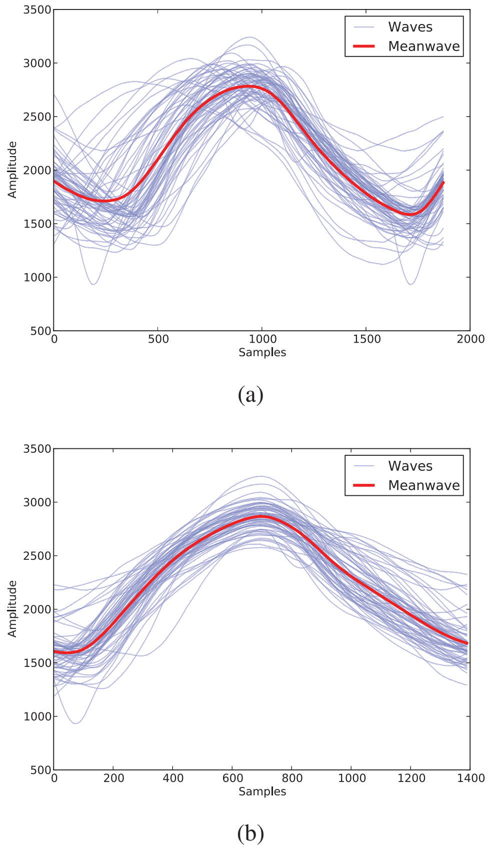
(a)

(b)

Figure 3: Illustration of the two alignment approaches: (a) events are aligned using the maximum value of the *meanwave* and (b) events are aligned using the maximum value of each wave.

One might observe the differences between the two alignment approaches in Figure 4. It is also interesting to notice that waves' (and *meanwave*) length differ due to the different method chosen to estimate the fundamental frequency, which were already exposed previously in this section.

The last improvement made on the *autoMeanWave* algorithm is the ability to run over long-term biosignals. In order to accomplish this goal, our al-



(a)

(b)

Figure 4: Waves and *meanwave* alignment using (a) maximum value of the *meanwave* and (b) maximum value of each wave.

gorithm divides signals into N parts and each part is processed individually. Hence, being L the length of the original signal, each part will have a length of $L_p = L/N$; it is important to notice that the last part might be smaller than the remaining ones. To guarantee that no information is lost among transition zones, we introduce a f_0-dependent overlap with length L_o, resulting in a total length for each part of $L_{pf} = L_p + L_o$. This will result in double detections and to remove them, the following logic was applied: if the event e_i, from part i (with $i = 2, \ldots, N$), belongs to the set

$$V(e_{i-1}) =]e_{i-1} - 0.3 \times winsize; e_{i-1} + 0.3 \times winsize[,$$

where e_{i-1} is the last event of the part's $i-1$ overlap, then all the detected events that precede e_i (inclusive) are eliminated.

To overcome the obstacle of dealing with large amount of data, the HDF5 format was used. Hence, the storage of large sized data and its fast access is possible, being these features the main goal of HDF5 files (HDF Group, 2012).

Once the events are correctly detected and aligned, distance measures are taken and clustering techniques are applied to obtain signal annotations.

2.2.2 Distance Measures

There are several distance measures that can be applied to one-dimensional arrays and more specifically, to time-series. In order to obtain inputs to our modified k-means algorithm, we use a set of different distance functions. First of all, the Minkowski-form Distance defined as (Chan et al., 2000)

$$L_p(P,Q) = \left(\sum_i |P_i - Q_i|^p \right)^{1/p} \quad , \quad 1 \leq p \leq \infty \quad (3)$$

with P and Q two one-dimensional arrays.

In this study, we will use the L_1, L_2 and L_∞ distance functions, which are defined as

$$L_1(P,Q) = \sum_i |P_i - Q_i|$$

$$L_2(P,Q) = \sqrt{\sum_i (P_i - Q_i)^2}$$

$$L_\infty(P,Q) = \max_i |P_i - Q_i|$$

The squared version of L_2, L_2^2, will also be used. Besides, the χ^2 histogram distance given by (Pele and Werman, 2010)

$$\chi^2(P,Q) = \frac{1}{2} \sum_i \frac{(P_i - Q_i)^2}{P_i + Q_i} \quad (4)$$

will also be utilized to obtain distance measures. Due to the suitability of our algorithm in long-term biosignals, these distance measures are not represented by a distance matrix. In fact, one can see biosignals as time-series which have an important feature that allows distance measures without building an extremely high computational cost distance matrix when dealing with long records: the order relationship between two consecutive samples. Hence, morphological comparisons between waves, w_i, result in a *distance array* where each element, d_i, is given by:

$$d_i = f(w_i, w_{i+1}), \, i = 1, \ldots, n-1 \quad (5)$$

being f the distance function and n the number of waves, representing also the number of events detected by the previous step of our algorithm. In the particular case where $w_i = mw$, where mw denotes the signal's *meanwave*, then each element of the distance array will be:

$$d_i = f(mw, w_i), \, i = 1, \ldots, n \quad (6)$$

Although the distance matrix carries richer information about waves' resemblance than the distance array, its high computational costs makes it impracticable in long records.

Using this set of distance functions, a comparison between the efficiency of each one as an input for our clustering algorithm will be made, allowing to state which is the most adequate distance function for morphological analysis in biosignals.

2.2.3 Clustering Algorithm

The final step of the presented algorithm consist of implementing a modified k-means, being able to perform unsupervised learning on long-term biosignals. The main concept of the k-means algorithm was kept (Forgy, 1965; MacQueen et al., 1967), which is a partitioning method for clustering where data is divided into k partitions (Warren Liao, 2005). The optimal partition of the data is obtained by minimizing the sum-of-squared error criterion with an interactive optimization procedure. Clustering algorithms can perform hard-clustering when each cluster can only be assigned to one partition; otherwise, they perform fuzzy-clustering. Our algorithm was designed to perform hard-clustering since it was based in the k-means hard clustering algorithm. As an additional modification to the original k-means, we also introduce n iterations to the algorithm in order to minimize one of the biggest drawbacks related to initial partition. In fact, different initial partitions usually converge to different cluster groups (Xu et al., 2005; Xu and Wunsch, 2009).

To proceed with our algorithm explanation, one might observe Figure 5. After the distance measures are obtained, the array of length M containing that information is divided into N parts; the last part might be smaller than the remaining ones. Then, k-means clustering algorithm will be run in each part and a set of centroids $[\mathbf{a}_i, \mathbf{b}_i, \ldots, \mathbf{k}_i]$ will be computed, with $i = 1, \ldots, N$ being the part number and k the number of partitions given as input for the k-means algorithm. If each observation's element is n-dimensional, therefore each centroid will be also n-dimensional. Assembling each part's set of centroids, we obtain

$$[[\mathbf{a}_1, \mathbf{b}_1, \ldots, \mathbf{k}_1], \ldots, [\mathbf{a}_N, \mathbf{b}_N, \ldots, \mathbf{k}_N]]$$

Considering this as a new set of observations, one can run one more time the k-means algorithm, obtaining the *global centroids* of the k partitions of the original array containing the distance measures information. With these centroids, the Euclidean Distance (defined as the L_2 distance) is computed between each centroid and each observation, resulting in a $k \times M$ matrix. Searching for the line where the minimum element of each row is located, the cluster which that observation will be assigned to is provided. Due to this approach, all of the k-means issues will be amplified but still, the obtained results are quite satisfactory, as it will be shown in the next section.

3 RESULTS AND DISCUSSION

Since our algorithm aims at extracting information in a broad perspective, two different approaches were taken, resulting in two independent types of clustering results. First, our algorithm verifies the time-samples difference, $\Delta t_i = t_{i+1} - t_i$; finally, it performs a morphological comparison between waves (see Equation 5 and 6).

A visual inspection for performance evaluation was taken and different criteria were used for the different types of clustering results. However, the concepts of *error* (when a cycle is wrongly identified or classified) and *miss* (when a cycle is not classified) are used in both types of results.

The validation process was taken only using the *meanwave* approach to obtain signals' events since its higher accuracy is necessary to obtain more reliable clustering results.

3.1 Clustering using time-samples difference information

Despite its conceptual simplicity, an almost perfect detection and alignment events can lead to a time-samples variability analysis between those events. An example is shown in Figure 6 where a electrocardiography (ECG) signal is represented.

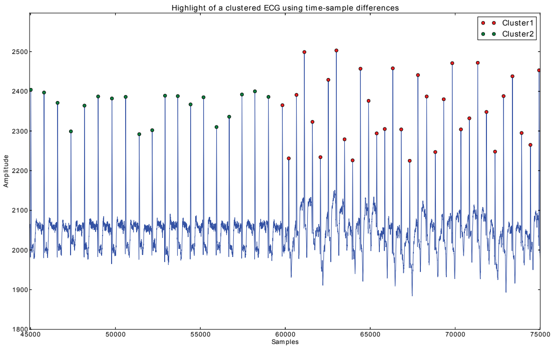


Figure 6: Clustering using events' time-samples variability.

With this information only, one can conclude that during the first minute (recall that it was used a sampling frequency of 1000 Hz) the subject was at rest due to signal's lower frequency represented by a

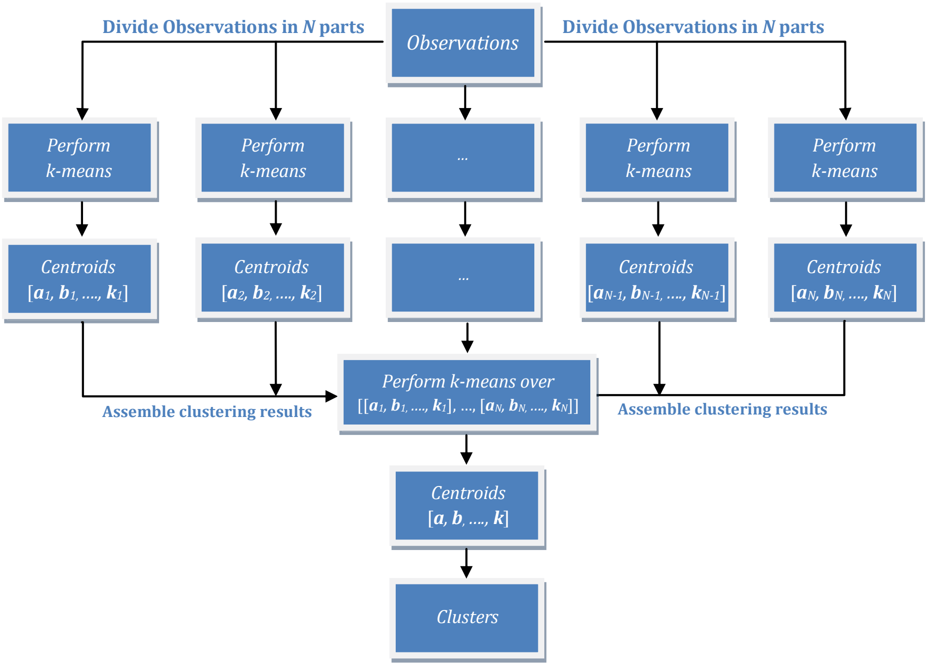

Divide Observations in N parts
Observations
Divide Observations in N parts
Perform k-means
Perform k-means
...
Perform k-means
Perform k-means
Centroids [a1, b1, ..., k1]
Centroids [a2, b2, ..., k2]
...
Centroids [aN-1, bN-1, ..., kN-1]
Centroids [aN, bN, ..., kN]
Perform k-means over [[a1, b1, ..., k1], ..., [aN, bN, ..., kN]]
Assemble clustering results
Assemble clustering results
Centroids [a, b, ..., k]
Clusters

Figure 5: Modified k-means algorithm schematics.

longer Δt_i intervals and afterwards, the subject started some activity or exercise verified by Δt_i intervals' reduction.

After running this clustering method in order to divide the data in k partitions, the obtained results are presented in Table 1.

The ECG$_2$ and BVP signals were acquired when the subject was at rest and then started some exercise. Thus, their major difference relies in frequency variations. Using Δt_i information, better results than using the morphological analysis were returned. On the other hand, on ECG$_1$, a heart rate variability (HRV) analysis should be taken in order to assess the clustering results.

3.2 Clustering using morphological comparison

Next, a morphological analysis was taken in order to obtain signals' annotations. As it was mentioned before, our algorithm uses a set of distance functions in order to study which one brings better results. The obtained results are presented in Table 2 and the respective accuracy obtained using each distance function in Table 3.

Firstly, the clustering results for ECG$_2$ and BVP are expected since there isn't any notable change in their morphology and, therefore, since k-means forces data to be divided into k partitions, we obtained poor results using any distance function. However, as it was shown previously, clustering using Δt_i information had a good performance.

In ECG$_1$, an approximately 7 hours signal, we were able to test our algorithm to perform events detection and clustering on a long record. The highest accuracy was obtained using the L_1 distance, although the L_2 and *meanwave* distances also led to high algorithm performance. It is also important to analyse the number of missed cycles; in fact, the signal's visual inspection was taken by dividing it into 26 parts and, therefore, we recorded each part's number of missed cycles. The mean value were 10.5 but the standard deviation was 14.4. This shows that in some parts of the signal, there were significant changes in fundamental frequency, resulting in a high number of missed cycles, while in other parts, the number of misses was extremely low. In order to minimize the number of missed cycles, smaller parts could be analysed allowing a more sensible perception of the fundamental frequency's temporal evolution. However, sensitivity for noise presence is also augmented, producing poorly results when determining cycles' sizes.

For the ACC$_1$ it was asked the subject to walk, jump, walk and jump again. The ACC$_2$ (jumping, leg flexion and single leg vertical jumping) represents a three mode signal. Analysing these signals, only *meanwave* distance resulted in high algorithm performance. These results are possibly related to the higher sensitivity of *meanwave* distance measures. In fact, when a signal is divided into n parts, it will be constructed n *meanwaves* that will be used to calculate distances between them and each cycle present among the n parts.

Analysing the results globally, the L_1 and L_2 distances returned a total of 318 and 344 errors out of 25008 cycles, achieving 97.6% and 97.5% of accuracy, respectively. Besides, the *meanwave* distance returned a total of 357 errors out of 25707 cycles, achieving 97.5% of accuracy. The other distance functions performed poorly and should not be considered as good distance functions for biosignals' clustering.

4 CONCLUSIONS AND FUTURE WORK

In this paper we presented a signal-independent algorithm for long-term signal's processing and time series clustering. First, an events detection step is taken and then clustering techniques are applied using a modified k-means capable of classifying large sized data. Our algorithm does not need any prior information of the signal and has a high speed performance due to the order relation between two consecutive samples, a key feature in time series that allows the computation of a distance array instead of a distance matrix. Besides, we also concluded that the L_1, L_2 and *meanwave* distance functions lead to better clustering results.

The presented algorithm proves to be an asset in biosignals processing research area since it as a high efficiency in performing biosignals' annotation, which can be used as a resource to aid and complement physicians' analysis.

In the future, we aim to automatically find the optimal length of each part of the divided signal that allows a better monitoring of the temporal evolution of the fundamental frequency. This would lead to a significant reduction in the number of missed cycles. Being events detection the more time consuming step of our algorithm, we also aim to improve this point by using parallel computing techniques.

ACKNOWLEDGEMENTS

This work was partially supported by National Strategic Reference Framework (NSRF-QREN) under projects AAL4ALL and wiCardioResp, whose support the authors gratefully acknowledge.

REFERENCES

Chan, F., Yang, Y., Lam, F., Zhang, Y., and Parker, P. (2000). Fuzzy EMG classification for prosthesis control. *Rehabilitation Engineering, IEEE Transactions on*, 8(3):305–311.

Ciaccio, E., Dunn, S., and Akay, M. (1993). Biosignal pattern recognition and interpretation systems. *Engineering in Medicine and Biology Magazine, IEEE*, 12(3):89–95.

Cuesta-Frau, D., Pérez-Cortés, J., Andreu-García, G., and Novák, D. (2002). Feature extraction methods applied to the clustering of electrocardiographic signals. A comparative study. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 961–964. IEEE.

Dhillon, I. and Modha, D. (2000). A data-clustering algorithm on distributed memory multiprocessors. *Large-Scale Parallel Data Mining*, pages 802–802.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769.

Gerhard, D. and of Regina. Dept. of Computer Science, U. (2003). *Pitch extraction and fundamental frequency: History and current techniques*. Dept. of Computer Science, University of Regina.

H. Duarte-Ramos, F. Coito, R. S. and Ortigueira, M. (2006). *Análise de Sinais em Engenharia Biomédica*. FCT-UNL.

Hansen, P. and Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical programming*, 79(1):191–215.

HDF Group (2012). The HDF Group. `http://www.hdfgroup.org/why_hdf/`. [Accessed on August, 2012].

Kantabutra, S. and Couch, A. (2000). Parallel K-means clustering algorithm on NOWs. *NECTEC Technical journal*, 1(6):243–247.

Lagerholm, M., Peterson, C., Braccini, G., Edenbrandt, L., and Sornmo, L. (2000). Clustering ECG complexes using Hermite functions and self-organizing maps. *Biomedical Engineering, IEEE Transactions on*, 47(7):838–848.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA.

Nunes, N., Araújo, T., and Gamboa, H. (2011). Two-modes cyclic biosignal clustering based on time series analysis.

Pele, O. and Werman, M. (2010). The quadratic-chi histogram distance family. *Computer Vision–ECCV 2010*, pages 749–762.

PLUX (2012). PLUX - Wireless Biosignals, S.A. `http://www.plux.info/`. [Accessed on August, 2012].

Scipy (2012). Scipy. `http://www.scipy.org/`. [Accessed on August, 2012].

Tsai, H., Horng, S., Tsai, S., Lee, S., Kao, T., and Chen, C. (1997). Parallel clustering algorithms on a reconfigurable array of processors with wider bus networks. In *Parallel and Distributed Systems, 1997. Proceedings., 1997 International Conference on*, pages 630–637. IEEE.

Warren Liao, T. (2005). Clustering of time series dataa survey. *Pattern Recognition*, 38(11):1857–1874.

Wu, C., Horng, S., Chen, Y., and Lee, W. (2000). Designing scalable and efficient parallel clustering algorithms on arrays with reconfigurable optical buses. *Image and Vision Computing*, 18(13):1033–1043.

Xu, R. and Wunsch, D. (2009). *Clustering*. Wiley-IEEE Press.

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

Zhang, Y., Xiong, Z., Mao, J., and Ou, L. (2006). The study of parallel k-means algorithm. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, volume 2, pages 5868–5871. IEEE.

Signal	Cycles	Misses	Correctly clustered cycles	Errors	Accuracy
ECG_1 $(k = 2)$	24551	272	24279	0	98.9%
ECG_2 $(k = 2)$	165	1	163	1	98.8%
BVP $(k = 2)$	165	1	162	2	98.2%

Table 1: Clustering results using Δt_i with k clusters.

Signal	Cycles	Misses		Correctly clustered cycles	Errors
ECG_1 $(k = 2)$	24551	272	L_1	24028	251
			L_2^2	23579	700
			L_2	23998	281
			L_∞	23447	832
			χ^2	23565	714
			Mw	24021	258
ECG_2 $(k = 2)$	165	1	L_1	140	25
			L_2^2	99	66
			L_2	146	19
			L_∞	138	27
			χ^2	98	67
			Mw	125	40
BVP $(k = 2)$	225	2	L_1	184	39
			L_2^2	192	31
			L_2	198	25
			L_∞	186	37
			χ^2	189	34
			Mw	184	39
Respiration $(k = 2)$	67	1	L_1	65	3
			L_2^2	64	4
			L_2	46	19
			L_∞	44	21
			χ^2	60	5
			Mw	52	13
ACC_1^* $(k = 2)$	672	1	Mw	667	5
ACC_2^* $(k = 3)$	27	1	Mw	25	2

Table 2: Clustering results using morphological comparison with k clusters.

Distance Function	All Cycles	Correctly clustered cycles	Accuracy
L_1	25008	24417	97.6%
L_2^2	25008	23934	95.7%
L_2	25008	30469	97.5%
L_∞	25008	29741	95.2%
χ^2	25008	23912	95.6%
Mw	25707	25074	97.5%

Table 3: Accuracy obtained using different distance functions for the clustering algorithm's input.