1997

Information gain study for cyclic voltammetry

Nahid Yazdani
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Yazdani, Nahid, "Information gain study for cyclic voltammetry" (1997). *Master's Theses*. 1615.
DOI: <https://doi.org/10.31979/etd.gazr-9yxv>
https://scholarworks.sjsu.edu/etd_theses/1615

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

**INFORMATION GAIN STUDY FOR CYCLIC
VOLTAMMETRY**

A Thesis

Presented to

The Faculty of the Department of

Chemistry

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Nahid Yazdani

December 1997

UMI Number: 1388233

UMI Microform 1388233
Copyright 1998, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

©1997

Nahid Yazdani

ALL RIGHTS RESERVED

**APPROVED FOR THE DEPARTMENT
OF CHEMISTRY**

Sam P. Perone

Dr. Sam P. Perone

Joseph J. Pesek

Dr. Joseph J. Pesek

Bradley M. Stone

Dr. Bradley M. Stone

APPROVED FOR THE UNIVERSITY

Serena H. Stanford

ABSTRACT

INFORMATION GAIN STUDY FOR CYCLIC VOLTAMMETRY

by Nahid Yazdani

This work describes an evaluation of information content of the electrochemical technique known as Cyclic Voltammetry (CV). The specific informational goal in this study was qualitative identification of redox couples and electrode processes based on voltammetric shape classification only.

The influence of experimental parameters, such as number of cycles, scan rate and switching potential on information content was studied. The voltammetric information content was related to the ability to distinguish among several different electrode reaction mechanisms. This evaluation was done using an unsupervised pattern recognition method called Hierarchical Cluster Analysis.

ACKNOWLEDGMENTS

I am deeply indebted to my Professor and research advisor, Dr. Sam P. Perone, for continuously encouraging and advising me to complete my research studies. I would also like to thank him for his teaching, guidance, and also his meticulous review of my thesis.

And, I would also like to thank Dr. Joseph J. Pesek and Dr. Bradley M. Stone for providing me valuable suggestions and help as my committee members.

Also, I would like to extend my appreciations and gratitudes to Dr. Daniel L. Severance for his enormous help on programming related to my research.

Finally, my special appreciations are extended to my parents and grandparents for all their love and support.

TABLE OF CONTENTS

Chapter 1	Introduction	
1.1	Previous Work.....	1
1.2	Objective of the Present Study.....	5
Chapter 2	Theory	
2.1	Voltammetry.....	6
2.2	Cyclic Voltammetry.....	8
2.3	Digital Simulation.....	11
2.4	Information Theory.....	15
2.4.1	Information Theory Concepts.....	15
2.4.2	Information Gain.....	18
2.5	Pattern Recognition.....	19
2.6	Fast Fourier Transform.....	26
Chapter 3	Experimental Approach	
3.1	Instrumentation.....	28
3.2	Data Base Management and Data Analysis.....	29
3.3	Digital Simulations.....	30
3.3.1	Experimental Mechanisms and Compound Classes for Cyclic Voltammetry.....	30
3.3.2	Experimental Parameters for Cyclic Voltammetry.....	37

3.4	Feature Definition and Extraction.....	39
3.5	Pattern Recognition Procedure.....	40
Chapter 4	Results and Discussions	
4.1	Summary of Procedures Used.....	53
4.2	Observations.....	54
4.2.1	Summary of Observations with Unsupervised Pattern Recognition.....	61
4.2.2	Summary of Observations with Supervised Pattern Recognition.....	62
4.3	Discussions.....	62
Chapter 5	Conclusions.....	66
References		67

LIST OF TABLES

Table 1a.	Chemical Parameters for EC Mechanism.....	32
Table 1b.	Chemical Parameters for CE Mechanism.....	33
Table 1c.	Chemical Parameters for ECE Mechanism.....	34
Table 2a.	Chemical Parameters for E (Quasi-Reversible) Mechanism.....	35
Table 2b.	Chemical Paramters for Reversible E Mechansim.....	36
Table 3.	Experimental Parameters for Simulated Voltammograms.....	38
Table 4.	Feature Selection for Experimental Conditions A-D and A2-D2.....	52
Table 5.	Observations of Quasi-Reversible (QR)Voltammograms.....	55
Table 6a.	Observations of 4-Cluster Voltammograms, 3 Mechanism Classes.....	56
Table 6b.	Observations of 4-Cluster Voltammograms, 4 Mechanism Classes.....	56
Table 7.	Observations of 5-Cluster Voltammograms, 4 Mechanism Classes.....	57
Table 8.	Observations of Various Types of Information mostly in a Single Cluster.....	58
Table 9.	Feature Set 1 (Features 2, 8, 9).....	64
Table 10.	Feature Set 2 (Features 2, 9, 22, 23).....	65

LIST OF FIGURES

Fig 1.	Potential Excitation Signals used in Voltammetry.....	7
Fig 2.	Quasi-Reversible Electron Transfer Case.....	10
Fig 3.	Analytical System.....	16
Fig 4.	Dendrogram.....	25
Fig 5.	Dendrogram with Similarity Index set for Two Clusters.....	41
Fig 6.	Dendrogram with Similarity Index set for Three Clusters.....	42
Fig 7.	Dendrogram with Similarity Index set for Four Clusters.....	43
Fig 8.	Range of Similarity Values for Specific Clustering Groups vs. # of Features.....	47
Fig 9.	Cluster Size vs # of Features (for 2 clusters)	48
Fig 10.	Cluster Size vs # of Features (for 3 clusters)	49
Fig 11.	Cluster Size vs # of Features (for 4 clusters)	50
Fig 12.	Cluster Size vs # of Features (for 5 clusters)	51

Chapter 1 Introduction

1.1 Previous Work

One of the most challenging tasks in chemical analysis is the classification of a chemical structure by using electrochemical techniques. Although many organic compounds of biological and commercial importance are electroactive, they generally produce electrochemical responses which consist of only a small number of peaks or waves. This lack of fine-structure makes “finger printing” difficult. The task of qualitative electroanalysis is made even more formidable from both a theoretical and experimental standpoint by the complex dependence of an electrochemical response on many variables.

An empirical approach such as computerized pattern recognition would seem to be the best solution for qualitative identification. Pattern recognition has already been used, for instance, in determination of biological activity, and in spectral interpretation. Within the field of electrochemistry, several pattern recognition studies have been performed which have implications for qualitative electroanalysis (1).

Schachterle and Perone (2) used pattern recognition to demonstrate that analysis of cyclic linear sweep voltammograms and cyclic staircase voltammograms produced information which could be used to accurately classify the mechanism of an electrode reaction. This approach should be helpful in qualitative analysis since similar structures can be expected to react by similar mechanisms.

Burgard and Perone (3) reported computerized pattern recognition for classification of 29 compounds using Cyclic Staircase Voltammetry (CSCV). The classes examined were aromatic-nitro, aliphatic-nitro, aromatic-aldehyde, and aromatic-aliphatic-ketone. These classes were almost completely separated on the basis of peak potential, but this feature alone cannot be considered sufficient for many identification problems. Therefore, the voltammograms were examined for any shape information which might characterize a particular electroactive group on the skeleton to which it was attached. It was found that the change in peak shape with scan rate produced good classification (70% correct) but that complete separation of the classes was not possible for the experimental conditions and compounds which were chosen. This result suggested that the information content of the electrochemical data base should be increased for more reliable structural classifications.

Byers, Freiser, and Perone (1,4,5) analyzed forty-five compounds using cyclic staircase voltammetry. Experimental variables studied were percent ethanol in the solvent, pH, surfactant concentration, number of cycles, scan rate, mercury drop hang time, and sampling time. The data set contained 19 nitrobenzenes, 9 nitrodiphenyl ethers, and 17 ortho-hydroxy azo compounds. Of the nitrodiphenyl ethers, 4 were strong herbicides and 5 were either weak or nonherbicides. It was possible with computerized pattern recognition to distinguish between these three structural classes with similar electrochemical characteristics. Even though nitrodiphenyl ethers, the nitroaromatics, and the azo compounds were reduced in the same potential region with the same number of electrons, 93.3% classification accuracy was achieved. Herbicidally active and inactive

nitrodiphenyl ethers were distinguished from each other with a classification accuracy of 100% over a range of experimental conditions. The most significant result of this study was demonstration that the information content of voltammetric data could be enhanced considerably by the systematic adjustment of experimental conditions.

Thomas and Perone applied pattern recognition techniques to the detection of severely overlapped voltammetric signals on both theoretical (6) and experimental data (7). In each of these studies, the training and prediction sets used for pattern classification and feature selection were derived from Stationary Electrode Polarographic (SEP) theory and were prepared to include a wide variety of both singlet and doublet curve shapes. Their results showed that a K-nearest neighbor (KNN) pattern recognition approach was the most effective selection and classification method. Also, features derived from the discrete Fast Fourier Transform (FFT) seemed to allow the most reliable and consistent classification of voltammetric doublets and singlets.

Zhou in 1995 (8) investigated the influence of several experimental parameters such as number of cycles (nCy) and scan rate (V) on information content of cyclic voltammetry. First, simulated cyclic voltammetric data were generated for hypothetical chemical compounds of five different classes. Then, pattern recognition techniques were used to identify compounds/classes from Fourier transform cyclic voltammetry data and to determine which features and/or experimental parameters were most useful for accurate classifications. Finally, information theory was used to provide a more quantitative measure of information content related to experimental parameters for qualitative identification of redox species.

Of the experimental parameters studied by Zhou, the results with variable number of cycles were the most surprising. That is, contrary to what was expected, there was no significant enhancement of information gain for two cycles compared to one. This is probably related to the fact that, for these particular electrode processes, no new redox couple is observed during the second cycle.

Enhanced information content was observed for lower scan rate, probably because for the particular electrode processes studied the effects of coupled chemical reactions are more pronounced at slower scan rate, allowing better discrimination among classes from voltammetric shape data. It is important to note that these observations were based on a very limited data set and some natural subsets of compounds were not incorporated in the study.

Pang in 1997 (9) used a more expanded database to further investigate the influence of number of cycles (n_{Cy}), scan rate (V), and switching potential (E_{sw}) on information content of cyclic voltammetry. Her study indicated that lower scan rate is a significant parameter for classification of compounds based on cyclic voltammogram shape. It also showed that increasing the number of cycles does increase the information gain under certain conditions. This study also suggested that switching potential is not an important parameter, at least under the conditions studied.

1.2 Objective of the Present Study

The long-range objective of this study is to quantify the information content inherent in the shape of cyclic voltammogram data. First, to understand which parameters in a cyclic voltammetric experiment influence qualitative identification of compounds and second to determine the quantitative impact on information content.

This study used the same expanded data base which was employed in the previous work (9). The study here used non-supervised pattern recognition to measure voltammetric information content as a function of experimental parameters such as number of cycles, scan rate, and switching potential. With the aid of information theory, the information content obtained from these experimental parameters can be quantitatively calculated. Therefore, it is possible to choose the most influential parameters affecting class related information content.

Chapter 2 Theory

2.1 Voltammetry

Voltammetry comprises a group of electroanalytical methods in which information about the analyte is obtained from the measurement of current as a function of applied potential. Voltammetry is widely employed by inorganic, physical, and biological chemists for nonanalytical purposes. Some of these applications include fundamental studies of oxidation and reduction processes in various media, adsorption processes on surfaces, and electron-transfer mechanisms at chemically modified electrode surfaces (10).

In voltammetry, a variable potential excitation signal is established upon an electrochemical cell which contains a microelectrode. This excitation signal elicits a characteristic current response upon which the method is based. Depending upon the applied potential, voltammetry can be classified into four groups: Linear Sweep Voltammetry, Differential Pulse Voltammetry, Square Wave Voltammetry and Cyclic Voltammetry. These four waveforms of excitation signals used in voltammetry are shown in Figure 1 (10).

	<u>Name</u>	<u>Wave form</u>	<u>Type voltammetry</u>
(a)	Linear scan		Polarography Hydrodynamic voltammetry
(b)	Differential pulse		Differential pulse polarography
(c)	Square wave		Square wave voltammetry
(d)	Triangular		Cyclic voltammetry

Figure 1. Potential Excitation Signals used in Voltammetry (11)

1. Linear Sweep Voltammetry

The classical voltammetric excitation signal is the linear scan shown in Figure 1-a, in which the dc potential applied to the cell increases linearly as a function of time usually over a 2- to 3-V range. The current that develops in the cell is then recorded as a function of time and thus as a function of the applied potential.

2. Two pulse-type excitation signals are shown in Figure 1-b and 1-c.

Currents are measured at various times during the lifetime of these pulses.

3. Cyclic Voltammetry

With the triangular waveform shown in Figure 1-d, the potential is cycled between two values, first increasing linearly to a maximum and then decreasing linearly with the same numerical slope to its original value. This process may be repeated numerous times, with the current being recorded as a function of time. A complete cycle may take several seconds or as about as a few microseconds or less. Cyclic Voltammetry is discussed in more detail below since it is the technique used in this study.

2.2 Cyclic Voltammetry

Cyclic Voltammetry (CV), while not used for routine quantitative analysis, has become an important tool for the study of mechanisms and rates of oxidation-reduction processes, particularly in organic and metal-organic systems. Often, cyclic voltammograms will reveal the presence of intermediates in oxidation-reduction

reactions. The primary use of cyclic voltammetry is as a diagnostic tool that provides qualitative information about electrochemical processes under various conditions.

In cyclic voltammetry, the current response of a small stationary electrode in an unstirred solution is excited by a triangular potential wave form. The potential is first varied linearly in one direction, whereupon the scan direction is reversed and the potential is returned to its original value. The excitation cycle is often repeated several times. The initial potential is E_{start} ; the potential at which reversal takes place is called the switching potential (E_{sw}). The voltage scan rate, v , is expressed in volts/s or mV/s. The range of potentials chosen for a given experiment is one in which a diffusion controlled oxidation or reduction of one or more analytes occurs. Depending upon the composition of the sample, the direction of the initial scan may be either negative or positive. A scan in the direction of more negative potentials is termed a forward scan, while one in the opposite direction is called a reverse scan (11).

Important parameters in a cyclic voltammogram are the cathodic peak potential E_{pc} , the anodic peak potential E_{pa} , the cathodic peak current i_{pc} , and the anodic peak current i_{pa} . For a reversible electrode reaction, anodic and cathodic peak currents are approximately equal and the difference in peak potentials is $0.0592/n$, where n is the electrons involved in the half-reaction. In cyclic voltammetry, the reverse scan is very useful for examining products of the forward scan and for detecting processes which might be occurring to remove these products from the solution. The shape of the voltammogram is a function of temperature, the experimental parameters (E_{start} , E_{sw} , v),

Figure 2. Quasi-Reversible Electron Transfer Case, when $k_s = 0.0005$, $\alpha = 0.5$
One Cycle, (a): $v = 0.05$ V/s, (b): $v = 1.00$ V/s

the heterogeneous electron transfer rate parameters (K_s , E° , α), the diffusion coefficients and concentrations of each species, and the electrode geometry. K_s is the heterogeneous rate constant which determines the reversibility of the charge transfer step. α is a symmetry factor which is important in determining the relative potential dependence of the cathodic and anodic steps. α is between 0 and 1. E° is the standard potential. Figure 2 gives an example of a cyclic voltammogram for the simple mechanism:

2.3 Digital Simulation

Since the publication of Nicholson and Shain's classic paper (12), cyclic voltammetry has become the most commonly used electroanalytical technique, widely applied by non-electrochemists as well as electrochemists. The computation of CV responses for even the simplest electrochemical system ($A + e^- \rightleftharpoons B$) requires significant, non-trivial, mathematical manipulation, and then, ultimately, numerical integration. Extension of the approach to more complicated systems involving coupled heterogeneous electron transfer and first- and second-order homogeneous chemical reactions always has been difficult, sometimes impossible, and never routine.

Because of these difficulties, researchers attempting to produce a general CV simulator have had limited success. Speiser has used orthogonal collocation as the basis of EASIEST, which is linked to a mainframe computer and is used to access a library of

algorithms to compute the cyclic voltammetric responses for specific mechanisms (13).

Gosser has devised CVSIM, a simulator based on an explicit finite difference algorithm that is useful for simulating responses for simpler systems (14). Bieniasz has devised ELSIM, a potpourri of mathematical and numerical techniques that can be used to simulate different electrochemical responses but requires a mathematically sophisticated user (15).

Rudolph developed a fast implicit finite difference (FIFD) algorithm, a CV simulation program, which meets this computational challenge effectively (16,17). Rudolph's CV simulator accepts any user specified mechanism such as any combination of one-electron heterogeneous electron transfers as well as first- and/or second-order homogeneous chemical reactions. This CV simulator efficiently simulates CV responses and demands no mathematical or computational skills from the user. As long as a few simple and invariant rules are followed, the FIFD algorithm is extremely robust and can be used for accurate simulation of virtually any electrochemical-chemical system over a wide dynamic range of parametric values. Therefore, it is possible to translate a straight forward description of an electrochemical mechanism into the mathematical expressions required to compute the cyclic voltammetric response. The user can simulate CV responses, compare simulations with each other and with the experimental data, perform least-squares fit simulations to other simulations or to experimental voltammograms with selected parameter optimization, and observe the concentration profiles of the various species as they change during cyclic voltammetric perturbation.

The commercial software DigiSim[®] 1.0 (Bioanalytical Systems Inc., W. Lafayette, IN) is used in this study to generate theoretical cyclic voltammograms. Once you enter a mechanism, using simple chemical notation, the DigiSim[®] program constructs the matrices required to simulate the mechanism. The program generates all the requests for data input associated with a given mechanism and computes CV responses in seconds even for complex electrochemical-chemical systems. DigiSim[®] offers a wide range of features that will not only expedite the analysis of experimental cyclic voltammograms, but combined with an appropriate text, will also offer a novel and effective approach to teaching the basics of cyclic voltammetric behavior in particular and electrochemistry in general.

Four mechanisms were studied in this study: electron transfer reaction (E), electron transfer reaction followed by a chemical reaction (EC), chemical reaction followed by a electron transfer reaction (CE), and electron transfer reaction followed by a chemical reaction then another electron transfer (ECE). Electron transfer reactions can be further divided into reversible and Quasi-reversible.

1. E mechanism

This is the basic electrode process without chemical complication. It is given by:

where the heterogeneous rate constant k_s determines the reversibility of the charge transfer step.

2. EC mechanism

It is given by:

where the following homogeneous reaction can be characterized by an equilibrium constant K and rate constants for forward and backward reaction. ($K = k_f/k_b$)

3. CE mechanism

It is given by:

where the preceding homogeneous reaction can be characterized by an equilibrium constant K , and rate constant k_f and k_b .

4. ECE mechanism

It is given by:

where the intermediate homogeneous reaction can be characterized by an equilibrium constant K , and rate constants k_f and k_b .

2.4 Information Theory

2.4.1 Information Theory Concepts

The use of information theory in analytical chemistry originates from the idea that each analysis is a process of obtaining information about the qualitative or quantitative composition of a sample, or information about the structure of the analyzed substance (18).

An analytical system possesses an input, an output, and an input-output relation (Figure 3). The system can be viewed as consisting of two relatively independent subsystems. In the first, the analytical device, information arises. In the second, for example a computer, it is decoded. The systems can be linked with a feedback channel.

The information content provided by particular analytical procedures or devices is important because it enables us to compare their relative suitability for different applications. In addition, optimization can be performed by employing an information quantity as a response function.

Information theory is concerned with the study of information and its transmission. Probability theory forms the basis for understanding and applying the information theory approach to chemical analysis. The first question that arises is how we can measure information at all. The uncertainties before and after an experiment and the difference between them play key roles.

Most of the present work in information theory is based upon probabilistic models of communication developed by C.E. Shannon in 1948 (19). The basic principle of this

Figure 3. Analytical System

theory is that a message with a high probability of random occurrence conveys little information. The most information is conveyed by the message that is least likely to occur spontaneously.

This principle is formalized by the concept of entropy which equates information and uncertainty. Entropy is a quantitative measure of the amount of information supplied by a probabilistic experiment. It is based upon classical Boltzman entropy from statistical physics. Shannon's formula, equation (1) defined the entropy or the average information to be equal to the weighted average of the specific information for each event in the system under consideration:

$$H_{av} = I_{av} = - \sum_{i=1}^N P(x_i) \log_2 P(x_i) \quad (1)$$

where the symbol $P(x_i)$ is the probability of occurrence of an event, and N is the number of possible events. H is the maximum when events are mutually exclusive and random; then $H_{max} = \log_2 N$.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used, the resulting units may be called binary digits, or more briefly bits. Specific information is information conveyed by the occurrence of a particular event and is quantified by the $-\log_2$ of the probability of the event ($P(x_i)$).

2.4.2 Information Gain

One of the most important concepts in information theory is that of informational gain or information content. The information content for a particular qualitative identification exercise can be expressed in terms of the classification accuracy. However, this expression is not readily comparable to another situation where different numbers of samples or different class distributions are involved.

Since information content can only be established after the analysis, these measures cannot be used as a quality criterion for selecting an analytical procedure. However, they can be used to evaluate the performance of a procedure. Measures of information content had been applied to information theory models of structural analysis, qualitative analysis, quantitative analysis, trace analysis and instrumental analysis (5).

The application of information theory concepts to analytical chemistry can illuminate methods to increase the efficiency of chemical analysis. It has been demonstrated that the informational goal(s) will dictate the most favorable choice of experimental conditions. The use of objective systematic information enhancement methods can highlight experimental parameters that are often traditionally overlooked.

The equation used to calculate the information gain is equation (2) below:

$$\Delta H = \sum_{R=1}^Z (C_R/J) \log_2 (n_o/n) \quad (2)$$

where ΔH is information gain

J = total No. of patterns

C_R = total No. of correctly classified compounds of class R

Z = total No. of classes

n_0 is the total number of possible identities that might be assigned to an “unknown” item, n is the number of possible identities which might be assigned to an “unknown” item after the pattern recognition procedure (see section 4.2 for example calculations).

2.5 Pattern Recognition

Modern analytical methods allow the determination of many substances simultaneously and computers provide facilities for storing or handling the large amounts of data obtained in this way. As many data are obtained at a time, they should be used together, particularly because the results obtained for the parameters are often related to each other (20).

In automated qualitative chemical analysis, one is often confronted with a large number of measurements, some of which are thought to be useful in characterizing unknowns. However, due to the large number of measurements or their complexity, it is difficult to decide which are the most useful. When this is the case, computerized pattern recognition may be the best means of selecting appropriate measurements and making classifications (21).

The general term “pattern recognition” refers to automatic procedures for classifying individual observations into discrete groups on the basis of a multivariate data matrix. Pattern recognition involves the detection and recognition of regularities among sets of measurements describing objects or events. It is concerned with the investigation of analytical techniques for processing large amounts of data, and extracting useful information to reduce the data.

Pattern recognition is concerned primarily with the description and analysis of measurements taken from a physical system. Data observed (d-dimensional pattern space) are the experimentally determined values used to characterize the objects. A pattern is defined as a d-dimensional vector composed of d calculated independent data items, and can be represented by:

$$P = W_1 X_1 + W_2 X_2 + W_3 X_3 + \dots + W_d X_d \quad (3)$$

$X_1, X_2, X_3, \dots, X_d$ = components (measurements) of the pattern vector

$W_1, W_2, W_3, \dots, W_d$ = components of the weight vector

d = number of dimensions

One of the difficulties in pattern recognition is what has been called “the curse of dimensionality”. If many variables are present, the classification problem may become too complex, and a reduction in the number of dimensions can help to make it more manageable. Further, if one is able to represent the data originally present in d dimensions in two or three dimensions in such a way that the similarities and dissimilarities between the data points are conserved, at least partially, this can be a valuable aid in arriving at a better understanding of the data.

Pattern recognition can be described by four sequential steps for the purpose of discussion: measurement, preprocessing, feature selection, and classification.

In the first step, one or more transducers produce an array of measurements for each compound or chemical system to be determined. This array of measurements is called a pattern. Since the measurements may be from different transducers, they could vary quite widely in magnitude and be in different units.

In the preprocessing step the measurements are usually normalized in magnitude, and mathematical transformations may be applied to produce new descriptors called “features” which are more useful in classification of the unknowns. The definition of features is the most crucial aspect of pattern recognition. Features must be defined so as to effect class separation on the basis of real class differences and not on the basis of trivial information. This is often a highly intuitive step but it must be based on the analyst’s best understanding of the data (6).

The transformation may be as simple as multiplying each measurement by a scale factor which reflects its importance (weighting), or it may make features from linear combinations of the original measurements. Fourier transform is commonly used in preprocessing.

With Fast Fourier Transform (FFT) preprocessing, the convoluted voltammogram, spectrum, or chromatogram is transformed from the time domain into a corresponding function in the frequency domain. In the Fourier domain, deconvolution is accomplished by separating out each of the pure components (22).

Sharper peaks produce higher frequency components in the frequency domain. Because noise also has high frequency components, these may be isolated in the frequency domain and removed from contribution (22).

It is desirable to have as few features as possible for computational simplicity and statistical confidence in classification. Thus, the third step is to eliminate as many features as possible without adversely affecting classification. The only way to be sure of picking the best subset of the original features is to try all possible combinations, which may not be feasible.

In the last step of pattern recognition, each compound or chemical system is identified by assigning it to some predefined class. This means that one tries to place the samples characterized by their individual patterns in the category to which they belong. The classifications are almost always made by considering the location of the patterns in a hyperspace formed by using each of the features as an axis. This classification is based on distances between points in the r -dimensional space. The smaller the distance between points, the more probable it is that they belong to the same category. The way in which the classification is carried out depends on whether one is concerned with supervised or an unsupervised learning problem.

The pattern recognition process is normally first carried out on a set of known samples called a training set. The true identity of each pattern is then compared to the identity assigned in the classification step, and the results are fed back into one or more of the previous steps for error correction. The feedback process may be totally automated or may allow for input from the analyst. This technique is called supervised learning.

Unsupervised pattern recognition deals with the interpretation of the structure in a data matrix. The data matrix consists of many features for many objects. The aim is to understand the relationship between either the objects or the features. This means that a prior division into classes such as in supervised pattern recognition is not assumed (23).

One of the basic problems in the interpretation of data is finding structure in large sets of empirically gathered information. An important and much used way of solving this problem is to group the objects being studied in a number of sets or subgroups. A general expression for a subgroup of a set of objects is a cluster.

In the process of finding clusters one is looking for a “natural” structure of the data set; this structure must make it possible to explain the cohesion of the objects of each cluster. A generally acceptable definition of natural structure is not simple to find and this is one of the reasons for the large number of clustering methods to be found in the literature. However, most of the methods are based on a simple basic idea: a measure of the quality of a grouping is found by comparing the “similarities” between the different clusters with the “similarities” between the object within each of the clusters (24). One of these methods is called Hierarchical Cluster Analysis which is part of the group of unsupervised pattern recognition techniques.

Multivariate data can be represented on an N-dimensional plot, where N is the number of variables measured on samples. Each sample then is a single point in the N-space, and unsupervised pattern recognition technique such as Hierarchical Cluster Analysis is simply looking for relationships among these points.

In Hierarchical Cluster Analysis (HCA), distances between the samples (or variables) in a data set are calculated. Samples whose distances are relatively small are likely similar, at least with respect to the measurements taken. Dissimilar samples will have larger relative distances.

A dendrogram (Figure 4) is a tree-shaped distance map which is constructed from the set of inter-sample distances, by linking samples and clusters of samples, as a function of distance. The dendrogram tree branches from right to left dividing into more and more branches until it terminates on the left side of the display with individual samples, called leaves. The point where a branch divides is called a node. The leaves are the terminal nodes of the tree. Non-leaf nodes always have two descendants, and continue branching in pairs until they terminate in individual leaves.

Initially, each sample is considered a cluster of a single sample. First, the two most similar samples clusters are linked. Once a cluster has been linked with another cluster, a new cluster is defined from this combination. The distances between this new cluster and all other existing clusters are computed, and the smallest inter-cluster distance is again searched and another linking formed. This process is contained until all samples/clusters have been linked. The result is displayed as the dendrogram, whose branch lengths are proportional to the distances between the connecting clusters.

The distances in the dendrogram plot are shown in the units of similarity, a scale conversion made such that identical samples are assigned a similarity value of one, while

Figure 4. Dendrogram

the two most dissimilar samples in the data set are defined as having a similarity value of zero. All other data points will fall in between.

An important concept associated with clustering is that of distance or similarity. If i and j are two of the objects of the set which is to be clustered it is necessary to quantify their relative position. In other words, based upon the measurements of N variables a similarity or distance between any pair of objects must be computed. The distance between two points in N -dimensional space, D_{ij} , is computed by comparing the differences at each of N variables. There are many ways to define a distance function on a set of objects. The most commonly used distance is the Euclidean distance defined by:

$$D_{ij} = \left[\sum_{n=1}^N (x_{in} - x_{jn})^2 \right]^{1/2} \quad (4)$$

2.6 Fast Fourier Transform

Many possible numerical transform techniques might be used to reduce data dimensionality while retaining desired information content. Of these, the Fourier Transform has been the most useful for voltammetry data, and will be discussed here.

Fourier Transform is a technique for solving problems in linear systems and analyzing periodic data (6,7,25,26). It is expressed in the following equations:

$$F(v) = \int_{-\infty}^{+\infty} f(t) e^{-2\pi i v t} dt \quad (5)$$

$$F(\tau) = \int_{-\infty}^{+\infty} f(v) e^{-2\pi i \tau v} dv \quad (6)$$

$F(v)$ is frequency domain signal, and $F(\tau)$ is a time domain signal. The algorithm which allows the efficient computation of a transform for discrete frequencies (or time intervals) is called the Fast Fourier Transform (FFT). A discrete frequency domain function transform is represented schematically by the equation:

$$F(v) = a_0f(v_0)+a_1f(v_1)+...a_nf(v_n) \quad (7)$$

FFT achieves dimension reduction and noise reduction when only lower frequency components are selected. It also emphasizes frequency content of data, which can be very useful if relevant to informational goal.

Thomas and Perone (6,7) have shown that the coefficients obtained from the discrete Fourier transform of voltammetric curves are useful for curve shape analysis. FFT features seem to allow the most reliable and consistent classification of voltammetric doublets and singlets. Features derived from the discrete FFT have been shown to be useful for the enhancement of intrinsic class differences in real voltammetric data. Thus, these features from FFT are apparently good enough for the representation of analytical waveforms for pattern recognition, when intrinsic wave shape information is sought. Burgard and Perone (3) studied computerized pattern recognition for classification of organic compounds from voltammetric data and found that curve shape features obtained by FFT were very useful for accurate classification of the voltammograms.

Chapter 3 Experimental Approach

3.1 Instrumentation

An IBM clone personal computer (PC) was used for database management, computerized digital simulation, fast Fourier transform and Hierarchial Cluster Analysis.

DigiSim[®] 1.0 software was used to generate synthetic cyclic voltammograms. DigiSim[®] 1.0 runs in the DOS environment, and source code is programmed in C++. The C++ compiler has effected a DOS extender which allows DigiSim[®] to utilize as much of the RAM as needed. At least 2 M-bytes of extended RAM is recommended. DigiSim[®] 1.0 will run only with DOS based configuration. The DOS configuration requires 2M byte RAM and 20 M byte hard disk on a 386 DX/33 MHz PC or a 486 DX.

In this study, the digital simulation run for one-cycle CV with step size $E = 0.0025$ V is about 11.5 sec for the E mechanism, 22.08 sec for the EC mechanism, 22.08 sec for the CE mechanism, and 69.15 sec for the ECE mechanism. There is no big difference in run time when α , k_s , K , k_f are changed. For two-cycle CV, the run time is about the same as the one-cycle CV because the step size is changed to 0.005 V.

Fast Fourier analysis can be performed under Microsoft EXCEL[®] 4.0 or 5.0 with a data analysis tool package. It takes about 2 minutes for 1024 data points on a 486 DX/25 MHz PC.

Ein*Sight 3.0 software (Infometrix, Seattle, WA.) was used for unsupervised pattern recognition, Hierarchical Cluster Analysis. Ein*Sight requires the following hardware: A PC compatible 286, 386 or 486 system with MS DOS (version 3.1 or higher), a hard disk drive (with at least 2 MB free) and a Microsoft mouse, a math coprocessor (80287, 80387), an EGA or VGA graphics monitor and 640kB RAM.

3.2 Data Base Management and Data Analysis

All raw data generated from DigiSim[®] are transferred to Microsoft EXCEL[®] spreadsheet data base management system. Rotated cyclic voltammograms (CV), basic statistical computations, Fourier feature transform calculation, normalized Fourier feature transform data and associated graphical procedures are conducted in Microsoft EXCEL[®]. Unsupervised pattern recognition analysis are conducted with Ein*Sight software.

To compare voltammograms, they must be collected under the same conditions, examine the same size voltage window, and be submitted to the same preprocessing. First, each simulated voltammogram was shifted so that the peak potential was equal to zero ($< \pm 0.01$ mv). This step avoids introducing information into the transform related to peak potential. Second, prior to the Fourier transform, each voltammogram was rotated so that the first and last data points were zero, and no high frequency artifacts were introduced. The raw CV data file was rotated by subtracting a linear term which was computed by connecting the first current value to the last. This method has been shown to

be an effective means of maintaining the fidelity of electrochemical data during Fourier transformation. Third, the fast Fourier transform is calculated by using the Microsoft EXCEL[®] analysis tool package, and only the real part of the complex numbers from the output data were used as Fourier coefficients. The 1024 voltammetric data result in 512 Fourier coefficients. Finally, the Fourier coefficients were normalized to the maximum absolute value. The 512 normalized magnitudes were used as Fourier features, whose values range from 0 to 1.0. These features were used for further pattern recognition study.

3.3 Digital Simulations

3.3.1 Experimental Mechanisms and Compound Classes for Cyclic Voltammetry

The purpose of this study was to determine what types of information might be contained in voltammetric data in addition to that related to identification of different classes of compounds. This requires representative cyclic voltammograms for a large number of different compounds. In this study, synthetic voltammograms were obtained by digital simulations representing five different classes of compounds where each class has a different electrochemical mechanism.

Four mechanisms were introduced in section 2.3. The “E” mechanism is further sub-divided into Reversible and Quasi-Reversible reactions. When the fundamental

properties of each electrode process are allowed to vary, a unique simulated voltammogram can be generated for each unique set of properties. Each unique set represents a hypothetical “compound”. As a result, 156 “compounds”, each of which is represented by a number in Tables 1 and 2, are distributed to these 5 classes. Cyclic voltammograms representing each of the “compounds” were obtained. Tables 1 and 2 define the unique set of electrochemical properties for each “compound”. These properties include chemical and electron transfer parameters, such as k_s , α , n , K and k_f (defined in sections 2.2 and 2.3).

When chemical or physical reactions are absent, k_s , α , and n affect the shape of voltammograms (section 2.2). k_s is rate constant which affects the reversibility of electron transfer. α is the symmetry factor. n is the number of electrons transferred in the reaction. In this study, n was always considered as 1, as to avoid conditions where voltammetric shape differences were due solely to different n values. The only exception to this was for the ECE case, where two successive reactions might appear as a single reaction with $n_1 + n_2$ electrons under some conditions.

When homogeneous chemical reactions are present, the reaction constants, K and k_f are considered. These will affect voltammetric shape in various ways depending on the overall mechanism.

Table 1a. Chemical Parameters for EC Mechanism

k_s	10^6				0.0005											
α	0.5				0.3				0.5				0.7			
K	0.05		2.5		0.05		2.5		0.05		2.5		0.05		2.5	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	6	47	28	9	50	31	12	53	77	100	120	80	34	15	56	37

k_s	0.005											
α	0.3				0.5				0.7			
K	0.05		2.5		0.05		2.5		0.05		2.5	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	103	131	83	106	122	86	109	125	89	112	136	92

k_s	0.05											
α	0.3				0.5				0.7			
K	0.05		2.5		0.05		2.5		0.05		2.5	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	18	59	40	21	115	127	95	118	62	43	24	65

Table 1b. Chemical Parameters for CE Mechanism

k_s	10^6				0.0005											
α	0.5				0.3				0.5				0.7			
K	0.005		0.1		0.005		0.1		0.005		0.1		0.005		0.1	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	26	7	48	29	10	51	32	13	98	78	101	121	54	35	16	57

k_s	0.005											
α	0.3				0.5				0.7			
K	0.005		0.1		0.005		0.1		0.005		0.1	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	81	104	132	84	107	123	87	110	134	90	113	137

k_s	0.05											
α	0.3				0.5				0.7			
K	0.005		0.1		0.005		0.1		0.005		0.1	
K_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	38	19	60	41	93	116	128	129	22	63	44	25

Table 1c. Chemical Parameters for ECE Mechanism

k_s	10^6				0.0005											
α	0.5				0.3				0.5				0.7			
K	1		100		1		100		1		100		1		100	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	46	27	8	49	30	11	52	33	119	99	79	102	14	55	36	17

k_s	0.005											
α	0.3				0.5				0.7			
K	1		100		1		100		1		100	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	130	82	105	133	85	108	124	88	111	135	91	114

k_s	0.05											
α	0.3				0.5				0.7			
K	1		100		1		100		1		100	
k_f	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	58	39	20	61	126	94	117	96	42	23	64	45

Table 2a. Chemical Parameters for E (Quasi-Reversible) Mechanism

k_s	0.0005						
α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	150	2	151	71	152	3	153

k_s	0.002						
α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	154	155	156	157	158	159	160

k_s	0.005						
α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	161	72	162	73	163	74	164

k_s	0.02						
α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	165	166	167	168	169	170	171

k_s	0.05						
α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	172	4	173	76	174	5	175

Table 2b. Chemical Parameters for Reversible E Mechanism

k_s	10^6
α	0.5
	1

3.3.2 Experimental Parameters for Cyclic Voltammetry

Three main experimental parameters were considered in this study: scan rate (V), number of cycles (nCy) and switching potential (E_{sw}). Each parameter is considered at two levels (Table 3). Thus, there were eight different combinations of experimental parameters. Experiments were conducted at each set of conditions with one-cycle and two-cycles and are referred to as type A-D or A2-D2 as listed in Table 3.

For each simulated voltammogram, 1024 data points were collected because Microsoft EXCEL[®] Fourier analysis has a 2^n (n = positive even integer number) data input range requirement. For one-cycle voltammograms, the step size $\Delta E = 0.0025$ v, gives a window for the cyclic voltammogram of $\Delta E * (\text{data points}/2) = 2.5 * (1024/2) = 1280$ mv. For a two-cycle voltammogram, the step size ΔE was increased to 0.005 v so that it had the same number of data points (1024) and the same window (1280 mv). Thus, both one-cycle and two-cycle voltammograms examine the same voltage window, with the experimental cathodic E_p of cycle one always set arbitrarily at 0.000 v, and the peak potential always occurs at either 50 mv or 300 mv prior to the switching potential E_{sw} .

Table 3. Experimental parameters for simulated voltammograms.

V (V/s)	E_{sw}/E_i (mV)	$nCy^- = 1$	$nCy^+ = 2$
$V^- =$ 0.05	$(E_{sw}/E_i)^-$	A	A2
	$(E_{sw}/E_i)^+$	B	B2
$V^+ =$ 1.00	$(E_{sw}/E_i)^-$	C	C2
	$(E_{sw}/E_i)^+$	D	D2

$(E_{sw}/E_i)^-$: $E_{sw} = -50$ mV, $E_i = 1230$ mV

$(E_{sw}/E_i)^+$: $E_{sw} = -300$ mV, $E_i = 980$ mV

3.4 Feature Definition and Extraction

Traditionally, peak potentials and peak current ratios of cyclic voltammograms are considered as the important parameters. In this study, the Fourier transform of each voltammogram is considered as a concise way of describing all shape factors.

To compare all voltammograms under the same conditions, the following procedures are followed:

1. The voltammogram curve is shifted so that the peak potential equals zero ($< \pm 0.01\text{mv}$).
2. The switching potential is set to either -50 or -300mv. Each curve is rotated so that the curves at the starting potential (E_s) and at the switching potential (E_{sw}) are at zero current.
3. The fast Fourier transform is calculated by using the Microsoft EXCEL[®] analysis tool package and the real part of the complex output (512 coefficients) is used as features.
4. The Fourier coefficients are normalized to the maximum absolute value. The normalized magnitudes are used as Fourier features for the further pattern recognition study.

3.5 Pattern Recognition Procedure

Voltammograms of all 156 “compounds” defined in tables 1 and 2 were analyzed by an unsupervised pattern procedure called Hierarchical Cluster Analysis. The Ein*Sight version 3.0 software package was employed to perform this cluster analysis procedure on these 156 “compounds”. The Ein*Sight software requires that all the Fourier transform data files be converted to text files before being read into the program. The first 80 Fourier coefficient data points from each “compound” were converted to text files and merged into a single input file for the Ein*Sight program and the cluster analysis procedure was performed.

Feature selection in this study was performed by sequentially adding features (from 3 to 80) according to increasing frequencies (feature 1 corresponds to 1 Hz and 1 Hz corresponds to 1 complete one- or two-cycle voltammogram period). This is a limited approach to feature selection. The best approach to feature selection would be to select from all possible combinations of features. However, this is an impossible task due to the amount of time required for human interpretation, much less for computer time. In any case, the feature selection approach in this initial study is considered a reasonable arbitrary approach, as it will at least demonstrate the frequency dependence of voltammetric information.

Figures 5, 6, and 7 represent dendrograms for the selection of 2, 3, and 4 clusters, respectively, using features 1 to 45. The dendrogram is used to show sample similarities, a metric derived from the N-dimensional distances between samples. The similarity

Figure 5

Figure 6

value of the “root” on the right side of the display is 0.00. The terminal nodes on the left side of the display, called “leaves”, are individual samples and have a similarity value of 1.00. All nodes between the root and leaves are “branches” and have similarity values between 0.00 and 1.00. Samples are considered all the more similar when they are near to each other in pattern space. The distance between samples is therefore a measure of similarity between these samples in the sense that a small distance means a high similarity and therefore a high “similarity index”.

A similarity line is a dotted line that runs through display region and allows the user to modify similarity, used for establishing sample clusters. The color bar (or gray-level) on the left side of the dendrogram is an aid to interpreting the relationships among samples, based on the clusters defined by a similarity threshold. When a dendrogram is generated by Hierarchical Cluster Analysis (HCA), the initial similarity value is 0, and the associated color bar is a single color, because all items are considered to be in the same cluster. As the cursor is moved to the left, the similarity value will change to a number between 0 and 1. Whenever the similarity line is moved to the left and cuts across horizontal branches of the dendrogram, the color bar will change to represent the number of clusters defined by that similarity value. Each time the similarity line is left at a position in which more branches of the tree are cut, the number of color bands increases correspondingly. The significance of identified clustering increases with the range of similarity index over which the cluster grouping is stable. Figures 5, 6, and 7 illustrate the selection of 2, 3, and 4 clusters, respectively; the minimum similarity index for each is 0.003, 0.437, and 0.602, and the range of similarity index for each is 0.434, 0.165, and

0.074, respectively. What was sought was the condition for the largest range of stability for a given type of clustering (2-, 3-, 4-, or 5-clusters).

Clustering analyses were performed for up to five clusters for eight experimental conditions of A-D and A2-D2 (Table 3). The results for experimental condition A were merged and are displayed graphically in Figure 8. This graph shows the range of similarity index for 2, 3, 4, and 5 clusters for features 1-3 through 1-80. For experimental condition A, the region over which the range of similarity index remains stable and large is from 1-40 to 1-49, for 2- to 5-cluster groups. The magnitude and stability of similarity index range is one of the two criteria which is used in this study for feature selection.

The second criterion for feature selection is to find a region where the size of each cluster (i.e, number of members of each cluster) also remains stable as well. For experimental condition A, this region of cluster size stability is in the range of features from 1-41 to 1-45 for each of the cluster groupings, 2, 3, 4, and 5. Figures 9, 10, 11, and 12 represent cluster size for features at 2-, 3-, 4-, and 5-cluster groupings respectively.

Thus, for experimental condition A features 1-45 were selected for evaluating cluster groupings, based on the fact that observed clustering best satisfies the two stability criteria defined above. Table 4 summarizes how features were selected for each experimental condition of A-D and A2-D2 based on the two criteria of stability defined above.

For comparison purposes, data from Fei Pang's studies (9) were also examined using Hierarchical Cluster Analysis. Her study focused on supervised pattern recognition of five different classes of compounds with different electrode mechanisms. The features

which indicated the highest overall classification accuracy were chosen for cluster analysis here. These included feature set 1 (Features 2,8,9) and feature set 2 (Features 2, 9, 22, 23) under experimental condition B2.

Range of Similarity Values for Specific Clustering Groups vs # of Features

Figure 8

Figure 9

Figure 10

Figure 11

Figure 12

Table 4. Feature Selection for Experimental Conditions A-D and A2-D2

Experimental Condition	Cluster SI Range Stability Feature #'s	Cluster Size Stability Feature #'s	Feature # Selected For Cluster Analysis
A	1 to 40 - 49	1 to 41 - 45	1 - 45
B	1 to 4 - 10	1 to 7 - 9	1 - 7
C	1 to 6 - 8	1 to 6 - 7	1 - 6
D	1 to 20 - 35	1 to 25 - 30	1 - 25
A2	1 to 4 - 10	1 to 7 - 8	1 - 7
B2	1 to 6 - 10	1 to 8 - 10	1 - 8
C2	1 to 20 - 35	1 to 25 - 30	1 - 25
D2	1 to 15 - 35	1 to 15 - 30	1 - 15

Chapter 4 Results and Discussions

4.1 Summary of Procedures Used

As an aid to interpreting results and understanding the observations in this section, the procedures used to acquire and evaluate data are summarized here. First, cyclic voltammograms for the hypothetical “compounds” were generated by digital simulation. Characteristic data for each “compound” were obtained (simulated) under eight conditions corresponding to different experimental parameters (Table 3) which were designated as A,B,C,D, and A2,B2,C2 and D2. Second, each cyclic voltammogram was converted to a frequency domain by fast Fourier transform. The first 80 Fourier coefficients were used in Hierarchical Cluster Analysis studies, *i.e.* unsupervised pattern recognition. Third, Hierarchical Cluster Analysis was employed to obtain equivalent clustering of data at various experimental conditions as a function of the number of features included. This was done by systematically increasing the number of features and looking for the point at which clustering appeared to stabilize. Fourth, information theory was applied to the results of the Hierarchical Cluster Analysis study to provide a more quantitative measure of information gain related to the experimental parameters.

4.2 Observations

The objective of this study was to learn what types of information related to compound properties could be observed with cyclic voltammetry in the frequency domain using Hierarchical Cluster Analysis. Tables 5-8 show the types of information that were obtained using cluster analysis in this study. These included information about mechanism (class of compounds), reversibility of charge transfer, and solution chemistry coupled to the charge transfer. Tables 5-8 also summarize the ΔH and $\Delta H_{\max}$ related to information gain for several types of information.

Table 5 lists the experimental conditions where most compounds exhibiting Quasi-reversible charge transfer ($k_s \ll 1$) are in one or two clusters. Tables 6 and 7 show 4-cluster and 5-cluster groupings which represent voltammograms based on electrode mechanism type (considering the four cases defined in section 2.3, where only case 1 mechanisms included were quasi-reversible). Table 8 lists various observed clusterings where a specific type of information was contained primarily in one cluster. Tables 5 to 8 also list the information gain associated with each observed clustering.

Equation (2) in section 2.4.2 was used to calculate information gain in Tables 5-8. In this equation, $n_o = J$ and n = the number of items in class R. The first entry in Table 5 is calculated using equation (2), where $n_o = J = 156$, the total number of compounds. There are two classes of 13 reversible and 143 quasi-reversible compounds each providing the values of n in each of two terms. Clustering separates 11 and 124, respectively, into

Table 5. Observations of Quasi-Reversible (QR) voltammograms falling mostly in one or two clusters, $\Delta H_{\max} = 0.42$ bits

Voltammetric Conditions	QR Clusters	Total # of Clusters	%QR in Cluster vs. Total QR	%QR in Cluster vs. Total in Cluster	ΔH , bits	% of $\Delta H_{\max}$
B2	1	2	87%	92%	0.35	83
A2	1	4	73%	91%	0.32	76
B2	1,2	5	29%	100%	0.32	76
D	2,3	3	49%	100%	0.31	74
D	3,4	4	49%	100%	0.29	69
D2	2	2	76%	96%	0.27	64
D2	2,3	3	36%	100%	0.27	64
D2	2,3	4	29%	100%	0.26	62
C2	2	2	78%	95%	0.25	60
A	2,3	5	30%	100%	0.24	57
C2	2,3	4	36%	100%	0.24	57
D2	3,4	5	29%	100%	0.24	57
C2	2,3	5	31%	100%	0.23	55

Table 6a. Observations of 4-cluster voltammograms with information regarding 3 mechanism classes*, $\Delta H_{\max} = 1.48$ bits

Voltammetric Conditions	Total # of Clusters	ΔH , bits	% of $\Delta H_{\max}$
B	4	1.07	72
A	4	0.89	60
B2	4	0.79	53
C	4	0.75	51
D	4	0.73	49

*classes separated were: E(QR), CE, EC

Table 6b. Observations of 4-cluster voltammograms with information regarding 4 mechanism classes*, $\Delta H_{\max} = 1.98$ bits

Voltammetric Conditions	Total # of Clusters	ΔH , bits	% of $\Delta H_{\max}$
C2	4	0.92	46
A2	4	0.88	44
D2	4	0.81	41

*classes separated were: E(QR), CE, EC, ECE

Table 7. Observations of 5-cluster voltammograms with information regarding 4 mechanism classes*, $\Delta H_{\max} = 1.98$ bits

Voltammetric Conditions	Total # of Clusters	ΔH, bits	% of $\Delta H_{\max}$
B	5	1.07	54
A2	5	1.00	51
C2	5	0.92	46
D2	5	0.92	46
B2	5	0.91	46
A	5	0.83	42
C	5	0.75	38
D	5	0.69	35

*classes separated were: E(QR), CE, EC, ECE

Table 8. Observations of various types of information mostly in a single cluster

Types of Information	% of type in Cluster	% of Cluster This Type	Clustering Conditions			ΔH	$\Delta H_{\max}$	% of $\Delta H_{\max}$
			Expt.	Total # of Clusters	Cluster #			
ECE	90	49	A	2	2	0.66	0.82	80
ECE	90	49	A	3	3	0.66	0.82	80
Low K_{eq} ^(a)	98	60	B2	2	1	0.61	0.96	64
High K_{eq} ^(b)	87	54	C2	2	2	0.53	0.96	55

^(a)Low K_{eq} = 0.05, 0.005, 1.0 for EC, CE, ECE mechanism respectively

^(b)High K_{eq} = 2.5, 0.1, 100 for EC, CE, ECE mechanism respectively

different clusters, providing the values of C_R in each of two terms; thus the information gain is calculated as:

$$\Delta H = (11/156) \log_2 (156/13) + (124/156) \log_2 (156/143)$$

$$\Delta H = 0.25 + 0.10 = 0.35 \text{ bits}$$

The maximum possible information gain would be obtained if all compounds were ideally separated into two clusters, $C_R = 13$ and 143 for each term.

$$\Delta H_{\max} = (13/156) \log_2 (156/13) + (143/156) \log_2 (156/143)$$

$$\Delta H_{\max} = 0.30 + 0.12 = 0.42 \text{ bits}$$

The information gain for the first entry in the Table 6a is calculated as follows. There are three classes containing 35, 40, and 40 compounds each; and 28, 27, and 27 compounds, respectively, are separated among four clusters (where one class occupies two clusters).

Thus, the information gain is calculated as:

$$\Delta H = (28/156) \log_2 (156/35) + (27/156) \log_2 (156/40) + (27/156) \log_2 (156/40)$$

$$\Delta H = 0.39 + 0.34 + 0.34 = 1.07 \text{ bits}$$

The maximum possible information gain is:

$$\Delta H_{\max} = (35/156) \log_2 (156/35) + (40/156) \log_2 (156/40) + (40/156) \log_2 (156/40)$$

$$\Delta H_{\max} = 0.48 + 0.50 + 0.50 = 1.48 \text{ bits}$$

The information gain for the first entry in the Table 7 is calculated as follows. There are four classes containing 35, 40, 40, and 40 compounds each; and 28, 11, 27, and 10 compounds, respectively, are separated among five clusters (where one class occupies two clusters). Thus, the information gain is calculated as:

$$\Delta H = (28/156) \log_2 (156/35) + (11/156) \log_2 (156/40) + (27/156) \log_2 (156/40) + (10/156) \log_2 (156/40)$$

$$\Delta H = 0.39 + 0.14 + 0.34 + 0.13 = 1.00 \text{ bits}$$

The maximum possible information gain is:

$$\Delta H_{\max} = (35/156) \log_2 (156/35) + (40/156) \log_2 (156/40) + (40/156) \log_2 (156/40) + (40/156) \log_2 (156/40)$$

$$\Delta H_{\max} = 0.48 + 0.50 + 0.50 + 0.50 = 1.98 \text{ bits}$$

For the first two entries in Table 8, the clustering appears to provide two classes of information, ECE mechanism vs all other mechanisms. Thus, the information gain for the first entry in Table 8 is calculated as follows. There are two classes of information containing 40, and 116 compounds each; and 36, and 78 compounds, respectively, are contained in each of two clusters. Therefore, the information gain is calculated as:

$$\Delta H = (36/156) \log_2 (156/40) + (78/156) \log_2 (156/116)$$

$$\Delta H = 0.45 + 0.21 = 0.66 \text{ bits}$$

The maximum possible information gain is:

$$\Delta H_{\max} = (40/156) \log_2 (156/40) + (116/156) \log_2 (156/116)$$

$$\Delta H_{\max} = 0.50 + 0.32 = 0.82 \text{ bits}$$

4.2.1 Summary of Observations with Unsupervised Pattern Recognition

From Tables 5 to 8, it is clear that several types of information can be obtained from the frequency domain analysis of cyclic voltammograms. Table 5 illustrates that many experimental and clustering conditions lead to an unsupervised discrimination between reversible ($k_s \gg 1$) and quasi-reversible ($k_s \ll 1$) charge transfer behavior, regardless of the overall electrode mechanism. Although the maximum possible information gain for this discrimination with this data set is small (0.42 bits), it is useful to know that this type of information is extracted (up to 83% of $\Delta H_{\max}$) with the relatively crude frequency domain feature selection procedure used here.

Tables 6 and 7 illustrate that the information regarding type of electrode mechanism (“compound type”) can also be detected with nonsupervised clustering. Here the maximum possible information content is greater (1.98 bits for 4-class separation), and as much as 1.07 bits were observed. The degree of discrimination is not as great as observed with supervised pattern recognition, which uses more involved and target-directed feature selection (see discussion of Fei Pang’s work below).

Table 8 lists observations of clustering where one type of information was found primarily in a single cluster. It is interesting to note that one type of information observed in this manner was that related to the magnitudes of the equilibrium constant for electrode processes involving a coupled chemical reaction.

4.2.2 Summary of Observations with Supervised Pattern Recognition

Tables 9 and 10 are summary observations from Fei Pang's studies (9). Table 9 is feature set 1 (Features 2,8,9) with B2 experimental condition and Table 10 is feature set 2 (Features 2,9,22,23) with B2 experimental condition. These sets of data were studied using the Hierarchical Cluster Analysis procedure. Tables 9 and 10 show the types of information are mostly in a single cluster.

4.3 Discussions

The effects of scan rate, number of cycles, and switching potential were examined for training performance by Fei Pang (9). She observed that there is a significant difference in information content between one-cycle and two-cycles, with the highest information content obtained with two-cycles. At lower scan rate she observed that the class-related information content is higher. She also observed, for the ranges of values studied, there is no significant effect of the switching potential on information content.

In Fei Pang's study, only one type of information was considered, namely compound class (type of mechanism). In this study, information about mechanism type, Quasi-Reversible vs. Reversible, and chemical properties was obtained. This type of information was obtained with the relatively crude frequency domain feature selection procedure used in this study. In contrast to the supervised pattern recognition procedure of Fei Pang (9), the unsupervised clustering study here does not allow feature selection

based on targetted class information. This explains why Pang's work demonstrated that greater information gain could be obtained when class distinctions are defined in advance, and the most effective features for specific class discrimination can be selected. On the other hand, this study has demonstrated what other types of information are contained in the frequency domain voltammograms. Subsequent studies can now be done with targetted feature selection based on newly defined information classes, with greater information gain using supervised pattern recognition.

Table 9. **Feature Set 1 (Features 2, 8, 9)**
Observations of various types of information mostly in a single cluster

			Clustering Conditions					
Types of Information	% of Type in Cluster	% of Cluster This Type	Expt.	Total # of Cluster	Cluster #	ΔH	$\Delta H_{\max}$	% of $\Delta H_{\max}$
E mechanism, Quasi-Reversible	60	30	B2	3	1	0.46	0.76	61
EC mechanism	58	30	B2	2	2	0.46	0.82	56
CE mechanism	50	28	B2	3	1	0.43	0.82	52
ECE mechanism	53	26	B2	2	1	0.40	0.82	49

Table 10. Feature set 2 (Features 2, 9, 22, 23)
Observations of various types of information mostly in a single cluster

Types of Information	Clustering Conditions					% of $\Delta H_{\max}$
	% of Type in Cluster	% of Cluster This Type	Expt.	Total # of Cluster	Cluster #	
E mechanism, Quasi-Reversible	60	28	B2	2	1	54
EC mechanism	58	28	B2	2	2	55
CE mechanism	65	32	B2	2	2	60
ECE mechanism	53	28	B2	2	1	52
						ΔH
						$\Delta H_{\max}$

Chapter 5 Conclusions

This study has demonstrated for the first time the value of unsupervised pattern recognition for characterizing the information content of frequency domain cyclic voltammograms. Results of this study indicate that voltammetric properties are indeed intrinsically related to several different types of information. Informative clustering occurs at various experimental conditions, and with various frequency components, although the most effective frequency components appear to be < 25 Hz (Table 4), where one complete one-cycle (or two-cycle) voltammogram corresponds to 1 Hz.

This study showed the type of information that can be learned from unsupervised pattern recognition includes mechanism type, reversibility of charge transfer, and properties of coupled chemical equilibria. This study was limited, however, by a very crude feature selection procedure. It is clear that more information can be obtained for specific target information with supervised pattern recognition, because one can direct feature selection. A future benefit of this study will be to define what other information might be sought with supervised pattern recognition. This study revealed the type of information that might be present with supervised pattern recognition. Thus, in the future we can use supervised pattern recognition with targeted feature selection for information that we now know is present.

REFERENCES

1. W. A. Byers, B. S. Freiser, and S. P. Perone, Anal. Chem., vol. 55, 620, 1983
2. S. D. Schachterle and S. P. Perone, Anal. Chem., vol. 53, 1672, 1981
3. D. R. Burgard and S. P. Perone, Anal. Chem., vol. 50, 1366, 1978
4. W. A. Byers and S. P. Perone, Anal. Chem., vol. 55, 615, 1983
5. S. P. Perone and C. L. Ham, Journal of Research of the National Bureau of Standards, vol. 90, 531, 1985
6. Q. V. Thomas and S. P. Perone, Anal. Chem., vol. 49, 1369, 1977
7. Q. V. Thomas, R. A. DePalma, and S. P. Perone, Anal. Chem., vol. 49, 1376, 1977
8. M. Zhou, Master of Science Thesis, 1995, San Jose State University
9. F. Pang, Master of Science Thesis, 1997, San Jose State University
10. D. A. Skoog, J. J. Leary, Principle of Instrumental Analysis, Saunders College Publishing, 1992, 535-578
11. A. J. Bard and L. R. Fualkner, Electrochemical Methods: Fundamentals and Applications, John Wiley & Sons: New York, 1980, 213 - 248
12. R.S. Nicholson and I. Shain, Anal. Chem., vol. 36, 706, 1964
13. B. Speiser, Comput. Chem., vol. 14, 127, 1990
14. D. K., Jr. Gosser, Cyclic Voltammetry: Simulation and Analysis of Reaction Mechanisms, VCH, Inc.: New York, 1993
15. ELSIM, Technical Software Distributors, 1016 Hartmont Rd., Baltimore, MD 21228
16. M. Rudolph, J. Electroanal. Chem., vol. 314, 13, 1991

17. M. Rudolph, J. Electroanal. Chem., vol. 338, 85, 1992
18. K. Eckschlager, Anal. Chem., vol. 54, 11, 1982
19. C. E. Shannon, Bell System Technical Journal, vol. 27, 379, 1948
20. D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier Scientific Publishing Company, 1978, 309-314
21. W. A. Byers and S. P. Perone, Anal. Chem., vol. 52, 2173, 1980
22. P. A. Boudreau and S. P. Perone, Anal. Chem., vol. 51, 811, 1979
23. D. Coomans and D. L. Massart, Analytica Chimica Acta, vol. 133, 225, 1981
24. B. R. Kowalski, Chemometrics, Mathematics and Statistics in Chemistry, D. Reidel Publishing Company, 1983, 393-401
25. R. A. DePalma and S. P. Perone, Anal. Chem., vol. 51, 825, 1979
26. R. A. DePalma and S. P. Perone, Anal. Chem., vol. 51, 829, 1979