1997

Information gain study on cyclic voltammetry

Fei Pang
San Jose State University

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

Recommended Citation

Pang, Fei, "Information gain study on cyclic voltammetry" (1997). *Master's Theses*. 1521.
DOI: <https://doi.org/10.31979/etd.9k7k-425j>
https://scholarworks.sjsu.edu/etd_theses/1521

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 ON 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

Fei Pang

August, 1997

UMI Number: 1386222

UMI Microform 1386222

Copyright 1997, 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

Fei Pang

ALL RIGHTS RESERVED

**APPROVED FOR THE DEPARTMENT
OF CHEMISTRY**

Dr. Sam Perone

Dr. Joseph J. Pesek

Dr. Brad Stone

APPROVED FOR THE UNIVERSITY

ABSTRACT

INFORMATION GAIN STUDY ON CYCLIC VOLTAMMOGRAM

by Fei Pang

This work describes an evaluation of information content of the electrochemical technique-cyclic voltammetry. The information goal in this study is qualitative identification of redox couples and electrode processes based on voltammetric shape properties only.

The effect of experimental parameters, including number of cycles, scan rate and switching potential, on information content is studied. The voltammetric information content is related to the ability to distinguish among several different electrode reaction mechanisms. This evaluation was done using K-nearest neighbor (KNN) computerized pattern recognition. The influence of training procedures on the KNN study was also studied.

ACKNOWLEDGMENTS

I would like to thank my academic advisor Dr. Perone for his guiding and support. I would not be able to fulfill my thesis without his help. I would also like to thank all of my committee members, Dr. Pesek and Dr. Stone.

Also, I would like to thank Daniel Severance for his great help on programming.

Finally, my special appreciation is given to my family.

Table of Contents

Chapter 1	Introduction	
1. 1	Previous Work.	1
1. 2	Objective of the Present Study.	2
Chapter 2	Theory	
2. 1	Voltammetry.	4
2. 2	Cyclic Voltammetry.	6
2. 3	Digital Simulation.	9
2. 4	Information Theory.	11
	2. 4. 1 Information Theory Concepts.	11
	2. 4. 2 Information Gain.	13
2. 5	Pattern Recognition.	13
	2. 5. 1 Classification Method: KNN.	16
2. 6	Fast Fourier Transform.	22
Chapter 3	Experimental Approach	
3. 1	Instrumentation.	23
3. 2	Data Base Management and Data Analysis.	24
3. 3	Digital Simulations.	25
	3. 3. 1 Experimental Mechanisms and Compound Classes for Cyclic Voltammogram.	25
	3. 3. 2 Experimental Parameters for Cyclic Voltammetry.	31
3. 4	Feature Definition and Extraction.	31
3. 5	Pattern Recognition Procedure.	32
	3. 5. 1 Effects of Experimental Parameters.	34
Chapter 4	Results and Discussions	

4. 1	Effects of Experimental Parameters.	45
4. 1. 1	Scan Rate.	46
4. 1. 2	Number of Cycles.	47
4. 1. 3	Switching Potential.	47
4. 2	Effects of Training Procedure.	47
4. 2. 1	Comparison between Class-Weighted Training and KNN Overall Training.	48
4. 3	KNN Study Applied to Prediction.	49
4. 4	Discussions.	55
4. 4. 1	Training Results.	55
4. 4. 2	Prediction Results.	56
Chapter 5	Conclusions.	57
Chapter 6	Further Study.	58
Reference		59

List of Tables

Table 1a. Chemical Parameters for EC Mechanism.	27
Table 1b. Chemical Parameters for CE Mechanism.	28
Table 1c. Chemical Parameters for ECE Mechanism.	29
Table 2a. Chemical Parameters for E (Quasi - Reversible) Mechanism..	30
Table 3. Experimental Parameters for Simulated Voltammograms.	35
Table 4. Factorial Design for KNN Study.	36
Table 5. KNN Study of Voltammograms Obtained under Condition A.	37
Table 6. KNN Study of Voltammograms Obtained under Condition B.	37
Table 7. KNN Study of Voltammograms Obtained under Condition C.	38
Table 8. KNN Study of Voltammograms Obtained under Condition D.	38
Table 9. KNN Study of Voltammograms Obtained under Condition A2.	39
Table 10. KNN Study of Voltammograms Obtained under Condition B2.	39
Table 11. KNN Study of Voltammograms Obtained under Condition C2.	40
Table 12. KNN Study of Voltammograms Obtained under Condition D2.	40
Table 13. Summary of Highest Information Gain for Each Experimental Condition (Extracted from Tables 5-12).	41
Table 14. Factor Effects on Information Gain.	41
Table 15a. Classification Training Accuracy from Class-Weighted Training and Overall Training.	42
Table 15b. Classification Training Accuracy from Class - Weighted Training and Overall Training.	43
Table 15c. Summary of Highest Training Information Gain for Each Class from Class - Weighted Training.	44

Table 16. Prediction Results, Overall Classification.	51
Table 17a. Prediction Results, Class - Specific Classification (Prediction Features Selected from Tables 15a. and 15b.).	52
Table 17b. Prediction Results, Class - Specific Classification (Prediction Features3 Selected from Tables 15a. and 15b.).	53
Table 18. Summary of Highest Information Gain for Each Class.	54

List of Figures

Fig 1. Potential Excitation Signals used in Voltammetry.	5
Fig 2. Quasi-Reversible Electron Transfer Case, $A + e \rightleftharpoons B$, $B \rightleftharpoons C$, $C + e \rightleftharpoons D$, when $K_s=0.0005$, $\alpha=0.5$, One Cycle, (a): $v=0.05\text{v/s}$, (b): $v=1.0\text{ v/s}$	8
Fig 3. 2-Dimensional KNN Example.	17
Fig 4. Euclidean Distance Calculation of Pattern Vectors in N-Space.	19
Fig 5. Supervised Pattern Recognition Procedure.	20

Chapter 1 Introduction

1. 1 Previous Work

Voltammetry is widely used by organic, physical, inorganic and biological chemists for non-analytical purposes, including fundamental studies of oxidation and reduction processes in various media, adsorption processes on surfaces and electron transfer mechanisms at chemically modified electrode surfaces.

In the 1970s, it was suggested that information extracted from the shape of voltammograms could be used to classify compounds (1 - 8).

In 1983, Byers, Freiser and Perone (9 - 10) analyzed 45 compounds using cyclic staircase voltammetry. The data set consisted of 19 nitrobenzene, 9 nitrodiphenyl ethers, and 17 ortho-hydroxy azo compounds. Of the nitrodiphenyl ethers, 4 were strong herbicides and 5 were either weak or nonherbicides. The information goals for this problem were the classification of the 45 compounds by their structural type and the classification of the 9 nitrodiphenyl ethers according to herbicidal activity. Seven experimental variables, percent ethanol in the solvent, pH, surfactant concentration, number of cycles, scan rate, mercury drop hang time, and sampling time were varied in a fractional factorial design to generate a complete data base of collected cyclic staircase voltammograms and cyclic differential capacity curves for subsequent analysis. They suggested that "The effect of each variable was unique, so that redundant information was not recorded. The information content appeared to be large. " (9)

In the same year, Barnes and Perone studied the enhancement of chemical process information through experimental design. A simple model of controlled potential electrochemical processed based upon the Cottrell equation was developed and implemented. The information goal was to determine the effects of input voltage sequence, data collection fraction to analyze, and preprocessing scheme upon the determination of the diffusion coefficient.

Ming Zhou in 1995 (11) studied the effect of several experimental parameters on the information content of cyclic voltammetry. The parameters she studied included scan rate and number of cycles. Synthetic voltammograms were generated representing a variety of “compounds” with different classes of electrochemical behavior. She found that voltammograms yielded greater class related information content at lower scan rates for the classes of compounds studied.

Zhou’s work was limited by the number of compounds and classes studied; some natural subsets of compounds were not incorporated.

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. The first step towards this goal would be to understand which parameters in a cyclic voltammetric experiment affect qualitative identification of compounds, and determine the quantitative impact on information content.

This thesis presents an expanded experimental design, expanded with respect to previous works (11). With the help of information theory, the information content

yielded from different experimental parameters, such as number of cycles, scan rate and switching potential can be quantitatively calculated. Thus, it is possible to select the most influential parameters affecting class related information content.

Chapter 2. Theory

2. 1 Voltammetry

Voltammetry is based upon the measurement of a current that flows in an electrochemical cell under conditions of complete concentration polarization. The information about the analytes is derived from the measurement of current as a function of applied potential (12).

In voltammetry, a variable potential excitation signal is impressed on an electrochemical cell containing 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 wave forms of excitation signals used in voltammetry are shown in Figure 1 (12).

1. Linear sweep voltammetry

The classical voltammetric excitation signal (Fig1(a)) is the linear scan, in which the dc potential applied to the cell increases linearly as a function of time. The current that develops in the cell is then recorded as a function of time.

2. Differential pulse polarography and square wave voltammetry

Two pulse-type excitation signals are shown in Fig1(b) and Fig1(c). Current is measured at various times during the lifetime of these pulses.

<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

Fig 1. Potential Excitation Signals Used in Voltammetry (12)

3. Cyclic voltammetry

With the triangular waveform shown in Fig1(d), the potential is cycled between two values, first increasing linearly to a maximum and then decreasing linearly to its original value. This process may be repeated numerous times, with the current being recorded as a function of time.

Since cyclic voltammetry is the technique used in this study, it will be discussed in detail.

2. 2 Cyclic Voltammetry

Cyclic voltammetry (CV) is an important tool for the study of mechanisms and rates of oxidation- reduction processes, particularly in organic and metal-organic systems.

The CV technique is relatively straightforward (13,14). A working electrode is immersed in a solution containing some electroactive analyte, an excess of electrolyte, a reference electrode, and a counter electrode. The potential between the working electrode and reference is controlled by a potentiostat which passes a current between the working and auxiliary electrodes to maintain or attain a desired potential.

The experimental objective is to choose a starting potential for the working electrode where the analyte is electroinactive and then to change the potential at a constant rate, $dE/dt = v$. For diffusion controlled processes a wide range of values of v have been used: 0. 01 V/s to $>2 \times 10^6$ V/s, moving from Estart through a potential range where oxidation or reduction of the analyte can occur, to a reversing potential, Erev(reverse potential), where the sweep direction is reversed and the potential returns to

Estart. This would complete a single cycle but the process can be repeated any number of times.

Fig 2. gives an example of cyclic voltammogram. ($A + e \rightleftharpoons B$)

The shape of the voltammogram is a function of overall mechanisms of the electrode process, the experimental parameters (E_{start} , E_{switch} , scan rate), the heterogeneous electron transfer rate parameters (k_s , E^0 , α), the diffusion coefficients and concentrations of each species, rates of coupled chemical reactions, temperatures, 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^0 is the standard potential.

Fig 2 Quasi - Reversible Electron Transfer Case, $A + e \rightleftharpoons B$, $B = C$, $C + e \rightleftharpoons D$, when $K_s = 0.0005$, $\alpha = 0.5$, One Cycle, (a): $v = 0.05 \text{ V/s}$, (b): $v = 1.0 \text{ V/s}$

2.3 Digital Simulation

The basic concepts of digital simulation of electrochemical phenomena have been discussed in detail by several authors (15,16). The validity and usefulness of the technique have been indicated by application to a variety of problems.

The outstanding virtue of this technique is the ease with which one can translate electrode phenomena into operational mathematical terms. Basic digital simulation techniques have been demonstrated to be most useful in predicting the effects of diffusion, convection, and simple chemical reactions on a wide variety of current-potential-time experiments. These methods have been employed to handle most of the elementary electrochemical problems that have been developed over the past two decades, and it will continue to be useful in solving the problems associated with mass transfer and chemical reaction.

Nicholson and Shain launched the modern era of theoretical characterization of electrochemical cyclic voltammetric responses and initiated the evolution of cyclic voltammetry into the most commonly applied electrochemical technique (17).

A cyclic voltammogram is a complicated, time dependent function of the large number of physical-chemical parameters which describe any electrochemical-chemical system. Unraveling the coupled dynamics of heterogeneous electron transfer, homogeneous chemical kinetics, and diffusion, presents a formidable experimental and computational challenge (18-20). Even the computation of the cyclic voltammetric response for a simple electron transfer ($A + e^- \rightleftharpoons B$) requires numerical integration. Previous attempts to produce a general CV simulator have had limited success (21-24).

Speiser has utilized orthogonal collocation as the basis of EASIEST which accessed a library of algorithms to compute the cyclic voltammetric responses for a

number of specific mechanisms, and which currently uses a mainframe computer. Gosser has devised DVSIM, a CV simulator based on an explicit finite difference (EFD) algorithm which is useful for simulating CV responses for simpler systems. Bieniasz has devised ELSIM, a potpourri of mathematical and numerical techniques which can be used to simulate various electrochemical responses and which demands a mathematically sophisticated user.

Now, this computational challenge is effectively met by Rudolph's recently developed fast implicit finite difference (FIFD) algorithm (25-27). This simulator is truly general and efficient. It accepts any user specified mechanism and efficiently simulates CV response for any systems.

The commercial software DigiSim® (Bioanalytical Systems, W. Lafayette, IN) 1. 0 is used in this study to generate theoretical CV voltammograms. The DigiSim® program is capable of generating voltammograms with a wide range of electrochemical-chemical mechanisms.

Four mechanisms were studied in this research: 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 reaction. (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

The homogeneous chemical 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

The chemical reaction is characterized by K , k_f and k_b .

4. ECE mechanism

It is given by $A + ne^- \rightleftharpoons B$

Two electron transfer reactions are characterized by rate constants k_{s1} , k_{s2} , the chemical reaction is characterized by K , k_f and k_b .

2. 4 Information Theory

2. 4. 1 Information Theory Concepts

Information theory is concerned with the study of information and its transmission. The earliest information theorists studied the encoding and decoding of secret codes. Modern work on information theory can be traced back to the 1920's beginning with

Carson's study of frequency modulation. Nyquist determined the minimum bandwidth required for the transmission of a definite quantity of information. Hartley established a definite bandwidth-time product required for the transmission of a definite quantity of information (1).

Most of the present work in information theory is based upon probabilistic models of communication developed by C. E. Shannon in 1948. The basic principle 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 equated information and uncertainty. Entropy is a quantitative measure of the amount of information supplied by a probabilistic experiment. It is based upon classical Boltzmann entropy from statistical physics. Shannon's formula defines the entropy or the average information to be equal to the weighted average of the specific information:

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

where the symbol $p(x_i)$ is the probability of occurrence of an event, and N is the number of possible events. The logarithmic base corresponds to the choice of a unit for measuring information. Since base 2 is used, the resulting units may be called binary digits, which was suggested by J. W. Tukey.

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

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

$$\Delta H = - \sum_{R=1}^Z (C_R/J) \log_2 (n_0/n) \quad (2)$$

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. For an incorrect result, $n=n_0$. Thus, in equation (2), $n_0=J$ and n =the number of items in class R . For example, if $J=100$ and there are two classes containing 40 and 60 compounds each, if pattern recognition identifies 35 and 45 correctly from each class,

$$\begin{aligned} \Delta H &= (35/100)\log_2(100/40) + (45/100)\log_2(100/60) \\ &= 0.463 + 0.332 = 0.795 \end{aligned}$$

The maximum information gain is:

$$\begin{aligned} \Delta H_{\max} &= (40/100)\log_2(100/40) + (60/100)\log_2(100/60) \\ &= 0.97 \text{ bits} \end{aligned}$$

2. 5 Pattern Recognition

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, it is difficult to decide which are the most useful. Computerized

pattern recognition is one of the best means to select appropriate measurements and resulting classifications (4).

Pattern recognition can be described by four sequential steps for the purpose of discussion: measurements, 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 and “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.

The transformation may be as simple as to multiply each measurement by a scale factor, or it may make new 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 time domain into a corresponding function in the Fourier or frequency domain. In the Fourier domain, deconvolution is accomplished by separating out each of the pure components (3).

Sharper peaks produce higher frequency components in the Fourier domain. Because noise also has high frequency components, these may be isolated in the frequency domain and removed from contribution (5).

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.

In the last step of pattern recognition, each compound or chemical system is identified by assigning it to some predefined class. 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.

Two common classifiers which work on this principle are the linear discriminant function and the k nearest neighbor (KNN) rule.

In the case of the linear discriminant, a hyperplane is drawn in space such that it defines class boundaries. In the case of the nearest neighbor rule, each pattern is assigned to the class which has the greatest representation among its k nearest neighbors. Euclidian distance is usually taken as the measure of nearness. The nearest neighbor rule often produces higher classification accuracies than the linear discriminant, but it is slower to implement.

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.

2. 5. 1 Classification Method: KNN

Pattern recognition assumes that the characteristic properties or individual identities of objects which distinguish one from another can be determined by examination of pattern vectors associated with each object. Classification involves making a proper decision to assign each pattern to one of the possible pattern classes. The decision is usually based on comparing pattern vectors for unknown objects with those of known objects. It is assumed that pattern vectors for items of the same class occupy the same region of space.

Fig 3. 2-Dimensional KNN Example

The K-nearest neighbor (KNN) classification is a standard method in pattern recognition. Fig3 shows an example of 2-dimensional KNN. In this method, each pattern can be considered as a point in the N - dimensional feature space where N = 2 for Fig 3.

The distance between any two pattern vectors, I and J, in the N-space is defined by Euclidean geometry, and Euclidean distance calculation of pattern vectors in N-space is shown in Fig4.

The equation for calculating the distance is:

$$D_{ij} = \left[\sum_{n=1}^N (x_{in} - x_{jn})^2 \right]^{1/2} \quad (3)$$

where, D_{ij} : distance between two pattern vectors in N space

i, j : refer to specific pattern vectors i and j

n : index for features

N : total number of features

x_{in} : value of feature n for pattern i

x_{jn} : value of feature n for pattern j

In order to find out the nearest neighbors for each pattern, it is necessary to compute the distance between each specific pattern and all other pattern vectors in the feature space. In the classification procedure, an unknown pattern is classified according to the majority vote of its K-nearest neighbors.

Supervised pattern recognition involves identifying an unknown item by including its pattern vectors in a set of pattern vectors for items of known class. Identification is based on the majority class of nearest neighbors by computerized calculation of inter-item distances in feature space.

Fig 5. gives an example of a supervised pattern recognition procedure(28).

Fig 4. Euclidean Distance Calculation of Pattern Vectors in N-Space

Fig 5. Supervised Pattern Recognition Procedure

Pattern recognition is usually first carried out on a training set to develop a classifier that recognizes the class membership of these patterns as well as possible(7). The procedure is called training. The true identity of each pattern is compared to the identity assigned to it in the classification step, the percentage of correctly classified training set patterns is called recognition rate. Also it is the classification accuracy of the classifier.

If the recognition rate is poor, then other measurements or alternative data transformations should be tried until an acceptably high accuracy is achieved. Once a sufficiently accurate decision rule has been obtained by the training procedure, the next step is to demonstrate that it is reliable. This can be done by observing how successful it is at classifying a different set of known patterns, the prediction set. These patterns have essentially the same origins as the training set and represent the same classes included in the training set, but are not part of the training set. This process of applying the trained classification rule to the prediction set is called prediction. The percentage of correctly classified prediction set patterns is called prediction ability. If patterns of the prediction set are also correctly classified with high accuracy, the classification rule can be considered valid and reliable. If the decision rule is successful for the prediction set, it can be concluded that the sought for information is indeed contained in the data matrix. Truly unknown patterns may then be analyzed as long as each is collected in the same manner as the training/prediction sets and belongs to one of the classes represented in the training/prediction set.

2. 6 Fast Fourier Transform

Fourier Transform is a technique for solving problems in linear systems and analyzing periodic data(4,5, 29, 30). It is expressed in the following equations:

$$F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \nu t} dt \quad (4)$$

$$F(\tau) = \int_{-\infty}^{\infty} f(\nu) e^{-2\pi i \tau \nu} d\nu \quad (5)$$

$F(\nu)$ 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 equation:

$$F(\nu) = a_0 f(\nu_0) + a_1 f(\nu_1) + \dots + a_n f(\nu_n) \quad (6)$$

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 goals.

Thomas and Perone have shown that the coefficients obtained from the discrete fourier transform of voltammetric curves are useful for curve shape analysis. (4,5) 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, FFT features are apparently good enough for the representation of analytical wave forms for pattern recognition, when intrinsic wave shape information is sought.

Burgard and Perone 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. (6)

Chapter 3 Experimental Approach

3. 1. Instrumentation

An IBM clone personal computer (PC) is used for database management, computerized digital simulation , fast fourier transform and K-nearest neighbor pattern recognition.

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. It is recommended that the RAM is at least 2M bytes. The DOS configuration requires 2M byte RAM and 20M 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 $E=0.0025V$ is about 11. 15 sec for E mechanism, 22. 08 sec for EC mechanism, 22. 08 sec for CE mechanism, 69. 15 sec for ECE mechanism. There is no big difference in run time when a , k_s , K , K_f are changed (see section 2. 2). For two cycle CV, the run time is about the same as the one-cycle because the step size is changed to 0. 005v.

Fast fourier analysis can be performed under Microsoft EXCEL® 4. 0 or 5. 0 with data analysis tool package. It takes about 2 minutes for 1024 data points on a 486 DX/25 MHz PC.

For pattern recognition, the minimum configuration requires 1 M byte RAM and 20 M byte hard disk, with a 4 MHz 286 microprocessor. For some more complicated applications, a similar system with a 12 MHz 386 microprocessor and 8087 math co-processor is preferred. The software KNNPRA was written in Microsoft compiled BASIC, version 1.1.

3. 2 Data Base Management and Data Analysis

All raw data generated from DigiSim® were transferred to Microsoft EXCEL® spreadsheet data base management system. Rotated cyclic voltammograms, basic statistical computations, fourier feature transform calculation, normalized fourier feature transform data and associated graphical procedures were conducted in Microsoft EXCEL®. Pattern recognition analysis are conducted with KNNPRA 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 (less than $\pm 0.01\text{mv}$). This step avoids introducing extraneous shape information into the transform.

Second, each curve was rotated and translated prior to the fourier transform to retain the most fidelity in the transform. Each voltammogram was rotated and translated in such a way that high frequency artifacts in the transform are avoided, so that the first and last data points were zero, and no discontinuities were introduced. The raw CV data file was translated (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 number from the output data was 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 1.0 to 1.0. Those features were used for further pattern recognition study.

3. 3 Digital Simulations

3. 3. 1 Experimental Mechanisms and Compound Classes for Cyclic Voltammogram

The purpose of this study is to determine voltammetric information content 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, 155 “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 these “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 (section 2. 2).

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

ks	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	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	<u>6</u>	47	28	<u>9</u>	50	<u>31</u>	<u>12</u>	53	<u>77</u>	100	120	<u>80</u>	34	<u>15</u>	<u>56</u>	37

ks	0.005											
α	0.3				0.5				0.7			
K	0.05		2.5		0.05		2.5		0.05		2.5	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	<u>103</u>	131	83	<u>106</u>	122	<u>86</u>	<u>109</u>	125	<u>89</u>	112	136	<u>92</u>

ks	0.05											
α	0.3				0.5				0.7			
K	0.05		2.5		0.05		2.5		0.05		2.5	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	18	<u>59</u>	<u>40</u>	21	<u>115</u>	127	95	<u>118</u>	62	<u>43</u>	<u>24</u>	65

Table 1b. Chemical Parameters for CE Mechanism

ks	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	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	<u>26</u>	7	48	<u>29</u>	10	<u>51</u>	<u>32</u>	13	<u>98</u>	78	101	<u>121</u>	54	<u>35</u>	16	57

ks	0.005											
α	0.3				0.5				0.7			
K	0.005		0.1		0.005		0.1		0.005		0.1	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	<u>81</u>	104	132	<u>84</u>	107	<u>123</u>	<u>87</u>	110	<u>134</u>	90	113	<u>137</u>

ks	0.05											
α	0.3				0.5				0.7			
K	0.005		0.1		0.005		0.1		0.005		0.1	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	38	19	<u>60</u>	41	<u>93</u>	116	128	<u>129</u>	22	63	<u>44</u>	25

Table 1c. Chemical Parameters for ECE Mechanism

ks	10 ⁶				0.0005											
α	0.5				0.3				0.5				0.7			
K	1		100		1		100		1		100		1		100	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	1
	<u>46</u>	27	8	<u>49</u>	30	<u>11</u>	<u>52</u>	33	<u>119</u>	99	79	<u>102</u>	14	<u>55</u>	<u>36</u>	1

ks	0.005											
α	0.3				0.5				0.7			
K	1		100		1		100		1		100	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	<u>130</u>	82	105	<u>133</u>	85	<u>108</u>	<u>124</u>	88	<u>111</u>	135	91	<u>114</u>

ks	0.05											
α	0.3				0.5				0.7			
K	1		100		1		100		1		100	
kf	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10	0.2	10
	58	<u>39</u>	<u>20</u>	61	<u>126</u>	94	117	<u>96</u>	42	<u>23</u>	<u>64</u>	45

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

ks	0. 0005						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	150	<u>2</u>	<u>151</u>	<u>71</u>	152	<u>3</u>	153

ks	0. 002						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	154	<u>155</u>	156	<u>157</u>	158	<u>159</u>	160

ks	0. 005						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	<u>161</u>	72	162	<u>73</u>	<u>163</u>	74	164

ks	0. 02						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	<u>165</u>	166	<u>167</u>	168	<u>169</u>	170	<u>171</u>

ks	0. 05						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	172	<u>4</u>	<u>173</u>	76	174	<u>5</u>	<u>175</u>

Table 2b. Chemical Parameters for Reversible E Mechanism

ks	10^6
α	0. 5
	<u>1</u>

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

ks	0. 0005						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	150	<u>2</u>	<u>151</u>	<u>71</u>	152	<u>3</u>	153

ks	0. 002						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	154	<u>155</u>	156	<u>157</u>	158	<u>159</u>	160

ks	0. 005						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	<u>161</u>	72	162	<u>73</u>	<u>163</u>	74	164

ks	0. 02						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	<u>165</u>	166	<u>167</u>	168	<u>169</u>	170	<u>171</u>

ks	0. 05						
α	0. 2	0. 3	0. 4	0. 5	0. 6	0. 7	0. 8
	172	<u>4</u>	<u>173</u>	76	174	<u>5</u>	<u>175</u>

Table 2b. Chemical Parameters for Reversible E Mechanism

ks	10^6
α	0. 5
	<u>1</u>

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

1. The voltammogram curve is shifted so that the peak potential equals zero ($\pm < 0.01$ 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) and at the switching potential (E_s) 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 “ compounds” defined in Tables 1 & 2 were analyzed by KNN pattern recognition procedure. They were separated into “training” group and “prediction” group. “Compounds” underlined belong to the training set and were studied by KNN first to extract the most useful features. (Compound #1 from Table 2B was the only member of class 1, reversible uncomplicated. Thus, it was represented twice in the compound set, so that KNN would be able to recognize that class.) Other “compounds” are treated as “unknowns” in the prediction set and were classified by KNN with the features extracted from the training procedure.

The information gain of each study can be calculated.

A K-nearest neighbor feature elimination approach was used to find an optimum set of features which would accurately classify a set of voltammetric data. The software requires all the fourier transform data files be converted to text files before running the program. It also limits the number of pattern descriptors to 17, of which the first two are reserved for pattern ID and class. In this study, the first fifty fourier features were

requires all the fourier transform data files be converted to text files before running the program. It also limits the number of pattern descriptors to 17, of which the first two are reserved for pattern ID and class. In this study, the first fifty fourier features were examined for each voltammogram. Prior to the pattern recognition analysis, all features are autoscaled. This involves the following transformation:

$$xs_{ij} = (x_{ij} - \bar{x}_j)/(s. d)_j \quad (7)$$

xs_{ij} = autoscaled j th feature for the i th compound

$\bar{x}_j$ = average value of j th feature

Several different pattern recognition procedures are involved in this study to achieve the goal of this research: systematic feature elimination, weight optimization and class - weighted training.

Systematic feature elimination: the KNN program examines all patterns in the training set using a leave-one-out (LOO) K-nearest neighbor pattern classification algorithm. This procedure assigns a class to each pattern as if it were an “unknown”, after all patterns are classified by the LOO procedure, the overall accuracies are determined and printed out. The operator can select which of all possible features are used for KNN classification . Also, the operator can select either forward or backward feature elimination. A feature is eliminated if its absence does not degrade the classification accuracy criterion.

Weight optimization: can be selected by the operator if it occurs during the feature elimination process. That is, the weight of each feature is varied from about 1 to 100 to evaluate the effect on the classification accuracy criterion.

Class - weighted training procedure: this procedure is similar to the KNN training with feature elimination, except that the operator selects one class for which class-specific

In this study, pattern recognition class accuracy was evaluated one class at a time, selecting a different set of features which maximized the class specific accuracy for each class. Class-weighted training was used at some point in the overall training process. The first step in the feature elimination procedure does not involve class weighting, but involves a first pass through the feature elimination process using overall class accuracy as the figure of merit.

3. 5. 1 Effects of Experimental Parameters

Eight KNN studies were performed corresponding to different experimental parameters (Table 4). The 2-level 3-factor factorial design is shown in Table 4. The “+” and “-” symbols refer to each factor at its high or low level, as defined in Table 3. This factorial design will allow extensive study and interpretation of parameter effects on information gain.

Fifty features were examined for each cyclic voltammogram. Due to the limitation of software, these fifty features were divided into 4 subsets. The first three subsets contain fifteen features, whereas the last subset contains five features. These four text files were examined separately by KNNPRA software. The most useful features which provide the highest classification accuracy from each of the subsets are combined to a new file and examined by KNNPRA program again. Once the most useful features are extracted, the weights of these features were optimized and the information gain was calculated based on the optimized overall accuracy.

Table 3. Experimental Parameters for Simulated Voltammograms

$\underline{v \text{ (v/s)}}$	$\underline{E_{sw}/E_i \text{ (mv)}}$	$\underline{ncv = 1}$	$\underline{ncv = 2}$
$v^- = 0.05$	$(E_{sw}/E_i)^-$	A	A ₂
$v^- = 0.05$	$(E_{sw}/E_i)^-$	B	B ₂
$v^- = 1.0$	$(E_{sw}/E_i)^-$	C	C ₂
$v^- = 1.0$	$(E_{sw}/E_i)^+$	D	D ₂

$(E_{sw}/E_i)^- : E_{sw} = -300\text{mv}, E_i = 980 \text{ mv}$

$(E_{sw}/E_i)^+ : E_{sw} = -50\text{mv}, E_i = 1230\text{mv}$

Symbols A - D, A₂ - D₂ refer to each of the eight different sets of parameters (Table 4)

Table 4. Factorial Design for KNN Study

Study	Experimental parameter		
	V	E _{sw} /E _i	ncy
A	-	-	-
B	-	+	-
C	+	-	-
D	+	+	-
A2	-	-	+
B2	-	+	+
C2	+	-	+
D2	+	+	+

Table 5. KNN Study of Voltammograms Obtained under Condition A

Features	Useful Features	Overall Accuracy	ΔH
1-15	2, 3, 10	67. 9	1. 414
16-30	16-18, 22, 28	52. 6	1. 361
31-45	32,38, 43	65. 4	1. 604
46-50	46, 49	70. 5	1. 323
1-50	2,32,46	71. 8	1. 532

Table 6. KNN Study of Voltammograms Obtained under Condition B

Features	Useful features	Overall accuracy	ΔH
1-15	1,3 10	70. 5	1. 581
16-30	16, 17,25,27	55. 1	1. 351
31-45	33,36,40	52. 6	1. 083
46-50	47,49,50	47. 4	1. 209
1-50	10,16,17,25,40,49	61. 5	1. 323

Table 7. KNN Study of Voltammograms Obtained under Condition C

Features	Useful Features	Overall accuracy	ΔH
1-15	1,4,7,13	50. 0	1. 328
16-30	20,22,29	38. 5	1. 208
31-45	34,36,37,45	43. 6	1. 333
46-50	46	35. 9	1. 138
1-50	1,20,22,45,46	53. 8	1. 168

Table 8. KNN Study of Voltammograms Obtained under Condition D

Features	Useful Features	Overall Accuracy	ΔH
1-15	3,5,9	46. 2	1. 348
16-30	18,22,23,26,30	46. 2	1. 303
31-45	35,43	38. 5	1. 101
46-50	48	30. 8	0. 828
1-50	5,9,23	44. 9	0. 984

Table 9. KNN study of Voltammograms Obtained under Condition A2

Features	Useful Features	Overall Accuracy	ΔH
1-15	1,4,5	80. 8	1. 830
16-30	28,29,30	56. 4	1. 323
31-45	35,40,42,43	52. 6	1. 322
46-50	46,47,49,50	42. 3	1. 237
1-50	1,4,5	80. 8	1. 717

Table 10. KNN Study of Voltammograms Obtained under Condition B2

Features	Useful Features	Overall Accuracy	ΔH
1-15	2,8,9	88. 5	1. 450
16-30	21,22,23,24,25,26,28	57. 7	1. 509
31-45	33,34,37,40	59. 0	1. 272
46-50	47,48,49,50	52. 6	1. 322
1-50	2,21,33,37	84. 6	1. 793

Table 11. KNN study of Voltammograms Obtained under Condition C2

Features	Useful Features	Overall Accuracy	ΔH
1-15	2,4	75. 6	1. 667
16-30	18,21,22,24,27,28	56. 4	1. 247
31-45	32,36,39,40	53. 8	1. 242
46-50	46,47,48,49	43. 6	1. 243
1-50	2,24,36	75. 6	1. 611

Table 12. KNN Study of Voltammograms Obtained under Condition D2

Features	Useful Features	Overall Accuracy	ΔH
1-15	1,2	83. 3	1. 793
16-30	17,19	87. 2	1. 419
31-45	37,38,39,40,41,43,44	52. 6	1. 428
46-50	48,49,50	47. 4	1. 138
1-50	2,40,41	82. 1	1. 753

Table 13. Summary of Highest Information Gain for Each Experimental Condition

(Extracted from Table 5-12)

Study	ΔH
A	1. 604
B	1. 586
C	1. 333
D	1. 348
A2	1. 830
B2	1. 793
C2	1. 667
D2	1. 793

Table 14. Factor Effects on Information Gain

Scan rate	-0. 17
No. of cycle	+0. 33
Switching Potential	+0. 06

•
:

Table 15a. Classification Training Accuracy from Class-Weighted Training and Overall Training (FP: False Positives)

Condition	Class	Class-Weighted Training				Overall Training
		%Accuracy	FP	ΔH	Features Set	%Accuracy
A	2	50.0	5	0.477	2,3,10,16,18	45.0
	3	60.0	10	0.566	2,3,10	50.0
	4	95.0	0	0.877	2,3,22	90.0
	5	75.0	3	0.681	2,3,10,16	75.0
B	2	43.8	8	0.435	1,3,10,16	37.5
	3	45.0	12	0.479	1,3,10,16	40.0
	4	85.0	4	0.725	1,3,10,40	75.0
	5	75.0	5	0.670	1,3,10,17	60.0
C	2	56.3	11	0.481	1,4,7,13,20	43.8
	3	40.0	15	0.471	1,4,7	35.0
	4	85.0	1	0.742	1,4,7,13	70.0
	5	50.0	6	0.538	1,4	40.0
D	2	50.0	8	0.464	3,5,9,22	37.5
	3	50.0	17	0.477	3,5,9	45.0
	4	65.0	4	0.624	3,5,9	60.0
	5	50.0	3	0.554	3,5,18,22	35.0

Table 15b. Classification Accuracy from Class-Weighted Training and Overall Training
(FP: False Positives)

Condition	Class	Class-Weighted Training				Overall Training
		%Accuracy	FP	ΔH	Features Set	%Accuracy
A2	2	68.8	5	0.565	1,4,5,28	75.0
	3	75.0	4	0.675	1,4,5	75.0
	4	90.0	2	0.760	1,4,5	85.0
	5	85.0	4	0.724	1,4,29	75.0
B2	2	75.0	5	0.594	2,8,9,21,22	50.0
	3	90.0	3	0.754	2,8,9	90.0
	4	100.0	1	0.816	2,9,22,23	90.0
	5	80.0	2	0.710	2,8,9	70.0
C2	2	62.5	8	0.522	2,4,18,21	62.5
	3	60.0	7	0.582	2,18,24,32	45.0
	4	90.0	2	0.836	2,4	90.0
	5	85.0	1	0.742	2,4,18	80.0
D2	2	68.8	5	0.565	1,2,17,19,38	56.3
	3	75.5	3	0.681	1,2,19	75.0
	4	85.0	0	0.747	1,2,19	85.0
	5	90.0	4	0.725	1,2,17,19	85.0

Table 15c. Summary of Highest Training Information Gain for Each Class from Class-Weighted Training

Class	No. of Compounds	Condition	ΔH	$\Delta H_{\max}$
2	16	B2	0. 594	0. 732
3	20	B2	0. 754	0. 821
4	20	B2	0. 816	0. 821
5	20	C2	0. 742	0. 821

Chapter 4

Results and Discussions

This study involves four parts. First, cyclic voltammograms for the hypothetical “compound” were generated by digital simulation. Characteristic data for each “compound” were obtained (simulated) under eight conditions corresponding to different experimental parameters (Table 4), which were designated as A, B, C, D, A2, B2, C2 and D2.

Second, each cyclic voltammogram was converted to a frequency domain by fast fourier transform. The first 50 fourier coefficients were used in the KNN studies.

Third, information theory was applied to the results of the KNN training study to provide a more quantitative measure of information gain related to experimental parameters and training procedure for qualitative identification .

Fourth, results of the KNN training study were applied to a prediction set .

Evaluation of results is based primarily on some classification accuracy criterion. However, the information gain quantity, ΔH , provides a more reliable evaluation, because it incorporates the influence of false positive identifications.

4. 1 Effects of Experimental Parameters

Tables 5 -12 give the results of each KNN study. It was observed that individual subset features sometimes gave more information than the set of features obtained from combined subsets (Table 6,8,10,12). The accuracy of class 1 compounds was not reported because it has an artificially higher classification accuracy (100%). The reason is there is only one shape for the uncomplicated reversible case. However, this case must be included in the training set because it is necessary to find features which will distinguish

other classes from the reversible case. To insure that there is only a minimal positive impact on ΔH calculations, only two examples of class 1 voltammograms are included in the training set. (Two is the minimum so that class 1 can be recognized in training.)

Table 13 summarizes the best results (highest information gain) for each of the different combinations of experimental parameters defined in table 4. The “ factor effect” for each experimental parameter can be calculated by finding the difference between results (information gain) obtained at high and low conditions, and divided by 4, half number of experiments. Thus, from table 13, the factor effect = $\frac{1}{4}(C + D + C2 + D2 - A - B - A2 - B2)$. Factor effect calculations for all three experimental parameters are summarized in Table 14.

4. 1. 1 Scan Rate

Scan rate is the most commonly considered variable to characterize electrode processes. In this study, it was examined at two levels, low (0. 05 v/sec) and high (1. 0 v/sec). ΔH was calculated for classification obtained with the optimized feature weights.

The factor effect of scan rate is -0. 17. This suggests that lower scan rates provide greater information content. The F-test regarding this factor effect gives a value of 1. 23 which is less than $F_c(0. 05, 1, 6) = 5. 99$. This F-test indicates that, for the range of scan rate tested here, the scan rate is not a significant experimental parameter. However, these results suggest that using a greater range of scan rate (about 10 to 100 fold increase) would lead to a statistically significant factor effect.

4. 1. 2 Number of Cycles

In this study , one-cycle and two-cycle voltammograms were studied. It was observed that a two-cycle voltammogram gives more information gain. The factor effect of number of cycles is +0. 33. The calculated value of F is 10. 6, which means that the number of cycles has a statistically significant effect on information gain.

4. 1. 3 Switching Potential

The effect of switching potential on information gain was not studied previously. To determine its effect on information gain, two switching potential levels were compared in this study. The factor effect, summarized in Table 14, is 0. 06. Based on the F-test, it can be concluded that for the range of this parameter studied here, switching potential does not have a statistically significant effect on information content. However, it was observed that when the number of cycles is 1, the lower switching potential leads to more information gain. With two-cycle voltammograms, the higher switching potential gives more information gain. (see Tables 9-12).

4. 2 Effects of Training Procedure

A class-weighted training procedure was performed in this study as well . This classification procedure finds optimum features for identifying one class at a time. These binary decisions can produce a maximum of 0. 732 or 0. 821 bit for two classes, depending on the number of compounds in different classes. Tables 15a and 15b summarize class-specific accuracy (section 3. 5)using the class-weighted training procedure (see P. 39 for details of procedure).

It is observed that the highest information gain was 0. 816 which was achieved under condition B2 for class 4 (Table 15c).

4. 2. 1 Comparison Between Class-Weighted Training and KNN Overall Training

From Tables 15a and 15b, it is observed that the accuracy for identifying a single class is improved with class-weighted training procedure under most experimental conditions. However, the performance is generally better with two-cycle voltammograms. The best class-specific training accuracy is achieved with class-weighted training are two - cycle voltammograms (75, 90, 100, 90% for classes 2-5). Class 2 could also be classified at 75% under condition A2 without class-weighted training.

The maximum information gain can be extracted from overall training procedure is 2. 115. The best result is 1. 830 under condition A2 which is about 87% of the maximum value(Table 10). However, in class-weighted training procedure, nearly 100% of the maximum information gain was achieved under condition B2 for class 4 (Table 15c).

4. 3 KNN Study Applied to Prediction

The relatively high accuracy of recognition of compound class within the training set described above confirm that class information is contained in the shape of cyclic voltammetric data. However, the most demanding validation of the ability of selected classifier features to identify compound class is by application of these features to a prediction set of compounds, whose voltammetric properties were not included in the training set.

The prediction set in our study included 78 compounds, which had the same class distribution as for the training set. (The specific compounds selected are the non-underlined items indicated in Tables 1 and 2.) The results of prediction using features selected for overall training accuracy are presented in Table 16. Clearly, there is considerable variability in the prediction ability of the various feature sets, and prediction accuracy is generally lower than training accuracy. This is not surprising, as the training feature elimination procedure may lead to feature sets selected for fortuitous reasons. Thus, we examined several different feature sets selected from training, even those with lower training accuracy, as they may be more reliable in prediction. As seen in Table 16, prediction accuracy as high as 61. 5% was achieved for a training feature set obtained under condition D. This observation suggests that the most overall class discriminating information is obtained under condition D.

An alternative prediction study involved the application of class-specific feature sets (given in Tables 15a and 15b). The results are given in Tables 17a and 17b. In this case the prediction accuracy are generally lower than training accuracy, as expected. However, class 5 could be identified with 85. 0% accuracy; and the average class-specific accuracy for the best prediction set features is 79. 1 under condition C2(Table 17b).

Thus, it appears that the best approach to compound identification by pattern recognition with voltammetric fourier transform features would involve using class-specific rather than multi-class procedures.

The highest information gain achieved for each class is summarized in Table 18. It is observed that the most useful condition is C2 which gives the highest information gain on all classes.

Table 16. Prediction Results, Overall Classification

	%Overall Accuracy	Features used
A	30. 8	2,3,10
	33. 3	2,32,46
B	28. 2	1,3,10
	25. 6	10,16,17,25,40,49
C	60. 3	1,4,7,13
	60. 3	1,20,22,45,46
D	61. 5	3,5,9
	53. 8	5,9,23
	64. 1	2,8,9,40,49
A2	27. 2	35,40,42,43
	30. 8	1,4,5
B2	28. 2	2,8,9
	29. 5	2,21,33,37
C2	37. 2	2,4
	39. 7	18,21,22,24,27,28
D2	23. 1	2,40,41
	28. 0	1,2

Table 17a. Prediction Results, Class - Specific Classification(Prediction Features
Selected from Tables 15a. and 15b.)

Condition	Class	%Accuracy	ΔH
A	2	0	0
	3	25.0	0.362
	4	20.0	0.331
	5	60.0	0.577
B	2	0	0
	3	55.0	0.546
	4	35.0	0.423
	5	10.0	0.270
C	2	30.0	0.330
	3	45.0	0.485
	4	60.0	0.577
	5	50.0	0.516
D	2	41.2	0.380
	3	65.0	0.607
	4	45.0	0.485
	5	85.0	0.731

Table 17b. Prediction Results, Class - Specific Classification (Prediction Features
Selected from Tables 15a. and 15b.)

Condition	Class	%Accuracy	ΔH
A2	2	0	0
	3	25. 0	0. 362
	4	50. 0	0. 516
	5	0	0
B2	2	76. 5	0. 598
	3	65. 0	0. 607
	4	75. 0	0. 67
	5	80. 0	0. 70
C2	2	76. 5	0. 60
	3	75. 0	0. 67
	4	80. 0	0. 70
	5	85. 0	0. 73
D2	2	23. 5	0. 30
	3	65. 0	0. 61
	4	60. 0	0. 58
	5	80. 0	0. 70

Table 18. Summary of Highest Information Gain for Each Class

Class	No. of Compounds	condition	ΔH	$\Delta H_{\max}$
2	17	B2, C2	0. 60	0. 73
3	20	C2	0. 67	0. 82
4	20	C2	0. 70	0. 82
5	20	C2,D	0. 73	0. 82

4. 4 Discussions

4. 4. 1 Training Results

The effects of scan rate and number of cycles were examined for training performance only by Ming Zhou. She observed that there was “no significant enhancement of information gain for two cycles compared to one.” However, this observation was from a relatively small pattern distribution and it was not a statistically valid result.

In this study, it was observed for training performance that there is a significant difference in information content between one cycle and two cycles. The statistical study confirms that the number of cycles is an important experimental parameter. For most “compounds” without coupled chemical reactions, the voltammograms of the first and the second cycles are not very different. However, for “compounds” involving coupled chemical reactions, the second cycle can be quite different than the first cycle, providing more class information.

At lower scan rate, it was observed that the class-related information content is higher. This conclusion is consistent with the previous result.

This observation of scan rate effect may be a reflection of the fact that lower scan rates allow greater impact of coupled chemical reactions on the overall voltammetric behavior, for the conditions selected in this study. Moreover, in the limit of the very highest possible scan rates, the properties of all electrode processes would approach those of the electron transfer limited step.

The effect of switching potential was also examined in this study. Unfortunately, for the range of values studied here, there is no significant effect on information content.

4. 4. 2 Prediction Results

Based on the prediction results, it appears that the most information for multi-class identification is obtained under conditions C and D. (Table 16).

For class-specific prediction, the condition providing the greatest information content is different for each class (Table 17). However, condition C2 is most useful.

Chapter 5 Conclusions

This study has provided further investigations on experimental parameters which affect the ability to obtain quantitative identification information from Fourier transform voltammetric data.

Based on training results, our study confirms that lower scan rate is a significant parameter for classifying compounds based on cyclic voltammogram shape. It also confirms that increasing the number of cycles does increase the information gain under certain conditions. In addition, it suggests that switching potential is not an important parameter, at least under the conditions studied here.

High classification accuracy for training or prediction set compounds was not achieved for all classes. However, the primary objective of this study was to learn how voltammetric parameters affected class information content, and that has been achieved.

Chapter 6 Further Study

Further studies of the information content of voltammetric shape factors should be on a more continuous pattern distribution . More reaction mechanisms should be examined. More parameters, including numbers of electron transferred can be considered. In addition, substantial changes of the range of experimental parameters should be considered.

Reference

1. S. P. Perone and C. L. Ham, Journal Research of the National Bureau of Standards, vol. 90, 531, 1985
2. C. E. Shannon, Bell System Technical Journal, vol. 27, 379, 1948
3. P. A. Boudreau and S. P. Perone, Anal. Chem., vol. 51, 811, 1979
4. Q. V. Thomas and S. P. Perone, Anal. Chem., vol. 49, 1369, 1977
5. Q. V. Thomas and S. P. Perone, Anal. Chem., vol. 49, 1376, 1977
6. D. R. Burgard and S. P. Perone, Anal. Chem., vol. 50, 1366, 1978
7. W. A. Byers and S. P. Perone, Anal. Chem., vol. 52, 2173, 1980
8. S. D. Schachterle and S. P. Perone, Anal. Chem., vol. 53, 1672, 1980
9. W. A. Byers and S. P. Perone, Anal. Chem., vol. 55, 615, 1983
10. W. A. Byers and S. P. Perone, Anal. Chem., vol. 55, 620, 1983
11. Ming Zhou, M. S. Thesis, 1995, San Jose State University
12. Douglas A. Skoog, James J. Leary, Principle of Instrumental Analysis
13. A. J. Bard and L. R. Fualkner, Electrochemical Method, Fundamentals and Applications, John Wiley & Sons: New York, 213 - 248, 1980
14. P. T. Kissinger and W. R. Heinemar, Laboratory Techniques in Electroanalytical Chemistry, Marcell Dekker, New York, 86 - 93, 1984
15. J. W. Dillard, J. J. Odea and R. A. Osteryoung, Anal. Chem., vol. 51, 115, 1979
16. J. W. Dillard, J. J. Odea and R. A. Osteryoung, Anal. Chem., vol. 51, 120, 1979
17. R. S. Nicholson and Shain, Anal. Chem., vol. 36, 706, 1964
18. D. H. Evans, Chem. Rev., vol. 90, 706, 1990
19. C. P. Andrieux, P. Hapiot and J. M. Savoant, Chem. Rev., vol. 90, 723, 1990

20. R. M. Wightman and D. O. Wipf, Electroanalytical Chemistry, vol. 15, 267 -353
21. Speiser, B. Comput. Chem. 1990, 14, 12
22. Gosser Jr. , D. K. Cyclic Voltammetry: Simulation and Analysis of Reaction Mechanisms, VCH, Inc. , New York, 1993
23. Bieniasz, L. Computers. Chem, in press.
24. ELSIM, Technical Software Distributors, 1016 Hartmont road, Baltimore, MD 21228
25. Rudolph, M. Electroanalytical Chem. 1991, 314, 13
26. Rudolph, M. Electroanalytical Chem. 1992, 338, 85
27. Bard, A. J. , Faulkner, L. R. , Electrochemical Method: Fundamentals and Applications, John Wiley & Sons: New York, 1980, 213 -248
28. H. C. Andrew, Introduction to Mathematical Techniques in Pattern Recognition, Wiley-Interscience, New York, 1972
29. R. D. Depalma and S. P. Perone, Anal. Chem. , vol. 51, 825, 1979
30. R. D. Depalma and S. P. Perone, Anal. Chem. , vol. 51, 829, 1979