UNIVERSITA' DEGLI STUDI DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE:

ELETTRONICA, INFORMATICA, TELECOMUNICAZIONI

DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE

Ph.D. THESIS

DEVELOPMENT OF METHODS BASED ON VOLTAMMETRY FOR THE CHARACTERISATION OF LIQUIDS

SUPEVISORS

Prof. Giuliano Manara

_____

CANDIDATE

Prof. Giovanni Corsini **Andrea Scozzari**

_____ _____

Dott. Massimo Guidi

_____

February 2007

Preface

The growing interest into multi-sensor systems able to determine general attributes of a process under monitoring, has recently involved the qualitative analysis of liquids; various methodologies to develop taste sensors, often referred to as “e-tongues” have been presented in the literature. The common concept in the different approaches presented by various authors, lies in the combination of signals originated by poorly specific sensors for the characterization of liquids.

The fundamental idea of this PhD work is to investigate how an adequate signal processing approach, applied to a mature and affordable sensor technique (voltammetry), can address the issue of extracting an aggregate chemical information, useful to characterize the liquid under measurement, and even to provide information useful for continuous monitoring and change detection purposes.

Among the various application fields of such systems, there are: food and beverage quality tests, pollution monitoring of superficial and ground water, drinking water quality tests and classification, and control of water distribution networks.

This work has been conducted in the framework of a research for innovative monitoring technologies for the assessment of water quality and for the fast characterisation of surface and ground water; nevertheless, the described methodology and the experiences gained are fully applicable to other areas concerning environmental monitoring, industrial control, and security.

What makes the approach based on “taste sensors” a particularly attractive technology, is its capability to provide direct information about the quality of the resource, instead of the complex data sets coming from traditional selective laboratory measurements. Such feature makes this concept particularly suitable to automatic quality surveillance purposes.

In fact, while the information coming from traditional field instrumentation, providing the main physical/chemical parameters, is poor to fully characterise a liquid sample, a complete laboratory analytical measurement is expensive and cannot provide a continuous control over the resource.

The goal of this PhD activity has been to find new and better solutions to improve the processing of the data acquired with respect to the existent, with a specific interest into a sensor technique based on pulse voltammetry, which is a mature analytical technique, here used in a non-traditional way. Moreover, a prototype has been developed and practical measurement experiences have been carried out, to evaluate and improve the capabilities of the processing chain designed.

Acknowledgments

This Thesis has been made possible thanks to the contribution and great support from a series of people. I wish to thank all the personnel and researchers of CNR IGG (Institute of Geosciences and Earth Resources) that have always been very available to provide great help and suggestions.

Moreover to the supervisors of this Thesis, I must be very thankful to Dr. Nicola Acito, which has been co-author of a relevant part of the works that I published in the recent years. Also, Dr. Roberto Cioni, for his tangible contribution to the development of the sensors system used in this work.

Many thanks go to the company Acque S.p.A (water Distribution Company for the area of Pisa) and Dr. Paolo Peruzzi in particular, for his availability and expertise in the studies carried out about drinkable water quality.

Last but not least, my parents, that have always supported me in the difficulties, and with which I very much want to share the good things.

Contents

Introduction

In many areas of the measurement technology the methods for getting information about a process are rapidly changing. Instead of measuring single parameters there is a need of measuring more general attributes like quality, availability, efficiency, etc. Many processes are complicated and many different physical as well as chemical parameters need to be measured in real-time.

On that account the concepts of electronic noses and tongues have been developed and attract today a wide interest in different applications such as environmental monitoring. Gas sensor arrays (e-noses) have been until now more studied than their wet counterparts (e-tongues). In fact, electronic noses have already been established for qualitative analysis in various fields in analytical chemistry, especially for the food industry and some environmental applications; there are also some prototypes already on the market. Similar concept; but for analysis in water and even food, have recently been described. The above mentioned devices are respectively related to the sense of smell and taste; thus, the terms electronic nose and electronic tongue (or even taste sensor) have been coined.

The bio-mimetic connection of these systems is due to the way that the human sense organs produce signal patterns to be qualitatively interpreted by the brain. Electronic tongues and noses are in fact normally used to give some qualitative answers about the samples studied and only in special cases to predict the concentration of individual species in the sample, in a similar fashion to what traditional analytical chemistry usually does.

The proposed approach allows an original monitoring method, based on distributed low cost sensors, making use of the technology of the e-tongue; this concept now represents an inexpensive and very attractive way to obtain overall quality indicators of a natural resource or, more in general, of a process under monitoring.

Several attempts are found in the literature regarding the concept of electronic tongue. A prototype of an electronic tongue based on voltammetry has been developed in the framework of this PhD activity. Most of the work has been focused on the development of signal processing algorithms to extract from the data series obtained the useful information for classifying liquid samples, and even to detect eventual changes, including the presence of pollutants.

Experiences made with this prototype and the proposed signal processing approach, have demonstrated their capability in classifying water samples of any kind, i.e. bottled mineral water, superficial and ground water from field sampling campaigns, even thermal water, tap water and bottled beverages.

A brief overview of the state of the art for “e-tongue” systems, a description of the prototype developed, the novel methodological approach to the processing of the signal, and the results obtained in several application experiments, are presented in the next Chapters.

In Chapter I, a general description of electronic taste sensor systems is given. A brief summary of the most used techniques in such devices is presented, followed by an overview of

the main possible applications. Finally, the advantages of pulse voltammetry with respect to other techniques will be highlighted.

Chapter II gives an overview of the working principles of e-tongues based on voltammetry, paying attention to the fundamentals of voltammetry and pulse voltammetry in particular, which is the proposed technique in this Thesis.

Chapter III describes the methodology that represents the core of this PhD Thesis work; the sensor device, the control software and the data processing approach are described in this sequence. The signal processing approach contains most of the novelty with respect to the known literature.

Chapter IV contains a series of case studies, selected according to their relevancy for the peculiarities of the approach described in this Thesis.

CHAPTER I

An overview about taste sensors

1.1 Human sense of taste

The sense of taste is due to the excitation of taste receptors; receptors for a large number of specific chemicals have been identified, that contribute to the reception of taste. Despite this apparent complexity, five types of tastes are commonly recognized by humans:

- Sweet - usually associated with energy rich nutrients
- Umami - the taste of amino acids (e.g. meat broth or aged cheese)
- Salty - allows spontaneous tuning of the diet for electrolyte balance
- Sour - typically the taste of acids
- Bitter - allows sensing of various natural toxins

None of these tastes are elicited by a single chemical. Moreover, there are thresholds for the detection of taste that differ among chemicals that taste in the same way.

The sense of taste affords an animal the ability to evaluate what it eats and drinks. At the most basic level, this evaluation is used to:

- prevent consumption of potential poisons or toxins,
- select food/drinks according to body needs.

In general, food selection and eventual rejection are mediated through the central nervous system. In fact, the complex pattern of signals transmitted to the brain from the taste receptor cells generates the taste sensations. In a similar fashion, an artificial system that provides a similar kind of information may be developed, with a series of foreseen applications such as the ones cited in paragraph 1.3.

Such a bio-mimetic device can be somewhat specialised for different purposes, and can exceed the human detection thresholds, especially for those substances which may be present or not (i.e. small quantities of pollutant in water), but generally undetectable by the human sense of taste, even if potentially harmful.

One of the main goals of the activity regarding this PhD, is a kind of extension of the artificial sense of taste, in order to be able to distinguish between different classes of water even with very slight differences in their constituents, including those cases in which such differences are due to pollution.

In fact, the aim of such devices does not necessarily reside in the ability to emulate the human senses, but is more focused on the capability to characterise the liquid under measurement, in a way that is inspired by biological systems.

1.2 “E-tongue”: the concept

1.2.1 Definition of “E-tongue”

Electronic tongues and taste sensors are devices used to detect and separate compounds. They are in fact systems able to determine general attributes of a process under monitoring. In some papers, a difference between “electronic tongue” and “taste sensor” is made. For instance in [7] a taste sensor system is described as an array of sensors used to classify the different gustatory sensations; those are saltiness, bitterness, sourness, sweetness, etc. An electronic

tongue is a more general device not necessarily related to some organolectic feature, which gives qualitative information (even taste) in food, drinks, water, process fluids, etc. However, in most papers both terms are used as synonyms.

Behind these somewhat misleading terms “electronic tongue” and “taste sensor”, one finds some kind of array of sensors. The response patterns of such sensors can be analyzed with pattern recognition routines and/or chemometrical methods, where possible. In general, the basic principle behind an electronic tongue is to make use of non-specific and overlapping signals, via the implementation of some pattern recognition technique, to obtain the necessary discrimination between different classes of liquid, for classification and change detection purposes.

In fact, the sensor array in these systems produces signals, which are not necessarily selective to any particular species in the environment, drinking water, etc. but a signal pattern (often a large dataset) is generated, which can be related to certain chemical/physical features of the sample.

The main advantage of electronic tongues and taste sensors lies in the fact that they can help classifying or detect changes of quality in a shorter time and with a far lower analytical cost than a complete set of laboratory analyses, although they are not intended to replace them. In fact, there is a wide range of applications where an overall information is needed, and, the fast response of the device, together with its low analytical cost, contribute to make the application attractive, without competing with traditional chemometrics in terms of selectivity and quantitative precision.

1.2.2 Working principles

Various approaches have been proposed in literature to make multi-sensor systems for wet-chemical applications such as: ion selective and non-selective (glass) electrodes, lipid/polymer membranes, SH-SAW (Shear Horizontal Surface Acoustic Wave) devices, spectrophotometry and voltammetry.

Both glass electrodes and lipid/polymer membranes are based on potentiometry as working principle and require an array of sensors with different selectivity properties to operate; SAW devices [1] are based on the measurement of the propagation speed and attenuation of an acoustic wave due to acoustoelectric interactions with the fluid; spectrophotometry is based on light spectral transmittance measurements, and has been recently experimented on wines [2]; voltammetry takes information about the concentration and diffusion coefficients of redox active species through a measurement of the current flowing in a metal electrode with an applied potential having a specific waveshape, depending on the kind of voltammetric technique employed.

As to potentiometric techniques, a high discrimination capability among different brands of mineral water using an array of glass ion-selective electrodes (ISE) has been shown in [3] and [4].

Results obtained with arrays of potentiometric sensors applied to different kinds of beverages have been shown also recently in [5] and [6].

The known drawbacks of potentiometric techniques [7] lie mainly in the limitation of the detection to charged species, sensitivity to electrical noise and some requirement in the

maintenance of the electrodes, such as preparation and conditioning; it is thus worth looking for simpler and more robust approaches, possibly offering similar performance. It is thus of interest to exploit the effectiveness of the technique based on pulse voltammetry and solid metal working electrodes. This approach offers several advantages, such as simplicity, robustness, and a wide range of analytical possibilities, according to several parameters, such as the shape and material of the working electrodes, the excitation signal chosen, and the particular processing which is performed over the data.

Voltammetry, among other possible techniques, is one of the most extensively used in analytical chemistry; in addition, pulse voltammetry, with the aid of a suitable signal analysis support, has proven to be one of the most attractive techniques for classification and change detection purposes [7, 8, 9, 10].

1.3 A selection of known applications

Among the various present and foreseen applications for e-tongue systems, the main industrial, civil and scientific uses are:

- quality control in food and beverage industry,
- environmental monitoring: pollution detection,
- drinkable water: quality assessment and contamination control,
- geochemical surveillance: monitoring of natural systems (study and identification of changings connected with events such as earthquake precursory alterations).

In [1] an e-tongue based on an SH-SAW device, in a dual-delay configuration, has been described. One line is shorted (metallised and electrically shielded), while the other is left free (electrically active). In this way, mass loading and viscosity are measured by the shorted delay line, while permittivity and conductivity are monitored by the free delay line. Such parameters of the liquid under test can be related to taste properties.

An experiment made on the filtering plant of a Swedish drinkable water network is reported in [9, 10]. Differences in water quality before (raw water) and after (clean water) the filters are well detected. Such discrimination capability is discussed in both the articles referenced. The sensor technique used in this works is based on pulse voltammetry; everything has been kept at a very preliminary level with respect to the application, thus a commercial device has been used as potentiostat and no custom development for the application is described.

In [3] an e-tongue based on an array of 29 potentiometric chemical sensors is described. The analysis of Italian bottled mineral water and dry red wines is discussed, particularly in terms of discrimination capability. Other works, based on array of non-specific potentiometric sensors, have been presented by the same research group also recently.

In [2] the fusion of three sensor modalities is discussed for the characterisation of red wines: a combination of e-nose (smell of head-space gas), potentiometry (liquid phase electrochemical characterisation) and colour analysis is proposed. Correlation between the different sensor techniques is discussed, and the enhancement of the discrimination capability is demonstrated.

Several experiences about the effect of pollutants dissolved in water has been done in this PhD work; the most relevant ones can be found in [11], where a slight contamination by trichloroethylene and tetrachloroethylene is detected in two different case studies, and detection of very small quantities of terbuthylazine (a diserbant) has also been proved in another case study. Said experiments can be found in Chapter IV, which contains a selection of the experiments made during this PhD work.

In general, there are two classes of applications into which the application concepts can be split:

- continuous control for asserting quality (such as the taste of some kind of food),
- sporadic measurement for classification by similarity.

The first category regards those problems often labelled as *change detection* ones, while the second represents a typical *supervised classification* application.

A supervised classification framework has been developed in this PhD work, and has been presented in [12]. As to the sensor technique, an electrochemical method based on pulse voltammetry has been proposed in this work.

Change detection applications require a high maturity in the sensor and signal processing technique; they actually represent one of the most attractive challenges in this family of methods. Both industrial processes in the food and beverage industry and water distribution networks, with their recent issues regarding security against contamination and terrorist threat, would benefit very much from the advancements of this technology.

REFERENCES FOR CHAPTER I

[1] M. Cole, G. Sehra, J. W. Gardner, V. K. Varadan, "Development of smart tongue devices for measurement of liquid properties", IEEE Sensors Journal, Vol. 4, No. 5, October 2004

[2] M. L. Rodriguez-Mendez, A. A. Arrieta, V. Parra, A. Bernal, A. Vegas, S. Villanueva, R. Gutierrez-Osuna, J. A. de Saja, "Fusion of three sensory modalities for the multimodal characterization of red wines", IEEE Sensors Journal, Vol. 4, No. 3, June 2004

[3] A. Legin, A. Rudnitskaya, Y. Vlasov, E. Di Natale, C. Mazzone, and A. D'Amico, "Application of electronic tongue for quantitative analysis of mineral water and wine," Electroanalysis, vol. 11, no. 10–11, pp. 814–820, 1999

[4] P. Ciosek, Z. Brzozca, W. Wroblewski, "Classification of beverages using a reduced sensor array", Sensors and actuators B, 103 (2004), 76-83

[5] L. Lvova, S. S. Kim, A. Legin, Y. Vlasov, J. S. Yang, G. S. Cha, H. Nam, "All-solid-state electronic tongue and its application for beverage analysis", Analytica Chimica Acta 468 (2002) 303–314

[6] A. Legin, A. Rudnitskaya, B. Seleznev, Yu. Vlasov, "Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie", Analytica Chimica Acta 534 (2005) 129–135

[7] P. Hauptmann, R. Borngraeber, J. Schroeder, J. Auge, "Artificial Electronic tongue in comparison to the electronic nose – state of the art and trends", 2000 IEEE/EIA International Frequency Control Symposium and Exhibition

[8] M. Lindquist, P. Wide, "Virtual water quality tests with an electronic tongue", IEEE IMTC Budapest (Hungary), May 21-23, 2001

[9] C. Krantz-Rulcker, M. Stenberg, F. Winquist, I. Lundstrom, "Electronic tongues for environmental monitoring based on sensor arrays and pattern recognition: a review", Analytica chimica acta, 426 (2001), 217-226

[10] F. Winquist, P. Wide, I. Lundstrom, "An electronic tongue based on voltammetry", Analytica chimica acta, 357 (1997), 21-31

[11] A. Scozzari, P. Peruzzi, R. Cioni, M. Guidi, "An innovative approach to urban water management based on taste sensors", EGU General Assembly, Wien (Austria), 2006

[12] A.Scozzari, N. Acito, G. Corsini, "A supervised algorithm for water classification by voltammetric measurements", Instrumentation and Measurement Technology Conference, Sorrento (Italy), 2006

CHAPTER II

Taste sensors based on voltammetry

2.1 Electrochemical fundamentals about voltammetry

2.1.1 Definition of voltammetry

Voltammetry is a group of electroanalytical methods in which information about the analyte is derived from the measurement of a current flowing in a polarised working electrode, as a function of the applied potential. That is, the working electrode potential is forced to adhere to a predetermined program, and the current is measured as a function of the applied potential, or as a function of time.

In the methods described in this Thesis the electrodes are supposed to be microelectrodes, and the solution volume is supposed to be sufficiently large to alter negligibly the concentration of electroactive species with the passage of current.

Mass transfer in solutions occurs either because of a gradient in electrochemical potential or because of convection. As a first step, we can consider an unstirred solution, thus with no convection.

According to the approach developed by Butler and Guggenheim [1, 2, 3], the electrochemical potential in a location P is defined as:

$$\overline{\mu_i^\alpha} = \mu_i^\alpha + z_i \cdot F \cdot \phi^\alpha \qquad \textbf{(2.1)}$$

Where:

z_i is the charge of species *i*.

F is the Faraday's constant.

ϕ^α is the electrical potential at the location P for phase α.

μ_i^α is the chemical potential of the chemical species *i* in phase α, defined as:

$$\mu_i^\alpha = \left.\frac{\partial G}{\partial n_i}\right|_{T,P,n_{j\neq i}} \qquad \textbf{(2.2)}$$

Where:

the quantity G is the Gibbs' free energy.

n_i is the number of moles of *i* in phase α.

Another way of writing the electrochemical potential can be:

$$\overline{\mu_i^\alpha} = \left.\frac{\partial \overline{G}}{\partial n_i}\right|_{T,P,n_{j\neq i}} \qquad \textbf{(2.3)}$$

Where $\overline{G}$ is the electrochemical free energy, and differs from G because it includes the influence of the large-scale (i.e. not due to the ions) electrical effects, as they appear in equation (2.1).

The difference of electrochemical potential between two points (A,B) of a solution can arise because of:

- a difference of concentration of electroactive species over the distance A to B,
- an electric field with a non-zero component along the direction A-B.

These two phenomena contribute to the mass transport by diffusion and migration, respectively; in addition, the transport flux of charged species is equivalent to a current density.

In the proximity of an electrode, in general, the electroactive substance is transported by both processes (migration and diffusion).

For non convective mass transfer the general flux equation is given by:

$$J_j(x) = -D_j \nabla C_j - \frac{z_j \cdot F}{R \cdot T} \cdot D_j \cdot C_j \cdot \nabla \phi \qquad \textbf{(2.4)}$$

Where:

D_j is the diffusion coefficient of species j,

C_j is the local concentration of species j.

Stirred solutions, which are mostly used in practice, involve forced convection. In this case, a model has been proposed [4], where it is assumed that convection maintains the concentrations of all the species uniform and equal to the bulk values, till a certain distance δ from the electrode. Within this layer having thickness δ, it is assumed that there's no solution movement, thus mass transfer is thought to be due by diffusion only.

Information about the controlled potential microelectrode techniques is given here at an introductory level, to allow the reader having a complete view of the approach presented in this Thesis, while a deeper explanation would go outside the scope of this work.

A deep theoretical treatment regarding the fundamentals of electrochemical methods, and specifically about voltammetry, can be found in [4, 5].

The basic component of the measurement system is called *potentiostat*. A potentiostat is device which injects current into an Auxiliary Electrode (AUX), closing the current loop via a Working Electrode (WE), in order to impose a known difference of potential between WE and a Reference Electrode (REF). Such voltage has to be measured with a high impedance differential amplifier, to make negligible the current going through REF.

The voltage imposed across WE and REF is supposed to be determined by a function generator placed at the input of the potentiostat. The experimental observable, that is obtained from each measurement session, is the response of the system potentiostat-cell to the excitation signal (voltage) used, that is, the loop current *i*(t).

Better detail about the structure of this device will be discussed in paragraph 3.1.1, where the prototype developed for the experimental setup of this Thesis is presented. A simplified connection diagram, which shows the above mentioned configuration, is reported in Fig. 2.1.

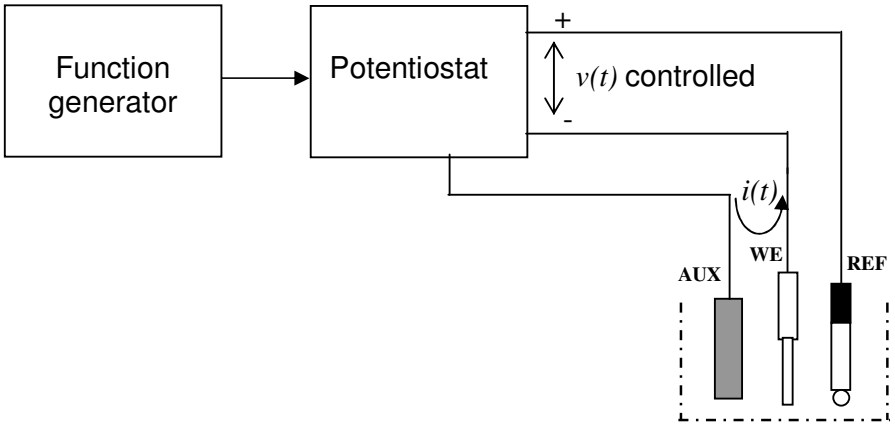

Fig. 2.1 Simplified schematic diagram of the experimental system

2.1.2 The basic step experiment

When the applied excitation signal is a step signal, such as the one shown in Fig. 2.2, the obtained response depends on the levels of the applied potential.

Under the hypothesis that there's only one dissolved species and that it is electrochemically inactive at potential E_1 but reduced at E_2, a typical response of the experimental system would be the one shown in Fig. 2.3a. The electroactive species eliminated (reduced) at the electrode (WE) produces a concentration gradient in the proximity of such electrode. Moreover, looking at equation (2.4), it comes clear that the flux (thus the current) is proportional to the concentration gradient.

The decay of *i(t)* is just due to the fact that the flux causes the depletion volume to thicken. In other words, the slope of the concentration profile $C_o(x)$ from the WE surface declines while current flows, the depletion volume gets higher, producing the described decay of *i(t)*. An example of the concentration profile behaviour vs. time is shown in Fig. 2.3b.

The time-dependent concentration and the shape of *i(t)* can be calculated by solving the diffusion equation under the applicable hypothesis for the problem considered. Details about the electrochemical and mathematical approach to the problem can be found in [4]. Since solutions in a closed form are almost always very difficult to obtain, some hard hypothesis, which give good approximation with a much simpler mathematics, are generally used. A brief overview of the special cases that are usually considered can be found in Paragraph 2.2.1.

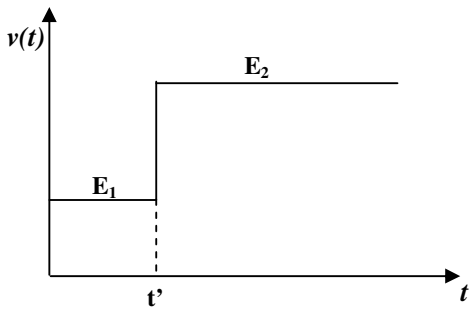

Fig. 2.2 Excitation signal waveform for the basic step experiment

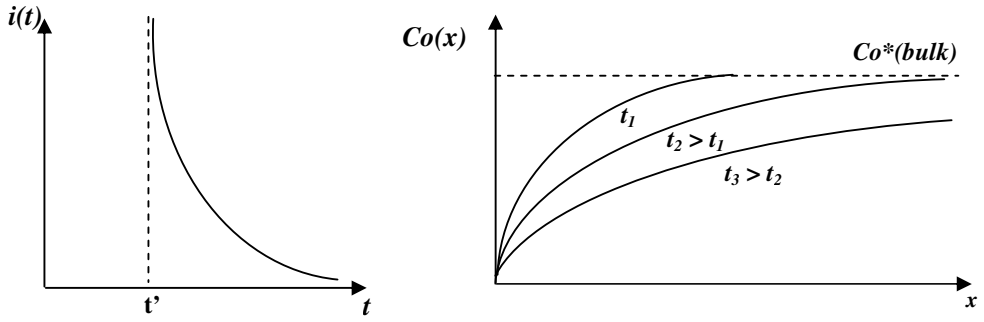

Fig. 2.3 a) *i(t)* for the step experiment. **b) Concentration profiles at various times**

If steps of different amplitudes are applied to the measurement system (Fig. 2.4a), the response *i(t)* may go through the following different conditions, depending on the varying amplitude:

- potentials at which the species is not electroactive, since there's no *overpotential* for reducing,
- potentials at which a flux is present and still incrementing with higher potential levels,
- range of potentials where the current value is clipped by a mass-transfer limitation, thus not incrementing with higher potential levels.

Such range of possible behaviours is clarified in Fig. 2.4b. Supposing that the current is sampled at instant τ, it is possible to plot $i(\tau)$ vs. the correspondent step potential, making what is called a *sampled-current voltammetry*.

The current-potential curve that is obtained (also called *voltammogram*), is shown in Fig. 2.5, which has been obtained by interpolation of seven samples.

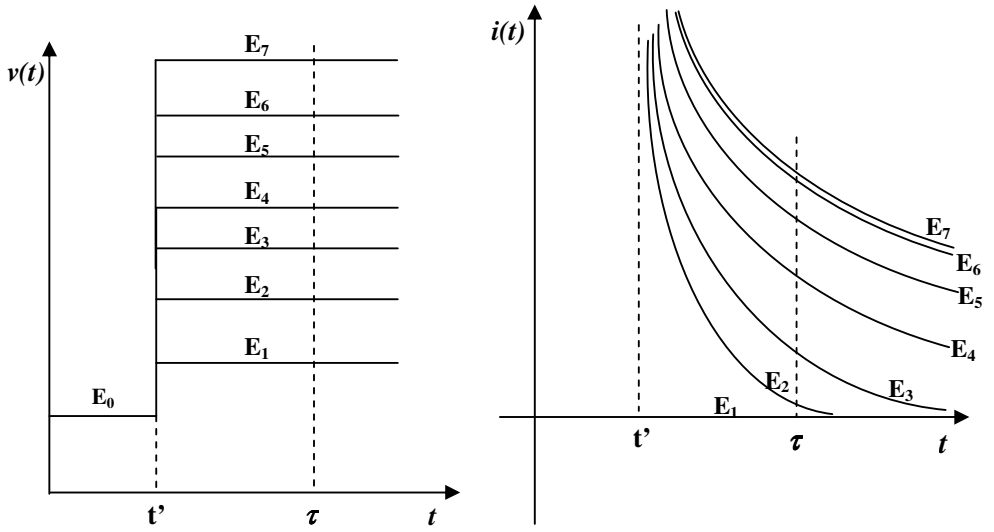

Fig. 2.4 a) Excitation signals for sampled-current voltammetry. b) Current vs. time for the different steps in potential

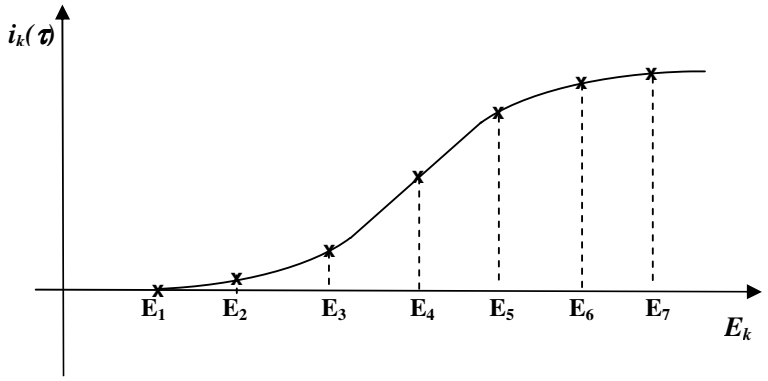

Fig. 2.5 Current-potential curve (voltammogram)

2.1.3 General remarks about voltammetry

The electrochemical behaviour of a system can be determined by imposing known varying potentials between WE and REF, recording the current-time curve that is obtained in each experiment.

The different *activation overpotential* levels, diffusion coefficients and mass-transfer limits that are exhibited by each species in a solution, generate a different contribution to the complex voltammogram obtained; the shape of such contribution depends on the waveform of the excitation signal.

It is also known that the complete electrochemical *signature* of a system can be obtained by applying a series of steps of different amplitude, always recording the current vs. time curve; the approach proposed in this Thesis derives from this concept.

Voltammetry has proven to be one of the most mature and affordable techniques for providing an aggregate information about the quality of water, beverage and food [6, 7, 8, 9, 10]; LAPV (Large Amplitude Pulse Voltammetry) measurements, derived from standard analytical laboratory techniques, has shown excellent sensitivity and discrimination performance for e-tongue applications [7], where, instead of analysing traditional voltammograms, the large multivariate space generated by the time series of the current signal is used to characterise the fluid under measurement, providing an overall quality information.

In fact, voltammetry, among other possible techniques, is one of the most extensively used in analytical chemistry, and appears to be very promising for classification and change detection purposes.

In the following Chapters a general overview of the available voltammetric methods will be given. Particular attention goes to those techniques which make use of large amplitude excitation signals.

2.2 Overview of the most common voltammetric techniques

2.2.1 Current-potential characteristics

As already said in Paragraph 2.1.2, there are a number of special cases according to which the excitation signals are generally classified. This is in order to make the calculation of current-time and current-potential curves easier, under some circumstances that lead to some approximate mathematical solution.

Methods can be classified according to the following special cases:

Large Amplitude Potential Step: current is limited by mass transfer, thus current is independent of potential for activated species and is negligible for the inactivated ones

Small Amplitude Potential Changes: such small variations may be of different shape, the relationship between current and overvoltage is simplified in a linear form

Reversible Electrode Process: when anodic and cathodic effects (oxidation and reduction) are almost simultaneous

Totally Irreversible Electrode Process: when only one of the two effects (oxidation or reduction) is appreciable

Most techniques, as discussed in Paragraph 2.2.2, make use of such approximations.

2.2.2 Chronoamperometry, chronocoulometry, polarography

The kind of experiments that can be made in the known voltammetric methods can be catalogued as follows:

Chronoamperometry: when current is recorded as a function of time in correspondence to a know excitation waveform

Chronocoulometry: integral approach equivalent to chronoamperometry, measuring the amount of charge which passed

Polarography: when samples of the current measurement are taken from a sequence of step experiments and placed on the current-potential plane. The curve obtained is said *polarogram*.

2.2.3 Techniques involving large amplitude excitation signals

This Paragraph presents a brief overview of the mostly used analytical techniques for voltammetry, where large amplitude excitation signals are used. A wider and more detailed explanation of the presented methods can be found in [11].

Linear sweep voltammetry: it is a chronoamperometry where the excitation signal is a ramp. Typical excitation waveform and response signal are shown in Fig. 2.6.

Polarography: it is classified in various versions, such as pure polarography, tast polarography and sweep polarography. Mostly used for quantitative measurements, it is generally used with renewable Working Electrodes, such as the dropping mercury ones (DME). Current values placed in the current-potential curve are the average or plateau (mass-transfer limited) values calculated for each measurement. The obtained curve is a kind of locus of the

calculated values for each current-time response obtained. Results are similar to the one shown in Fig. 2.5 Current-potential curve (voltammogram).

Triangular wave voltammetry: as for Linear Sweep Voltammetry, makes use of triangular excitation signals, eventually in a cyclic form.

Normal pulse polarography: the excitation waveform is made by successive pulses with linearly increasing amplitude. Again, a polarographic current-potential response is obtained, when sampled current is plotted against potential pulse value. Usually applied to dropping Working Electrodes. Excitation signal and a typical response curve are shown in Fig. 2.7.

Double potential step chronoamperometry: a double step pulse as described in Paragraph 2.2.1, with opposite edges. In [11] this technique is distinguished from pure chronoamperometry by the fact that E_1 in Fig. 2.8 differs from the open-circuit potential.

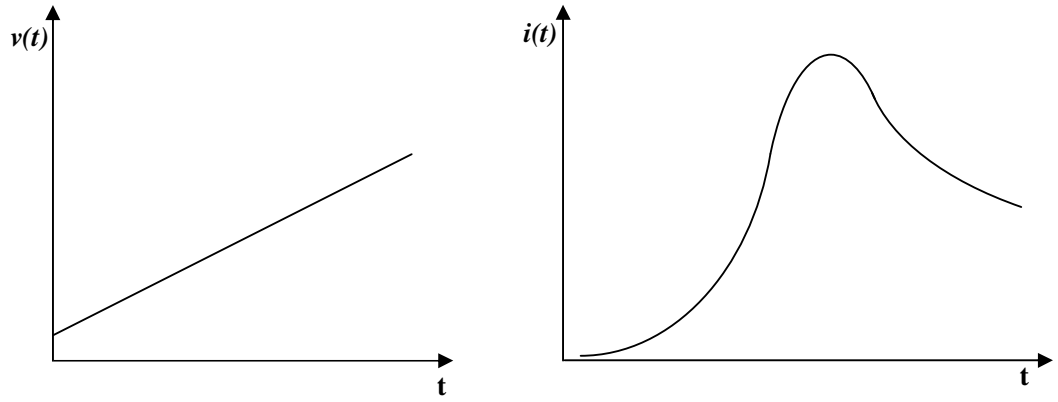

Fig. 2.6 Excitation waveform and typical response for linear sweep voltammetry

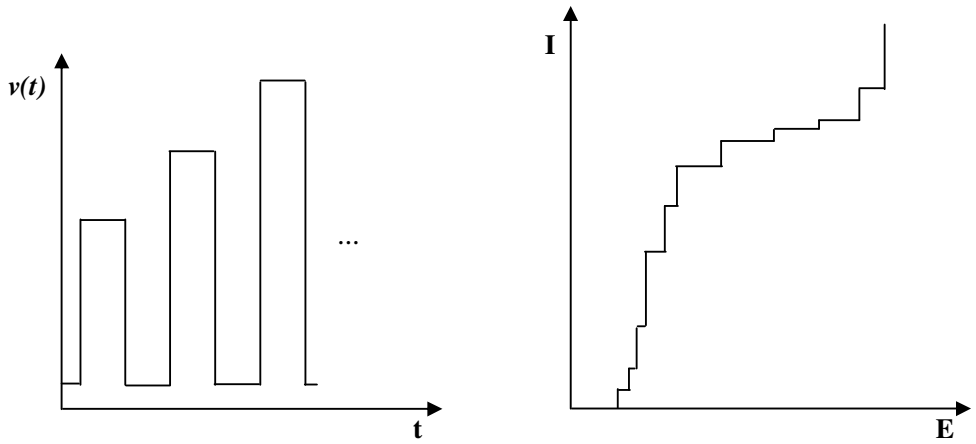

Fig. 2.7 Excitation waveform and typical voltammogram for normal pulse polarography; sampled values are usually represented in a stepwise manner. E is the step amplitude, I the sampled current

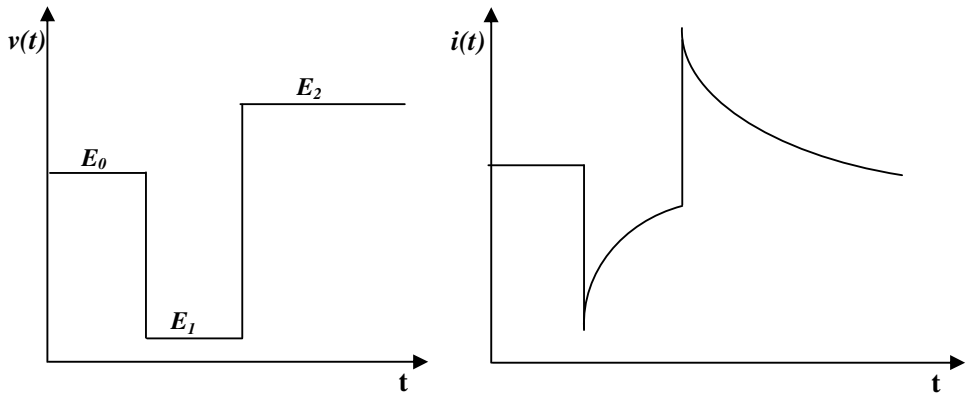

Fig. 2.8 Example of an excitation signal and a current-time curve for a double potential step chronoamperometry

2.3 Signal processing aspects relating to pulse voltammetry

The fundamental idea of this work is to investigate how an adequate signal processing approach applied to a mature and affordable sensor technique (voltammetry) can address the issue of extracting an aggregate chemical information, useful to characterize the liquid under measurement.

The measurement method proposed here is based on the following idea: instead of analysing traditional voltammograms, samples of the current signal are collected, making a kind of chronoamperometry. The shape of the excitation signal and the configuration of the acquisition system are discussed in Paragraph 3.1.1 and 3.1.2.

The large multivariate space, generated by the time series of the current signal acquired, is used to characterise the fluid under measurement; the work is thus focused on the signal analysis techniques useful to make a measurement system capable to:

- detect changes in a continuous monitoring framework,
- perform an automatic classification.

When dealing with data series coming from voltammetric measurements, especially when data come from a high sampling rate pulsed voltammetry, the crucial signal processing issues regard the reduction of dimensionality of the dataset and the extraction of suitable features to enhance the discrimination between different classes.

In fact, the efficiency of the dimensionality reduction algorithms has a beneficial impact on the hardware requirements for a system which has to implement them. The possibility of embedding the computational power in a low-cost monitoring device, strongly depends on how

efficiently datasets having large dimensionality can be reduced with an acceptable loss of information.

Consequently, one important aspect is to study techniques to extract useful information from the large multivariate space generated by such measurements.

This task can be accomplished in two steps:

- dimensionality reduction
- feature extraction.

Both classification and change detection applications require optimising the performance of the data processing system in terms of discrimination capability.

More in detail, both the hardware and the software of the equipment have to be designed taking into account the following basic requirements:

- easy implementation on embedded devices having relatively poor resources,
- light computational load,
- high discrimination performance.

The efficiency of the dimensionality reduction and feature extraction algorithm becomes an essential element in the development of such devices, according to the said guidelines.

Thus, one essential aim of this work is to define a classification system that can learn by a suitable training set (eventually acquired in the laboratory) to be used on the field for successive measurements after the tuning process; both electrolytic solutions made in the laboratory and water samples coming from the real world can be used for this purpose. In fact, depending on the target application, different approaches can be used in building up the training set, such as using water samples belonging to different chemical classes for characterisation purposes, or simulating different concentration levels of known pollutants, for change detection and surveillance uses.

Once the classification system has been trained, it is then ready to be used for characterising unknown samples in the successive measurements. This approach is described in 3.2.1, going into further detail in 3.5, finally an example is presented in 4.3.

Measurement repeatability is a fundamental aspect in the development of such an approach, and it involves the design of the measurement cell, the analog front-end and the acquisition strategy as well.

2.4 Signal processing approach

In this Paragraph, a brief overview about the signal processing approaches proposed up to now for pulse chronoamperometric data series, is given. There's actually not much literature of this kind, essentially because most of the research is focused on innovative sensor techniques, more then better exploiting known ones, with innovative signal processing concepts. Moreover, comparisons with respect to the known literature will be found along the discussion of the proposed signal analysis methodology, in Chapter III.

Most of the efforts are essentially focused on the aspects regarding the reduction of dimensionality of the collected data.

The main motivations for dimensionality reduction can be listed as follows:

- to reduce the complexity of classifiers,
- to improve classification accuracy with respect to known problems, such as Hughes effect [12, 13] due to insufficient number of observations, or poor generalisation,
- to reduce the computational cost of measurement extraction
- to better assess class separability.

In [6, 7, 8] PCA (Principal Component Analysis) is applied to the time-domain samples, where each sample is represented by a component of the observed vector; the dimensionality of the space may be limited before applying PCA, by taking selected samples from the time domain signals; in [7] a factor 30 down-sampling from the original 900Hz sampling rate is proposed, but no criteria for selecting a sub-set of optimal samples is given.

In general, the approach based on PCA implies quite a high computational load. In fact, for spaces having large dimensionality PCA is very intensive computationally; alternative techniques to determine a suitable representation space can be used, as discussed in Paragraph 3.4.

In [9] a Wavelet transform is applied to a high sampling rate time series, to reduce the dataset and to optimize both change detection (monitoring) and classification (discrimination) capabilities. The proposed data processing chain, applicable when a training set of known observations is available, is constituted by:

- Discrete Wavelet Transform (DWT) of the acquired data,
- sort of the transformed coefficients according to the following uni-variate discrimination measure:

$$\frac{V_{between}}{V_{within}} = \frac{\left(\frac{1}{q-1}\right)\cdot\sum_{i=1}^{q}(m_i - m_{mean})^2}{\frac{\sum_{i=1}^{q}(n_i-1)\cdot V_i}{\sum_{i=1}^{q}(n_i - q)}} \qquad \textbf{(2.5)}$$

Where:

q is the number of classes,

m_i is the mean of the selected coefficient in class *i*,

m_{mean} is the mean of the selected coefficient for the total of the observations,

n_i is the number of observations in class *i*,

V_i is the variance of the observations in class *i*,

- selection of the first L sorted coefficients according to their discrimination value,
- select the subset of coefficients (if it exists), which maximizes the Mahalanobis distance between clusters,
- apply PCA to said subset and take the first two Principal Components as a representation space.

This processing scheme is prone to the following considerations:

- there's no strategy to determine an optimal value for L,
- in practical experiments L is taken so large (order of thousands) that subset selection becomes slow,
- the first L sorted coefficients are not the best L coefficients. An approach closer to exhaustive optimality can be considered.

The signal processing scheme proposed in this Thesis, among other peculiarities, aims at finding an optimal (or sub-optimal) dimensionality of the feature space, by using sub-optimal search strategies that do not necessarily group coefficients sorted according to an uni-variate discriminability measure.

A review of data compression methods aimed at saving memory space and computational time, essentially based on the Wavelet transform and Hierarchical PCA is presented in [14].

Two different experiments are presented in the paper, both making use of pulse chronoamperometry, made sequentially with four different Working Electrodes. Aim of such experiments is to compare the compaction performance and class discriminability that it is possible to obtain with different approaches.

To reduce the datasets derived from voltammetric e-tongue measurements, the following three procedures have been experimented:

- Two-step HPCA, first on data coming from each WE, then on the combined data,
- Selection of Wavelet transformed samples according to their variance,
- PCA applied to electrochemical parameters extracted from time domain samples (time constant, Faradaic current, peak current).

The weak aspects of this approach can be individuated in:

- data selection according to variance does not imply a direct estimation of discriminability,
- as a consequence, Wavelet transformed data seem to have less performance then the PCA of the electrochemical parameters,
- only information about discrimination capability is given, no classification framework is presented.

In conclusion, especially when developing a device for field use, there's still the need to make further developments on specific methodologies to:

- save computational time and data storage,

- optimise the discrimination performance,
- develop dedicated supervised classification schemes,
- develop affordable change detection algorithms.

REFERENCES FOR CHAPTER II

[1] J. A. V. Butler, Proc. Royal Society, London, 112A, pg. 129, 1926

[2] E. A. Guggenheim, J. Phys. Chem., 33, pg. 842, 1929

[3] E. A. Guggenheim, J. Phys. Chem., 34, pg. 1540, 1930

[4] A. J. Bard, L. R. Faulkner, "Electrochemical methods, fundamentals and applications", John Wiley & Sons, 1980.

[5] G. Korthum, Treatise on electrochemistry, 1965

[6] M. Lindquist, P. Wide, "Virtual water quality tests with an electronic tongue", IEEE IMTC Budapest (Hungary), May 21-23, 2001

[7] C. Krantz-Rulcker, M. Stenberg, F. Winquist, I. Lundstrom, "Electronic tongues for environmental monitoring based on sensor arrays and pattern recognition: a review", Analytica chimica acta, 426 (2001), 217-226

[8] F. Winquist, P. Wide, I. Lundstrom, "An electronic tongue based on voltammetry", Analytica chimica acta, 357 (1997), 21-31

[9] T. Artursson, M. Holmberg, "Wavelet transform of electronic tongue data", Sensors and actuators B, 87 (2002), 379-391

[10] L. Robertsson, P. Wide, "Improving food quality analysis using a wavelet method for feature extraction", IEEE IMTC Ottawa (Canada), May 16-19, 2005

[11] J. Inezedy, T. Lengyel, A. M. Ure, "Compendium of analytical nomenclature", IUPAC analytical chemistry division, 3rd edition, 1997

[12] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, Vol. IT-14, No. 1, January 1968.

[13] L. Jimenez, D. Landgrebe, "Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate Data", IEEE Transactions on Geoscience and Remote Sensing,Vol 37, No. 6, November 1999

[14] S. Holmin, P. Spangeus, C. Krantz-Rulcker, and F. Winquist, "Compression of electronic tongue data based on voltammetry - a comparative study," Sensors and Actuators, vol. B 76, pp. 455–464, 2001

CHAPTER III

A novel approach to water characterisation based on voltammetry

The design of devices based on the principles illustrated in this Thesis requires a specific methodological approach, such as the one described in this Chapter. A description of the experimental setup and details about the proposed signal processing methodology are given here.

3.1 The prototype developed

3.1.1 Hardware

The equipment used to generate the datasets discussed in this Thesis has been custom designed and built, including the potentiostatic device; this makes possible to take advantage of the many analytical possibilities that voltammetry offers by modifying key working parameters of the instrument (i.e. number and type of the electrodes, shape, duration and amplitude of the excitation waveform). A simplified block diagram of the device is shown in Fig. 3.1.

Both the synthesis of the excitation signal and the collection of the current measurements are controlled by a Personal Computer (PC), where a commercial ISA card (Axiom model AX5621H+ [1]) provides all the necessary I/O (ADC, DAC and digital signals) plus an independent XTAL rate generator used as a timebase for the whole system, separated from the PC system clock.

Said ISA card hosts a SAR A/D converter with S/H stage, 12 bit ADC, 8 TTL compatible digital outputs and an 8254-like configurable counter used as timebase generator; the main specifications of such devices are found in Tab. 3.1.

Parameter	Typ. @25°C	Unit
SAR A/D resolution	16	bit
No missing codes	15	bit
S/H aperture delay	40	ns
A/D conversion time	8	µs
D/A resolution	12	bit
D/A linearity error	±1/4	LSB
D/A full scale range change settling time	4.5	µs

Tab. 3.1 Main specifications of the A/D and D/A devices used

With the configuration of Fig. 3.1 the potential on the working electrode is tied to the analog ground by the I/V converter (1); the differential electrometer (2) is connected to the negative input of the control amplifier (3) and closes the feedback loop, which ensures that the

voltage difference between the reference electrode and the selected working one is equal to the voltage set by the PC via the A/D converter. All this is true until both the voltage assumed by the Counter-Electrode (CE, also said Auxiliary Electrode) and the current which flows through it, lie in the output swing capability of (3).

One or more relays, directly controlled by the PC card digital I/O lines, select the proper working electrode for each phase of the acquisition process.

The current flowing through the selected Working Electrode (WE) is converted into a voltage and acquired via the A/D stage; collected data are stored by the PC and form one sampling sequence, such as those shown in Fig. 3.4.

The current signal conditioning circuit has to handle current levels down to the nA; it is based on a DIFET (dielectrically insulated FET) OPAMP (OPA129B), selected for its low bias current and small offset drift and noise, suitable for the application (see the data sheet [14]). A block schematic, that gives better detail of such stage, is reported in Fig. 3.2. To better exploit the dynamic range of the converted signal, two identical stages with different conversion resistors ($R_{I1,2}$) are selected via a low loss switch (SW2, ADG432B), to get the maximum unsaturated signal between the two paths; to even reduce switch and layout current losses at the input, SW1 has been chosen to be a manual mechanical switch. Full scale signal levels determine the switches' position, and are dependent on the nature of the water under measurement. Cc capacitors are placed to reduce the consequences of high frequency noise picked up by the cabling and the electrodes, such as switches' crosstalk and OPAMPs' oscillation and distortion. Their band limiting effect is neglectible with respect to the following filter, being their pole higher than 1KHz.

To improve the frequency response and linearity of the front end, gain is partially allocated to the later stage, after filtering with a first order filter. The cut-off frequency of said filter, as indicated in Fig. 3.2, is the one chosen for the experiments described in this Thesis; experience has shown that with the excitation signal used, there's no significant improvement in cutting at higher frequencies; only the usage of smaller electrodes and of faster excitation signals would imply the need for a wider bandwidth.

The four Working Electrodes are made of different solid metals, having different redox properties. Various metals and different quantities of WEs have been presented in the literature. As an example, in [8] the Working Electrodes are four, while in [9] a probe with six WEs has been used.

A first order filter at the input of the control amplifier (3) is necessary to prevent electrode-fluid system nonlinearities from inducing self oscillations in the potentiostat loop, when fast signals are applied at its input. Depending on the electrolyte under measure, the cut-off frequency may range between 5 and 50 Hz for the configuration used.

The shape of the excitation signal is programmable, via the implementation of a virtual waveform generator in the software which runs on the controlling PC. The voltammetry control software has been completely custom developed for this work; it is a DOS application written in PowerBasic language, suitable to run on any x86 platform. Signal analysis software has been developed in MATLAB environment and runs on a separate machine. Data between the two PCs are transferred through their parallel ports.

The main guidelines that have driven the design of the hardware are: sampling rate stability, small S/H aperture time (in the A/D converter), precision in the phase relationship between the excitation signal steps and the sampling times. This was to ensure a proper repeatability of the data sequences obtained, in the effort of better evaluating the subsequent signal processing results.

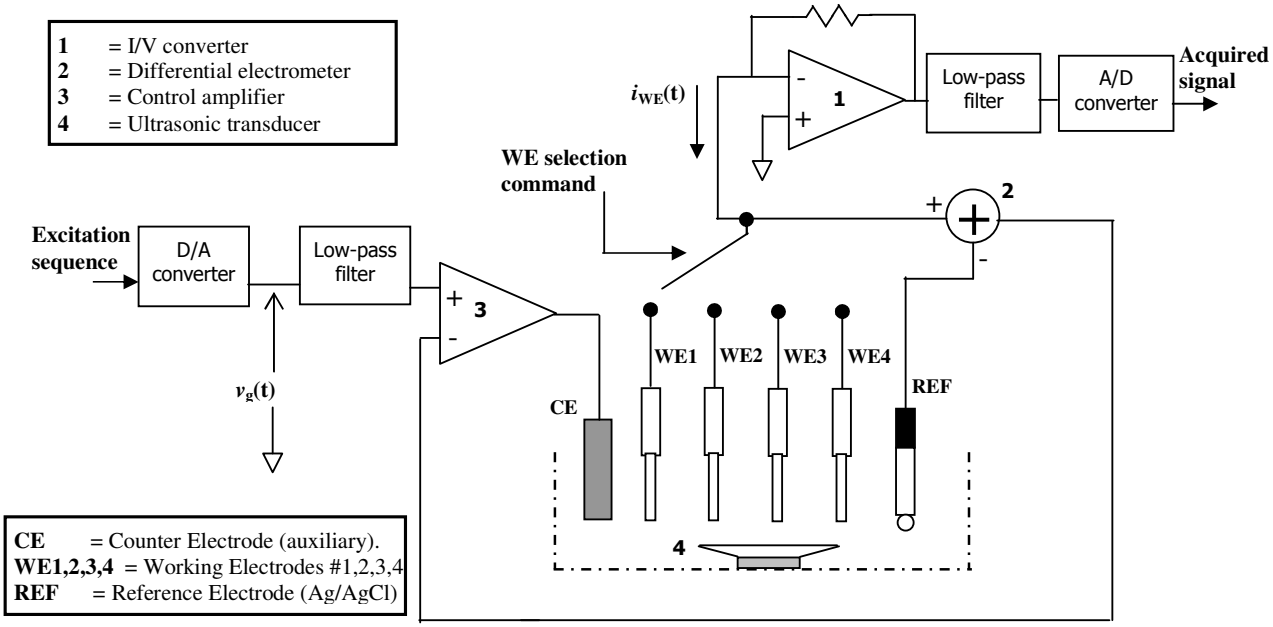

Fig. 3.1 Simplified block diagram of the test set used

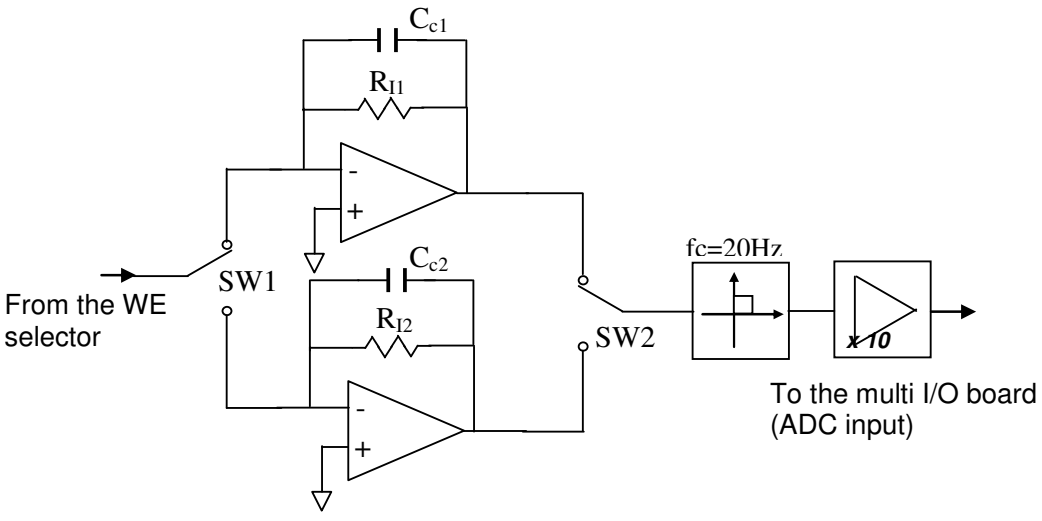

Fig. 3.2 Simplified block diagram of the current signal conditioning circuit

3.1.2 Sensor system

The excitation signal used in the presented experiments corresponds to the Large Amplitude Pulse Voltammetry (LAPV) method, and the waveshape is the one proposed in [10]. The signal is made by successive pulses of varying amplitude and constant duration (470 ms), starting from +900 mV and stopping at –130 mV, with 71 amplitude steps and a time interval between pulses equal to the pulse duration (470 ms). The excitation function, taken as the instantaneous voltage $v_g(t)$ applied to the input of the low-pass filter, is shown in Fig. 3.3.

The acquired data is constituted by the collected sequence of samples of the current signal measured by the I/V converter, thus making a kind of chronoamperometry generated by a sequence of pulse excitations.

In a first version of the sensor system, the Working Electrodes were two wires made of Au and Re, about 5mm long; the Reference Electrode was an Ag/AgCl one, while the Counter-Electrode was a stainless steel plate having an area of about 2 cm^2. The first preliminary experiments have been made in this configuration.

In a more recent version, the measurement cell comprises four Working Electrodes, made of Au, Re, Pt and Pd; all of them are encased in a Plexiglas tube filled by an epoxy sealant. The wet surface of each electrode has an area of 0.2 mm^2 and it is flattened to make a uniform surface with the epoxy compound.

The usage of Working Electrodes having a small wet surface (even microelectrodes) allows a better performance than by using traditional ones, thanks to their lower capacitance and a better exploitation of the diffusion effect. This implies that the analog front-end must handle smaller currents than with larger electrodes, with the same potentiostat bandwidth, and is prone to higher noise pick-up.

Each water sample under measurement is continuously stirred with a magnetic stirrer inside the measurement cell, during all the time of the measurement session. The volume of water contained in the cell is about 50 ml.

Different cleaning techniques have been experimented: chemical cleaning with H_2SO_4 1M, followed by rinsing in distilled water, has been performed in the first experiments between any two measurements of liquids belonging to different classes.

In the most updated version, the measurement cell is completed by an ultrasonic transducer placed at its bottom, used for the periodic washing of the wet parts, without making use of any aggressive chemicals or direct mechanical scrubbing, thus preventing any modifications to the electrodes' surface which may somewhat alter the repeatability of the measurements. Thanks to that, ultrasonic washing appears to be a useful procedure for long-term field use. Chemical cleaning is now used in a sporadic way, especially after long periods of inactivity.

Fig. 3.5 shows the configuration of the electrodes on their mounting probe, which has been designed and built for the measurement cell.

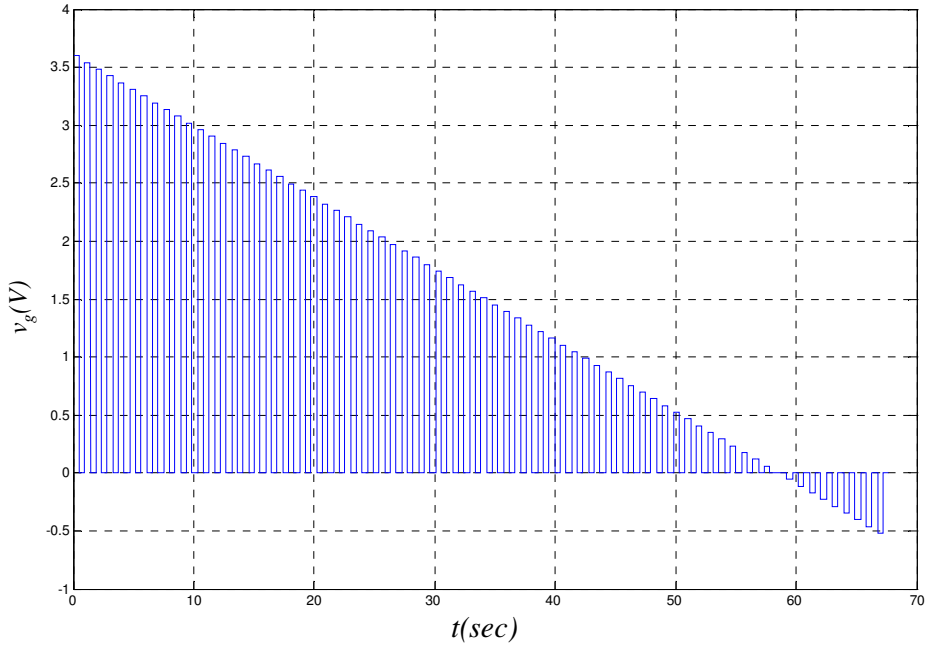

Fig. 3.3 Waveshape of the excitation signal used

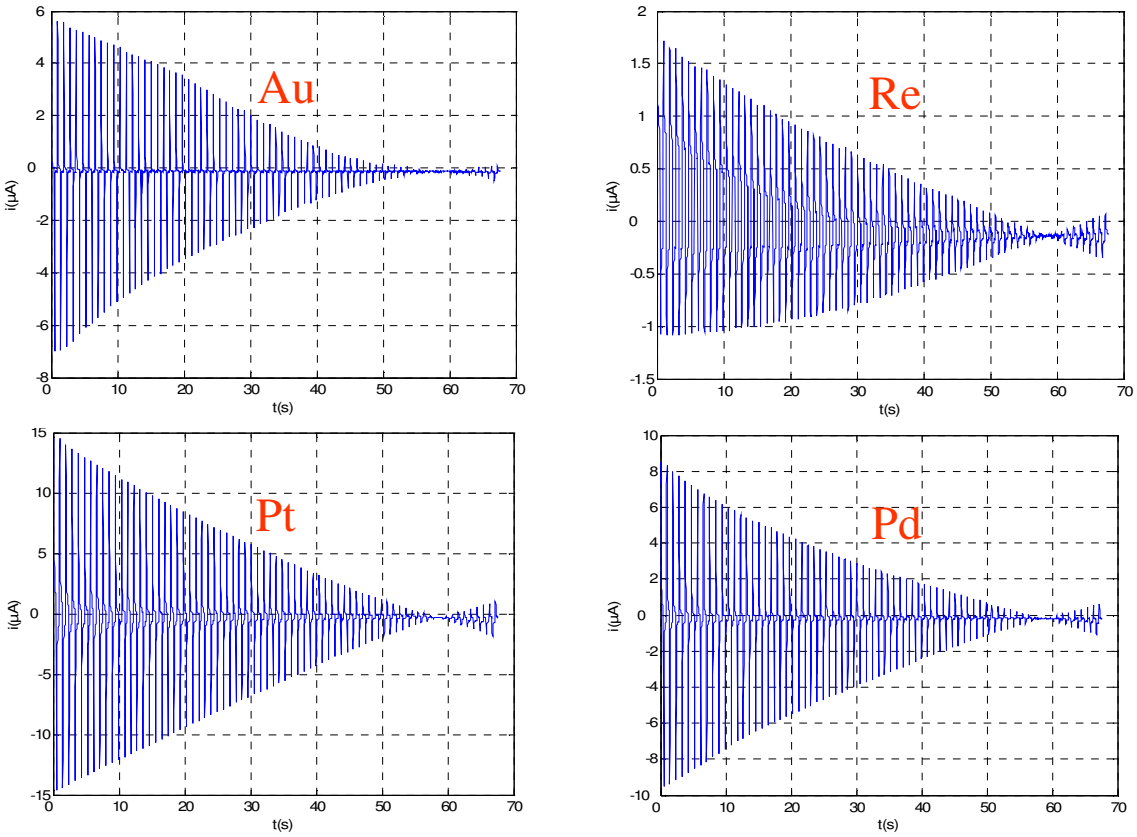

Fig. 3.4 An example of the acquired signals. The correspondent WE material is shown for each of the time series

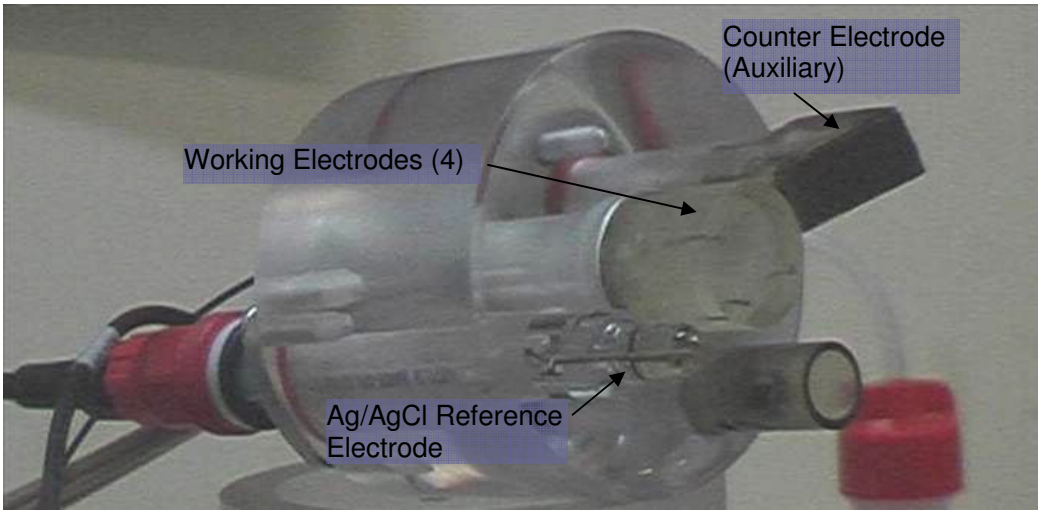

Fig. 3.5 Electrodes' mounting probe used for the experiments

3.1.3 Software

As already mentioned in Paragraph 3.1.1, the shape of the excitation signal is programmable, via the implementation of a virtual waveform generator in the software which runs on the local (controlling) PC. The voltammetry control software has been developed in two versions:

- a DOS application written in PowerBasic language, as a first version,
- a Windows CE 5.0 application written in C++, suitable to run on any x86 platform.

Signal analysis software has been developed in MATLAB environment and runs on a separate machine. Data between the two PCs have been transferred through their parallel ports, due to the availability of very affordable and fast enough applications for the interconnection between DOS and Windows platforms.

DOS for the acquisition PC in the experimental setup has been chosen for its complete absence of time critical issues due to the underlying Operating System; mere experiments without special needs of user interface and connectivity are very easy to handle, in this way. In addition, the development of custom low-level drivers for the direct access to the I/O space is immediate, as well as porting the drivers to microcontroller environments (i.e. simple machines that perform only the acquisition task, sending raw data remotely).

The described solution used for the experimental setup is now being ported to a Windows Mobile (CE 5.0) platform, in the view of providing the field device in its final version with the typical user interface and connectivity (i.e. TCP/IP, FTP, PPP), that are generally available nowadays. In this case, also the MATLAB software, re-written for the Windows CE environment in C++ language, is intended to run on the same target machine. Also, the hardware of the final field version differs from the experimental setup, especially in the acquisition front-end, but it keeps an identical functionality thanks to the high availability of register-compatible devices in the embedded market, which are compliant with the Keithley DAS-16 standard.

In fact, the components of the experimental setup and the software architecture are determined in the view of the development of a portable field instrument with fully embedded and autonomous functionalities.

The functional structure of the software embedded in a portable prototype must take into account two general application frameworks:

- stand-alone operation for continuous monitoring,
- sporadic laboratory or field use for characterization and supervised classification.

Fig. 3.6 shows a complete schema of the functionalities of the prototype.

The embedded device is thought to be connected to a local or remote workstation PC. The first step is therefore to determine the functionalities of the embedded device. These are: acquiring data (continuously or not), signalling alarms upon detection of changes, providing remote connectivity to retrieve data, managing a communication protocol to update working parameters and excitation functions and, finally, visualizing on a local display a representation of the acquired and processed time series.

The second step is to determine which operations have to be carried out by the workstation PC, which implies the implementation of the entire communication protocol and user interface that are not required at the development stage, but are necessary at a certain degree of availability in order to perform field applications. Nevertheless, a more detailed description of the application software framework goes outside the scope of this Thesis.

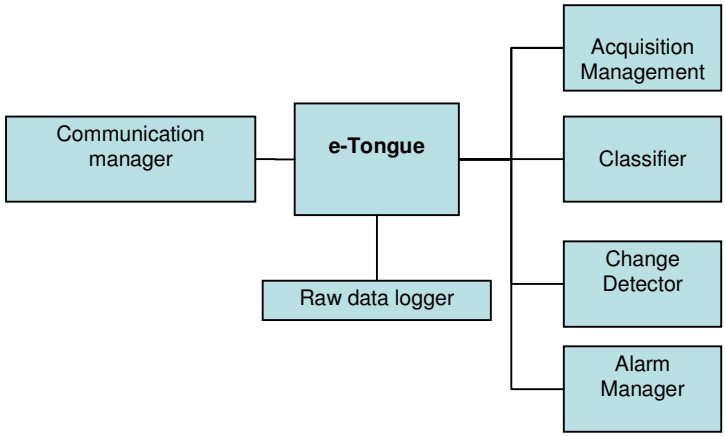

Fig. 3.6 Complete outline of the functionalities of the embedded prototype

Inside the "e-Tongue" module shown in Fig. 3.6, there are the functionalities of the control software that implements the algorithm for the voltammetric measurement.

Since the sampling interval is set to 1 ms, suitable sampling rate stability and synchronisation between pulse steps and sampling times have been obtained using a monolithic (uninterruptible) software routine that sets the D/A and reads the A/D, each time the sampling

interval elapses; this routine is triggered by the down counter contained in the multi I/O board, which is scaled down from an independent 10MHz source. In such a way, software delays are kept far smaller than the D/A settling time (see Tab. 3.1), which is more than two orders of magnitude below the sampling interval. Also, the time interval between D/A settling and A/D acquisition becomes highly repeatable with respect to the timings of the application, being essentially due to D/A settling time fluctuations. The obtained performance exceeds the potentiostat bandwidth, which is of the order of the tens of Hz, for the experiments presented.

Such an approach is made possible with specifically written drivers which have direct access to the I/O address space of the PC, where the registers' space of the AX5621H+ is statically located.

3.2 Structure of the signal processing chain

3.2.1 General description

Both the sampling rate of the acquired signal (current signal from the I/V converter) and the update rate of the excitation signal have been set to 1KHz, to allow for analyzing the useful bandwidth all over the range of the tested classes of liquid, sensors' geometry and measurement conditions. Also, the sampling rate cannot be straightly determined by the filters' bandwidth of the analog front-end, due to the nonlinear behaviour of the system potentiostat-cell.

The general structure of the signal processing chain is shown in Fig. 3.7. The excitation waveshape coming from the 12 bit DAC drives the potentiostat, already illustrated in Fig. 3.1. The resulting time series of the current signal obtained by the chronoamperometric measurements goes through a pre-processing step, consisting into a pre-filtering and down-sampling stage, as shown in Fig. 3.7.

The pre-filtering stage consists into a median filter with variable taps, to remove eventual glitches picked up by the potentiostat system, treating them as outliers in the filtering window; instead, average filtering is already implemented in the hardware with the illustrated low-pass filters.

In the experiments presented in this Thesis, the acquired sequence has been down-sampled to a rate of 50Hz, to obtain a pre-reduction of dimensionality adequate to the band of the system, estimated in the experiments that have been done.

Each measurement sequence per each one of the Working Electrodes (WEs) is made by M=3384 data points, after down-sampling. Thus, one observation is formed by four sequences of M data points, each one generated by the LAPV scan for each of the four WEs.

The dimensionality reduction and feature selection process go through a transformation into the frequency domain, according to the following steps:

- the data series obtained with different WEs are collected into n sequences made of M data points, where n is the number of WEs,
- DCT (Discrete Cosine Transform) or DWT (Discrete Wavelet Transform) of each sequence is calculated,
- the most informative (according to some criteria) transformed components are selected (feature selection).

The proposed approach [2, 3] resorts to the frequency domain transform of the data series in order to exploit the energy compression properties of the DCT, which has never been proposed in literature for this kind of application.

The use of the DCT is also motivated by the fact that such a transform can be efficiently performed with commercial hardware devices. On the other hand, the proposed usage of DWT for voltammetric time series [4, 5] suggests the experimentation of both techniques, in order to compare their performances and to determine whether the DCT can be fruitfully used in this framework. Besides, the feature selection procedure can also be applied to the original time-domain data (VLT), to compare results with those obtained in a transformed domain. Such comparative results are shown and discussed in Chapter 0.

Fundamentals of DCT are explained in many books and articles about signal and image processing; a good reference can be found in [6].

The concept that lies behind the usage of a transformation domain is to work in a domain, if it exists, where the same separability between classes is achieved with a smaller amount of features, with respect to the original domain of the collected signals.

Feature selection represents a critical issue in multidimensional data processing having the dual objective of delivering the most accurate answers with a minimum amount of data and with the maximum computational efficiency; it is thus fundamental to design suitable algorithms having optimal performance under this point of view. The feature selection consists into a two steps (not necessarily distinct) procedure, where the optimum number of data components and the best subset of them are determined.

The approach presented here consists into defining an index that provides a quantitative assessment of the effectiveness of a given subset of features, in terms of its ability to discriminate between different classes of liquid. Then, an exhaustive analysis of the feature space can be accomplished by varying the number of data components and by considering all the possible combinations between them, in order to find the optimal selected set.

The best feature subset can be obtained analytically by solving the optimization problem based on the defined index. Nevertheless, in most cases, a solution in a closed form is not achievable and a numerical approach becomes the only feasible one. In addition to that, a sub-optimal approach may be the only practically acceptable, due to the high computational time that an exhaustive (combinatorial) procedure may require.

Selected features (i.e. DCT, DWT coefficients or VLT samples) can be forwarded to three possible processing paths:

- an automatic classifier
- a change detection procedure
- a display procedure based on a further reduction of dimensionality.

All the three processing possibilities are explored in the next Paragraphs, particularly focusing on supervised classification systems and the individuation of a suitable representation sub-space. As to change detection methods, they are essentially based on the discrimination capability of data fed to the classifier, applying a suitable metric to assert the distance between observations taken and some reference vector.

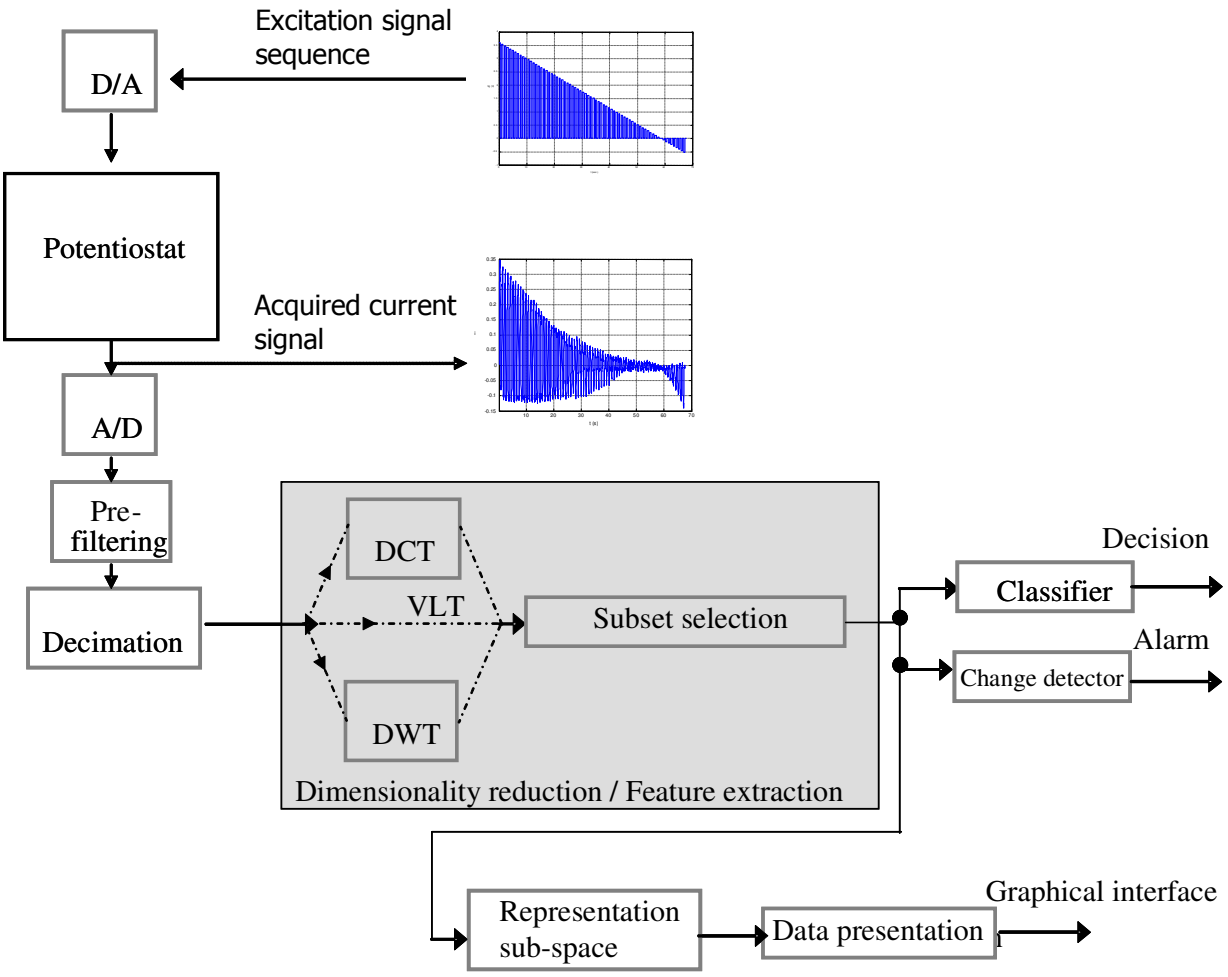

Fig. 3.7 General structure of the signal acquisition and processing chain

3.3 Study of sub-optimal strategies for feature selection

As already stated, neither an automatic classifier nor a change detector work directly on the M-dimensional measurements' representation space, but on a reduced one. A specific methodology is defined here, to determine the data representation space according to the two conflicting requirements characterising the application of interest:

- high discrimination capability,
- low dimensionality.

Such algorithm is based on a two steps procedure. First, a supervised feature selection strategy, based on K training samples, is applied. Then, a preliminary dimensionality reduction

of the feature space is done, by selecting the best subset of the DCT components for classification purposes (again it is clarified that the same procedures have been tested with DWT and VLT components, as well). Then, a further reduction is achieved by searching for the transformed training samples, spanning the feature subspace, which preserve most of the data information content. This last step may also be not applied, and is mainly oriented to generate 2D-3D data representations or to feed classifiers having input vectors with fixed dimensionality.

As concerns feature selection, it can be summarised as follows:

- the discrimination capability of each DCT component is quantitatively assessed by computing a suitable class separability index on the *K* training samples,
- the DCT components are sorted according to a decreasing order of their discrimination capability,
- a selection strategy is applied, to determine the optimal dimensionality for the feature space (*l'*), thus selecting the *l'* best components to generate the feature space.

3.3.1 Definition of a discriminability index

As stated in the previous paragraph, we are interested into selecting the subset of DCT components characterized by the maximum discrimination capability. As a consequence, it is important to assess a metric with respect to which such discriminability is evaluated.

Initially, it must be supposed that each time series obtained by the chronoamperometric measurements is equivalent to a vector having dimensionality M, whose components correspond to the samples taken in the time domain. Such vectors, from now on, will be called *observations* or *observed vectors*.

Hence, it is necessary to define an index that quantifies the separability between classes, for each subset of DCT components having dimension *l*. That is, such an index must provide a measurement of class separability on a generic dataset, formed by vectors obtained by grouping *l* components of the observations in the transformed domain, selected according to some criteria that will be seen later.

At a first attempt, said class separability index can be defined as:

$$S(l)=\frac{\sigma_W^2(l)}{D^2(l)} \qquad l\in[1,L] \tag{3.1}$$

where:

$$\sigma_W^2(l)=\frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N_j}\left[\mathbf{v}_i^j(l)-\mathbf{c}^j(l)\right]^T\left[\mathbf{v}_i^j(l)-\mathbf{c}^j(l)\right] \tag{3.2}$$

$$D^2(l)=\frac{1}{C}\sum_{j=1}^{C}\left[\mathbf{c}^j(l)-\mathbf{c}(l)\right]^T\left[\mathbf{c}^j(l)-\mathbf{c}(l)\right] \tag{3.3}$$

and:

L=nM is the whole dimensionality of the observations made with the *n* WEs,

l is the number of DCT components taken,
N is the total number of observations,
C the number of classes,
Nj is the number of observations belonging to class j,
$\mathbf{v}_i^j(l)$ the i-th transformed observation vector belonging to class j,
$\mathbf{c}^j(l)$ the centroid of class j,
$\mathbf{c}(l)$ the centroid of the whole observations' set.

In this approach, the target is the minimisation of the ratio between the two measures (a within-class distance and a between-classes distance). This is equivalent of saying that a specific algorithm must determine:

- the l-dimensional combination of the observations' components that minimizes the separability index for each value of l
- the value of l that gives the absolute minimum for S.

Since dimensionality of the observations' space is nM (where n is the number of WEs), in the practical case proposed in this Thesis this corresponds to nM=13536 components.

The combinatorial exploration of all the possible feature sets would be very intensive, under the computational point of view, if made exhaustively (i.e. for each group of combined l coefficients and each value of l up to nM). The usage of sub-optimal techniques for selecting a suitable feature set is discussed in Paragraph 3.3.2.

One natural approach to assess the efficacy of a discrimination index is to analyse its behaviour in a mono-dimensional space; that is, evaluating the uni-variate index $S(1)$ to determine how the score given to each component is reflected by its actual ability to discriminate between different classes.

Under this point of view, the index defined in (3.1), (3.2) and (3.3) is prone to the superposition of data pertaining to different classes, particularly when one class has a particularly "far" centroid with respect to the remaining classes of the dataset.

An example of such a situation is presented here, obtained by the following procedure:

- a dataset made of 40 measurements (4 classes, 10 measurements for each class) has been generated (experiment "long01", details can be found in Paragraph 0),
- DCT has been applied to the observation vectors made by M=3384 time domain samples,
- the n=4 observation vectors obtained by each measurement are appended in the transformed domain, so to form a set of 40 vectors having dimensionality L=13536,
- the uni-variate discriminability index $S(1)$ is calculated for each of the L components,
- the L components are sorted in reverse order according to the said index (the best in discriminability comes first),
- calculated values for $S(1)$ are sorted in increasing order and plotted.

The best components according to said index give results as shown in Fig. 3.8 and Fig. 3.9.

The "stem plot" in Fig. 3.8 shows the first component in ranking. On the x-axis the observation number is reported, that is, the progressive number of the measurement that generated the set of *n* observed vectors whose first DCT component in ranking is plotted. The 10 measurements belonging to each class are placed in a consecutively, independently on the measurement order; as a consequence, classes are clearly distinguishable, so the efficacy of the component selected according to this metric, to discriminate between classes, is clear.

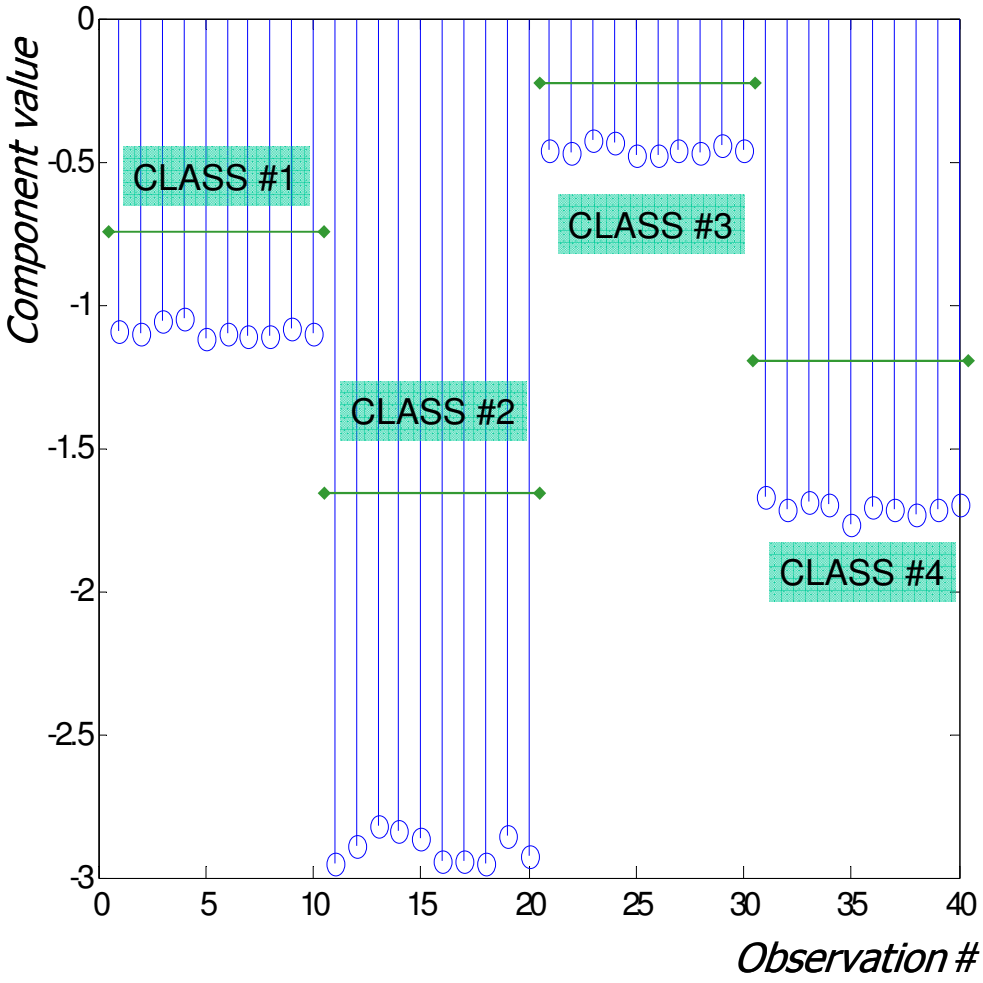

Fig. 3.8 Uni-variate discrimination capability for the 1st DCT component in ranking. Experiment: "long01".

Fig. 3.9 shows a comparison between the first three components in ranking according to the defined metrics. It is apparent how the second one is affected by a substantial superposition of classes (classes numbered 1 and 4). The averaged distance calculation found in

(3.3) may lead to an high ranking of components where one class is far from the others' centroids with respect to the distances between the others. This may lead to give an high rank to components which are somewhat poor in discriminability.

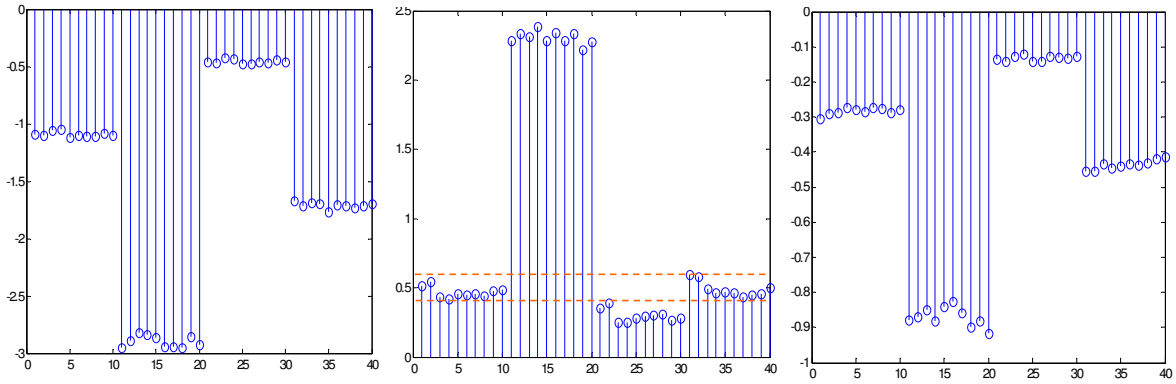

Fig. 3.9 Uni-variate discrimination capability for the 1st, 2nd and 3rd DCT components in ranking. Index S_2. Experiment: "long01".

A more efficient index under this point of view can be obtained according to the following modification:

$$D^2(l) = \min\left(\sum_{\substack{j,k=1 \\ j \neq k}}^{C} \left[\mathbf{c}^j(l) - \mathbf{c}^k(l)\right]^T \left[\mathbf{c}^j(l) - \mathbf{c}^k(l)\right] \right) \tag{3.4}$$

Where (3.4) is used in (3.1) instead of (3.3).

From this moment, the index obtained by the revised metric will be indicated as $S_x(l)$, while the index defined by equations (3.1), (3.2) and (3.3) will be labelled as $S_2(l)$. Notation $S(l)$ will be used when referring to a generic index.

Results obtained by using said modified index $S_x(l)$ are shown in Fig. 3.10. The 2nd component in ranking is now apparently efficient in discriminability. $S_x(l)$ will be the preferred index for the application discussed in this Thesis.

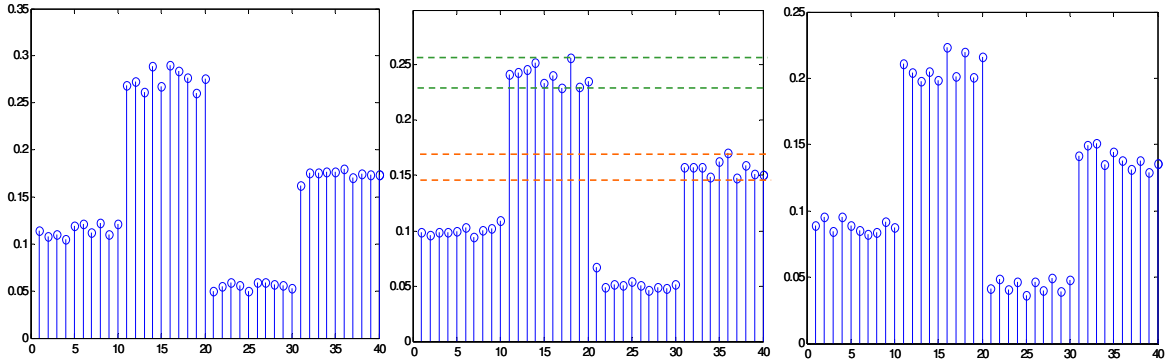

Fig. 3.10 Uni-variate discrimination capability for the 1st, 2nd and 3rd DCT components in ranking. Index S_x. Experiment: "long01".

3.3.2 Selection strategies

Generally speaking, feature reduction is performed by checking how separable different classes are when a reduced set of features (components) is used.

The optimal configuration of l components must be determined by looking for the following two information:

- the l-dimensional combination of the observations' components that minimizes the separability index for each value of l
- the value of l that gives the absolute minimum for $S(l)$.

As stated in Paragraph 3.3.1, an optimal search of the feature set, consisting into a combinatorial exploration of all the possible combinations of components, would be very intensive; a more practical sub-optimal search has to be determined.

Two possible strategies have been tested and are discussed here. Both of them, at their first step, make use of the uni-variate discrimination capability index $S(1)$.

Such index, which has to be plotted according to the series of actions listed in Paragraph 3.3.1 (page 34), for the experimental example used ("long01"), behaves as shown in Fig. 3.12 and Fig. 3.13. Components of the appended observations (n vectors appended per each measurement), having dimensionality L, are sorted in increasing order of $S(1)$.

Three diagrams are presented in each chart, corresponding respectively to:

- sorted components of the time-domain signal,
- sorted components of its DWT (COIF5 mother wavelet),
- sorted components of its DCT.

The Coiflet of order 5 (COIF5) [7] is a discrete wavelet designed by Ingrid Daubechies to be more symmetrical than the Daubechies wavelet.

Whereas Daubechies wavelets have $N / 2 - 1$ vanishing moments (where N is the order), Coiflet scaling functions have $N / 3 - 1$ zero moments and their wavelet functions have $N / 3$. In the proposed experiments are considered as features both the decomposition coefficients (detail) and the approximation ones (coarse). The comparison with the DCT and VLT based feature selection procedure is carried out by evaluating the value of the S index in the same experiment in all the domains.

Calculations of the transformed domain versions of the VLT signals are done for each observed vector separately, before appending; VLT signals are appended directly to form the whole vector of the observations pertaining each measurement; such concept is better explained in Fig. 3.11, where the practical case (n=4) corresponding to experimental conditions used in this work, has been presented.

Moreover, values for variables M and L are those related to the real experiments which have been conducted inside this PhD Thesis work.

Diagrams are presented both for the S_2 metric and for the S_x metric; notice how, in case of the S_x metric, the time-domain signals appear to have more discriminant single components then the transformed counterparts. Situation changes when dealing with vectors formed by grouped components, as it will be seen further in this Chapter.

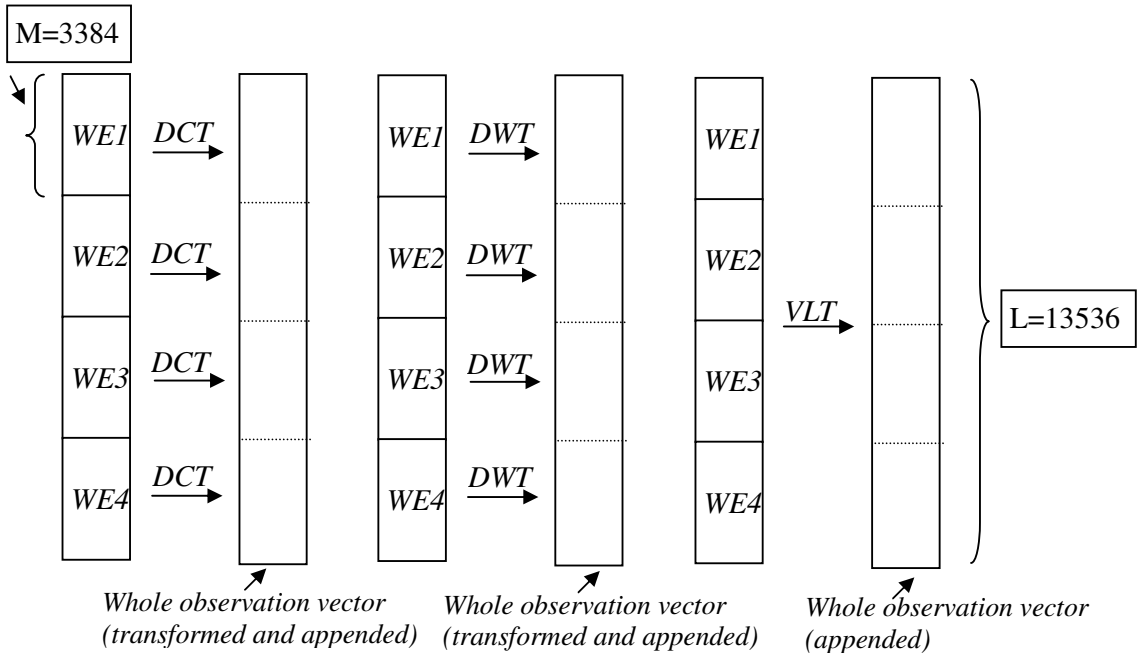

Fig. 3.11 Structure of the whole observation vectors for the calculation of discriminability.

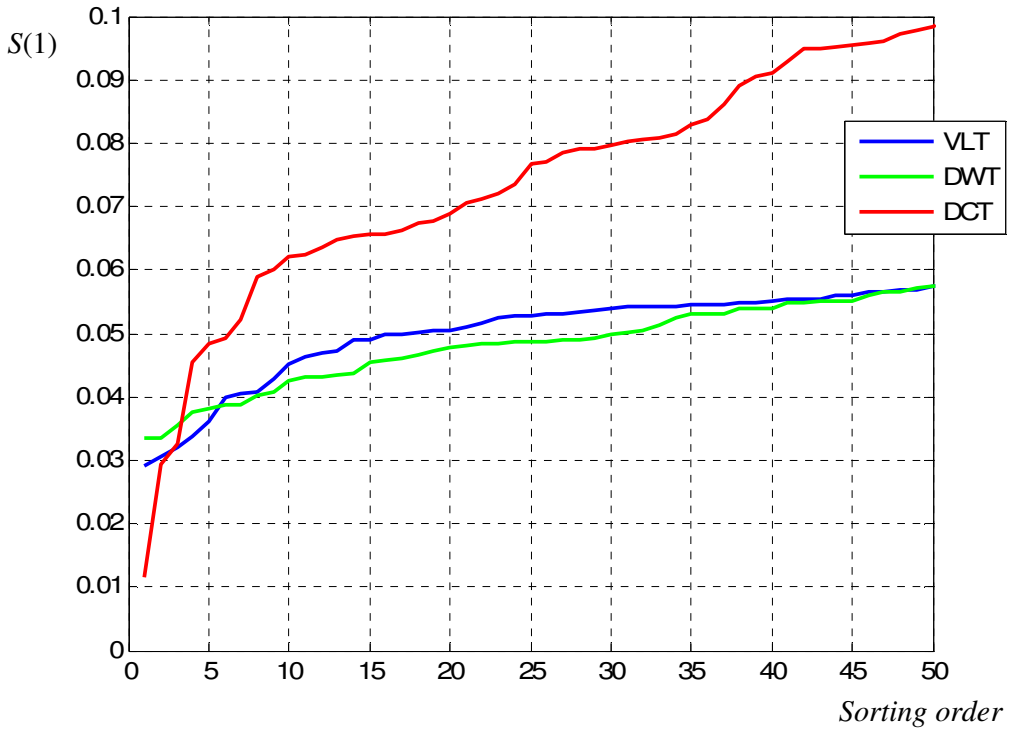

Fig. 3.12 Uni-variate discriminability index S_2 for the experiment "long01 ".

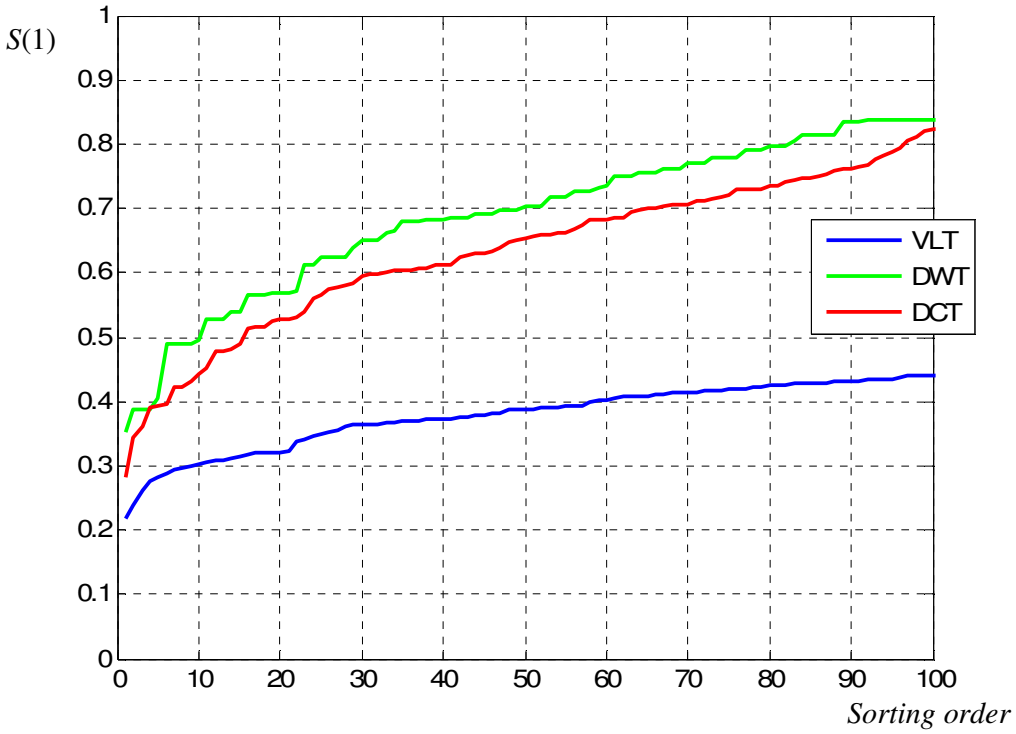

Fig. 3.13 Uni-variate discriminability index Sx for the experiment "long01 ".

Two different search strategies to select the feature subset are proposed here. They essentially consist into the individuation of an optimal (or sub-optimal) set of components of the observations' space that best preserves separability between classes according to the selected metric.

The first method, which can be named *simple sequential strategy*, consists into the following steps:

- evaluation of $S(1)$ for each component (either DCT, DWT or VLT domain),
- sorting of the components according to $S(1)$,
- evaluate the L values that $S(l)$ takes by grouping the first l sorted components, that is:
- $S(l), l \in [1, L]$
- find the l' value, which corresponds to the absolute minimum of the calculated $S(l)$ series.

The obtained $S(l)$ values for the experimental example used in this Chapter ("long01" dataset) are shown in Fig. 3.14, Fig. 3.15, Fig. 3.16 and Fig. 3.17.

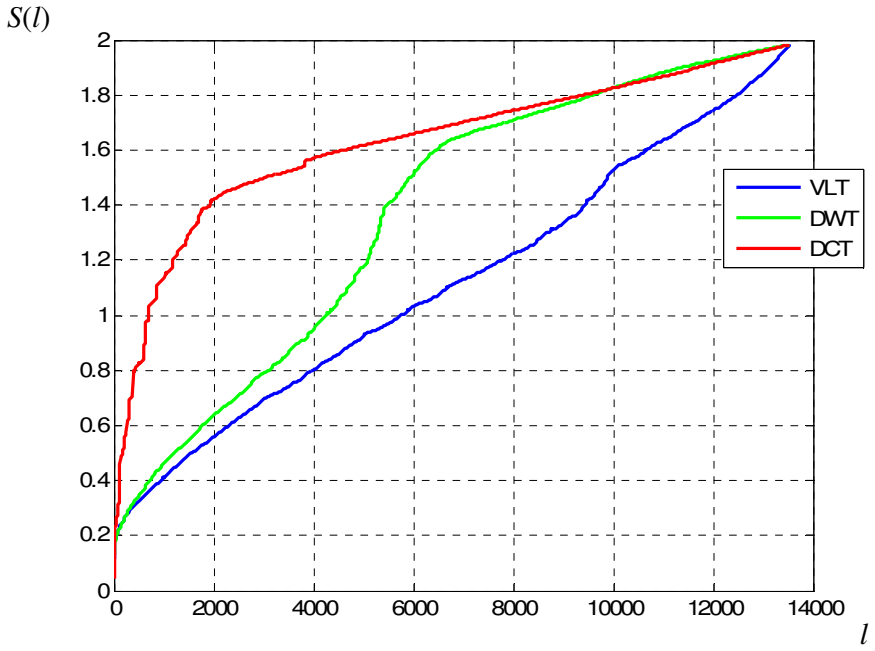

Fig. 3.14 Discrimination capability index S_2 for the l-dimensional subspace generated by grouping the first l sorted components: all the L values. Experiment: "long01".

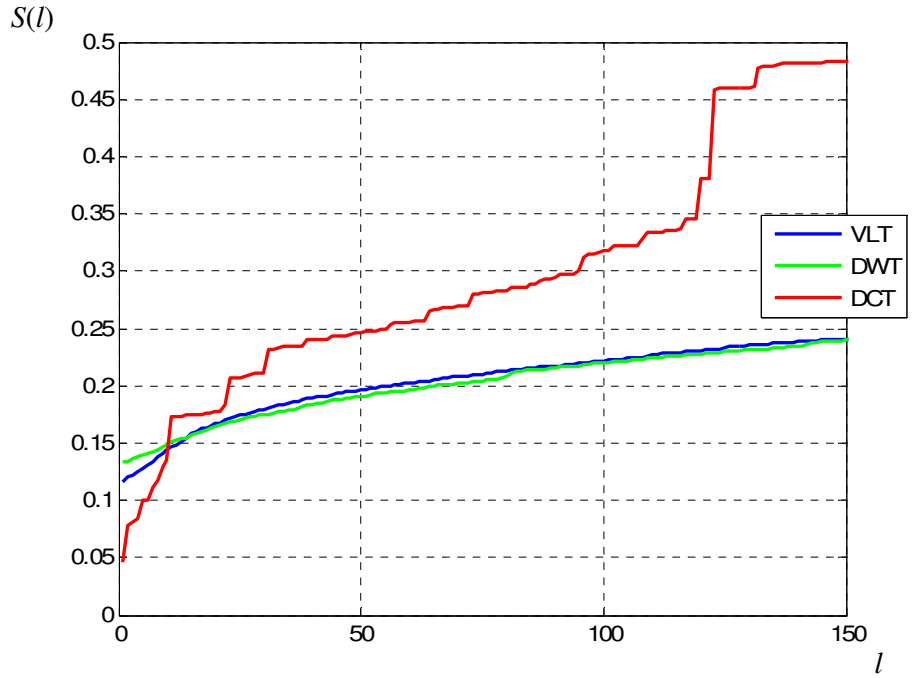

Fig. 3.15 Discrimination capability index S_2 for the l-dimensional subspace generated by grouping the first l sorted components: detail of the first 150 values. Experiment: "long01".

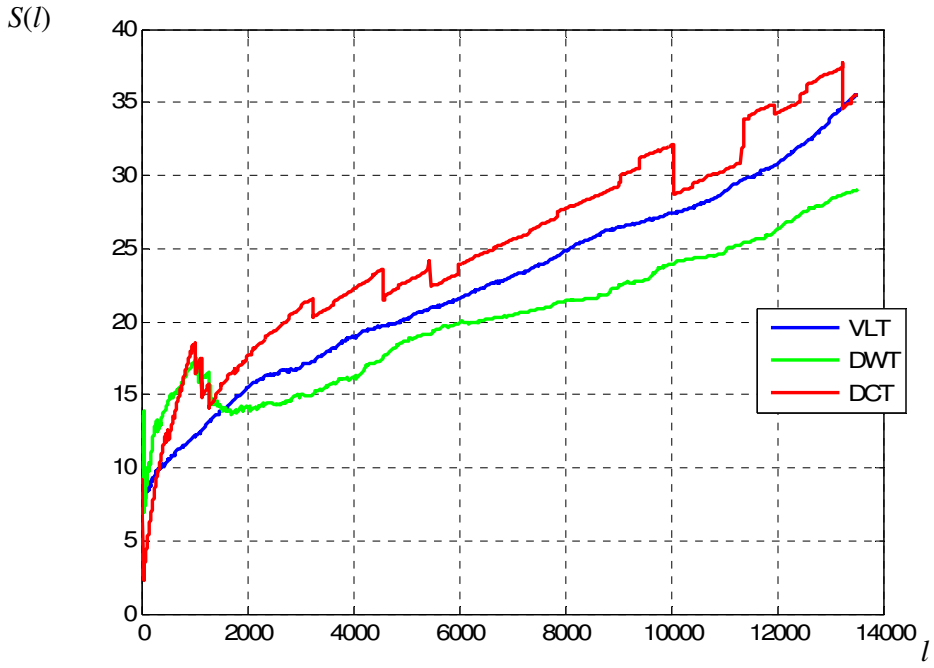

Fig. 3.16 Discrimination capability index S_x for the l-dimensional subspace generated by grouping the first l sorted components: all the L values. Experiment: "long01".

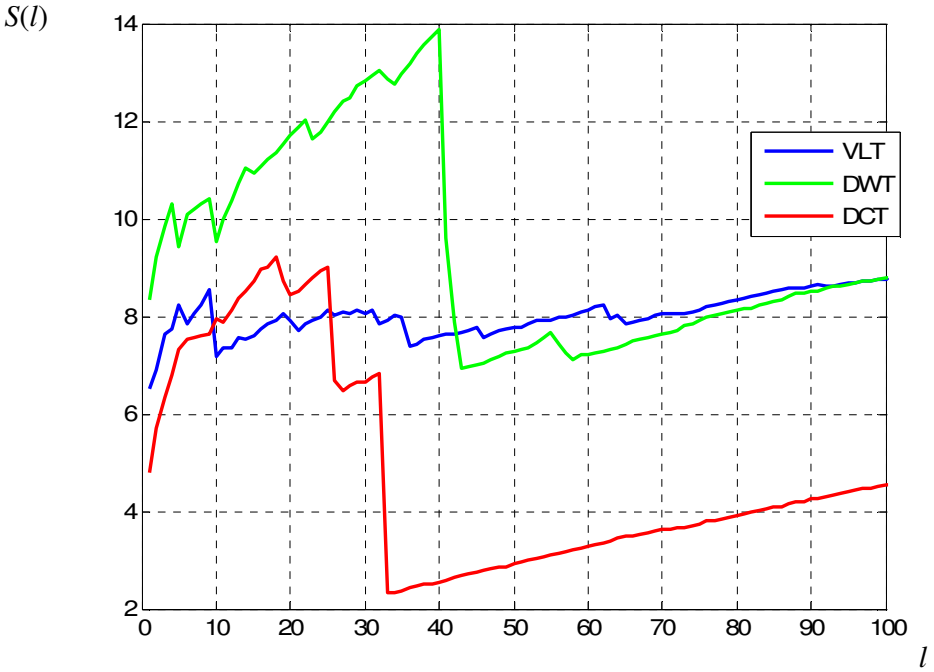

Fig. 3.17 Discrimination capability index S_x for the l-dimensional subspace generated by grouping the first l sorted components: detail of the first 100 values. Experiment: "long01".

The diagrams above suggest the following considerations:

- S_x can actually be effective in individuating an optimal dimensionality for the feature subset, by locating an absolute minimum of $S_x(l)$,
- S_2 has a monotonic behaviour, which would lead to find optimality in a mono-dimensional space; this is not particularly interesting since the preferred index will be S_x, for the reasons discussed in Paragraph 3.3.1.
- DCT is the most performing domain with respect to both metrics.

Moreover the simplest uni-variate selection strategy discussed here, discrimination may be enhanced by a sub-optimal search method, where the discriminability index is calculated in a multidimensional space, generated by selected groups of components. This is a very different approach to the one discussed above, since in the simple sequential strategy the feature subspace is formed by grouping components according to an uni-variate index The same metric is applied to vectors of the generated subspace only in order to determine its dimensionality, never to select the components to add.

A sub-optimal strategy, having a lower computational complexity then an exhaustive combinatorial evaluation of all the possible groups of components, is illustrated here; it will be named *sequential forward strategy*.

Its structure is explained in the next block diagram.

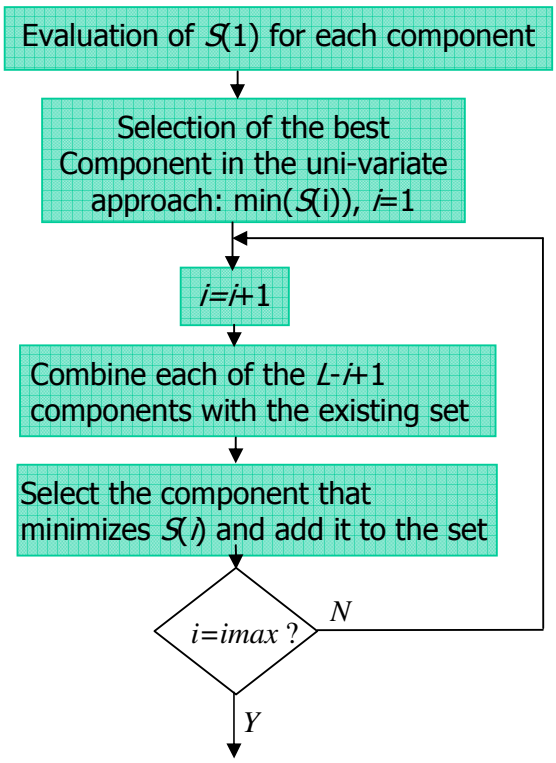

Fig. 3.18 Structure of the sequential forward strategy approach.

The concept that lies behind this algorithm is to combine the selected subset at a certain step with all the other components, one by one, spanning all the observations' space. Calculation of the discriminability index is performed for each of the said combinations. The component that led to the minimum of $S(l)$ is selected and added to the feature set.

This algorithm may continue till a predetermined maximum dimensionality is reached, then an absolute minimum of the calculated series may be searched, in order to determine an optimal dimensionality with respect to this criteria.

For applications implying a supervised classification scheme, such maximum dimensionality may be determined by considering the Hughes effect [8, 9, 10, 11, 12], which limits the dimensionality of the feature space depending on the number of observations taken as a training set. Such limitation has a particular impact on the classification accuracy.

The sequential forward strategy, applied to the simple example dataset used in this Chapter ("long01"), has given the result shown in Fig. 3.19. To obtain this result, the algorithm has been applied to the observations in the DCT domain by using the S_x index.

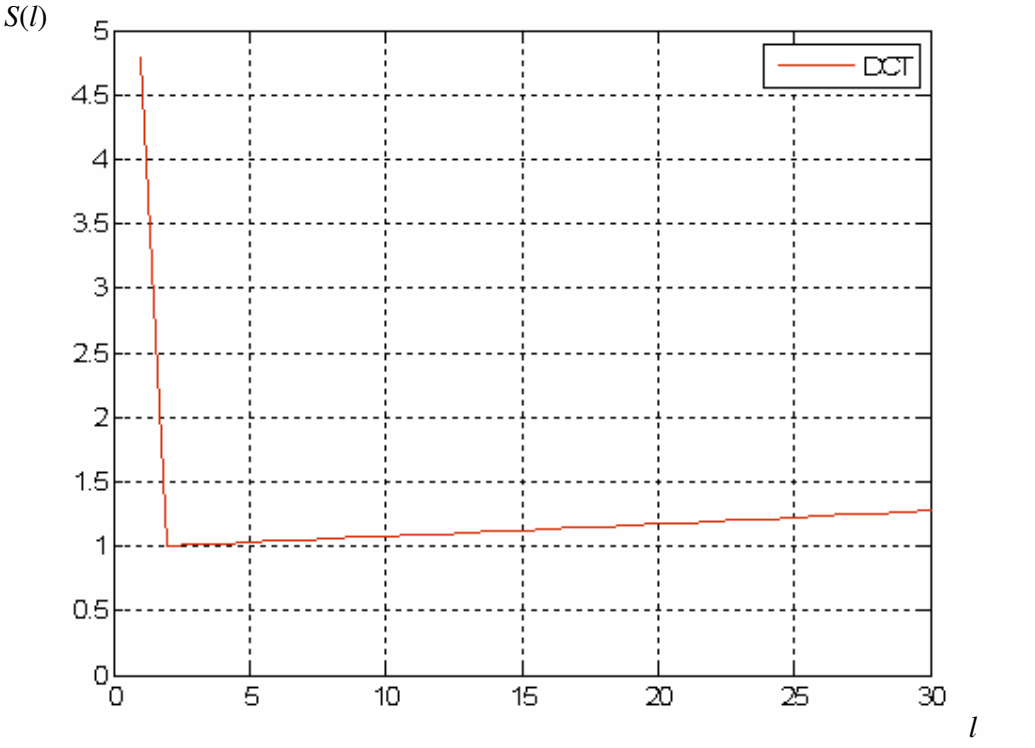

Fig. 3.19 Discriminability index obtained by using a sequential forward search strategy. S_x index. Experiment "long01".

This result takes to the following considerations:

- optimal dimensionality for the dataset "long01", according to the simple sequential strategy, is 33,
- optimal dimensionality for the same dataset", according to the sequential forward strategy, is 2,

- apart from the efficacy of the optimal dimensionality assessment, the number of training examples (collected observations) is well adequate for a feature subspace of dimensionality 2, while it is not enough in case of dimensionality 33, since it would be prone to the Hughes effect. In fact, the simple experiment "long01" consists into a total of 40 measurements of 4 classes of water (10 measurements each class), see Paragraph 4.2.1.

The sequential forward selection strategy combined with the S_x index is the preferred method chosen to approach real world measurements, according to the aspects discussed in this Paragraph and to the experimental work which has been done. A series of applications, including details about the measurement context, are discussed in the next Chapter.

3.3.3 The impact on real measurements

The described feature selection approach applied to the dataset coming from voltammetric measurements, may take meaningful information about the useful bandwidth of the collected signals, the most informative WEs, and the compaction properties of the used transformations.

Under this point of view, a scatter plot showing how each component is ranked according to one of the criteria above, would be very informative.

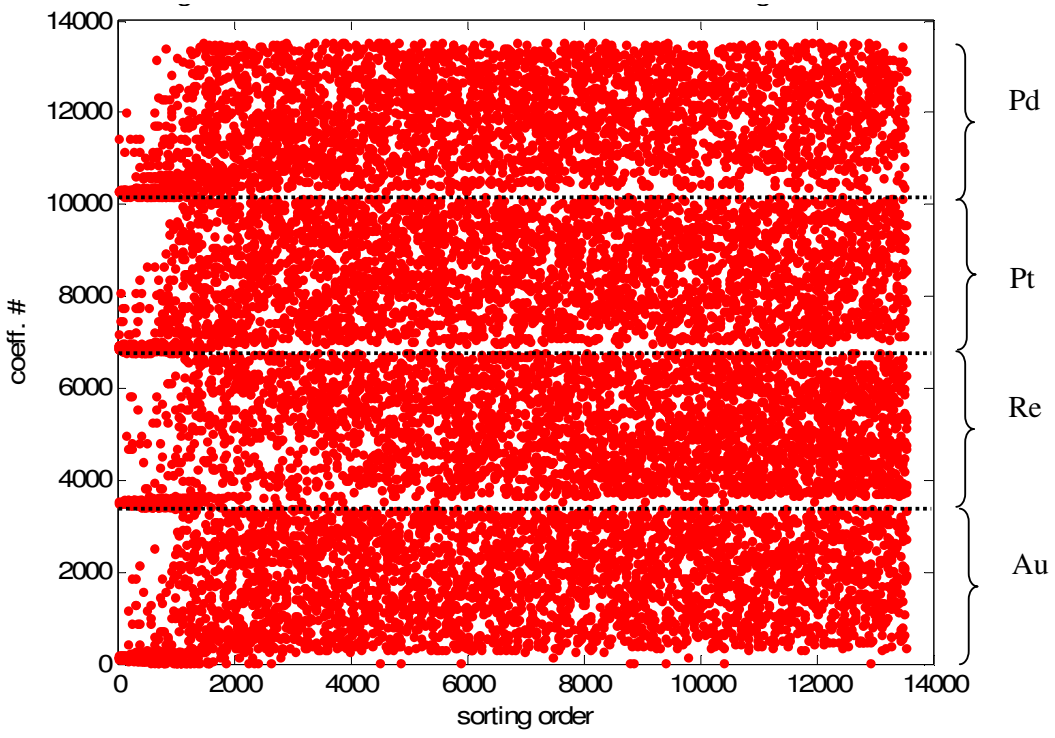

Fig. 3.20 Ranking of the components of the observation space in the DCT domain.

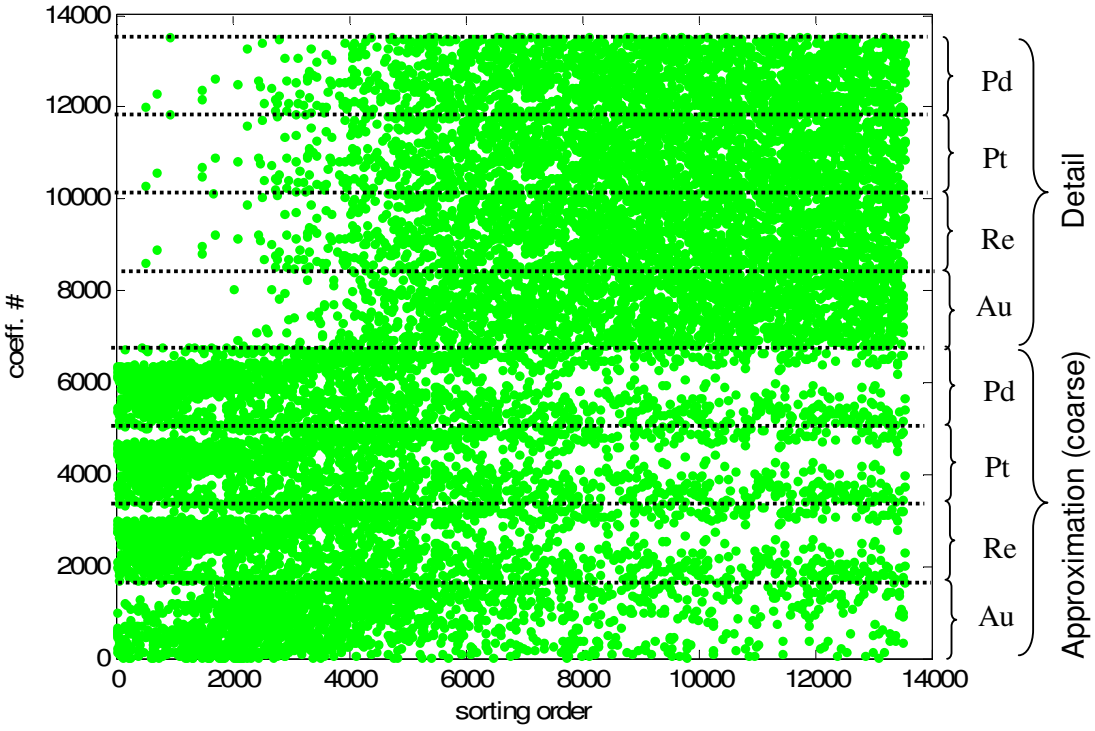

Fig. 3.21 Ranking of the components of the observation space in the DWT domain.

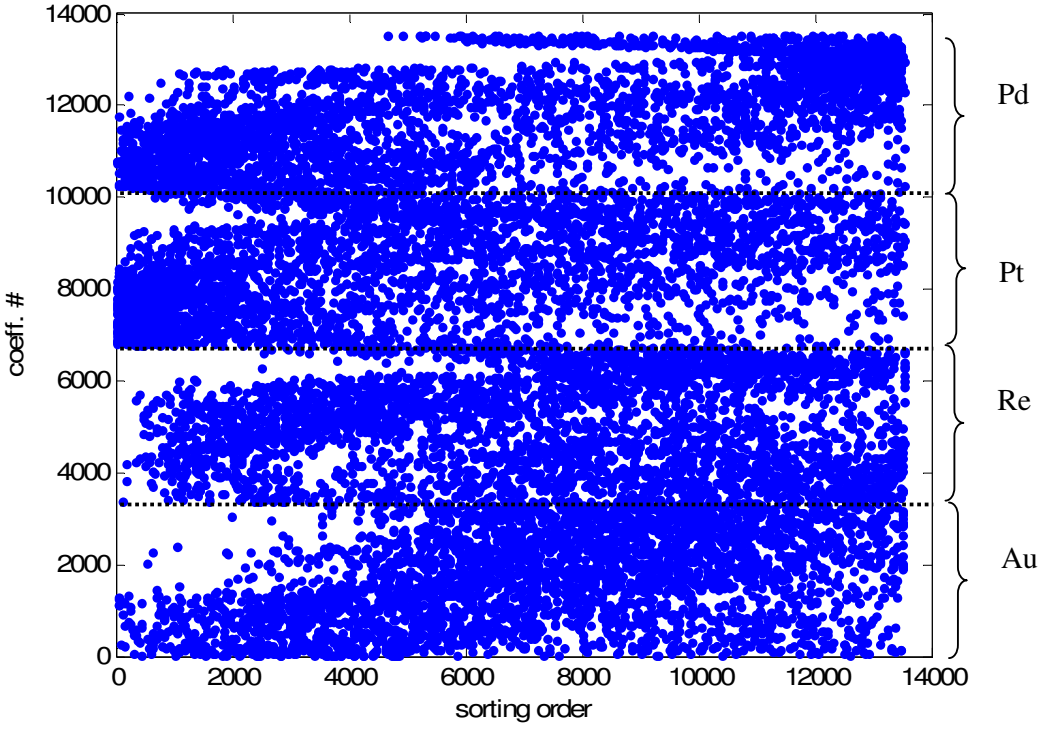

Fig. 3.22 Ranking of the components of the observation space in the VLT domain.

In Fig. 3.20, Fig. 3.21 and Fig. 3.22 an example of distribution of components in the observations' space is shown in the form of a scatter plot. The x-axis represents the sorting order, while on the y-axis is located the number assigned to each one of the component of the whole observation space. Data are split into n ranges in the DCT and VLT domain plots, according to the different WEs used, and are appended as described in Fig. 3.11. As to the DWT domain plot, an additional splitting is due to the fact that coefficients obtained after one transformation step are separated into coarse and detail coefficients. Coarse coefficients are appended first.

The scatter plots shown in said figures have been obtained in the following conditions:

- experimental dataset "long01",
- search strategy "simple sequential",
- discriminability index S_x.

In the DCT domain (Fig. 3.20) there's an evidence of the high ranking given to the low frequency components for all the four WEs; in the DWT domain it is apparent how the coarse coefficients give most of the contribution to discriminability, while detail coefficients are poorly ranked, being cumulated to the right side of the chart. In the VLT domain, it is easier to locate which WEs contribute most to discriminability. In the dataset used, Pt and Pd electrodes appear to be more informative under the point of view of discriminability, according to the chosen index and search strategy.

All the results illustrated up to now have no general applicability; experimental results discussed in Chapter 3 show how these problems must be approached on a case-by-case basis, due to the different peculiarity that each dataset may exhibit. Other datasets and different signal processing algorithms may lead to particular results, specific of the experiment considered. It is thus clear that the concept, which lies behind the presented case studies, is to provide a general guideline about the data processing methodology, without any general applicability of some peculiarities that may be eventually individuated.

3.4 Data representation subspace

Dimensionality reduction is accomplished by an iterative procedure aimed at determining the subspace of R^L where the dataset preserves most of its information content, according to some criteria, which will be defined later. This concept may be directly applied to the whole observation space, having dimensionality L, as shown in [2], but this is not the preferred method in this context. According to the approach proposed in Fig. 3.7, this procedure is applied to the space spanned by the extracted feature set, having dimensionality $R^{l'}$, in order to determine a sub-space with fixed dimensionality suitable for data representation purposes.

This two-steps approach relies on a feature extraction strategy that enhances separability between classes and makes use of a further reduction, for representation purposes, based on the maximisation of the energy content of the signals projected onto the representation subspace.

To quantify such projected energy content, a particular index, called Information Content (IC) is now defined.

By denoting with $\mathbf{s_i}$ the generic observation vector in the *l'*-dimensional feature space (thus a whole observation vector in the DCT/DWT/VLT domain reduced to a subset of its components), a matrix of the observation vectors in the feature space can be defined as:

$$\mathbf{S} = \left[\mathbf{s_1} \cdots \mathbf{s_i} \cdots \mathbf{s_N}\right] \tag{3.5}$$

where N is the number of observations and **S** has dimensionality *l'* x *N*.

An orthonormal basis:

$$\mathbf{B} = \left[\mathbf{b_1} \cdots \mathbf{b}_j\right] \tag{3.6}$$

of a subspace having dimensionality $j < l'$ can be built with an iterative procedure, which is first of all initialised in this way:

$$\mathbf{B}_0 = \mathbf{b}_1 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i \tag{3.7}$$

that is, by taking as the first vector of the base the centroid of all the *N* observations in the feature space. Now the procedure presented in the block diagram of Fig. 3.23 is performed.

In this procedure, the Information Content index (*IC*) is defined as:

$$IC(i) = \left(\frac{\left\| \mathbf{s}_i^{\mathbf{p}} \right\|^2}{\left\| \mathbf{s}_i \right\|^2} \right) \tag{3.8}$$

where $\mathbf{s}_i^{p}$ is the projection of the i-th observation onto the base generated at step j; such base at step j=0 is uni-dimensional and corresponds to the centroid of the whole observations stated in (3.7).

IC is thus defined as the ratio between the energy of the reconstructed signal in the subspace generated by **B** and the energy of the original signal.

In the described algorithm, at any generic step, the direction of the vector that is worst represented by the base is added to the base itself. This is equivalent to take the observation which has the minimum *IC* according to (3.8) and complete the base with its normalised version.

The iterative process is stopped at the first occurrence of one of two events:

- the projected energy of the observation which is worst represented by the base is above a threshold T,
- dimensionality has reached a predetermined maximum, K_{max}.

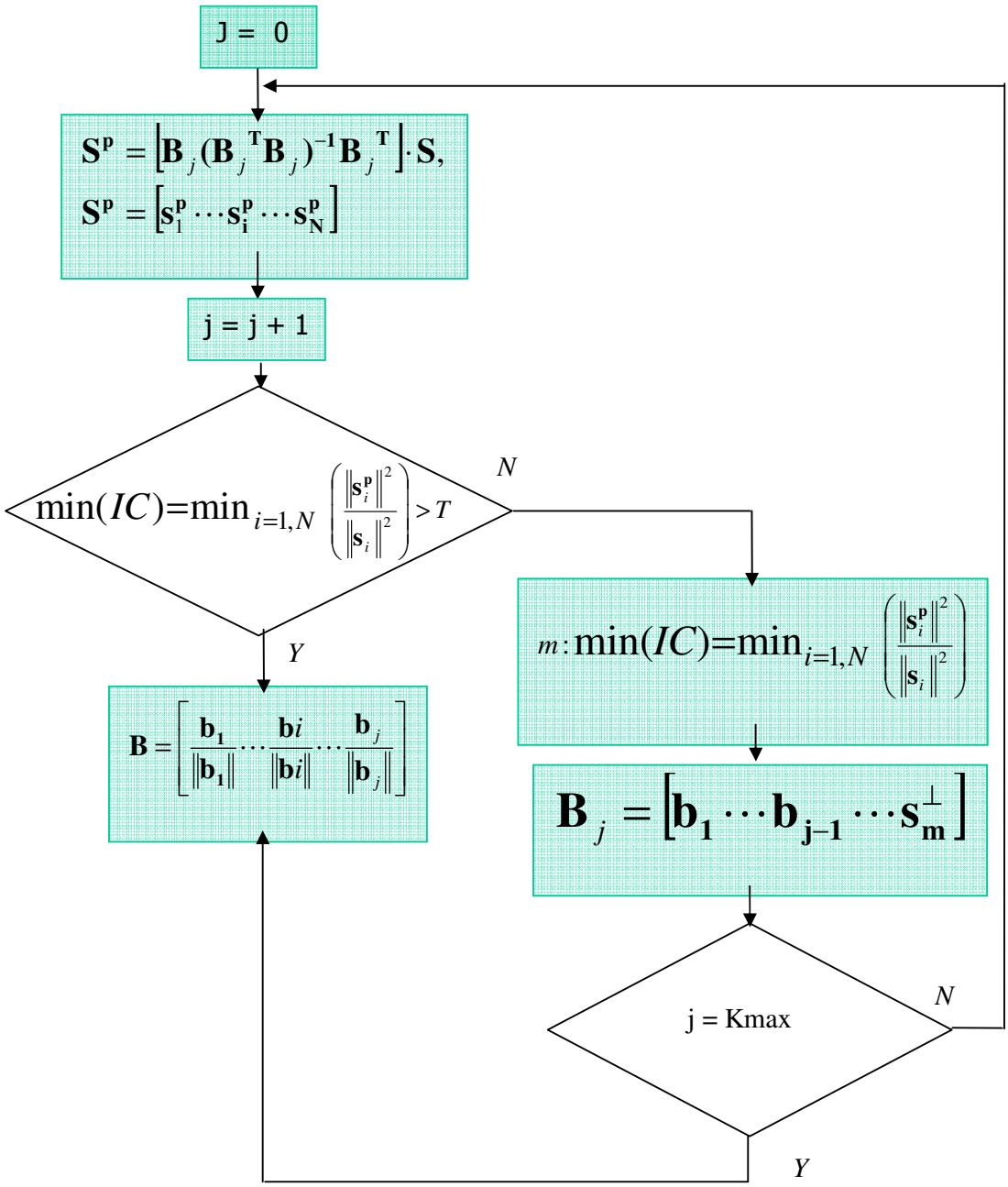

Fig. 3.23 Block diagram of the iterative orthogonalisation procedure to determine a representation subspace.

To provide an example of a typical result obtained by this algorithm, it has been applied to the dataset "long01" in the following conditions:

- feature set in the DCT domain,
- "simple sequential" search strategy,
- S_x discriminability index,

- feature space dimensionality l'=33,
- K_{max}=2

The obtained result is shown in Fig. 3.24. Separation between classes for this simple application is excellent. As it may be seen from the legend, each class has assigned a symbol differently coloured. The axes of the chart represent the components b_i of each projection onto the corresponding normalized vector $\hat{\mathbf{b}}_i$ of the base **B**. In this notation, $\hat{\mathbf{b}}_i$ is the column vector obtained by extracting the i-th column from **B**, and the scalar quantities on the axes (b_i) are the components along the respective vector.

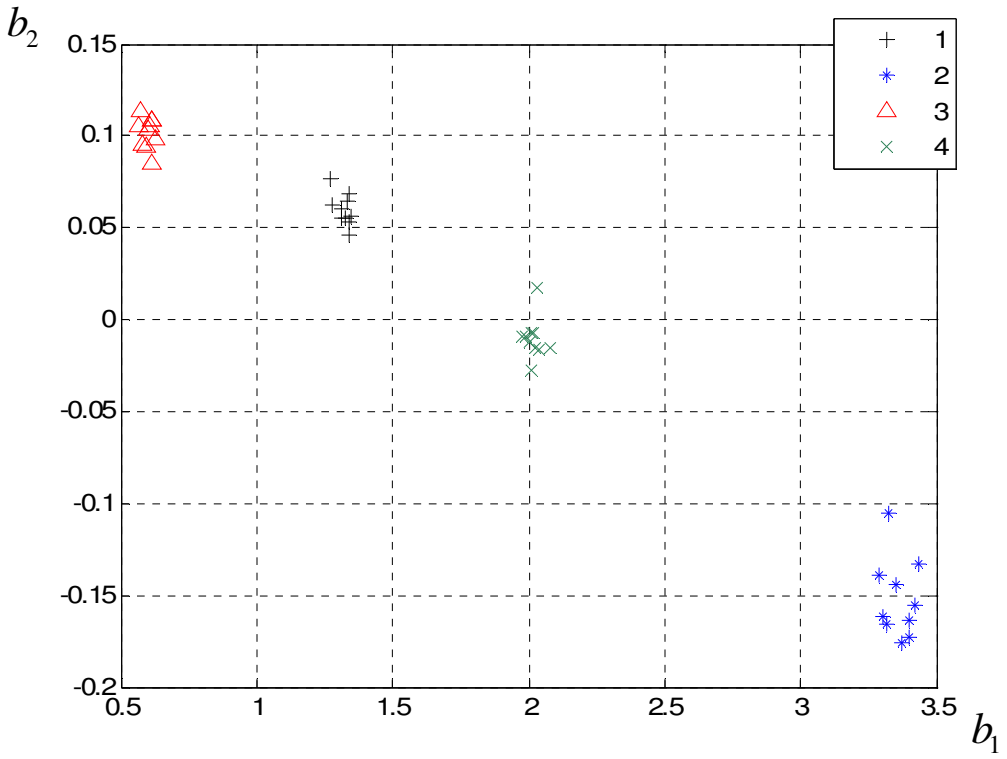

Fig. 3.24 Projection of the observations of the dataset "long01" onto the bi-dimensional representation space.

3.5 One proposed technique for supervised classification

An essential aim of this work is to define a classification system that can learn by a suitable set of examples (eventually acquired in the laboratory) to be used on the field for successive measurements after the tuning process; both aqueous solutions made in the laboratory and water samples coming from the real world can be used for this purpose. In fact, depending on the target application, different approaches can be used in building up the training set, such as using water samples having different chemical classification for characterisation purposes, or

simulating different concentration levels of known pollutants, for change detection and surveillance uses.

Once the classification system has been trained, it is then ready to be used for characterising unknown samples in the successive measurements.

Measurement repeatability is a fundamental aspect in the development of such an approach, and it is the reason for which the design of the measurement cell, the analog front-end and the acquisition system have been dedicated to the present application.

As to the signal processing procedure, the approach presented here is based on a supervised water classification algorithm that relies on different kinds of classifiers. In fact, the proposed scheme is applicable to substantially any kind of classifier, depending on the peculiarity of each application.

We assume that a set of measurements for each class of water is available to allow the algorithm learning (*training set*). Moreover to the training of a classifier in a supervised classification scheme, the training data set can be used to determine the base of the feature space and eventually of the representation space, to be applied at later experiments without repeating such calculations. Experiments would be made by simply extracting the already selected components from the observation vectors in the appropriate domain, eventually projecting vectors of the feature space onto the reduced dimensionality representation base.

This supervised classification scheme would allow the user to classify liquid samples belonging to an already known class or even to train a change detection mechanism based on a suitable metric, to assess the distance of the real-time observations from the training ones.

The fundamental structure of the signal processing chain is the one shown in Fig. 3.7. The supervised classification scheme, according to the illustrated structure, is based on the following steps:

- perform a batch of measurements to generate a training set,
- determine the feature set to extract according to the strategies discussed in Paragraph 3.3 and its sub-Paragraphs,
- train a classifier with observations represented in the feature sub-space,
- perform a new batch of measurements to generate a test set,
- extract the previously determined features (i.e. take the selected components),
- feed the classifier with test set and check the classification accuracy.

This sequence of actions has been used in the two example experiments illustrated in Paragraphs 4.2 and 4.3, respectively. The representation subspace (2-D or 3-D) has been determined making use of the training set vectors only, so vectors belonging to the test set have been projected onto the same base.

To ensure classification accuracy in the reduced dimensionality space (feature space), the quantity of available example observations (training set) plays a relevant role.

In fact, the size of the training set and the number of features (thus the dimensionality of the feature space) are in a relationship which has been studied by several authors, such as Jain & Chandrasekaran and Kalayeh & Landgrebe in the middle ‘80s [9, 10, 11].

As a general guidance, the suggested ratio, indicated by many authors, between the size of the training set and the dimensionality of the feature space lies in the range between 5 and 10;

this means that the number of example measurements per class should be at least 5 to 10 times the number of features, as suggested in [12] for remote sensing applications.

In this class of problems, the number of observations is frequently limited by practical aspects, such as the slowness of the electrochemical measurements. Also, limitations due to the high measurement cost may be possible in certain applications.

The feature reduction approach proposed in this Thesis work relies on a sub-optimal determination of the feature space dimensionality. This means that there's no *a priori* knowledge of the adequate number of observations to collect into the training set.

Moreover, a scheme with a fixed dimensionality for the feature space, where the search strategy stops when the desired number of feature is reached, is anyway possible, making use of the methods discussed in this Chapter.

The practical examples shown in Chapter 4 will clearly highlight the advantage of the forward search strategy over the simple sequential one, estimating lower sub-optimal dimensionalities for the feature space. Nevertheless, the small amount of observations cannot always guarantee the said requirement about the training set size to fulfill. In simple cases, where class separation is well appreciable even in single components of the feature space, there's substantially no consequence of this problem. More critical with a less apparent clustering would imply paying more attention to this rule.

REFERENCES FOR CHAPTER III

[1] Multifunction data acquisition board AX5621H+ manufacturer's website (Axiom Ltd.): http://eservice.axiomtek.com.tw/attach_files/K0602-0326/ax5621h+.pdf

[2] A. Scozzari, N. Acito, G. Corsini, "Signal analysis of voltammetric data series for water quality tests and classification", IEEE IMTC 2005, Ottawa (Canada), volume 1, pp. 89-92

[3] A. Scozzari, N. Acito, G. Corsini, "A novel method based on voltammetry for the qualitative analysis of water", IEEE Transactions on Instrumentation and Measurement, In press

[4] T. Artursson, M. Holmberg, "Wavelet transform of electronic tongue data", Sensors and actuators B, 87 (2002), pp. 379-391

[5] L. Robertsson, P. Wide, "Improving food quality analysis using a wavelet method for feature extraction", IEEE IMTC 2005, Ottawa (Canada), volume 1

[6] K. Ramamohan Rao, P. Yip, "Discrete Cosine Transform: Algorithms, Advantages, Applications", Academic Press, 1990

[7] I. Daubechies, "Ten Lectures on Wavelets," CBMS-NSF Lecture Notes nr. 61, SIAM, 1992

[8] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, Vol. IT-14, No. 1, January 1968.

[9] A. K. Jain, B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice", in "Handbook of statistics", North Holland, Amsterdam 1982, pp. 835-855

[10] H. M. Kalayeh, D. Landgrebe, "Predicting the required number of training samples", IEEE Transactions on Pattern Analysis and Machine Intelligence, 5 (1983), pp. 664-667

[11] Webb, "Statistical pattern recognition", John Wiley & Sons, 2nd edition, 2002

[12] P. H. Swain, S. M. Davis, "Remote sensing: the quantitative approach", McGraw-Hill, 1978

CHAPTER IV

Experimental results

4.1 Classification and change detection approach

4.1.1 General remarks about the experimental conditions

The measurement procedure applied for generating the datasets discussed in this Chapter, has undergone several modifications and updates during this PhD work, according to new results and thoughts carried out by the experience.

All the measurements have some common aspects in their experimental condition:

- stirred solution,
- REF electrode in Ag/AgCl (see Sensor system3.1.2),
- AUX electrode in stainless steel plate (see 3.1.2),
- LAPV excitation signal (see Fig. 3.3).

Moreover, the first experiments [1] have been carried out in the following conditions:

- 2 Working Electrodes (Au, Re), wires about 5mm long,
- chemical washing with H_2SO_4 1M between each set of measurements pertaining each class, followed by rinsing in de-ionized water,
- electrochemical washing between any two measurements.

As to the configuration of the analog front-end (potentiostat) and the acquisition strategy (sampling rate and pre-filtering) parameters they've been kept the same for the experiments conducted; the essential parameters are listed as follows:

- sampling frequency Fs=1KHz,
- median filter cells Nc=11,
- downsampling factor Ds=20.

The most updated experimental conditions [2], used for the most recent experiments, are as follows:

- WEs with a wet surface of 0.2 mm^2, made of Au, Re, Pt, Pd,
- no chemical washing,
- no electrochemical washing,
- ultrasonic washing with a transducer at the bottom of the cell between every set of measurements belonging to the same class.

The measurement cell and the potentiostat are shown in Fig. 4.1 and Fig. 4.2, where essential parts are illustrated.

Fig. 4.1 The measurement cell in its latest version

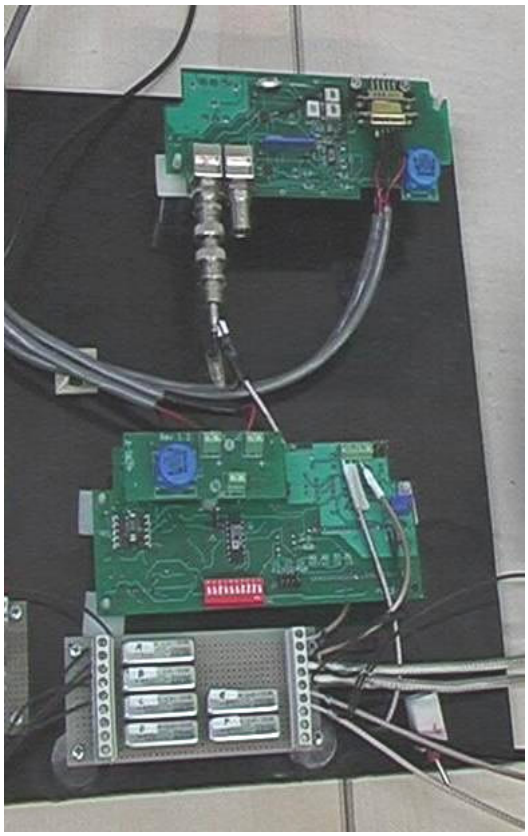

Fig. 4.2 The potentiostat.

4.1.2 Brief overview of the experiments

Five different case studies are presented, each of them representative of some peculiarities which worth to be shown, such as the different nature of liquids, the capability to detect pollutants in water, or the capability to discriminate between different classes of water.

Moreover, different aspects of data processing are presented, depending on the relevant remarks led by such examples.

The following table summarizes the relevant aspects of the presented experiments:

Paragraph title	Dataset name	Description
Bottled water discrimination	*Long01*	Four different bottled mineral water brands, discrimination capability is assessed with a total amount of 10 measurements for each class, distributed on 10 different days
Supervised classification	*Day004*	More complex experiment with 6 classes of water, part of them being laboratory solutions with different electrolytes; half of the observations form the training set. The complete dataset has been checked for classification accuracy
Characterisation by standard analytical techniques	*Flegrei*	Eleven water point in Southern Italy (Naples area). E-tongue results are compared to the PCA of standard analytical measurements
Detection of pollutants	*LaGabella*	Discrimination between raw and filtered water affected by a slight pollution due to trichloroethylene and tetrachloroethylene
Drifts in long term measurements	*Testt0102*	Two 24-hours measurements with continuous water stirring and data acquisition. Two different classes have been used.

Tab. 4.1 Summary of the case studies discussed in Chapter IV.

4.2 Experiment #1: bottled water discrimination ("Long01")

4.2.1 Description of the experiment

The classification system proposed in this Thesis is based on a supervised procedure where the parameters' tuning is obtained by using a training set of suitable measurements (see Paragraph 3.5). This experiment is focused on the discrimination capability of the system. In practice, the training set is collected during one or more specific measurement sessions; the system designed by means of such training set is then used in the subsequent measurements. This example concerns the typical use of the proposed system, where the repeatability after 5 different measurement sessions distributed over 10 days has been checked.

The essential information about the measurement procedure followed in this experiment is given in Tab. 4.2, while the configuration of the experimental setup is briefly described in Tab. 4.3.

Time coverage	10 days
Number of classes	4
Kind of classes	Bottled mineral water
Training set	8 measurements, 2 for each class
Total dataset	40 measurements, 10 for each class
Notes	2 measurements per each brand on 5 different days distributed in 10 days
	Measurement sequence of different water brands have been randomly determined in each measurement session

Tab. 4.2 Measurement procedure of the experiment ("long01").

Working Electrodes	Au, Re, Pt, Pd
Washing	Ultrasonic
Notes	Ultrasonic washing made in distilled water between each batch of measurements

Tab. 4.3 Experimental setup configuration ("long01").

The training set has been used to determine the base of the feature space and the representation subspace to make the projections of the observations belonging to the whole dataset in the following step.

4.2.2 Results

Both results obtained with a simple sequential search strategy and results obtained with a forward strategy are shown. The next table resumes the parameters used for the simple sequential strategy:

Discriminability Index	S_x
Search strategy	Simple sequential
Feature space dimensionality	$l' = 33$
Notes	Shown projections regard the observations in the DCT domain

Tab. 4.4 Feature extraction parameters for the dataset "long01" (simple sequential strategy).

The trend of the discriminability index Vs. the dimensionality of the feature space, for the chosen search strategy, is shown in Fig. 3.17.

Projections onto the first 2 vectors of the representation base, generated as described in Paragraph 3.4 by using the 8 vectors of the training set, are shown in Fig. 4.3.

Projection of the full dataset (10 measurements for each class) onto the same base previously generated, is shown in Fig. 4.4.

Even having distributed the measurements on a period of 10 days, measurement repeatability has allowed an high discrimination between clusters. Such discrimination is easily reachable, since the very reduced training set (8 measurement on 40) used in this context, has been sufficient for obtaining a clear cluster separation in the reduced data space, where observations are very well grouped according to the water brand they come from.

Water brands corresponding to each of the analysed classes are associated as follows:

Class #	Description
1	"Vera" bottled mineral water
2	"San Benedetto" bottled mineral water
3	"Levissima" bottled mineral water
4	"Tesorino" bottled mineral water

Tab. 4.5 Description of classes for the dataset "long01".

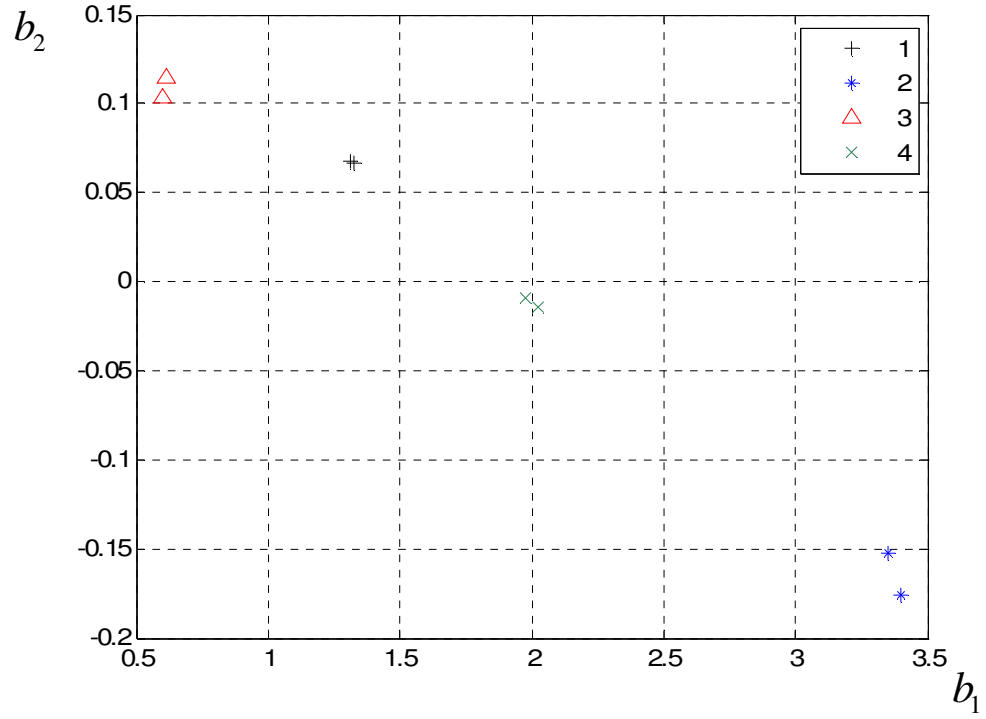

Fig. 4.3 Projections of the observations of the training set ("long01") onto the bi-dimensional representation space (simple sequential search strategy).

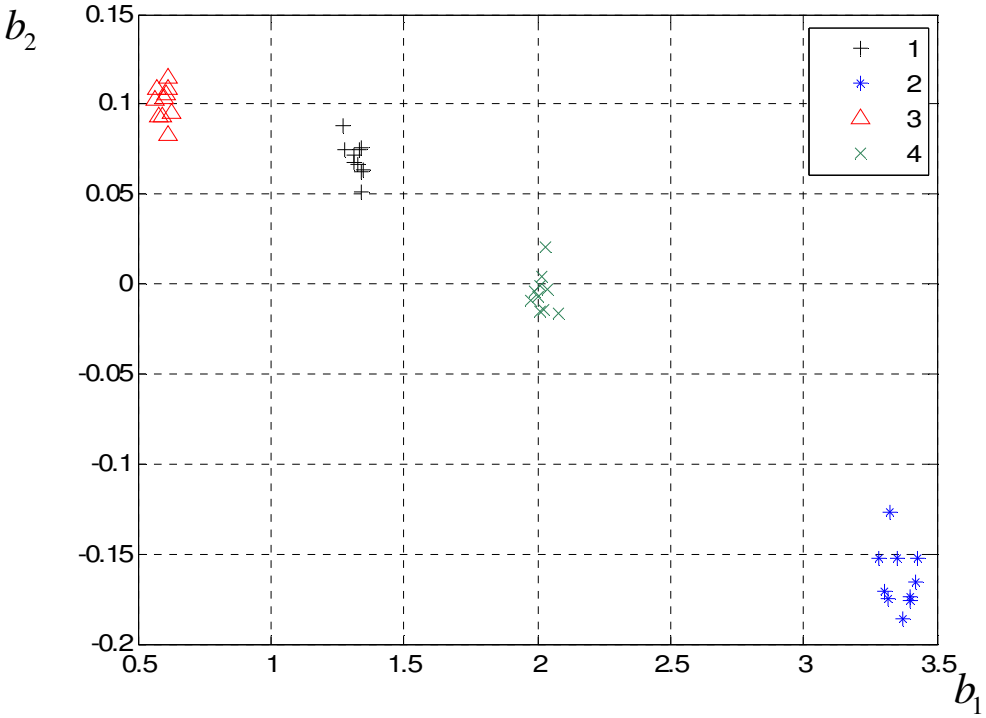

Fig. 4.4 Projections of the observations of the full dataset ("long01") onto the bi-dimensional representation space (simple sequential search strategy)..

A forward search strategy has been tested on the same dataset. Relevant parameters are as follows:

Discriminability Index	S_x
Search strategy	Sequential forward
Feature space dimensionality	$l' = 2$
Notes	

Tab. 4.6 Feature extraction parameters for the dataset "long01" (sequential forward strategy).

Again, projections for the training set and the complete dataset are presented in Fig. 4.5 and Fig. 4.6. While separation is absolutely apparent also in this case, an automatic procedure to determine the dimensionality of the feature subspace would lead to a sub-optimal bi-dimensional space, instead of the l'=33 found with the simple sequential strategy.

Since the nature of the experiments implies a relatively small number of observations per class, with respect to other areas of research where pattern recognition techniques are applied, it is expected that a search strategy that gives a feature space of the lowest dimensionality has a very positive impact on the classification accuracy, as the Hughes effect [3] indicates.

Under this point of view, it is worth noting that the typical requirements about the number of training examples with respect to the dimensionality of the feature space [4] is not respected here, in both of the two strategies. Details about this aspect can be found in Paragraph 3.3.2. In more complex experiments, where separability is more critical, such requirements have reasonably to be taken into account.

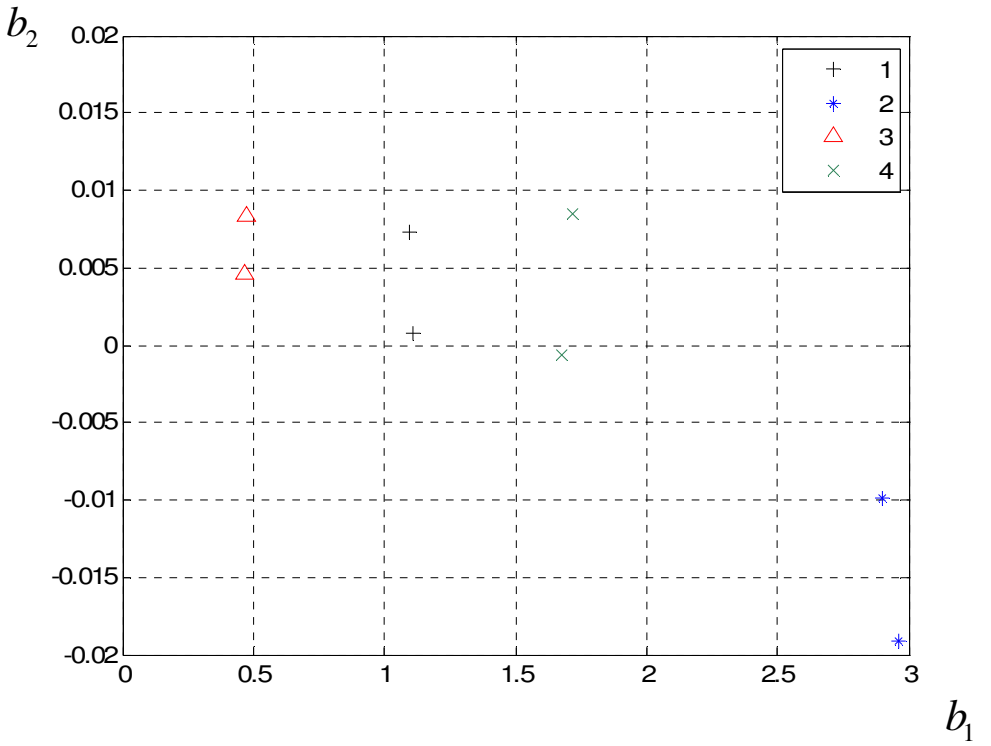

Fig. 4.5 Projections of the observations of the training set ("long01") onto the bi-dimensional representation space (sequential forward strategy).

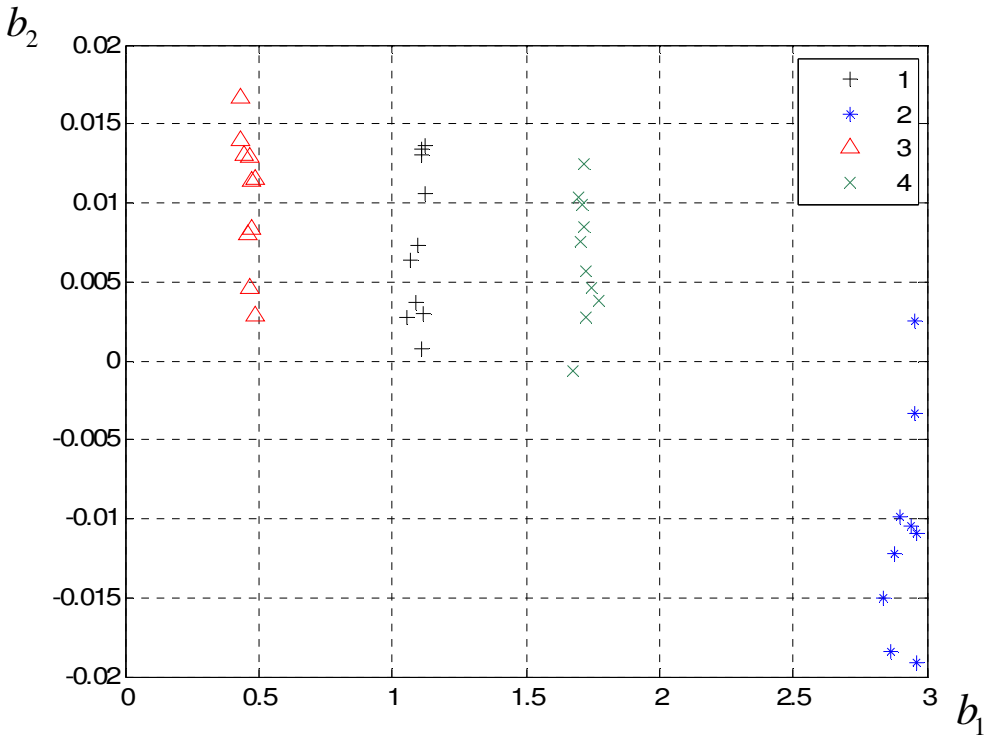

Fig. 4.6 Projections of the observations of the full dataset ("long01") onto the bi-dimensional representation space (sequential forward strategy).

An automatic classifier based on an MLP-NN (Multi Layer Perceptron – Neural Net) has been experimented on the feature set extracted by the simple sequential strategy.

As concerns the MLP-NN, a one hidden layer with four neurons architecture has been chosen. The selected transfer function for the hidden layer neurons was a log-sigmoid one, while the output neuron had a linear transfer function. In this configuration, the network trained by the said 8 examples achieves 100% of classification accuracy on the full dataset, consisting into 40 measurements.

4.3 Experiment #2: supervised classification ("Day004")

4.3.1 Description of the experiment

The supervised classification scheme proposed Paragraph 3.5 is tested in this experiment. In practice, the training set is extracted from the full dataset collected during one whole measurement session; the system designed by means of such training set is then used in the subsequent measurements. This example concerns 6 classes consisting into 3 different brands of bottled mineral water and 3 laboratory solutions made with different electrolytes, characterised by having the same electrical conductivity (200 μS/cm).

The use of different classes of water having the same conductivity is very useful to check the efficacy of chronoamperometric measurements to take information from the diffusion currents instead of the migration effects of bulk charges. Diffusion-limited currents are the only available resource to explain the difference between samples having the same bulk conductivity.

The essential information about the measurement procedure followed in this case study is given in Tab. 4.7, while the configuration of the experimental setup is briefly described in Tab. 4.8.

Time coverage	Few hours
Number of classes	6
Kind of classes	Bottled mineral water plus reference solutions with pure electrolytes
Training set(s)	24 measurements, 4 for each class 30 measurements, 5 for each class 42 measurements, 7 for each class
Total dataset	60 measurements, 10 for each class
Notes	

Tab. 4.7 Measurement procedure of the experiment "Day004".

Working Electrodes	Au, Re, Pt, Pd
Washing	Ultrasonic
Notes	Ultrasonic washing made in distilled water between each batch of measurements

Tab. 4.8 Experimental setup configuration of the experiment "Day004".

The training set has been used to determine the base of the feature space and of the representation subspace, to make the projections of the observations belonging to the whole dataset in the following step. In addition, the training set has been used to determine the class boundaries of a KNN (K Nearest Neighbours) classifier, in an Euclidean metric.

4.3.2 Results

Shown results have been obtained with a forward search strategy. The next table resumes the parameters used:

Discriminability Index	S_x
Search strategy	Sequential forward
Feature space dimensionality	$l' = 8$
Notes	The first 5 measurements per class have been used as a training set for the feature extraction procedure. The classifier has been experimented with 4,5,6 and 7 measurements per class. The feature set has been searched for in the DCT domain

Tab. 4.9 Feature extraction parameters for the dataset "day004" (sequential forward strategy).

The trend of the discriminability index Vs. the dimensionality of the feature space, for the chosen search strategy, is shown in Fig. 4.7, where a detail regarding dimensionality values up to $l = 13$, is plotted.

Projections onto a tri-dimensional representation base, generated as described in Paragraph 3.4, by using the 30 vectors of the training set for the feature extraction procedure (see Tab. 4.9), are shown in Fig. 4.8 for the training set of 30 vectors. Fig. 4.9 shows the full dataset projected onto the previously determined base.

A KNN (K-Nearest Neighbours) classifier has been tested on this dataset. A subset of the 60 measurements has been used to train the classifier (i.e. to make distance measurements from unlabeled observations); all the remaining dataset has been used to test the classifier in a rigid manner, that is, new labelled measurements did not contribute to increment the set of known examples (training set).

All the classification process has been performed in the reduced dimensionality space, according to the supervised classification scheme discussed in Paragraph 3.5.

The number of errors on the test set (formed by the full dataset with the training measurements removed) is plotted against the number of nearest neighbours to be evaluated to determine the class to assign (K).

The above mentioned plots are shown in the case of datasets formed by the first 4,5,6 and 7 measurements respectively. Results are presented in Fig. 4.10, Fig. 4.11, Fig. 4.12 and Fig. 4.13, respectively.

In this example, a minimum of 6 measurements in the training set have been necessary to have an error-free classifications. The exploration of all the possible training sets (combination of measurements) being not so meaningful, the really important information that comes from this experiment is that a training set formed by 7 measurements offers an error-free classification with a reasonable number of K=3 neighbours.

Class assignment for this case study is shown in Tab. 4.10.

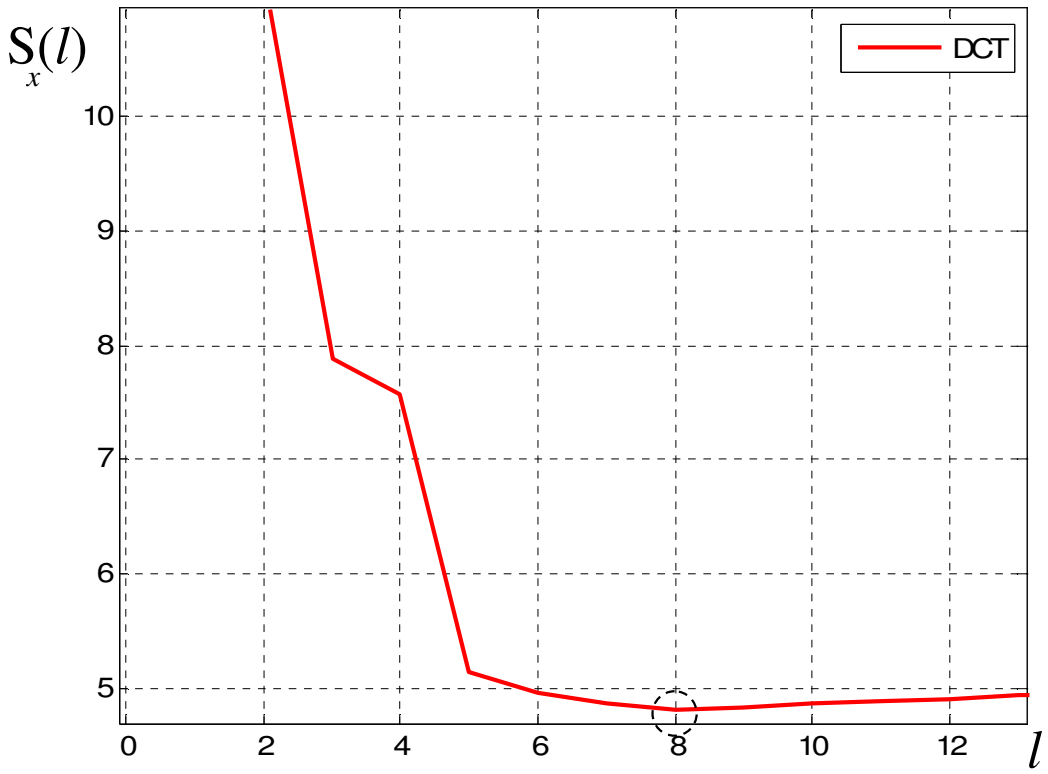

Fig. 4.7 Discriminability index trend obtained by using a sequential forward search strategy. Sx index. Experiment "day004".

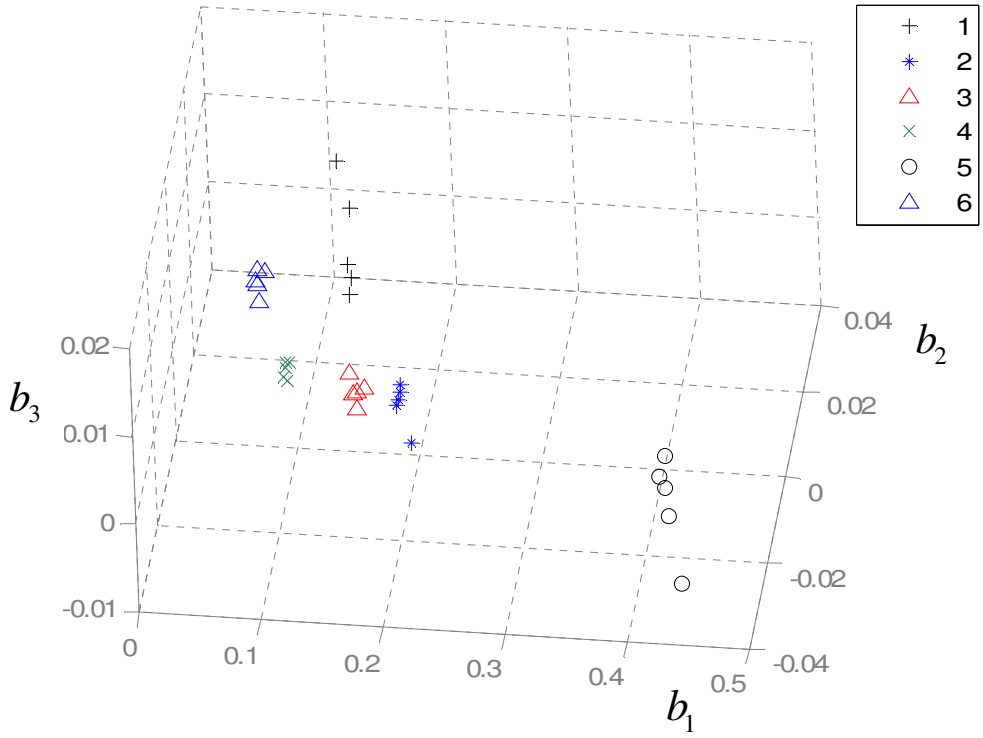

Fig. 4.8 Projections of the observations of the training set (subset of dataset "day004") onto the tri-dimensional representation space (sequential forward strategy).

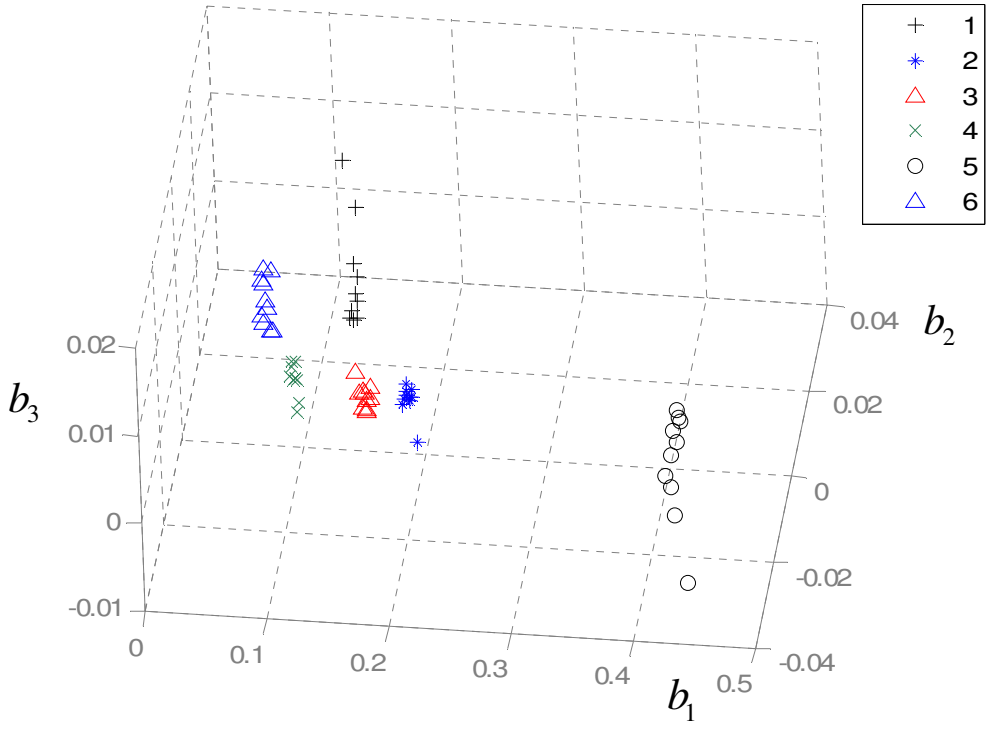

Fig. 4.9 Projections of the observations of the full dataset ("day004") onto the tri-dimensional representation space (sequential forward strategy).

Class #	Description
1	"Graziosa" mineral water
2	"Gaia" mineral water
3	"Lora Recoaro" mineral water
4	NaCl 200±5 µS·cm^{-1} @ 25°C
5	KCl 200±5 µS·cm^{-1} @ 25°C
6	$NaHCO_3$ 200±5 µS·cm^{-1} @ 25°C

Tab. 4.10 Description of classes for the dataset "day004".

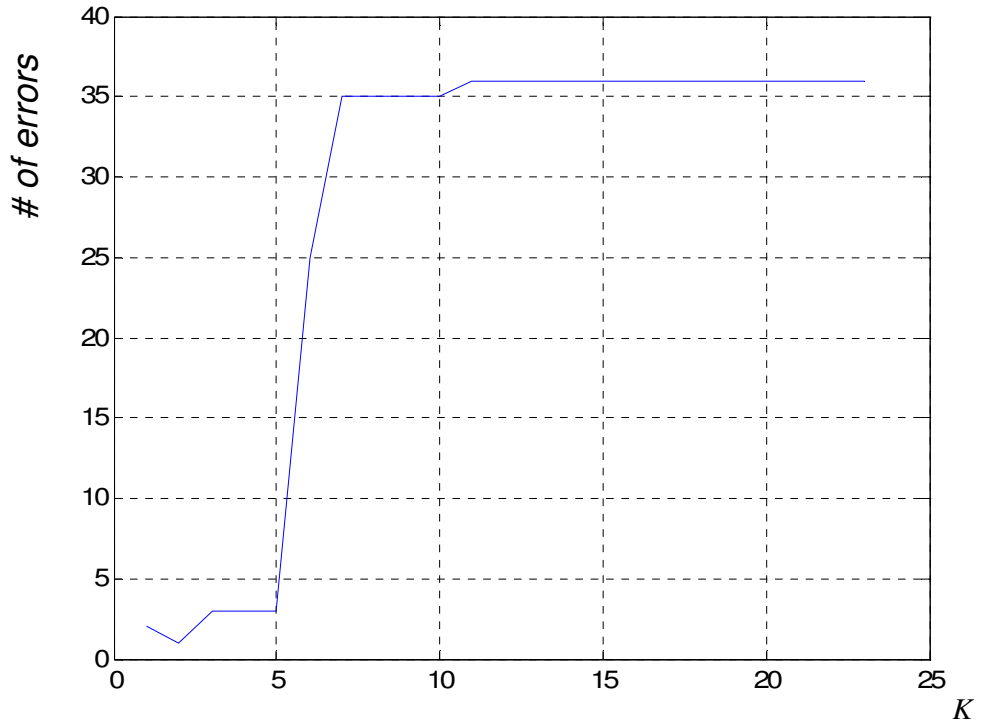

Fig. 4.10 Classification errors Vs. K for a KNN classifier applied to the dataset day004 (training set: 4 measurements per class).

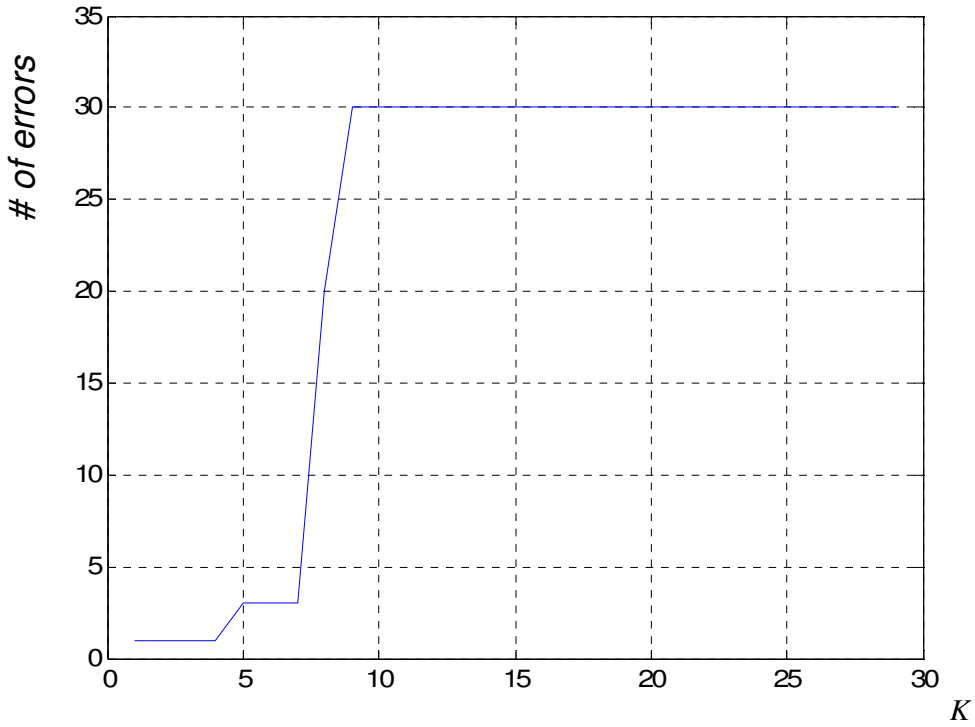

Fig. 4.11 Classification errors Vs. K for a KNN classifier applied to the dataset day004 (training set: 5 measurements per class).

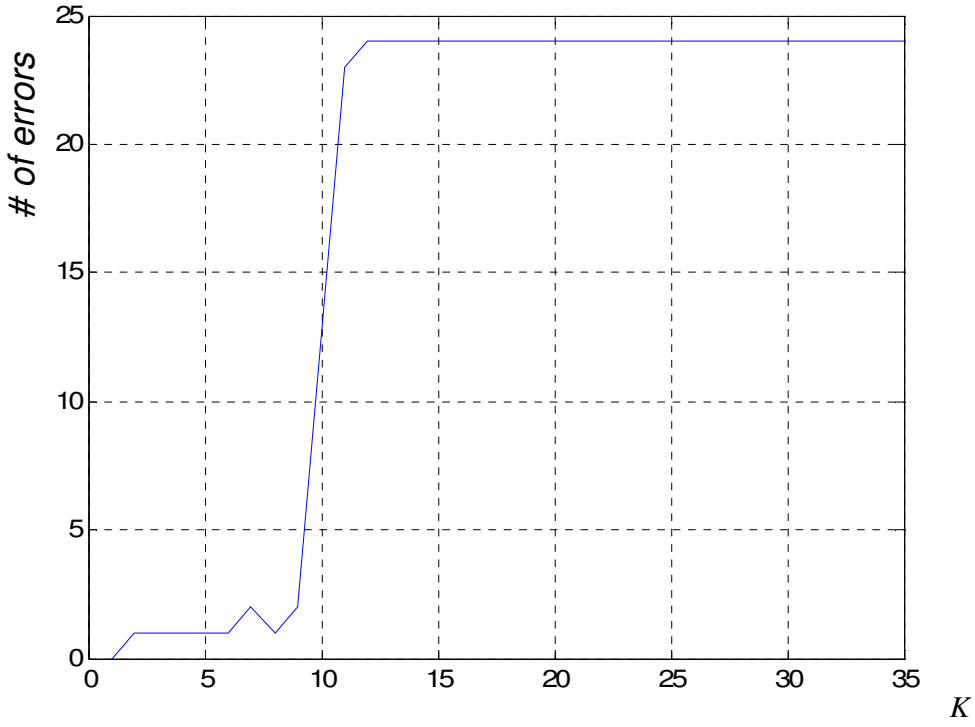

Fig. 4.12 Classification errors Vs. K for a KNN classifier applied to the dataset day004 (training set: 6 measurements per class).

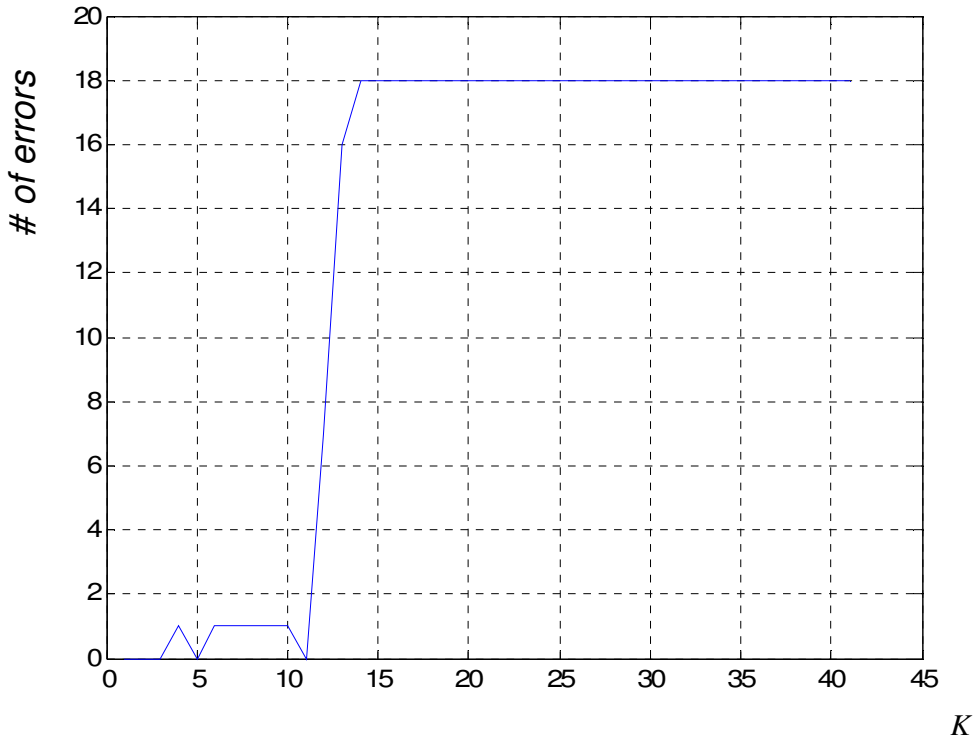

Fig. 4.13 Classification errors Vs. K for a KNN classifier applied to the dataset day004 (training set: 7 measurements per class).

4.4 Experiment #3: characterisation by standard analytical techniques ("Flegrei")

4.4.1 Description of the experiment

An experiment made on 11 water samples taken from 11 different sites, in the area of Campi Flegrei (Naples, Italy) is described here. The interesting aspect, that makes this case study a kind of natural benchmark coming from the real world, lies in the fact that different kinds of water with different chemical characteristics can be found in a small area of a few sq. Km. Their different peculiarities well suit the concept of making a general characterisation of the water by giving an aggregate chemical information, as it is one of the major strength of an e-tongue system [5].

A check of the discrimination capability has been performed with the above said water samples, by comparing the e-tongue results with classical classification/clustering based on traditional chemical analytical measurements.

Projections of the observed vectors (in the feature space) onto a tri-dimensional representation space are compared with the PCA (Principal Components Analysis) of the analytical data, by taking the first 3 Principal Components and representing them.

The part of Italy where the experiment has been conducted is shown in the map of Italy in Fig. 4.14. A detailed map showing the location of the sampling points can be found in

Fig. 4.15.

The essential information about the measurement procedure followed in this case study is given in Tab. 4.12, while the configuration of the experimental setup is briefly described in Tab. 4.13.

Fig. 4.14 The area of Italy where water samples of the dataset "Flegrei" have been collected.

Fig. 4.15 Location of the sampling points in the experiment "Flegrei".

	Sample	HS	CO2	NH4	HCO3	Cl	SO4	Na	K	Ca	Mg	Br	NO3	F	Li	B
1	Rist. Sabatino	0.88	0.11	0.11	372	118	60	199	23	33	7.9	0.15	0.95	8.62	<0.01	0.49
2	Fondi di Baia	0.72	0.11	0.06	201	182	250	102	43	154	29	0.59	32.5	1.92	0.01	0.28
3	Serre Mauriello	0.72	0.21	0.08	342	118	52	256	29	95	4.1	0.55	4.1	8.25	0.20	0.71
4	Ascolese Cuma	0.70	0.24	0.1	305	64	51	341	27	20	2.9	0.25	1.75	8.75	0.07	0.68
5	Via Licola Km52	0.69	0.23	0	333	65	68	169	15	33	1	0.18	1.12	7.75	0.12	0.34
6	Neapolis Procida	0.65	0.15	0.07	500	79	102	120	50	111	22	0.25	5.6	1.95	0.01	0.51
7	Tirrenia Marmi	0.70	0.15	0	275	42	43	55	40	54	12	0.12	13	1.62	0.01	0.36
8	Puteolane	0.83	0.81	0.26	2061	2125	650	2281	205	32	24	7.15	8.7	2.55	1.17	16.00
9	Averno	2.15	0.21	0.15	342	600	170	507	59	28	12	1.72	0.15	10.5	0.48	1.50
10	Castello di Baia	1.38	0.33	0.25	817	530	575	624	88	173	68	1.75	40	3.38	0.29	1.11
11	Agnano	0.62	17.14	4.64	1415	2900	375	1956	287	273	70	6.44	0.51	1.9	1.38	13.00

Tab. 4.11 Measurement locations and chemical analyses for the experiment "Flegrei".

Time coverage	Two days for taking samples on the field, half day for the measurements
Number of classes	11
Kind of classes	Superficial and downwell water samples carried in sealed containers
Training set	No
Total dataset	33 measurements, 3 per each class
Notes	Water samples have been collected on the field and taken to the laboratory after one week

Tab. 4.12 Measurement procedure of the experiment "Flegrei".

Working Electrodes	Au, Re, Pt, VC
Washing	H_2SO_4 1M + rinsing in distilled water
Notes	VC = Vitreous Carbon electrode. All the Working Electrodes in this experiment are in the form of wires about 5mm long, with the exception of the VC electrode, which is a standard laboratory electrode (rod) having 3mm diameter x 6mm long

Tab. 4.13 Experimental setup configuration of the experiment "Flegrei".

4.4.2 Results

One interesting aspect that emerges in analysing the results of this case study, lies in the different ordering of the DCT components as a function of the discrimination index used to determine such sorting order.

As it may be seen in Fig. 4.16 and Fig. 4.17, the usage of the two uni-variate discriminability indexes described in Paragraph 3.3.1 leads to different results in the ranking of the DCT components of the 4 observation vectors pertaining each Working Electrode.

The most apparent thing, lies in the fact that with the S_2 index the Au WE appears to be the most contributing to discrimination, due to the high density of DCT coefficients having a low sorting order. A general description about how the plots discussed are built can be found in Paragraph 3.3.3.

With the S_x index the Re WE shows to give the prevailing contribution to discrimination, as it may be well seen in Fig. 4.18 and Fig. 4.19, where the first 100 sorted DCT components are shown in detail.

This simple example gives a clear idea about the substantial practical importance of having multiple Working Electrodes. Even in situations where a certain WE appears to be the most informative one, a change in the discriminability index chosen, makes another WE the preferred one, according to the metric chosen.

Again, the feature selection domain chosen influences the contribution of the different WEs to discriminability. According to the metric asserted by the discriminability index S_x, the ranking of the components in the DWT and VLT domains has been calculated, with the results shown in Fig. 4.20 and Fig. 4.21. Once more, looking at these plots it comes clear how the signal analysis procedure may enhance the importance of different WEs to the discrimination capability of each single component; in fact, in the DWT domain the Re and Au WEs give a prevailing contribution, while in the VLT domain the Au and VC ones would be preferred.

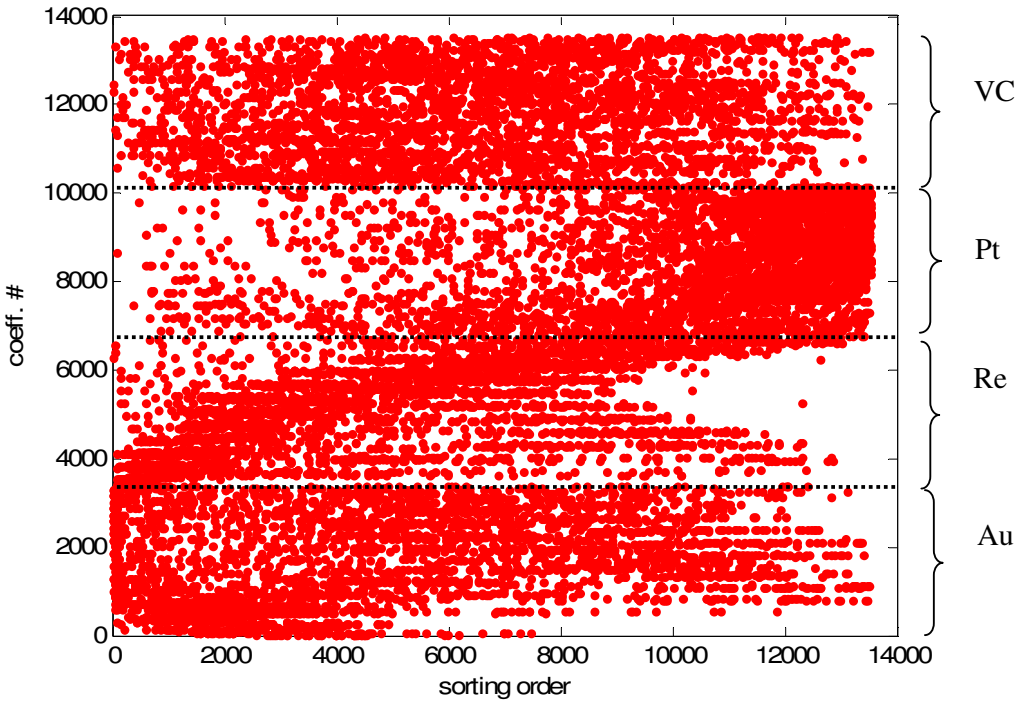

Fig. 4.16 Ranking of the components of the observation space in the DCT domain. Experiment "Flegrei". Index S_2.

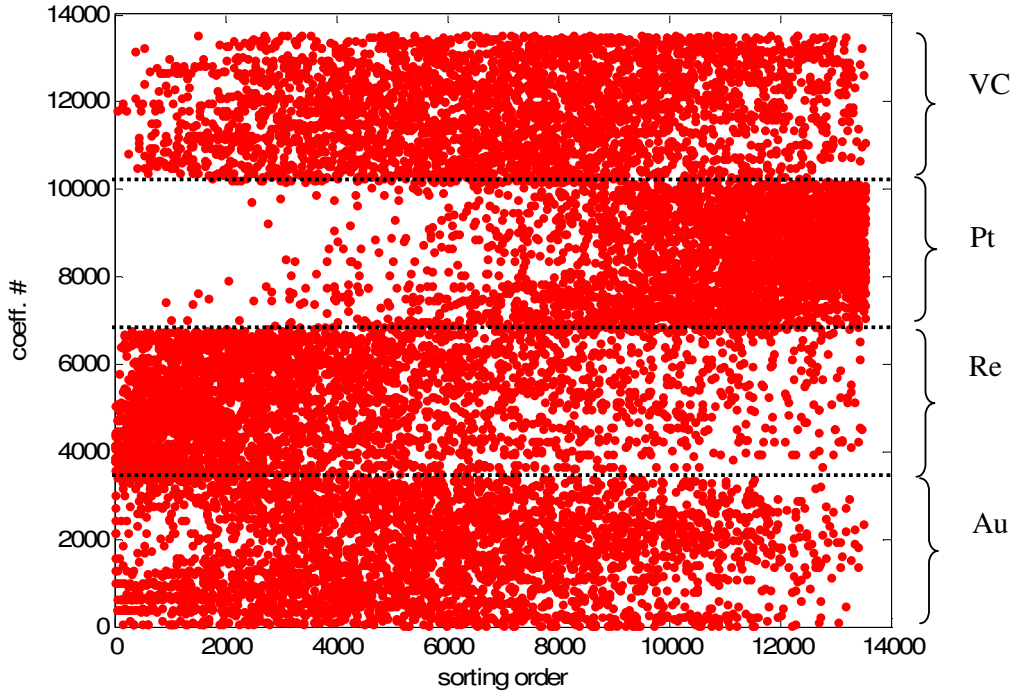

Fig. 4.17 Ranking of the components of the observation space in the DCT domain. Experiment "Flegrei". Index S_x

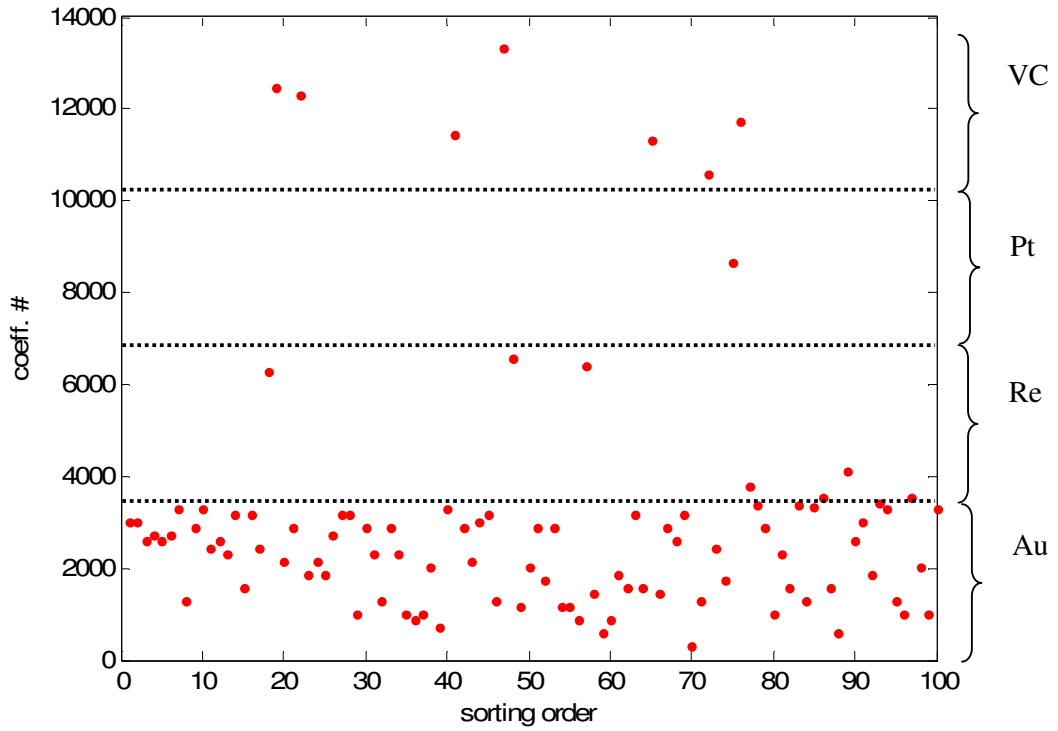

Fig. 4.18 Ranking of the first 100 components of the observation space in the DCT domain. Experiment "Flegrei".

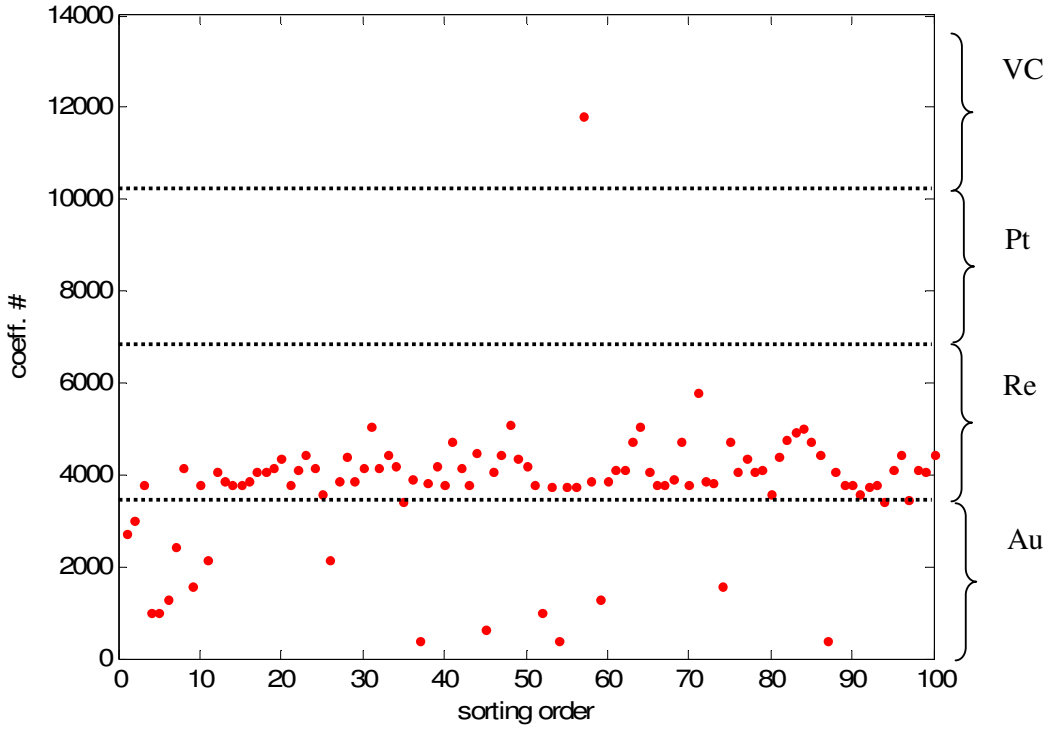

Fig. 4.19 Ranking of the first 100 components of the observation space in the DCT domain. Experiment "Flegrei".

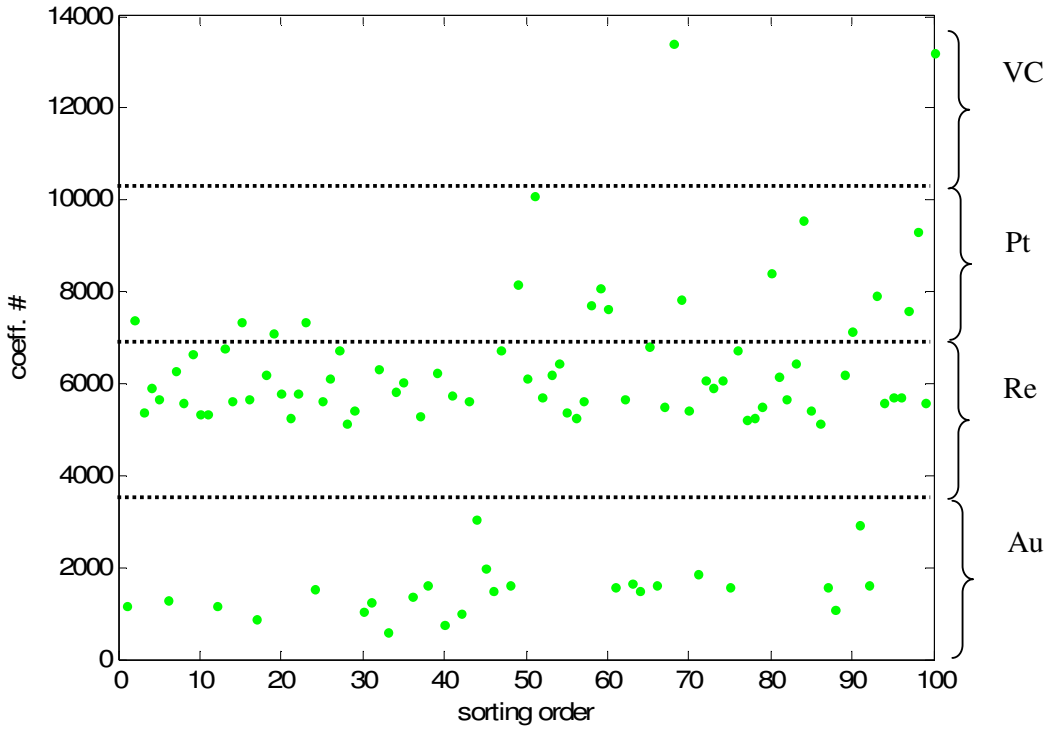

Fig. 4.20 Ranking of the first 100 components of the observation space in the DWT domain. Experiment "Flegrei".

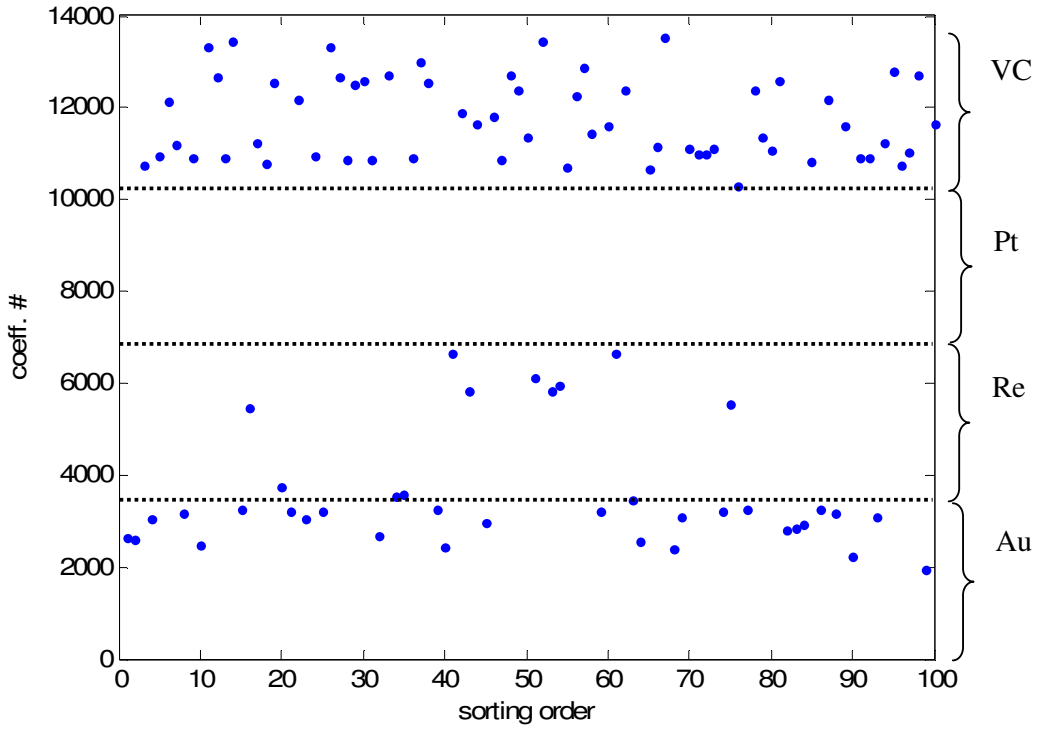

Fig. 4.21 Ranking of the first 100 components of the observation space in the VLT domain. Experiment "Flegrei".

Using the same approach presented in the previous case studies, a sequential forward search strategy has been tested on the whole dataset, making use of the S_x index in the DCT domain.

Tab. 4.14 and Fig. 4.22 show the feature extraction parameters and the trend of the discriminability index S_x for grouped coefficients in the forward search strategy. Fig. 4.23 shows the projections of the 33 measurements onto a tri-dimensional representation subspace, after reducing the dimensionality of the whole observations to the feature space, having dimensionality l'=4.

Discriminability Index	S_x
Search strategy	Sequential forward
Feature space dimensionality	$l' = 4$
Notes	The feature set has been searched for in the DCT domain

Tab. 4.14 Feature extraction parameters for the dataset "Flegrei" (sequential forward strategy).

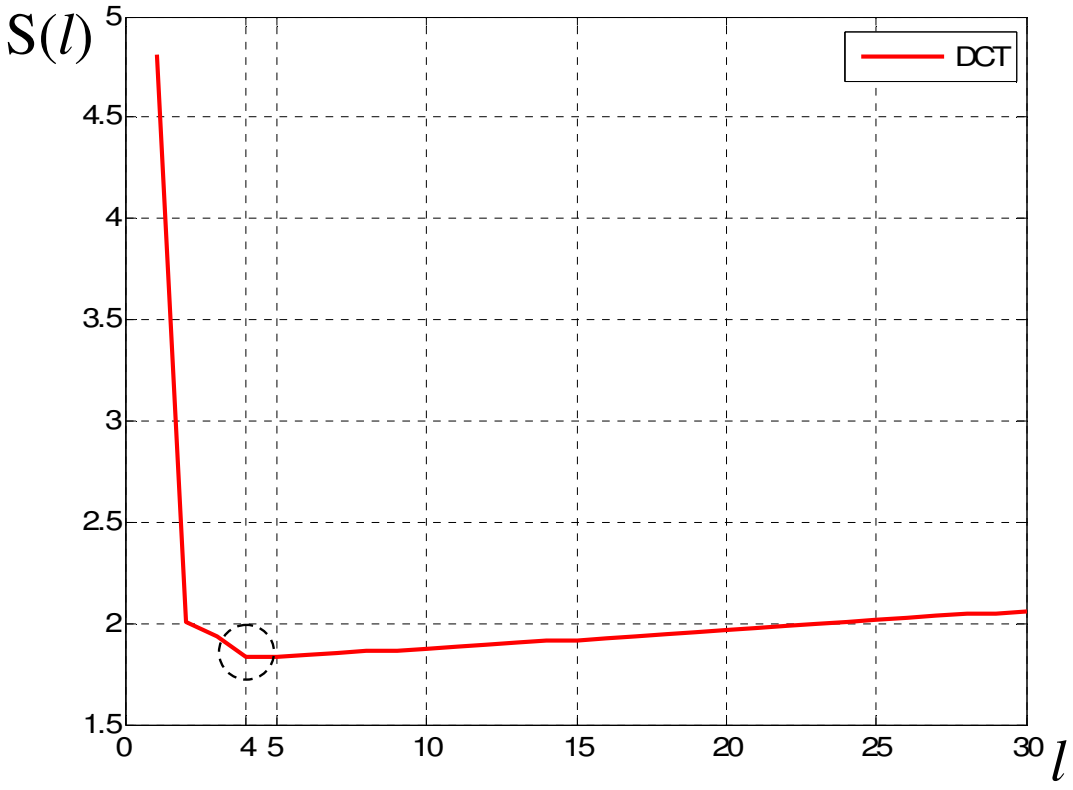

Fig. 4.22 Discriminability index trend obtained by using a sequential forward search strategy. Sx index. Experiment "Flegrei".

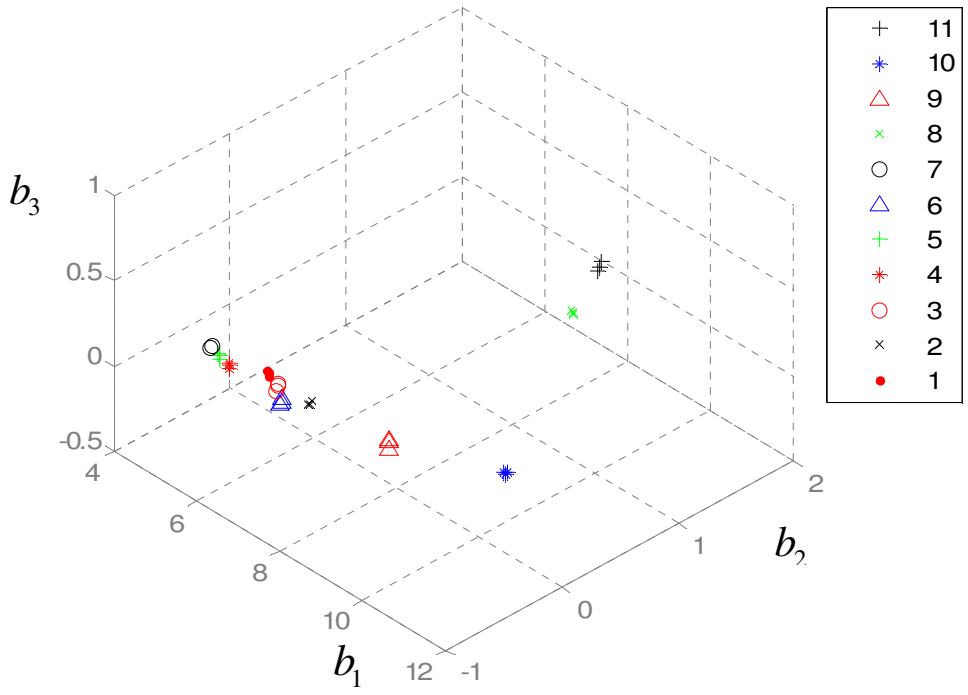

Fig. 4.23 Projections of the observations of the full dataset ("Flegrei") onto the tri-dimensional representation space (sequential forward strategy).

On the same dataset a preliminary experiment to determine how the characterisation capability of the e-tongue can be compared to a classical characterisation scheme based on traditional chemical measurements. To make this comparison, a very simple approach based on the following two steps has been performed:

- the first 2000 DCT components of each measurement, determined according to the S_2 metric in the uni-variate space, have been taken,
- data have been projected onto the tri-dimensional representation space, built as described in Paragraph 3.4.

The two plots shown in Fig. 4.24a,b allow the comparison between the classical analytical classification results and the e-tongue results.

Principal Components Analysis has been applied to the chemical analytical measurements; data of such analyses are expressed in *mg/l*, as it can be found in Tab. 4.11. No conversion has been performed, since the Principal Components can be freely scaled also after transforming. The effect of scaling on the components' variability, in general, changes the distribution of the measurements represented in the Principal Components' domain. The distribution of the observations coming from the voltammetric measurements, after DCT transform and dimensionality reduction, is obtained by projection of the observed transformed vectors onto the base formed by the first 3 vectors determined in accordance to the procedure described in Paragraph 3.4.

It must be remarked that e-tongue results are chemically altered with respect to what would be obtained by on-site measurements, due to the fact that Redox potentials [6] do not carry meaningful information, after the contact of the sampled water with the atmospheric gas. As a consequence, the main discriminating factor for the e-tongue measurements appears to be the chemical composition of the sample, more than its Redox properties, which are basically lost.

It appears immediately clear how the two thermal water samples (Terme Puteolane and Agnano) are strongly discriminated in both the representations. It is also interesting that the "Averno", "Castello di Baia" and "Fondi di Baia" samples are discriminated in both domains, and are actually distinguishable through the chemical parameters that mainly contribute to the Principal Components' variability.

The preliminarity of this investigation and the fact that the e-tongue measurements have not been made on-site, do not allow the interpretation of these representations with greater detail, thus it becomes difficult to explain the clustering of samples #1,3,4,5,6,7, which are split into two groups of three measurements according to the e-tongue, while they are grouped in one whole cluster in the PCA representation.

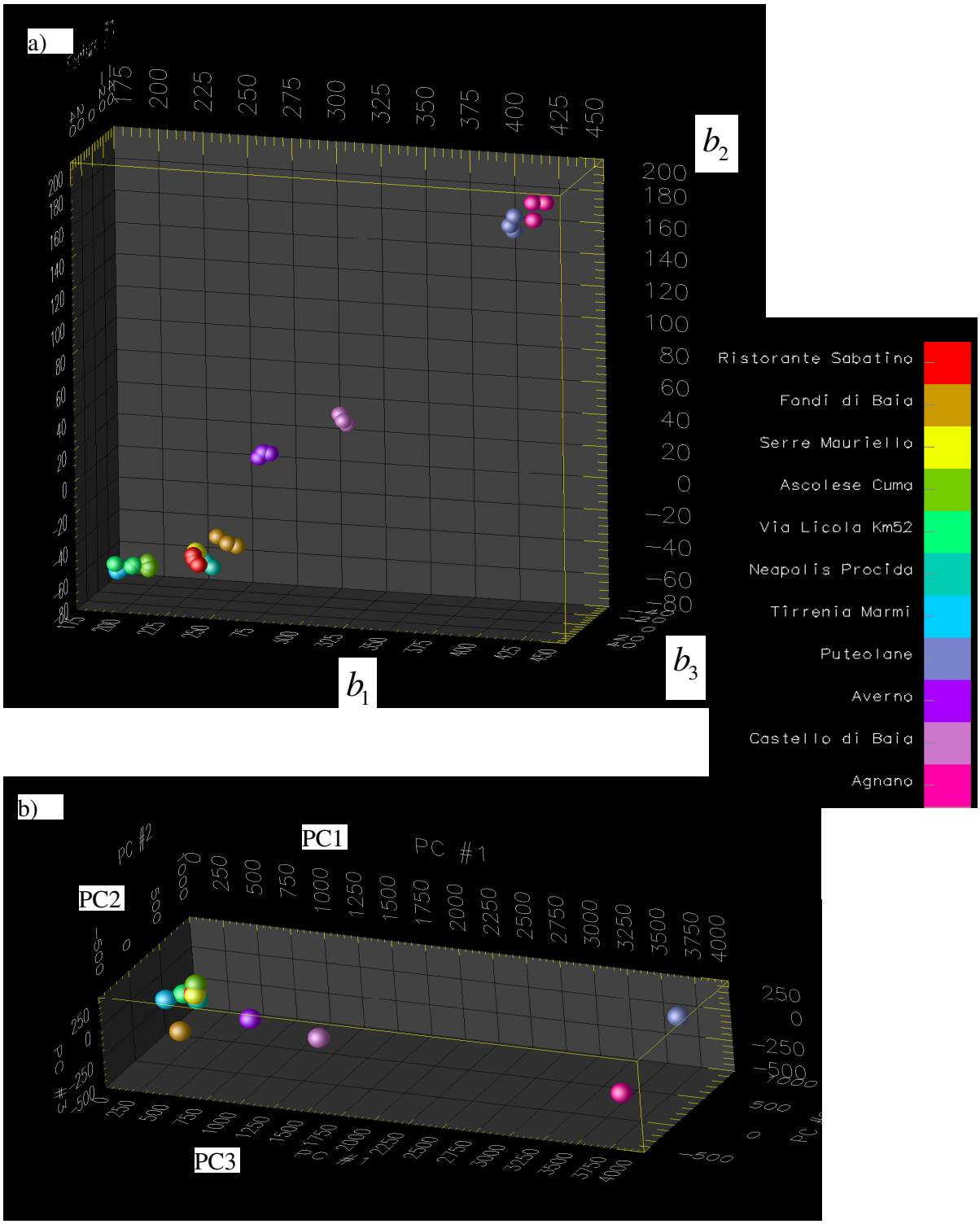

Fig. 4.24 a) Projection of the 33 meaurements (11 water samples) of the experiment "Flegrei" onto a tri-dimensional representatiotn space. b) Representation of the same 11 water samples in the space spanned by the first three Principal Components.

4.5 Experiment #4: detection of pollutants ("LaGabella")

4.5.1 Description of the experiment

This is an experiment intended to check the discrimination between water samples taken at the input and the output of an Active Carbon filter pack, as a possible technique for detecting changes with respect to the presence of some pollutant or toxic contaminant, which undergoes a known abatment by the filter.

This case study regards a slight pollution by an old industrial activity in the proximity of Pisa (Tuscany, Italy), at "La Gabella" wells, where water is extracted by the local drinkable water distribution network.

Four wells are in close proximity of three industrial activities now closed, where the pollution due to trichloroethylene and tetrachloroethylene appears to come from the area of two old oil production plants. Moreover, a small foundry was located in the area. In Fig. 4.25, industrial areas are in red and groundwater flow direction is in blue.

Pollution by trichloroethylene and tetrachloroethylene is the only parameter outside the limits for drinkable water. The test on sampled water, both filtered and not, has proven the possibility to detect the presence of pollutant with an approach based on the difference between measurements taken at the inlet and the outlet of the filter, respectively.

Analytical measurements made on water samples, taken both at the inlet and at the outlet of the carbon filters, have been done; results are presented in Tab. 4.15, where the values of each parameter at the input of the filter (raw water) and at its output, are compared. The only parameter affected by said slight pollution is shown on an orange background. The effect of the filters on this parameter is substantial, since it lowers the concentration of the contaminants (trichloroethylene and tetrachloroethylene) below the lower detectable limit of the analytical instrumentation used, and far below the limit imposed by the Italian regulation about water quality.

This test is very demanding for all the measurement and data processing chain, due to the very low concentration of the chemical contaminant to be seen: 22.4 $\mu g/l$ means 22.4 *ppb* (parts per billion) in mass ratio.

The measurement procedure is described in Tab. 4.16. The sequence, to void some kind of "memory effect" on the electrodes, has been done by measuring the filtered water first, then the raw water, then the filtered one again.

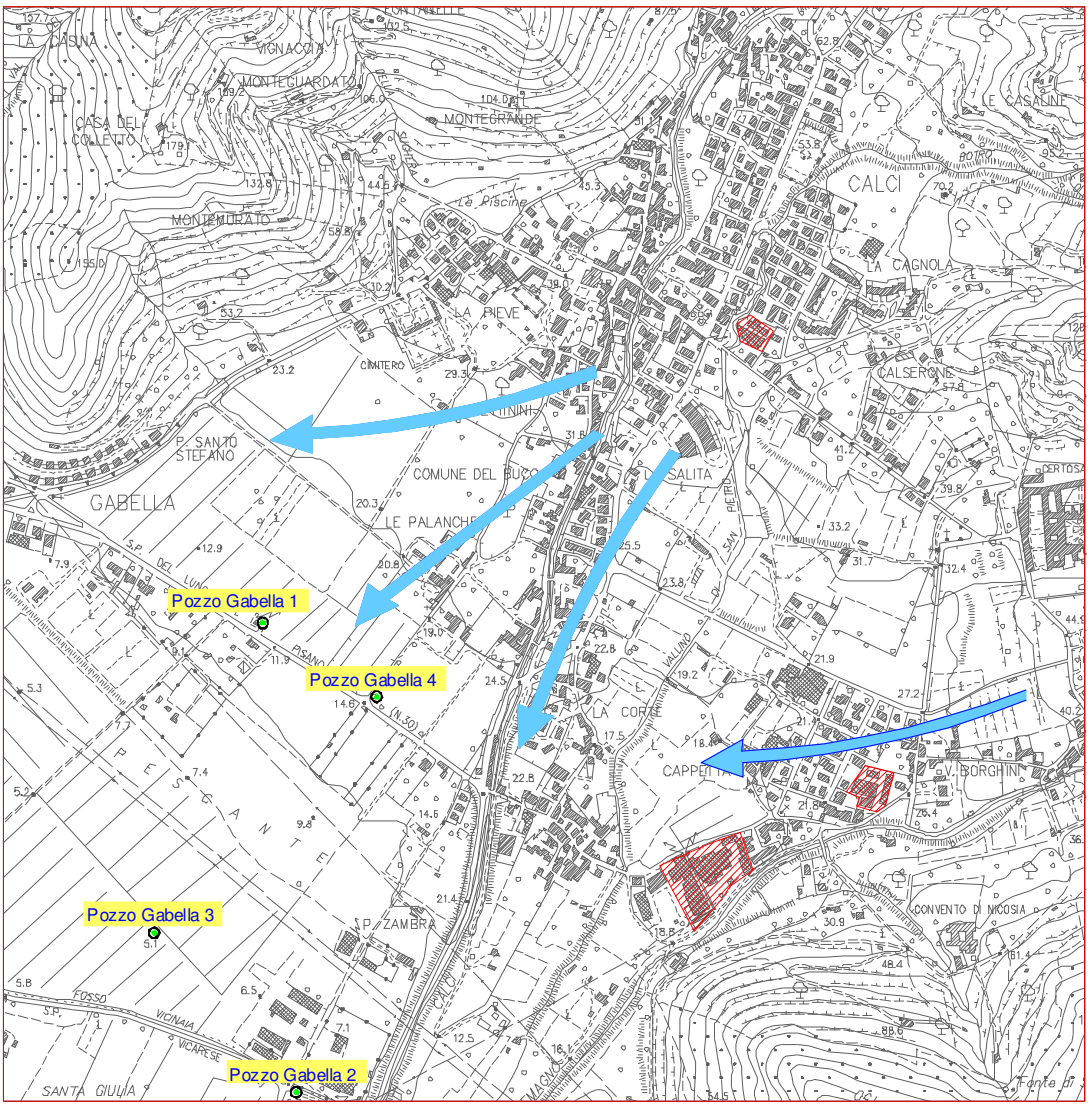

Fig. 4.25 Map of the area where water samples for the dataset "LaGabella" have been taken.

Parameter	Unit	Before GAC filters	After GAC filters	D. Lgs n° 31/01 Limit values (Italian regulation)
Hydrogen ion concentration	pH	6.8	6.9	6.5≤ pH ≤9.5
Conductivity	*µS/cm* @ 20°C	334	330	2500
Turbidity	NTU	0.54	0.37	-
Fluorides	*mg/l* F	0.09	0.06	1.50
Chlorides	*mg/l* Cl	37.8	35.9	250
Nitrates	*mg/l* NO_3	15.1	16.8	50
Sulphates	*mg/l* SO4	12.8	15.3	250
Sodium	*mg/l* Na	28.2	26.1	200
Total Hardness	°F	10.7	11.1	-
Ammonium	*mg/l* NH_4	<0.05	<0.05	0.50
Nitrites	*mg/l* NO_2	<0.05	<0.05	0.10
Nitrates and Nitrites relationship	-	0.32	0.35	{[NO3]/50+[NO2]/0.1≤}1
Boron	*mg/l* B	<0.05	<0.05	1.0
Iron	*µg/l* Fe	17	23	200
Manganese	*µg/l* Mn	<5	<5	50
1,2 Dichloroethane	*µg/l*	<0.2	<0.2	3.0
Trichloroethylene and tetrachloroethylene	***µg/l***	**22.4**	**<0.1**	**10**
Total Trihalomethanes	*µg/l*	<0.1	<0.1	30
Benzene	*µg/l*	<0.1	<0.1	1.0
Cadmium	*µg/l*Cd	<0.5	<0.5	5.0
Chromium	*µg/l* Cr	<1	<1	50
Lead	*µg/l* Pb	<3	<3	25
Nickel	*µg/l* Ni	<5	<5	20
Copper	*mg/l* Cu	0.001	0.001	1.0
Mercury	*µg/l* Hg	<0.1	<0.1	1.0
Arsenic	*µg/l* As	<3	<3	10
PAH	*µg/l*	<0.01	<0.01	0.10
Pesticides and related products	*µg/l*	<0.01	<0.01	0.10 (0.50)
Dry residues (calculated)	*mg/l*	187	192	-

Tab. 4.15 Chemical analytical measurements of raw water for the "LaGabella" case study.

Time coverage	One day, half for taking samples, half for performing the measurements
Number of classes	2
Kind of classes	Filter outlet (filtered water), filter inlet (raw water)
Training set	-
Total dataset	4 measurements each class
Notes	The measurement sequence has been as follows: 2 samples of outlet water 4 samples of inlet water 2 samples of outlet water

Tab. 4.16 Measurement procedure of the experiment "LaGabella".

Working Electrodes	Au, Re, Pt, Pd
Washing	Ultrasonic
Notes	Ultrasonic washing made in distilled water between each batch of measurements

Tab. 4.17 Experimental setup configuration of the experiment “LaGabella”.

4.5.2 Results

The first 4 measurements (2 on filtered water and 2 on raw water) have been used to determine the components of the feature space and the base of the representation space. The following 4 measurements (performed in reverse order with respect to the first batch) have been projected onto the same representation base, after extracting the previously determined components.

A simple sequential procedure with the S_x index, calculated for grouped coefficients of increasing l dimensionality, has been applied to this dataset. As a general guideline, in a case study with a small number of observations such this one, a very limited optimal dimensionality is expected; under this respect, a simple sequential strategy has given sufficient reduction in dimensionality ($l' = 3$). The essential parameters are listed in Tab. 4.18, while the behaviour of the S_x index is plotted in Fig. 4.26 and Fig. 4.27. Calculation of the index has been performed both in the DCT, the DWT (again COIF5) and time (VLT) domain.

From the detailed plot shown in Fig. 4.27, it comes clear how both DCT and DWT domain lead to a sub-optimal dimensionality $l' = 3$, even with a similar value for the index.

Discriminability Index	S_x
Search strategy	Simple sequential
Feature space dimensionality	$l' = 3$
Notes	The feature set has been searched for in the DCT, DWT and VLT domains

Tab. 4.18 Feature extraction parameters for the dataset "LaGabella" (simple sequential strategy).

In Fig. 4.28 the projections of all the observations for this case study are plotted. Feature extraction has been applied to observations in the DCT domain, grouped according to Fig. 3.11. Larger size of the markers has been given to the first 4 measurements, used to determine the feature extraction and representation parameters. Smaller size has been given to the markers relating to second batch of measurements, which are directly projected onto the representation space after having extracted the previously determined features.

In a similar fashion to what happened with the dataset "Day004", after training the feature extraction procedure with half of the dataset, successive measurements are still represented with a noticeable discrimination between classes, when reduced in dimensionality and projected according to the training parameters.

Feature extraction has also been applied to observations in the DWT domain, again grouped according to Fig. 3.11. In Fig. 4.29 the projections of all the observations for this case study in the DWT domain are plotted.

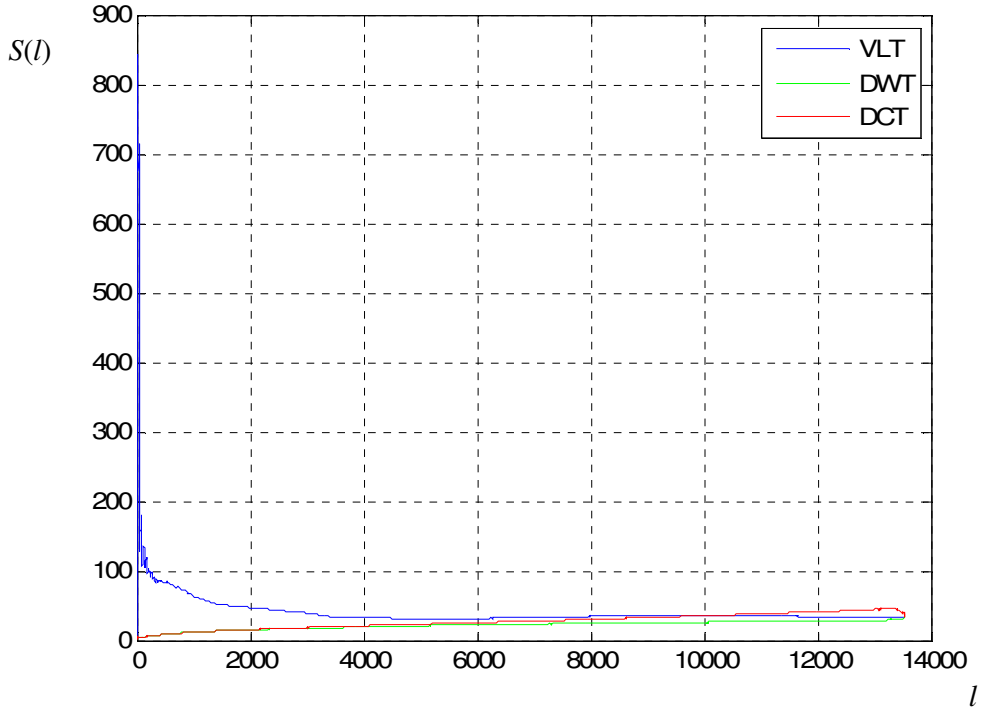

Fig. 4.26 Discrimination capability index Sx for the *l*-dimensional subspace generated by grouping the first *l* sorted components. Experiment: "LaGabella".

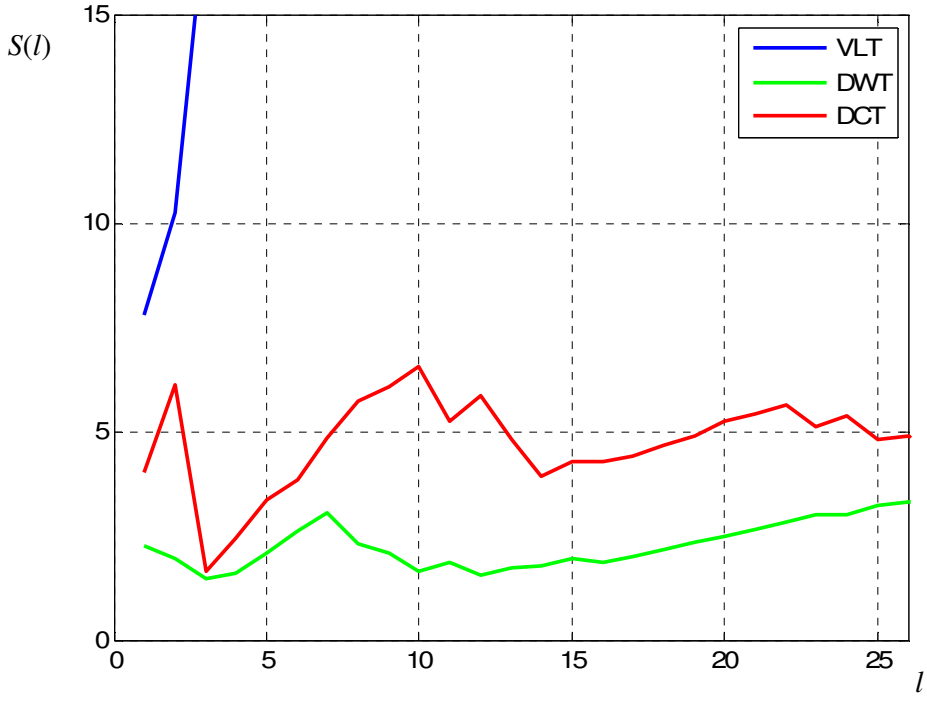

Fig. 4.27 Discrimination capability index Sx for the *l*-dimensional subspace generated by grouping the first *l* sorted components: detail of the first 26 values. Experiment: "LaGabella".

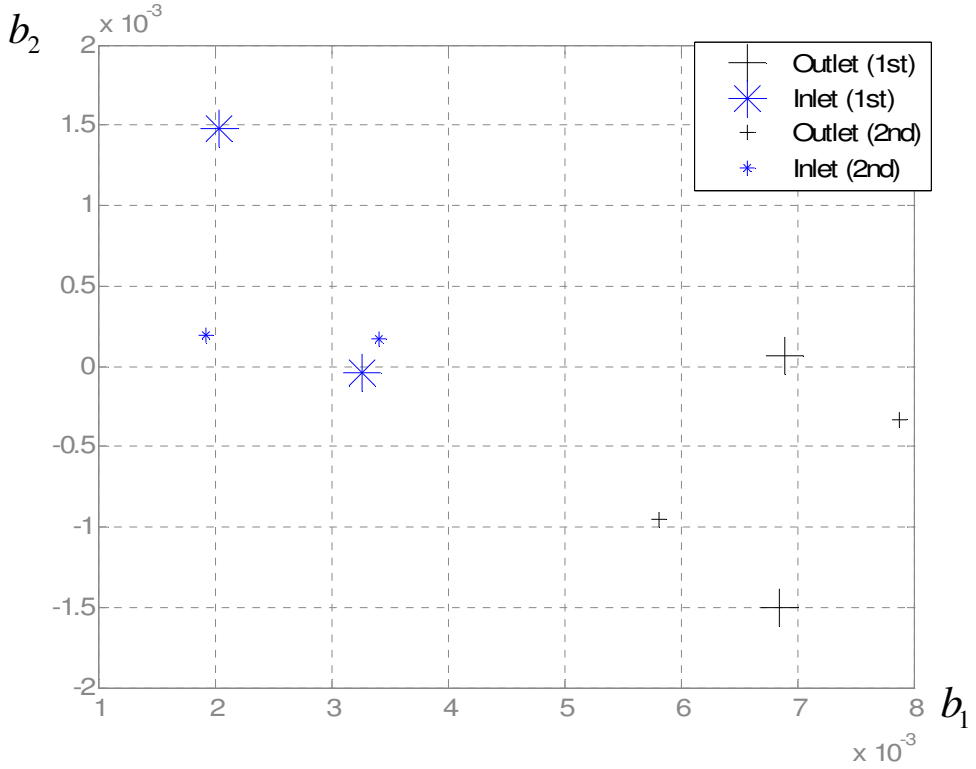

Fig. 4.28 Projections of the observations of the full dataset ("LaGabella") onto the bi-dimensional representation space (simple sequential strategy, DCT domain).

Again, also in the DWT domain representation, larger size of the markers is given to the measurements used to set-up the feature extraction procedure, smaller size is given to the second batch of measurements, that have been simply projected onto the previously determined base, after extracting the selected components in the DWT domain. The three representation planes (1-2, 1-3 and 2-3) are shown; discrimination between inlet and outlet samples is apparent in all the three plots.

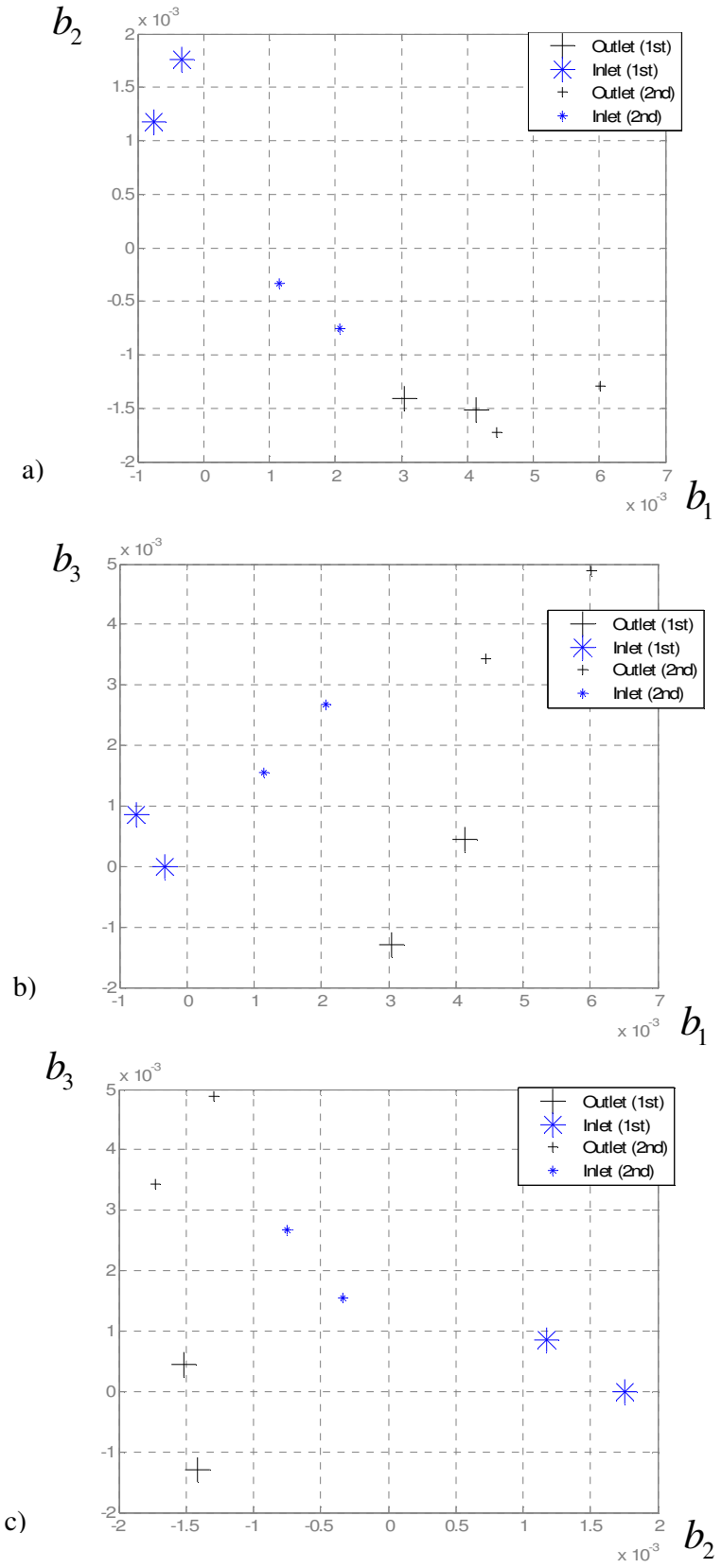

Fig. 4.29 Projections of the "LaGabella" dataset onto the representation space in the DWT domain. a) Base vectors 1-2. b) Base vectors 1-3. c) Base vectors 2-3.

In conclusion, the test on sampled water, both filtered and not, has proven the possibility to detect the presence of a contaminant with an approach based on the difference between measurements taken at the inlet and the outlet of the filter, respectively. This is a feasible technique to detect changes on a long term basis.

The same approach has been used in another case study ("Zamponi" dataset [7]), that is not presented in this Thesis due to the high similarity with this one, obtaining very promising results.

4.6 Experiment #5: drifts in long term measurements ("Testt0102")

4.6.1 Description of the experiment

The case studies seen up to now have not taken into account the long term drifts that may arise when performing measurements continuously on the same sample of water, which may be flowing or not. As a preliminary check of the possible trends of the signals considered, two measurement sessions lasting 24 hours each, performed on two different water samples, have been done.

Moreover, since the comparison between a reference liquid and the water under monitoring is a reasonable procedure to perform practical measurements on a long term basis, this case study provides information about the feasibility of such an approach.

Time coverage	Two measurement sessions, 24 hours each. One measurement every 10 minutes
Number of classes	2 (each water sample continuously measured during its 24-hours session)
Kind of classes	Mineral water brands (bottled) 144 measurements per each class
Training set	-
Total dataset	288 measuements, 144 per each class
Notes	Water temperature has been logged during the experiment to check day-night effects due to thermal drifts

Tab. 4.19 Measurement procedure of the experiment "Testt0102".

Working Electrodes	Au, Re, Pt, Pd
Washing	Ultrasonic
Notes	Ultrasonic washing made in distilled water between the two batches of measurements

Tab. 4.20 Experimental setup configuration of the experiment "Testt0102".

4.6.2 Results

For this dataset, having two classes and a relatively high number of observations per class (144 per each WE), the chosen metric for the discriminability assessment has been the "Euclidean" S_2 index. Feature selection has been done with a simple sequential search strategy.

The feature extraction parameters for the dataset "Testt0102" can be found in Tab. 4.21. Plots of the values taken by the index S_2 vs. dimensionality, according to the chosen search strategy, are shown in Fig. 4.30 and Fig. 4.31, with a detail on the first 50 values, in all the three domains investigated (DCT, DWT, VLT).

Discriminability Index	S_2
Search strategy	Simple sequential
Feature space dimensionality	$l' = 1$
Notes	The feature set has been searched for in the DCT, DWT and VLT domains

Tab. 4.21 Feature extraction parameters for the dataset "Testt0102" (simple sequential strategy).

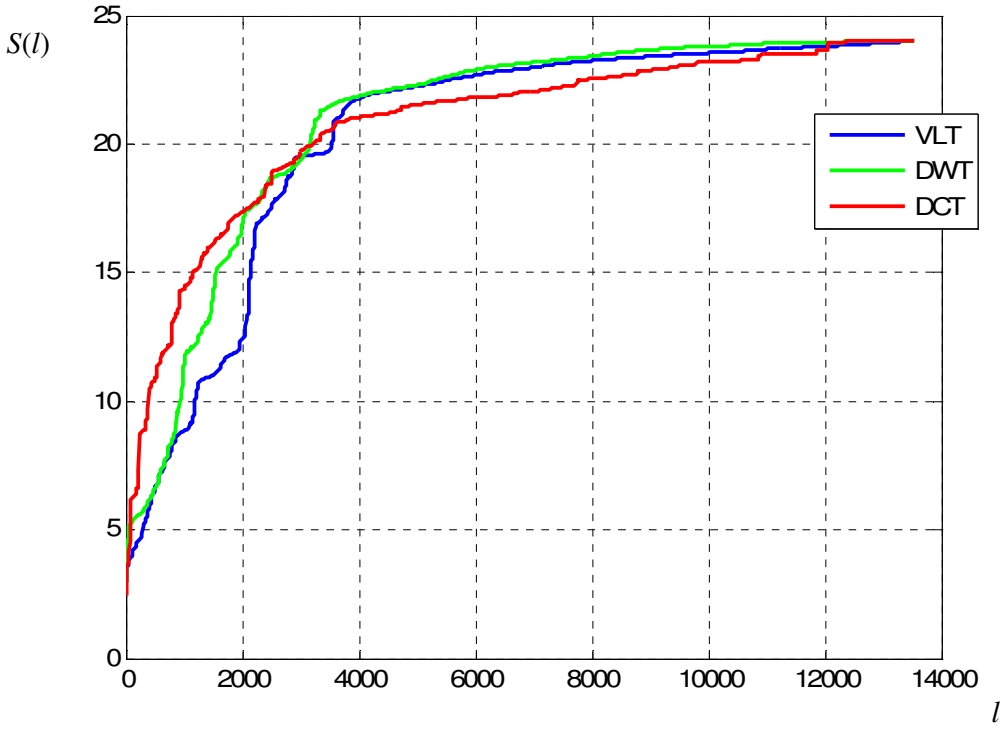

Fig. 4.30 Discrimination capability index S2 for the *l*-dimensional subspace generated by grouping the first *l* sorted components. Experiment: "Testt0102".

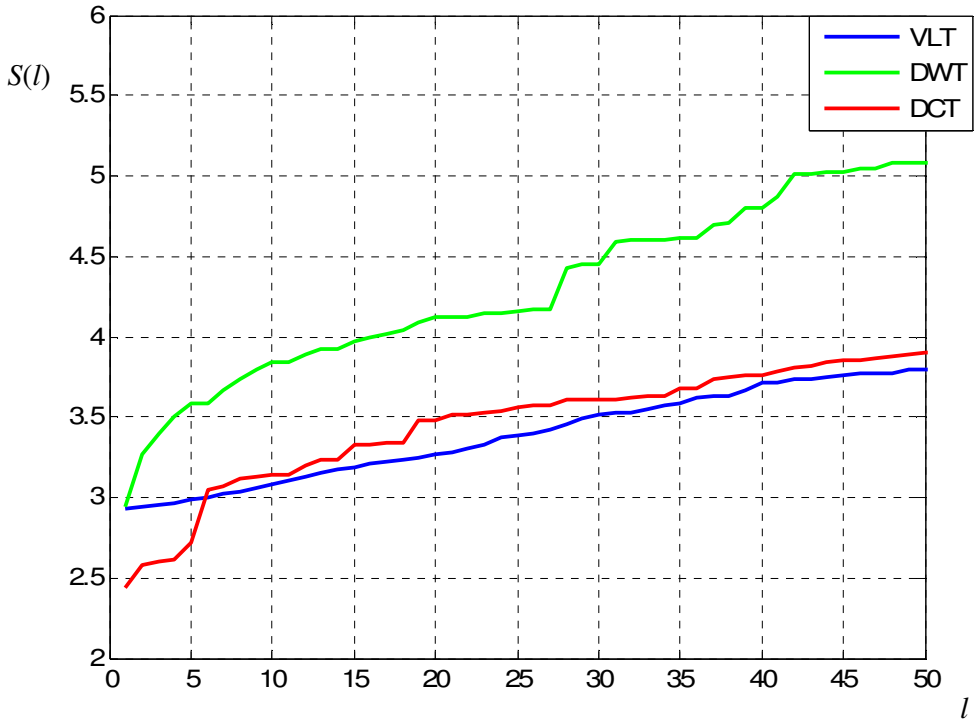

Fig. 4.31 Discrimination capability index S_2 for the *l*-dimensional subspace generated by grouping the first *l* sorted components: detail of the first 50 values. Experiment: "Testt0102".

Two aspects come immediately clear by analysing the plots in Fig. 4.30 and Fig. 4.31:

- DCT has the best performance according to the chosen metric,
- sub-optimal dimensionality is *l*'=1 (extraction of the first sorted DCT component is the best feature set)

It is interesting that the first DCT component in ranking corresponds to the coefficient #141 of the DCT transform of the observations from the Pt Working Electrode. The corresponding frequency is 1.0417 Hz, substantially identical to the pulse repetition frequency of the LAPV signal.

This happens not surprisingly, due to the fact that this is an experiment with two classes only, but this does not mean that the set-up of an efficient change detector would be done by extracting a mono-dimensional feature subspace.

Since one of the aims of this approach is to track eventual changes in the chemistry of water, a continuous monitoring device should not be reasonably trained to work in a mono-dimensional feature space, but, being trained with a number of classes corresponding to a set of known possible pollutants, an higher dimensionality would be expected.

The ranking of the components of the observation space in the DCT, DWT and VLT domain takes also meaningful information. Plots obtained according to the conditions listed in Tab. 4.21, are shown in Fig. 4.32, Fig. 4.33 and Fig. 4.34 respectively for the DCT, DWT (COIF5) and VLT (time) domain. Note that, differently from Fig. 3.21, DWT approximation and detail components are appended for each WE, instead of being plotted separately.

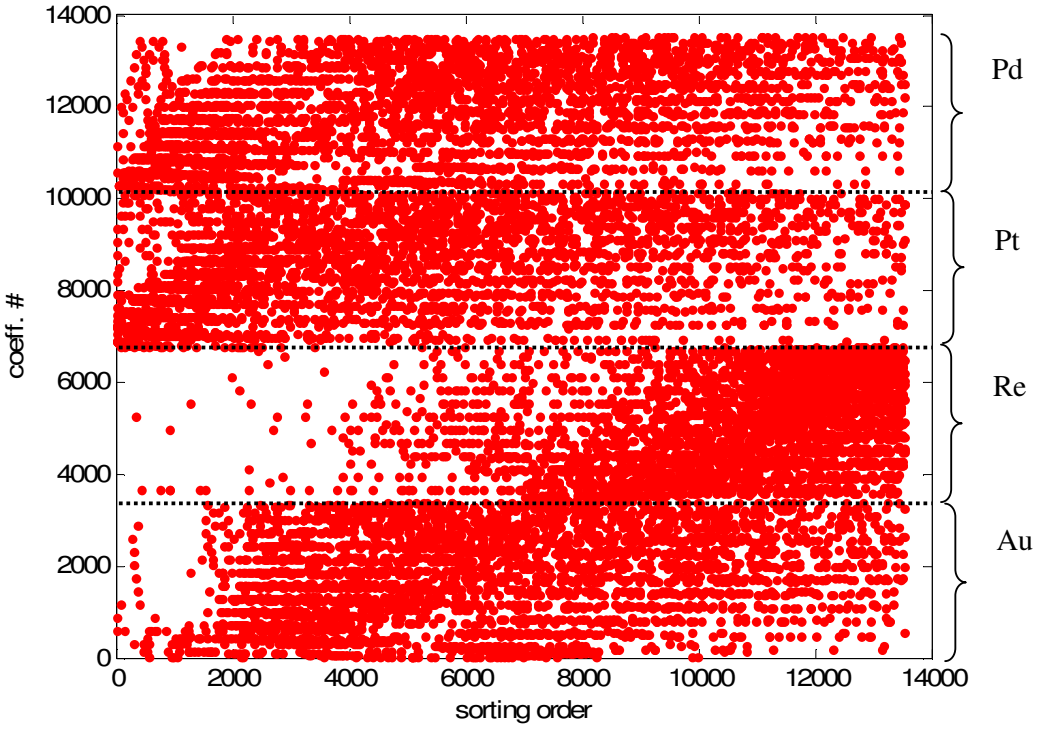

Fig. 4.32 Ranking of the components of the observation space in the DCT domain. Experiment "Testt0102".

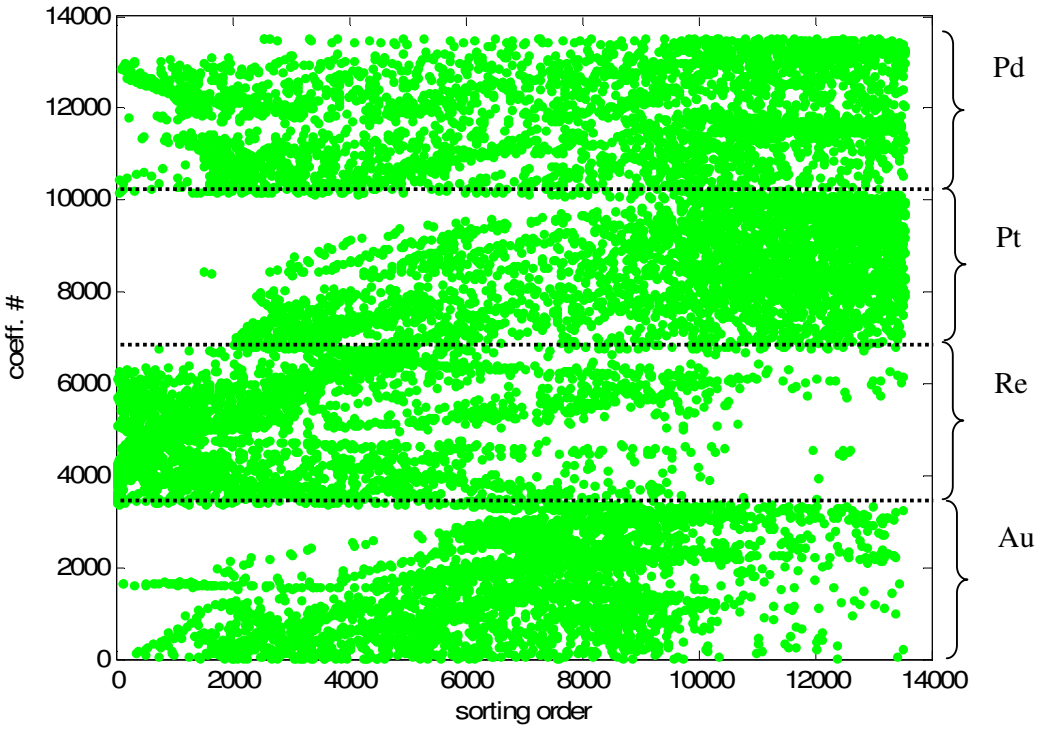

Fig. 4.33 Ranking of the components of the observation space in the DWT domain. Experiment "Testt0102".

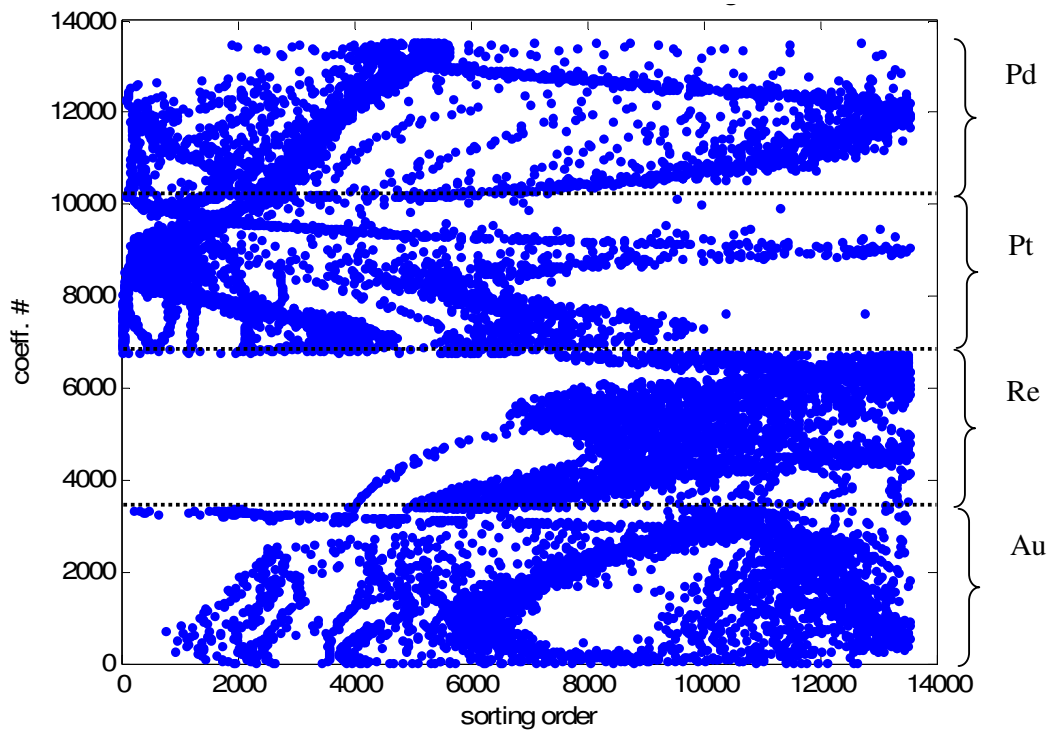

Fig. 4.34 Ranking of the components of the observation space in the DCT domain. Experiment "Testt0102".

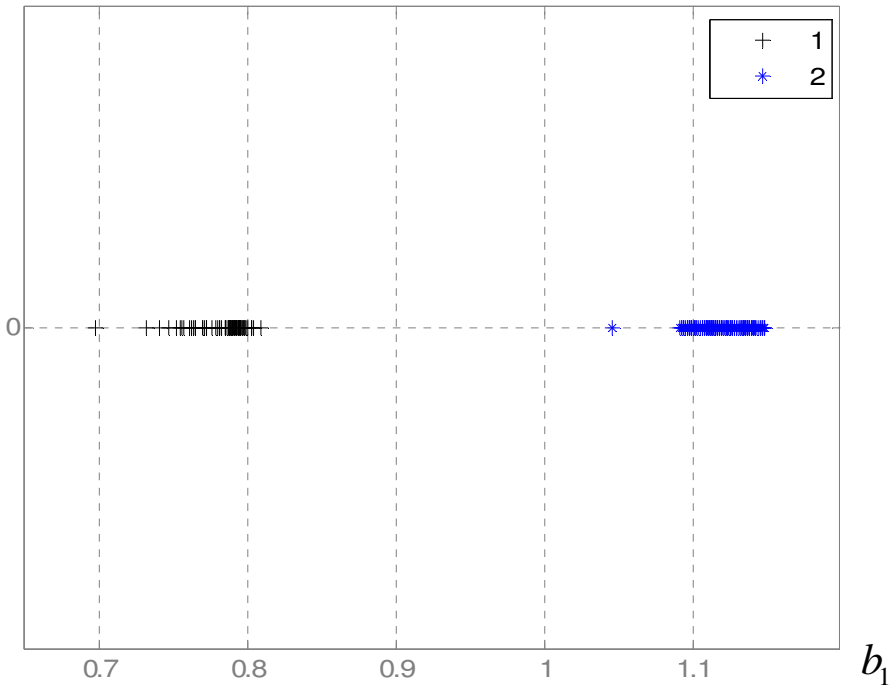

Fig. 4.35 Projections of all the 288 measurements of the dataset "Testt0102" onto the mono-dimensional space corresponding to the first feature in ranking.

Class #	Description
1	"Lora Recoaro" mineral water
2	"Gaia" mineral water

Tab. 4.22 Description of classes for the dataset "Testt0102".

Inspection of Fig. 4.32, Fig. 4.33 and Fig. 4.34, leads to a number of interesting considerations:

- working in different domains, different WEs play a different role,
- Re electrode, which is the least in ranking in the DCT domain, comes to be the most important in the DWT domain, confirming similar situations in the previous examples,
- in the time (VLT) domain, the final values of the current signal of the Au WE are high in ranking
- also, high contribution to discrimination in the time domain comes from most of the signal coming from the Pt electrode

The plots discussed here, especially for the observations in the time domain, give meaningful information about the electrochemical peculiarities of these experiments. As an example, in this case study, the above mentioned considerations about the VLT signals lead to the conclusion that the Au Working Electrode is the one which better enhances the effect of the diffusion-limited current, having a relatively high contribution to discriminability in the final values of the Au WE current transients.

Fig. 4.35 shows the projections of all the 288 measurements of the dataset "Testt0102" onto the mono-dimensional space spanned by the DCT component extracted by the feature selection strategy. In other words, the value of the coefficient #141 in the DCT expansion of each of the 288 observations relating to the Pt WE is plotted onto the x-axis of Fig. 4.35.

For mere representation purposes, the projection onto a representation subspace obtained according to the procedure described in Paragraph 3.4 after the extraction of the first two feature in the DCT domain, is shown in Fig. 4.33.

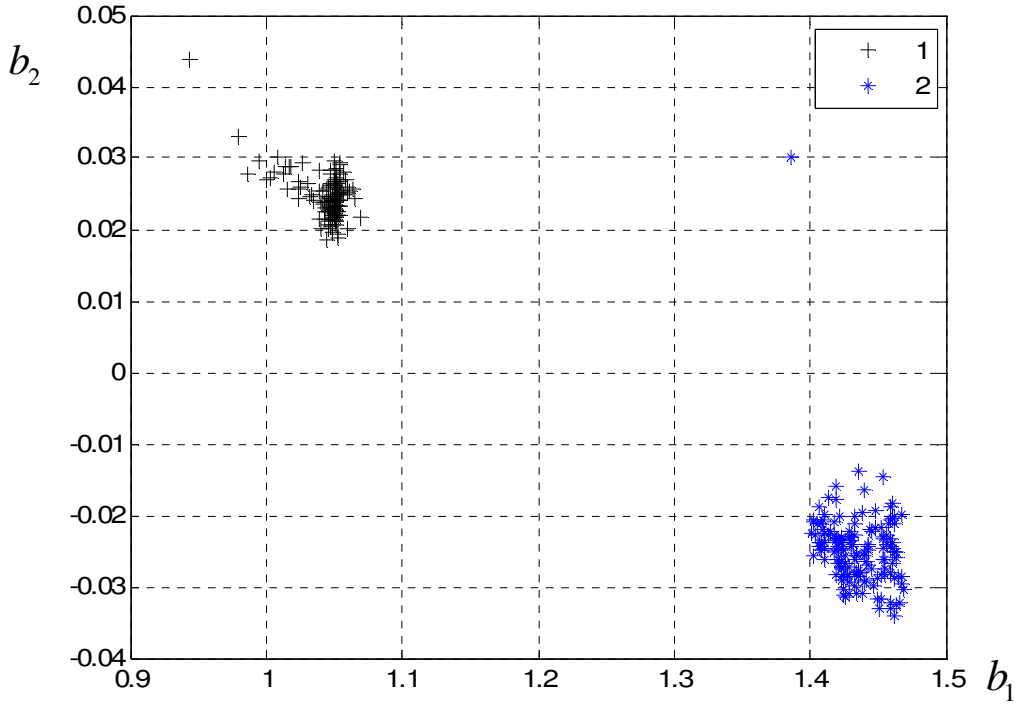

Fig. 4.36 Projections of all the 288 measurements of the dataset "Testt0102" onto the bi-dimensional representation space.

To simulate a day-night effect in a controlled environment, water temperature has been varied in the range between 23.5°C and 26°C with one cycle having a period of 24 hours. Results are shown in Fig. 4.37, where larger markers are assigned to later measurements and smaller ones to the first measurements.

Moreover to the increasing width of the markers as a function of time, markers are coloured as a function of temperature measured at the time the voltammetric measurement has been done. The colormap, common to both the measurement sessions, is also shown in Fig. 4.37.

Investigation of the plot reported in Fig. 4.37 leads to some preliminary considerations:

- drifts of the projections are both temperature and time dependent,
- part of such drifts are due to electrochemical modifications; in addition, the first measurements of both sessions are outliers of their respective clusters.

While the effect of temperature over the current contributions described in eq. (2.4) is known [6], the effect on the observation vectors in the feature subspace is substantially impossible to be determined in an analytical form. In order to compensate for the temperature effects, one possible way would be to take the temperature information as a variable that takes part to the training phase of a classifier. This approach opens the possibility to make further advancements in this technique, particularly for long-term monitoring applications.

It is worth noting that an alternative approach to the study of long term drifts, based on the integral of each pulse response, has been recently presented in [8].

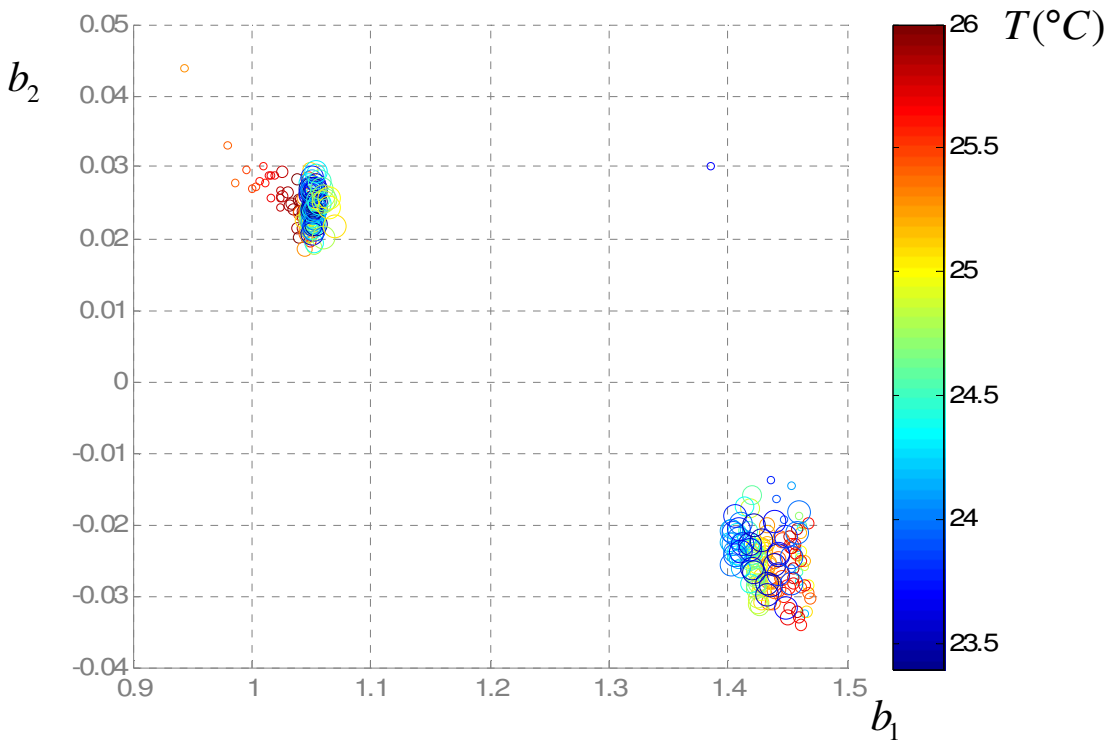

Fig. 4.37 Projections of the full dataset "Testt0102" onto the bi-dimensional representation space. Width of markers increasing with time. Colour of markers assigned as a function of temperature.

4.7 Conclusions

The monitoring technology is an essential aspect in the implementation of innovative and integrated techniques for the sustainable management of ground water and surface water. In particular, the management and control of the entry-points of water into the distribution network, as well as specific nodes of the network itself, can make use of continuous monitoring systems to assess the quality of water.

In this Thesis, a methodological approach to the development of e-tongue devices based on voltammetry is presented. This work has been mainly focused on the signal processing aspects.

Promising results have been obtained in the experiments made during the study of the methodology, a selection of such experiments has been presented in this PhD Thesis, according to their relevancy.

With respect to the known literature, novel signal processing aspects for this applications have been introduced:

- DCT of the time series obtained with pulse voltammetry has been used and presented for the first time in the literature, with proven benefits, in comparison with other techniques;
- sub-optimal search strategies to search for a suitable dimensionality of the feature space, to extract the best features according to the chosen metric, have been applied;
- an algorithm to generate a representation sub-space, which does not rely on PCA, voiding its known drawbacks with large dimensionality datasets, has been presented.

Moreover, even if the sensor technology used here is quite mature, this original application suggested a specific development of the hardware and of the control software in the device used. This aspect has surely contributed to the results obtained, and it must be stressed how the sensor system development, together with the signal processing chain, have experienced one step ahead in the roadmap that leads to a practical field application, during this work.

More work has to be done in the subject of e-tongue devices; also in the recent literature several challenges appear to be still open, such as:

- data fusion with other sensor techniques has still to be more investigated,
- characterisation of long term behaviour of the measurement chain and responsivity to slow changes in the liquid under monitoring is a very critical subject to be faced with; there's still not much literature about drift analysis.

Focusing on the application area regarding water quality and water distribution monitoring, this work has shown a methodological approach that may lead to more efficient, inexpensive, and distributed monitoring systems. In fact, while the information coming from the traditional field instrumentation, providing the main physical/chemical parameters, is poor to fully assess the quality of drinkable water, a complete laboratory analytical measurement is expensive and does not provide a continuous control over the resource.

The proposed approach may lead to a new monitoring method, based on distributed low cost water quality sensors, which make use of the technology of the taste sensors based on voltammetry, that now represents a mature enough, inexpensive and very attractive technology in the framework of innovative solutions for obtaining overall indicators of quality of a natural resource.

To conclude, if this Thesis had given some contribution to fill the gap between research about e-tongue devices and their practical implementation to the real world, this would be interpreted as the most desirable form of success for this work.

REFERENCES FOR CHAPTER IV

[1] A. Scozzari, N. Acito, G. Corsini, "Signal analysis of voltammetric data series for water quality tests and classification", IEEE IMTC 2005, Ottawa (Canada), volume 1, pagg. 89-92

[2] A.Scozzari, N. Acito, G. Corsini, "A supervised algorithm for water classification by voltammetric measurements", Instrumentation and Measurement Technology Conference, Sorrento (Italy), 2006

[3] A. K. Jain, B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice", in "Handbook of statistics", North Holland, Amsterdam 1982, pp. 835-855

[4] A. Webb, "Statistical pattern recognition", John Wiley & Sons, 2nd edition, 2002

[5] A. Scozzari, A. Caprai, R. Cioni, M. Guidi, "Application of voltammetric techniques for water classification and monitoring in a volcanic environment", EGU General Assembly, Wien (Austria), 2005

[6] A. J. Bard, L. R. Faulkner, "Electrochemical methods, fundamentals and applications", John Wiley & Sons, 1980

[7] A. Scozzari, P. Peruzzi, R. Cioni, M. Guidi, "An innovative approach to urban water management based on taste sensors", EGU General Assembly, Wien (Austria), 2006

[8] M. Lindquist, "Electronic tongue for water quality assessment", Doctoral Dissertation, Orebro University (Sweden), 2007

List of figures

List of tables