Winter 1990

Development and application of artificial intelligence strategies to solve infrared spectroscopic problems

Barry J. Wythoff
University of New Hampshire, Durham

Order Number 9119148

Development and application of artificial intelligence strategies to solve infrared spectroscopic problems

Wythoff, Barry J., Ph.D.

University of New Hampshire, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

DEVELOPMENT AND APPLICATION
OF ARTIFICIAL INTELLIGENCE STRATEGIES
TO SOLVE INFRARED SPECTROSCOPIC PROBLEMS

BY

BARRY J. WYTHOFF
B.A., Rutgers University, 1982
M.S., Rutgers University, 1987

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

in

Chemistry

December, 1990

This dissertation has been examined and approved.

Sterling P. Tomellini
Dissertation Director, Sterling Tomellini
Assistant Professor of Chemistry

Howard Mayne
Howard Mayne
Associate Professor of Chemistry

Gary R. Weisman
Gary Weisman
Associate Professor of Chemistry

Rudolf Seitz
Rudolf Seitz
Professor of Chemistry

Filson H. Glanz
Filson Glanz
Professor of Electrical Engineering

November 1, 1990
Date

DEDICATION

This work is dedicated to my parents, Willem and Antoinette Wythoff. Their warmth, gentle support, and tireless patience have given me the strength to persevere through what have often seemed like insurmountable difficulties in my life.

ACKNOWLEDGEMENTS

I would like to thank Dr. Sterling Tomellini for his support and consideration which he generously gave to me throughout my studies. Dr. Tomellini's very high standard for scientific work has had a great deal of influence on my career. In addition I would like to give thanks to the many friends whom I've had the pleasure of knowing during my stay at the University of New Hampshire. These people provided the support and the laughter which made the difficult times lighter and easy to forget now that it's over, and made the good times brilliant and memorable. To name them is of course to risk leaving some people out, but I will fondly remember time spent with Paul Franz, Michael Gousie, Semra Kilic, Osman Okyay, Svein Sagatun, Jose` Torres, Yu Chen, Charlotte Wilkinson, Jamey Clothier, John Beauregard, Dan Cervi, and Linda Olson.

I owe special thanks to Linda Olson for her warmth, patience and her sense of humor, without which my last several months at the university would have been very difficult to endure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

DEVELOPMENT AND APPLICATION OF ARTIFICIAL INTELLIGENCE STRATEGIES TO SOLVE INFRARED SPECTROSCOPIC PROBLEMS

by

Barry Wythoff
University of New Hampshire, December, 1990

The ever-increasing power of modern infrared instrumentation, coupled with the decreasing number of experienced spectroscopists has created an imbalance between information generation and interpretation capabilities. At the same time, digital computers are being developed which continue to grow in storage and processing capabilities, and shrink in cost. Clearly, the computer may serve as a valuable tool to aid the analytical chemist in interpreting spectroscopic information. This dissertation deals with the development of new approaches to exploiting computer technology to interpret infrared spectroscopic data.

A large existing expert system for functional group analysis, PAIRS, has been modified to transfer the maximum amount of information to the chemist. Two closely coupled knowledge based systems, IRBASE and MIXIR, have been created to identify major components of condensed phase mixtures. A second version of MIXIR has been developed to identify major components of vapor phase mixtures. Finally, a neural network approach to peak detection in analytical data has been developed.

INTRODUCTION

While research in artificial intelligence has been going on for decades, developments in the 1980's brought these systems from the academic laboratory to the industrial workplace. Artificial intelligence includes such diverse topics as computer vision systems, natural language processing, theorem proving, and modelling mammalian learning processes. The majority of recent interest, however, has been in the area of so-called "expert systems". These systems are computer programs which attempt to emulate the logical problem solving approach of a human expert in a limited problem domain. Such systems are potentially useful in any area which requires a good deal of expertise, that is, any area that is well understood, and requires a significant amount of knowledge and/or complex logic to solve problems. The creation of an expert system may allow a company to capture much of the expertise of a valuable employee approaching retirement. The resulting system may then be used to help train future "experts", or to provide assistance to personnel lacking extensive training to solve a difficult problem which they encounter at a later date.

Analytical chemistry is a highly developed science. The traditional goals of an analytical chemist are the identification and/or quantitation of chemical species. These substances may be of known or unknown origin, they may be pure substances, or complex mixtures. The task facing the analyst often requires a great deal of knowledge in the area, and the application of deductive reasoning. Chemical analysis is therefore a natural area for the application of expert system technology. Indeed, one of the first successful expert systems to be developed was designed to identify organic compounds from mass spectral data (1). Before looking at the application of knowledge based systems to analytical chemistry, a brief examination of the nature of this technology is in order.

Anatomy of An Expert System:

A number of excellent texts exist which detail the theory and practice (2-4) of expert system technology, therefore, only a summary of basic concepts is presented below. Knowledge Based Systems (KBS), as their name implies, differ from conventional computer software in that their primary function is to reach a conclusion by applying deductive logic to problem data. Unlike conventional algorithmic software, these systems often do not specify exactly how a problem is to be solved. Rather, only a basis for selecting and applying "appropriate" rules to the problem is specified, and the state of the problem then determines the action to be taken.

There are three major components in a typical knowledge based system. The **interface** is the portion of the program which handles all interaction with the user. While not a factor in program task performance, the quality of the user interface largely determines how useful the system will be in practice. The **knowledge base** is the heart of the expert system. It is composed of the rules and facts which constitute the domain specific knowledge for a given application. An example of a fact is "The upper limit on pump pressure is 5000 psi". A rule defines some condition (the "antecedent"), and the corresponding action to be taken or fact to be deduced (the "consequent") if the condition is satisfied. An example of a simple rule is "IF the sample is a nonvolatile liquid, THEN high pressure liquid chromatography is the separation method of choice".

The **inference engine** contains the overall control strategy. It is responsible for selecting and retrieving appropriate rules and facts, carrying out the consequent actions, and adding any deduced facts to a "dynamic knowledge base" in working memory. Two common control strategies are forward chaining and backward chaining. Forward chaining involves selecting rules which match known facts, and working forward, successively "deducing" new facts and selecting new rules, until no new facts may be deduced. This process involves matching existing facts to rule antecedents. Forward chaining allows the system to "deduce" everything which can be learned, given the available information and rules. Backward chaining involves selecting some desirable fact to be proven, and attempting to work backward through the rule chain to the initial known data. This process involves matching facts to be proven to rule consequents. Backward chaining is often the most efficient means of arriving at a specific conclusion.

As indicated earlier, in theory, the order of the rules and facts in the knowledge base is unimportant, since the inference engine should simply select the necessary rules as dictated by the state of the problem. In practice, the knowledge base is ordinarily highly structured. This provides more efficient program execution, and greatly simplifies program maintenance.

Applications:

There are many potential areas of application for expert system technology in analytical chemistry, and numerous reviews of these applications have been published (5-12). The majority of current systems, however, perform one of two general tasks. The first of these is the analysis of complex data resulting from nuclear magnetic resonance (NMR) (13-14), infrared (IR) (15-37), or mass spectrometric (MS) experiments (1,38-41), or some combination (42-45). The goal of such systems is ordinarily to assign the functional groups (structural fragments) present in a molecule, to assign the total structure of a molecule, or to identify the components of a mixture. Information from one or more of the above spectral techniques may be used, along with supplemental physical/chemical information on sample state, molecular weight, empirical formula, etc. The second major category of application is analytical methods development. Here, a KBS will assist in determining which specific procedures and techniques are best for a particular analysis, and/or what experimental conditions are desirable (46-50).

Another area in artificial intelligence research which has enjoyed rapid growth in the 1980's is the creation of artificial neural networks (51-59). While knowledge based systems are founded on a high-level, conscious model of human cognition, artificial neural networks are based on a low-level, physical model of cognition. They are a very simple hardware or software simulation based on a physical model of the brain. A number of fascinating applications of neural network technology to solve analytical chemistry problems have been reported (60-65).

This dissertation concerns the development and application of new computer software technology to solve infrared spectroscopic problems:

i) The identification of functional groups present in a condensed or vapor phase analyte.

ii) The automated creation of a compound-specfic knowledge base for a system to identify the likely components of condensed phase mixtures.

iii) The identification of the likely component identities in condensed phase mixtures.

iv) The detection of peak-shaped signals in a digitized infrared spectrum.

v) The identification of the likely component identities in condensed phase mixtures.

CHAPTER 1

DESCRIPTIVE, INTERACTIVE COMPUTER-ASSISTED INTERPRETATION OF INFRARED SPECTRA.

Introduction:

PAIRS (17-18,21-23,35), the Program for the Analysis of Infrared Spectra, is a rule-based expert system which interprets an IR spectrum by mimicking the thought process used by a spectroscopist. The only output of the original version of the system was a numerical indication, or "expectation value", of the likelihood of presence or absence for a particular functionality or subfunctionality, based on an interpretation of the spectral data entered. While this system provided useful information in a time efficient manner, it quickly became apparent that a single number cannot adequately convey all that can be learned from the spectral interpretation process. A subsequent modification of PAIRS attempted to address this problem by allowing the user to trace the decision making process (23). This improved version provided the user with a way to see the rules which were used by the interpreter to arrive at the expectation values, if so desired. While the resulting version of PAIRS was a major improvement over earlier versions, it still did not transfer the knowledge behind the rules.

The first goal of our present work was to develop a system, based on PAIRS, which would allow the knowledge in the rule base to be easily transferred to the scientist using the system. The result can be truly categorized as an expert system. Users of the system are able to ask why a decision is made and obtain a descriptive explanation of the decision making process. This advanced system raises the interpreter to the level of a "smart assistant" for researchers involved in infrared spectral interpretations.

Saperstein (26) recognized that the user could be a valuable resource if allowed to participate in the interpretation process. He developed a program, based on PAIRS, which allowed the user to evaluate and optimize the interpretation results by comparing the original spectrum with a synthetic spectrum created from the characteristic absorptions of the functionalities determined likely to be present by a preliminary PAIRS interpretation. The role of the user was thus limited to determining the best visual match between the synthetic spectrum and the actual spectrum.

The present modifications to PAIRS allow the user to actively participate in the spectral interpretation process, since the interpreter is able to explain the reasons for decisions as they are being made. This allows the chemist to overcome some of the limitations inherent in a static, rule-based system. For example, absorptions due to carbonyl stretches are generally expected to be strong. The interpretation rules, therefore, will return low expectation values for carbonyl containing functionalities for a spectrum containing only weak bands in the carbonyl stretching region. If the user suspects that the sample is a high molecular weight compound or a mixture, then a more reasonable interpretation might result if the user overrules decisions being made on the basis of intensity, for bands in the carbonyl stretching region. The ability to modify decisions intelligently during the interpretation clearly adds a new dimension to the interpretation process.

An important coproduct of the advancements described in this paper is the potential that PAIRS may be used as an instructional aid. MacDonald (27) suggested that an earlier version of PAIRS might be used as a "learning tool" for instruction of IR interpretation. Textbook discussion and demonstration of actual problem solving is often very limited in treatments of spectral interpretation. Merely looking at correlation tables cannot provide the skills necessary to solve actual interpretation problems. In real situations, the researcher must contend with multiple interpretations of experimental data, many of which may seem equally valid based on the information at hand. It is desirable, therefore, for a student to be able to see the approach to problem solving, along with the explicit thought process employed, in order to learn how to approach more complex problems. PAIRS, by providing the user with a detailed explanation of the knowledge and logic employed during the interpretation process, can now provide the basis for a Computer-Assisted-Instruction (CAI) approach to teaching spectral interpretation.

Equipment and Materials:

The minicomputer version of PAIRS was transferred from a Nicolet 640 (Nicolet Analytical Instruments, Madison, WI) to an AT&T 6300 microcomputer (AT&T, Bedminster, NJ) via a DEC 8650 (Digital Equipment Corporation, Maynard, MA). An AT&T 6300 (IBM XT compatible) consisting of: an 8 MHz Intel 8086 CPU, 640 kilobytes of RAM, and a 20 megabyte Winchester hard disk drive, was used for all program development.

Spectra were acquired and processed using a Nicolet 640 computer and a Nicolet MX-1 FTIR spectrometer bench at a nominal 1 cm^{-1} resolution.

Program Description:

The original PAIRS program, and subsequent modifications, have been described adequately elsewhere (17-18,21-23), and will be covered only briefly here. PAIRS consists of two FORTRAN programs, an interpreter and a rule compiler. The interpretation rules are written in an English-like language, CONCISE, and transformed by the compiler prior to use by the interpreter. Interpretations of the likelihood of 195 functionalities and subfunctionalities are performed, where, for example, ketone is considered a functionality, and 5-membered-ring ketone a subfunctionality.

A decision was made to develop this advanced version of PAIRS for IBM compatible microcomputers. This decision was based on the rapidly growing role of microcomputers in the industrial and academic environments. The starting point for this work was the minicomputer version of PAIRS previously developed for the Nicolet 1180 (18). The minicomputer version of both the interpreter and rule compiler were modified to run on the AT&T 6300. Modifications were then made to upgrade the interpreter to the level of the latest VAX version (23), which allows the user to trace the decision making process. This new version of the program can be run on IBM PC/XT/AT compatible microcomputers with a minimum of 512 kilobytes of RAM.

Extensive modification of the interpreter, rule compiler and CONCISE interpretation rules was required to provide an explanation of the rationale for the major decisions made during the interpretation process. The PAIRS rule base includes over 2000 IF-THEN-ELSE rules, which comprise over 16000 lines of text. Interpretation rules have been written for 195 functionalities and subfunctionalities. The first step in the process was to diagram the Boolean decision trees corresponding to the entire PAIRS rule base. The queries corresponding to band positions/shapes were then researched using a number of references to determine the appropriate vibrational assignments (66-70). This information was then used to help determine the reasoning used at the time of rule development to define both the overall knowledge base structure, and the reasoning behind individual queries and actions. The result is a very detailed explanation of query/action logic and content. In addition to providing the band assignments, the majority of the added comments provide explanations of the interpretation strategy. Comments concerning phenomena which influence infrared band position, shape and multiplicity may also be provided. Also, since the rule structure was required to be reviewed in such detail during the incorporation of explanatory comments, knowledge was gained which allowed the actual rules to be improved by modification. The overall process resulted in the inclusion of over 2200 explanatory lines in the CONCISE interpretation rules.

Program changes were required for both the rule compiler and interpreter to utilize the additional comment lines. The rule compiler was modified to recognize comments, encode their presence in the appropriate place in the compiled rule output and save all comment lines in a file for use by the interpreter. The interpreter was modified to allow the user to perform an interpretation with or without tracing the decision making process and to trace the interpretation with or without viewing the appropriate explanatory comments.

The interpreter was further modified to allow user interaction during the interpretation process. These modifications work with the explanation facilities. Together, they provide a trace of each query as it is processed by the interpreter, the resulting answer, and any comments explaining query/action rationale. The user is then allowed to agree or disagree with the decision, thereby playing an interactive role in the interpretation process. The user is given the option to participate interactively in the interpretation process for an individual functionality or the entire rule base.

Results and Discussion:

Examples demonstrating the use of the newly developed system in both the interactive and non-interactive modes for two different compounds are presented below.

The first step of any interpretation is entry by the user of spectral peak information. The spectral data can be entered via the keyboard or read into the program from a previously generated data file. The interpreter also requires information concerning the sample matrix. Providing the interpreter with empirical formula information is optional, and can significantly enhance program performance, since it provides an additional, independent data filter.

Once the experimental data are entered, the user chooses an interpretation option from an options menu. There are three levels of information which the interpreter is capable of providing. In general, as the amount of information transferred to the user increases, the time required for an interpretation increases. As with previous versions of PAIRS, the first option is to have the interpretation results presented as a table of the expectation values for each functionality, sorted according to decreasing likelihood. The second option allows the user to trace the decision making process of the interpreter. This allows the user to see which queries are made, the answers corresponding to the data as entered, and the actions taken, during the interpretation process. The user is given the choice of tracing the interpretation of any single functionality or all functionalities in the rule base, if so desired. The third option provides the trace output of the second option, and includes comments explaining the rationale for the decisions being made. This additional information can help the user to understand and evaluate the interpretation results better. All interactive interpretations are performed in this mode. The reason for this is simply that the user must be informed of the reason for the decisions being made to participate intelligently in the interpretation process.

The interpretation of a spectrum of maleic acid (Figure 1-1) will be used to demonstrate the results of a non-interactive interpretation. The spectrum was acquired from a pressed KBr pellet sample matrix. The peak data obtained from the spectrum and entered into the interpreter are presented in Table 1-1. The intensities are integral values, normalized to 10, and there are three width codes ("1", "2", and "3"), corresponding to sharp, medium and broad peaks, respectively. No empirical formula information was entered into the program. As a first step, the user will usually allow the system to perform an interpretation and return the results as a table of expectation values (option 1). This information provides a starting point for more focused inquiry. Such interpretation

Figure 1-1. Baseline Corrected Spectrum of Maleic Acid.

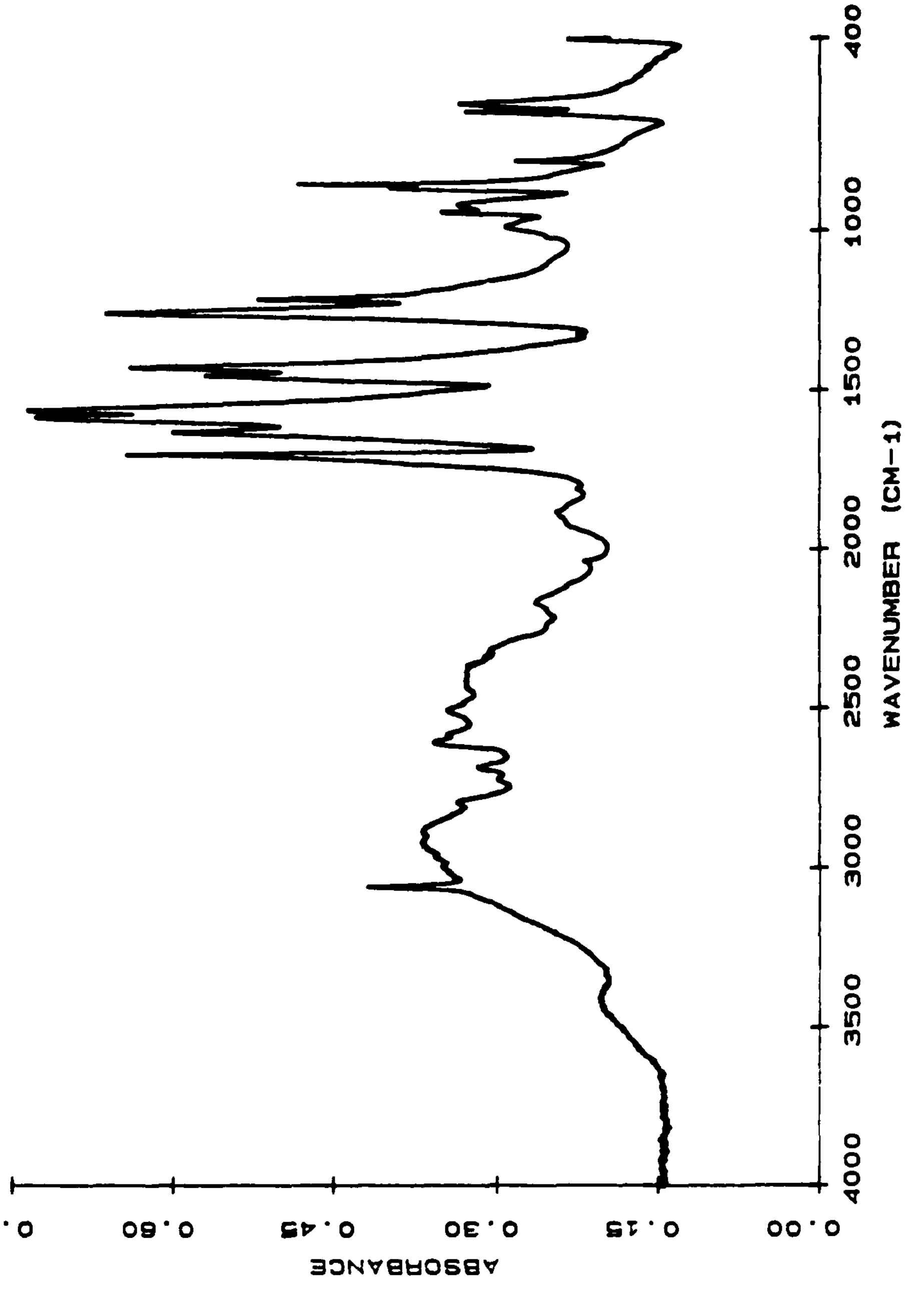


Table 1-1. Peak Data for Maleic Acid

	POSITION CM^{-1}	RELATIVE INTENSITY	WIDTH
1)	609	5	2
2)	634	5	2
3)	786	4	2
4)	863	9	1
5)	874	8	2
6)	923	6	2
7)	949	6	1
8)	990	5	2
9)	1220	8	2
10)	1263	10	2
11)	1434	9	2
12)	1459	9	2
13)	1569	10	2
14)	1590	10	2
15)	1636	9	2
16)	1706	9	2
17)	2480	4	3
18)	2611	4	2
19)	2915	5	2
20)	2920	4	3
21)	2985	4	2
22)	3060	5	2

results for maleic acid are given in Table 1-2. Expectation values may range from 0.01 to 0.99. The higher the expectation value, the stronger the likelihood of a given functionality or subfunctionality being present. Questions that the user might ask at this point include how the program determined that an acid was likely to be present, and why the subclasses for the acid functionality (i.e., acid-unsaturated) are reported at lower expectation values than the parent functionality (i.e., acid). The interpretation can be repeated with full trace and explanatory comments (option 3) to understand better the reasons for the expectation values reported. The results of such an interpretation for the "acid" functionality are presented in Appendix A.

The trace and explanatory comments inform the user that the expectation value for the parent class, "ACID", is determined first by considering the more general questions which apply to all acids. Speciation is then accomplished by focusing on that band which can be used to discriminate between the subclasses, the carbonyl stretching absorption. This band position is found to match that expected for both saturated and unsaturated subclasses of the acid class (of the four subclasses discriminated by the condensed phase rules). The result is that these two subclasses are set to the class value, and then reduced by 20 percent, to indicate the ambiguity of the subclass assignment. Both the explicit path to these results, as well as the local program flow (here, a form of hierarchical classification) are evident when the

Table 1-2. Interpretation Results for the Spectral Data Contained in Table 1-1.

	Group Name	Expectation value
1)	ACID	0.90
2)	AMIDE	0.75
3)	MERCAPTAN	0.75
4)	ACID-SATURATED	0.72
5)	ACID-UNSATURATED	0.72
6)	ETHER	0.70
7)	ETHER-EPOXIDE	0.70
8)	NITRAMINE	0.70

interpretation process is presented in this manner.

The interpretation of a spectrum of 1,1-diethoxyethane will be used to demonstrate the use of the interactive interpretation process. The spectrum of 1,1-diethoxyethane, Figure 1-2, was obtained for a neat sample taken as a liquid between potassium bromide plates (KBr). The corresponding peak data, given in Table 1-3, were entered into the interpreter. The interpreter was also provided with empirical formula information indicating the compound contained only carbon, oxygen and hydrogen.

As in the previous example, a non-interactive interpretation was performed to arrive at the results presented as a simple list of functionalities with assigned expectation values. The results for the interpretation are presented in Table 1-4. Once presented with these results, the user can begin to ask such questions as, "What data are the system using to arrive at 0.58 as the expectation value for "acetal"? and "Why is the expectation value for "ketal" lower than that for "acetal"?. A trace of the interpretation for the functionality "acetal", including explanatory comments, is presented in Appendix B. The information present in such a trace can often help to answer both questions. The trace indicates clearly that a band which appears between 1101 and 1110 cm^{-1} was used to differentiate between the "acetal" and "ketal" functionalities. Upon further inspection of the interpretation trace, the user notices that the first question asked is whether or not there are four or more peaks between

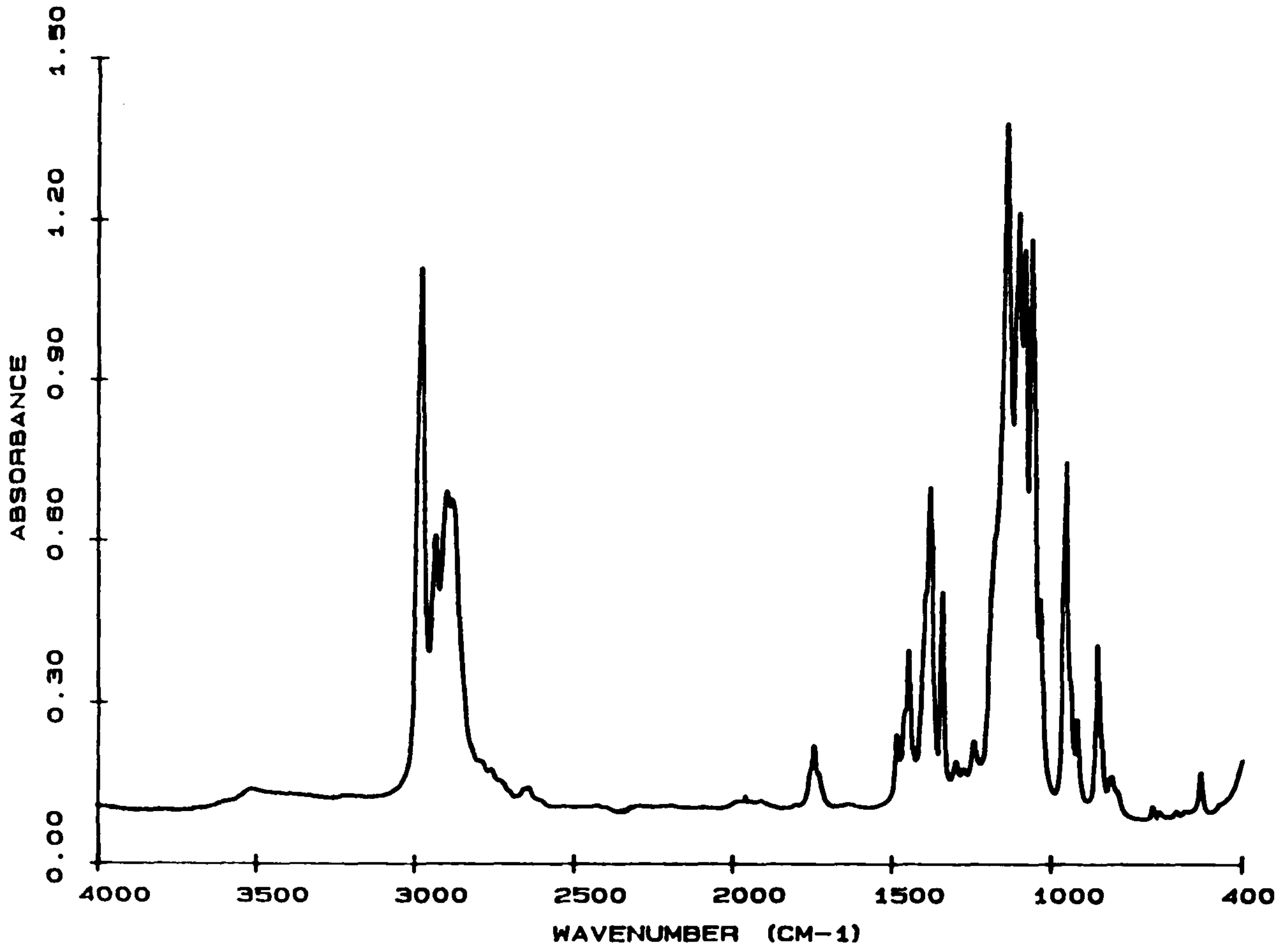


Figure 1-2. Infrared Spectrum of 1,1-Diethoxyethane.

Table 1-3. Peak Data for 1,1-Diethoxyethane.

	POSITION CM^{-1}	RELATIVE INTENSITY	WIDTH
1)	852	2	2
2)	952	5	2
3)	1030	3	2
4)	1061	8	2
5)	1082	8	2
6)	1101	8	2
7)	1138	10	2
8)	1339	3	2
9)	1380	4	2
10)	1444	2	2
11)	2881	4	2
12)	2898	4	2
13)	2933	4	2
14)	2978	8	2

Table 1-4. Results for Interpretation of Spectral Data for 1,1-Diethoxyethane.

Non-Interactive Interpretation:

	GROUP NAME	EXPECTATION VALUE
1.)	ETHER	0.75
2.)	ETHER-SATURATED	0.75
3.)	ACETAL	0.58
4.)	METHYL	0.50
5.)	KETAL	0.26

Interactive Interpretation:

	GROUP NAME	EXPECTATION VALUE
1.)	ETHER	0.75
2.)	ETHER-SATURATED	0.75
3.)	ACETAL	0.71
4.)	METHYL	0.50
5.)	KETAL	0.26

1035 cm^{-1} and 1210 cm^{-1}. The answer based on the spectral data entered is "yes". The following question asks if there are at least 5 peaks in the same region, the answer to which is "no" based on the data. The user at this point might be inclined to review the spectrum and ask the "What if....." type question. For example, "What if there are two overlapping peaks making a total of 5 bands in this region, how would this affect the interpretation for acetal?" The user can answer such questions by performing an interactive interpretation for the functionality in question.

A portion of the interactive interpretation for the "acetal" functionality is presented in Appendix C. (The user's responses are presented as lower case, underlined characters.) In this case, the user decides that there are or could possibly be 5 peaks between 1035 cm^{-1} and 1210 cm^{-1} and is interested in what if any effect this will have on the interpretation results. The user, by participating in the interpretation process can quickly determine the effect of such a change on the results. In this case, the change results in the interpreter assigning a higher expectation value for the "acetal" functionality. The results of the interpretation are presented in Table 1-4.

The ability to participate interactively in the interpretation process is not, however, without peril and the expectation values obtained should be viewed with caution. The user must realize the potential exists for generating erroneous results if decisions based on the spectral data are changed or peak data are inserted or deleted arbitrarily. Since the tracing facilities show only the path dictated by the experimental data input, they cannot show what results would have been obtained, had the data been different. In the example given above, the user must be aware that an additional peak in the above region may also affect the expectation values assigned to other functionalities. Further, as indicated by the comments included in the decision trace for acetal, other functionalities can be determined with higher reliability. If the additional peak causes the expectation value for one of these functionalities to increase, it is conceivable that the expectation value assigned for "acetal" can actually decrease. It is for this reason, that the user must perform a total non-interactive re-interpretation after making appropriate additions or deletions to the spectral data, to insure the reliability of the results. Used properly, the ability to interact during the interpretation process can significantly enhance the information transferred to the informed user.

Conclusion:

The goal in incorporating the explanatory comments was to provide the user with as much information as possible, while retaining a reasonably compact format. It was sought to provide comments which presented both an explanation of the interpretation, and provided information of an instructive nature. These new capabilities significantly extend the usefulness of computer assisted IR spectral interpretation, and further expand the applications of such systems to include instruction, by exploiting the knowledge gained during ten years of refinement. Since the developed system provides extensive resident explanation of the knowledge base, the comments included should also further facilitate the evolution of computer-assisted IR spectral interpretation.

CHAPTER 2

GENERATION OF COMPOUND-SPECIFIC DESCRIPTIONS FOR INTERPRETING INFRARED SPECTRA OF CONDENSED-PHASE MIXTURES

Introduction:

A rapid, simple, cost-effective method for identifying the likely components of unknown mixtures has many potential applications. These include identification of components in hazardous wastes, environmental screening for hazardous compounds and industrial process control. Infrared measurements on the intact mixtures, followed by interpretation of the complex spectra acquired can be used to solve such problems. Interest in the use of computer-assisted spectral interpretation techniques for such analyses is increasing due to their ability to interpret complex spectra quickly and reproducibly. Both statistically based (71-80) and knowledge based systems (1,13-45) have been used. The practical use of knowledge based systems has increased dramatically during the past few years, although the high expectation and promise of such systems remains largely unfulfilled. One of the major issues concerning large systems is the time and expense required for system development and maintenance (81). Many of the ideas for the present work resulted from an examination of the strengths and weaknesses of the Program for Automated Waste Mixture Identification

(PAWMI), including subsequent modifications (24-25,30), and the table driven approach used by Trulson and Munk (20).

The PAWMI system made use of an automated rule generator which yielded fixed (or static), compound-specific interpretation rules which were used by the PAIRS interpreter (17-18,21-23,35). Intensity and width criteria were not incorporated in the rules. The shifting of band positions (due to matrix interactions, etc.) was treated with three concentric position windows which were centered about the peaks in the spectrum of the pure compounds. The same position windows were used for all spectral bands, without regard to the functionality giving rise to the band. A major drawback arises from the disparity in the magnitude of band shifts, thus, a single position window for all features in a spectrum is likely to be inadequate.

As in the approach taken by Trulson and Munk, it was decided that the knowledge base produced by IRBASE should be composed of information or facts only, not rules. This separation of logic from data simplifies revision of both control and information modules. More importantly, it allows unlimited flexibility in the use of the information during a spectral interpretation. MIXIR, the system which uses the IRBASE output, approaches spectral problems dynamically and in an iterative fashion (37). A dynamic approach is difficult to implement when the information contained in the knowledge base is locked into predetermined, static rules as are used in systems such as PAIRS.

Rapid and consistent generation of the knowledge base, which contains the spectral descriptions of the suspected components, was judged to be critical for success. The accuracy of the compound descriptions depends on two key factors: (1) the ability to extract the important features from the spectra of the pure compound, and (2) the ability to decide when a corresponding relationship exists between a feature observed in the spectrum of the mixture, and a known feature from the spectrum of a pure compound.

The design of IRBASE is sufficiently flexible to allow the use of heuristic rules for other types of spectral data (i.e., Raman, ESCA, etc.). Similarly, the logic used in MIXIR is suitable for any compound data which can be represented as a set of peaks. For example, to build a Raman interpretation system for a small (approximately 50 compound) data set, one would need to perform the following tasks. A new knowledge base for IRBASE would have to be constructed, to account for the expected spectral features and their shifts. This would be performed using a utility program designed for this purpose. The reference spectra would be reduced to peak tables, and the peak tables entered into IRBASE, along with the corresponding functional group information. MIXIR could then be left intact to operate on the knowledge base produced. If the information to be entered into the knowledge base were available, the entire system would be ready for preliminary testing in perhaps two man-weeks.

Experimental:

A Nicolet 3600 FT-IR instrument, including a Nicolet 640 workstation (Nicolet Analytical Instruments, Madison, WI), was used for all spectral processing. The spectral data were processed using Nicolet 1180 and 620 minicomputers. The pure compound spectra were obtained from the 2 cm^{-1} resolution Nicolet-Aldrich FT-IR spectral library.

A VAX 8650 superminicomputer (Digital Equipment Corp., Maynard, MA) was used for program development. Once developed, the programs and data tables were downloaded to an AT&T PC 6300 personal computer (AT&T Information Systems, Bedminster, NJ).

Data Pre-Treatment:

All spectra were subjected to two successive nine point Savitsky-Golay smoothing routines (82), to avoid detecting noise-induced false peaks. This treatment was obtained through a compromise between decreased spectral information and increased peak data reliability with greater smoothing. The peak picking threshold was set to 5% of the intensity of the largest band in the spectrum. The spectra were reduced to peak tables containing position, intensity and width values. Positions were rounded to integral wavenumber values. Intensities were normalized to the largest peak in the spectrum and scaled to integral values from 0 to 20. Peak widths were classified empirically as being either sharp, medium or broad, corresponding to integer values of 1, 2, and

3, respectively. An additional integer code was assigned to peaks which needed "special consideration". These codes corresponded to split, shoulder and poorly formed peaks. The code for "poorly formed" was used to mark bands having poorly defined maxima.

Program Description:

The programs contained in the IRBASE system were written entirely in ANSI standard FORTRAN 77. This language was chosen primarily for its excellent portability. Other important considerations were the high performance provided by a compiled procedural language and the large number of scientific programmers familiar with FORTRAN.

IRBASE consists of two main programs. A database program generates the initial spectral descriptions. A processing program evaluates the "quality" of each component of these descriptions to arrive at the reduced description used by MIXIR. Together these two programs are comprised of approximately 2400 lines of code. Additional modules include a correlation table, a list of defined functionalities, and routines for maintaining them. The two programs of IRBASE are run independently, with the larger requiring approximately 70 kilobytes of system memory in the microcomputer version. The correlation and band shifting information are retained as separate files, and the logic used to process them is encoded in a generalized "meta" format in the FORTRAN code, using modular subroutines. The use of this generalized meta format circumvents the problems associated with manual rule generation. The "metareasoning" used by IRBASE can best be illustrated by describing so-called "metarules" (3,4).

Rules define some condition and the corresponding action to be taken when that condition exists. Metarules are rules which control other rules. One example of a metarule is a rule which directs the selection of other rules to be used by an inference engine, such as "If a band appears in the carbonyl region, then examine the rules concerning ketones". Another example of a metarule is a rule which defines a generalized approach to the formulation of more specific rules, such as: "If feature A appears in examples of X, but not in Y, then A can be used to discriminate X from Y". Although IRBASE uses metareasoning and uses meta-type rules, they are not, strictly speaking, metarules since the files they produce contain spectral descriptions, not rules.

The use of meta-reasoning has many advantages. First, creation and updating of the rule base may be accomplished very quickly using metarules. Generalization of rule strategies may allow the application of a set of metarules to several different problems. Further, the metarules provide a manageable, understandable perspective on any knowledge base which is created. Finally, the incorporation of metarules in the inference engine of an expert system can provide the capability to formulate rules as they are needed during a "consultation session". This reduces the storage requirements of a large rule base, and, more importantly, has fundamental implications on the ability of a knowledge based system to approach a so-called "intelligent" system.

A key feature of the IRBASE design is the separation of logic from data, or "rules" from "facts". This allows the program strategy to be developed and updated independently of the data that is being manipulated. Finally, the generalized rule format yields a program logic which can be readily examined, understood and rapidly changed.

Problem Description:

Many factors complicate the correlation of the spectral features of condensed-phase mixtures with those observed in the spectra of pure compounds. Peak position, intensity, and width can all be altered in mixtures, due to sample dilution and matrix effects. Some features which are found in the spectra of the pure compounds may be hidden in the spectrum of the mixture, and previously hidden features may be observed, due to these same effects. Finally, few spectral features are unique to a particular compound, even when only a limited number of compounds are in the database. Simple efforts to draw a one to one correspondence between known and unknown spectral features are, for these reasons, of limited value.

One approach to overcoming these problems is to use a library or database containing the spectral information for known mixtures, thereby providing information on the effects of peak shifting. The physical interactions leading to band shifting and distortions are, however, complex and often both compound and concentration dependent. Thus, the number of spectra required, even for a small number of compounds, would be prohibitive. Alternately, computer-generated spectra of mixtures can be produced by coadding of spectra of the pure components. These synthetically generated spectra would, however, not exhibit the matrix effects which are often observed in mixtures. Rather than attempting to extract information from an exhaustive database, IRBASE uses the pure compound spectra, along with functionality information provided by the user to assign the bands in the spectrum of the pure compound. Information describing reasonable peak shift ranges for various polar functionalities is used to predict the behavior for the bands in a mixture.

Generating Compound-Specific Spectral Descriptions:

The information flow for IRBASE is presented in Figure 2-1. The user, to create a spectral description, provides IRBASE with the compound's name, spectral information in the form of a peak table and the functional groups which are present in each compound. The program output consists of a total peak table, a major component knowledge base, a minor component knowledge base, and a compound dictionary. The total peak table contains the information necessary to evaluate the band significance. This information includes the position window determined by IRBASE, along with the relative intensity of the spectral features of the pure compound, for all bands of all compounds included in the data base. As will be explained, two different spectral descriptions are created, one corresponding to the features expected for the compound as a major component of the mixture and the other for the compound as a minor component.

A detailed flow chart for the IRBASE system is given in Figure 2-2. The strategy used by IRBASE is to assign bands in the spectra of the pure compounds, knowing the functional groups which are present. Once the band origins are known, reasonable predictions may be made concerning band shifting expected in mixtures. Spectral regions used in band assignment for the pure compounds were kept fairly conservative, (i.e., narrow), to reduce the number of broad spectral windows assigned. The user is informed if a band expected for a given functionality is not observed in the spectrum of the pure compound. IRBASE then displays the spectral region sought, along with any nearby peaks, and the user is given the opportunity to make the assignment.

The spectral regions in which the pure compound bands are sought and the corresponding spectral windows are stored in a "correlation table" by functional group name. This table is the knowledge base for the database program of IRBASE. The 39 common polar functionalities currently in the correlation table are given in Table 2-1. Nonpolar functional groups are not specifically included, since these are unlikely to experience strong matrix effects and can be treated with default windows. Discrimination among subfunctionalities is rarely performed, as little information existed to allow

Figure 2-1. A Diagram of the Information Input/Output for IRBASE.

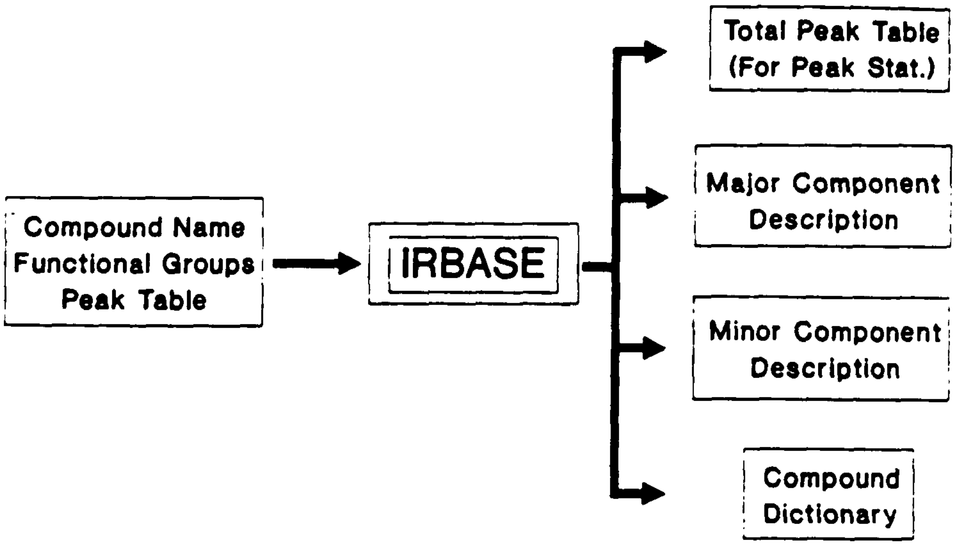


Figure 2-2. Overall Flow Diagram for IRBASE.

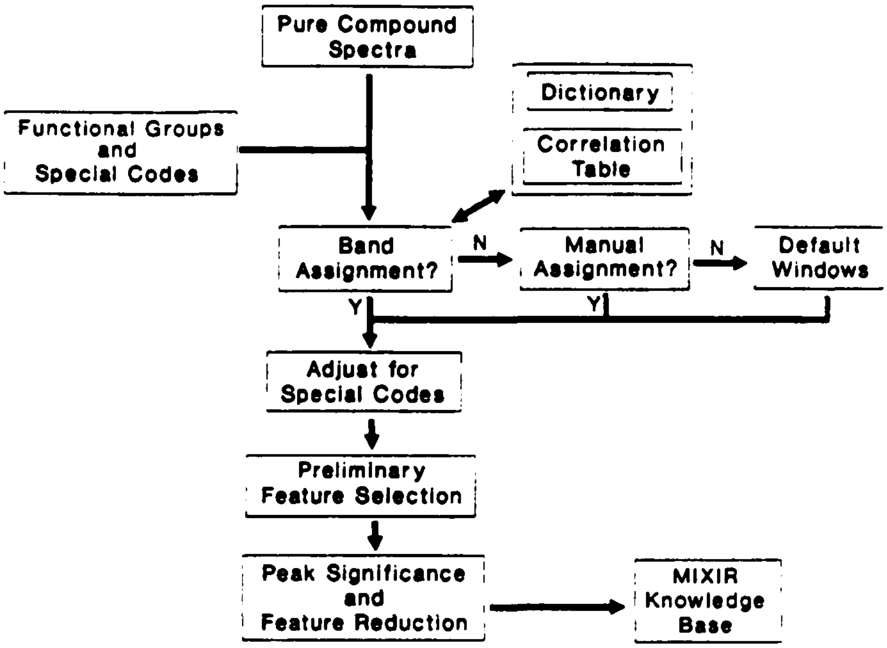


Table 2-1. Functionalities Available In the Knowledge Base.

ACETAL

ALCOHOL-PHENOL

ALCOHOL-PRIMARY

ALCOHOL-SECONDARY

ALCOHOL-TERTIARY

ALDEHYDE-SATURATED

ALDEHYDE-UNSATURATED

AMIDE-PRIMARY-SATURATED

AMIDE-SECONDARY-SATURATED

AMIDE-TERTIARY-SATURATED

AMINE-PRIMARY-AROMATIC

AMINE-PRIMARY-SATURATED

AMINE-SECONDARY-AROMATIC

AMINE-SECONDARY-SATURATED

AMINE-TERTIARY-SATURATED

BROMO

CHLORO-α-DI

CHLORO-AROMATIC

CHLORO-MONO-PRIMARY

CHLORO-MONO-SECONDARY

ESTER-ACETATE

ESTER-BENZOATE

ESTER-FORMATE

ETHER-PROPRIONATE

ETHER-SATURATED

ETHER-UNSATURATED

FLUORO-DI

FLUORO-MONO

KETONE-α,β-UNSATURATED

KETONE-ARYL

KETONE-DI-UNSATURATED

KETONE-SATURATED

NITRILE

NITRO-AROMATIC

NITRO-PRIMARY

NITRO-SECONDARY

NITRO-TERTIARY

OLEFIN

PHENYL

separate prediction of the band shifting effects among subfunctionalities. Table 2-2 gives some examples of the spectral windows. Initial values for these windows were empirically derived using information from Bellamy (70), along with references contained therein. Trial mixtures were prepared to test some of these values, and the windows adjusted accordingly. Finally, testing of the knowledge base prepared by IRBASE with the MIXIR interpretation system provided some additional information. These windows could possibly be improved with additional testing of the MIXIR system, however preliminary results indicate good performance (37). The spectral windows used for band assignment were obtained from published correlation tables (66-68).

IRBASE generates two corresponding descriptions from the pure compound data. One description corresponds to the spectral features expected for the compound as a major component of the mixture, the other is for the compound as a minor component. Major components are those compounds present in the mixture at approximately 30 percent by volume or greater. Minor components are those present at less than approximately 30 percent by volume. These definitions are not intended to be rigid or exclusive, however, as the bands which are observed in the spectrum of the mixture for any component depends on both the location and intensity of features observed for other components. Only bands with an intensity at least 50% of the largest in the spectrum of the pure compound are included in the minor components table. The

Table 2-2. Examples of Spectral Windows.

	POSITION (cm^{-1})		INTENSITY		WIDTH	
Band Origin	HI	LO	HI	LO	HI	LO
ALCOHOL-PRIMARY						
O-H in-plane-bend	+30	-15	+4	1	3	2
O-H out-of-plane	+20	-50	+4	1	3	3
O-H stretch	+200	-50	+4	1	3	2
(Essential)	+250	-250	20	1	3	1
(Alternately)	3700	3600	+4	1	2	1
C-O stretch	+6	-6	+4	1	3	1
ESTER-ACETATE						
C-O stretch	+8	-8	+4	1	3	1
(Essential)	+50	-50	20	1	3	1
C=O stretch	+5	-12	+4	1	3	1
(Essential)	+50	-50	20	1	3	1

Notes:

- Position (HI,LO), Intensity (HI,LO), Width (HI,LO) refer to upper and lower position, intensity and width bounds, respectively.
- Intensities are normalized from 1 to 20.
- Widths are 1= Sharp, 2= Medium, 3= Broad.
- Signed values represent offsets from the values found in the spectrum of the pure compound, unsigned values are absolute.

reason for this lower limit is that smaller bands are more likely to be obscured by spectral interferants and noise, since they will be diluted by the major components of the mixture.

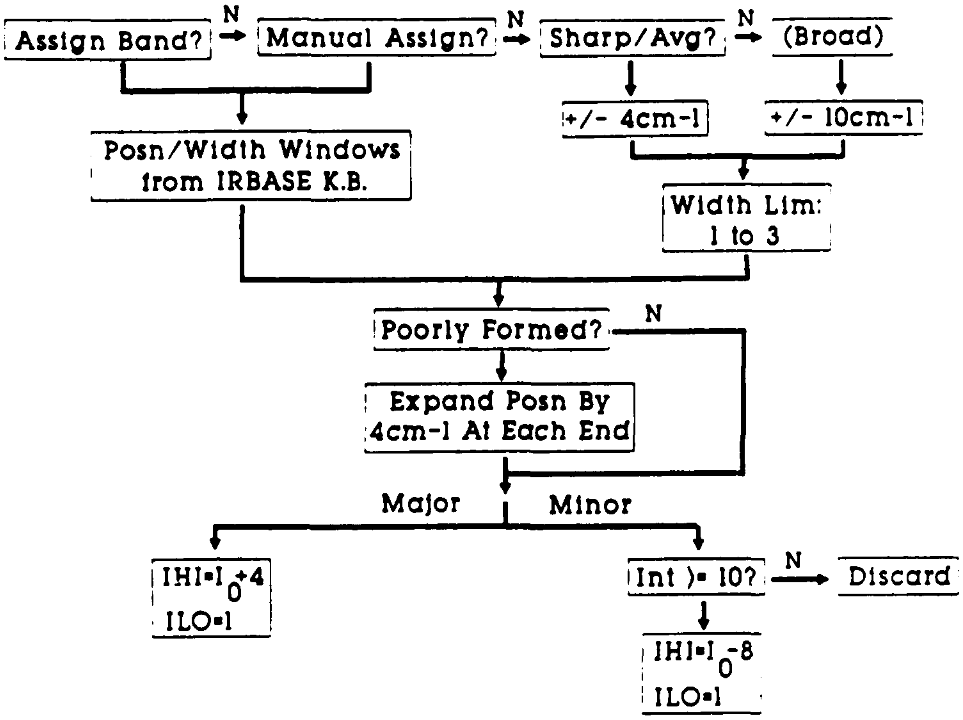
As in most knowledge based systems, a number of system parameters needed to be established during program development, with no a priori knowledge of the "correct" values. Therefore, scientific judgement was used to determine the initial settings. These were then adjusted by observing the results produced on trial mixtures by the analysis system, MIXIR. Generally, little or no adjustment was required in most cases.

A diagram of the basic flow used by IRBASE to assign the spectral windows is presented in Figure 2-3. Creating a spectral description requires the determination of both spectral position windows and the corresponding significance of the regions defined by those windows. The process used by IRBASE to assign the position windows and determine their significance is described below.

Spectral Window Criteria:

(A) Band Position Limits - Bands arising from polar functional groups contained in the knowledge base are assigned first. The position windows describing shifts are then established. Bands arising from nonpolar functionalities receive an initial "default" position window of +/- 4 cm^{-1} for bands having narrow and medium widths, and +/- 10 cm^{-1}, for

Figure 2-3. Window Assignment Scheme for IRBASE.



broad bands. Broad bands receive a larger window, due to the greater uncertainty in the assignment of band position. Once the initial limits have been established, the windows for those bands marked by the user as being "poorly formed" are increased by an additional +/- 4 cm^{-1}. These limits were deemed reasonable for 2 cm^{-1} resolution spectral data, and would require modification for optimum results at other resolutions. Bands indicated as being "shoulders" are not included in either the major or minor component table since they will likely be obscured in the spectrum of a mixture.

(B) **Band Intensity Limits** - The upper intensity limit, for major components, is set at $I_o + 4$, where I_o denotes the normalized intensity of the band in the spectrum of the pure compound. Although extinction coefficients may be larger for the band in the mixture than the pure compound, it is expected that such effects will be offset by dilution. For minor components, the upper intensity limit is set at $I_o - 8$, since such components are expected to exhibit only weak bands.

(C) **Band Width Limits** - Width criteria are rarely used, for several reasons. While it is possible to define classes for band width, determining these values experimentally is a difficult and time consuming process, particularly for overlapping bands. Even if the width values can be measured, it is expected there will be little variance among the band widths that is significant relative to the accuracy with which the widths can be determined. For these reasons, width discrimination is only attempted for very broad bands, such as hydroxyl stretching bands. Such bands can, in general, be unambiguously classified as broad, and this value is often significant.

The correlation table also contains information on alternative regions where spectral peaks may be found in the mixture, and on "essential peaks". The former is included to provide the ability to detect alternative species which may be present in the mixture. Hydroxyl stretches, for example, may be found as a broad band in the 3350 cm^{-1} region under the influence of hydrogen bonding, while free hydroxyl stretches appear as a sharp band at approximately 3650 cm^{-1}. These secondary regions were included to provide a more complete description of the spectral properties of the compound. In an effort to use "not" information, many of the more intense bands are marked as being "essential" when assigned, and a second, conservative spectral region is established about them, for use as an initial screening query for the compound. The presence of a given compound is considered to be extremely unlikely in the absence of such an "essential" band. Only bands with an intensity of 14 or greater may be considered as essential, to reduce the likelihood of false negative results.

Determination of Band Significance:

The reduction of the number of features sought in the spectrum of a mixture is important for two reasons: 1.) To prevent dilution of the significance of the more important features, and 2.) To speed the spectral analysis. At present, the 15 most significant spectral features are selected during the peak weighting process. If at least ten bands remain, exclusive of the C-H stretching bands, then the C-H bands are discarded, since the vast majority of organic mixtures will contain C-H stretching bands. The determination of such bands, therefore, will generally add little selectivity, relative to other spectral regions. If fewer than 10 bands exist without the C-H bands, however, they are included, since the use of too few peaks in the spectral description of the compound may also cause poor selectivity. The overall scheme for feature reduction is presented in Figure 2-4.

Three factors are considered to arrive at the value of the overall band significance. First is the uniqueness of a given spectral feature, which is taken as:

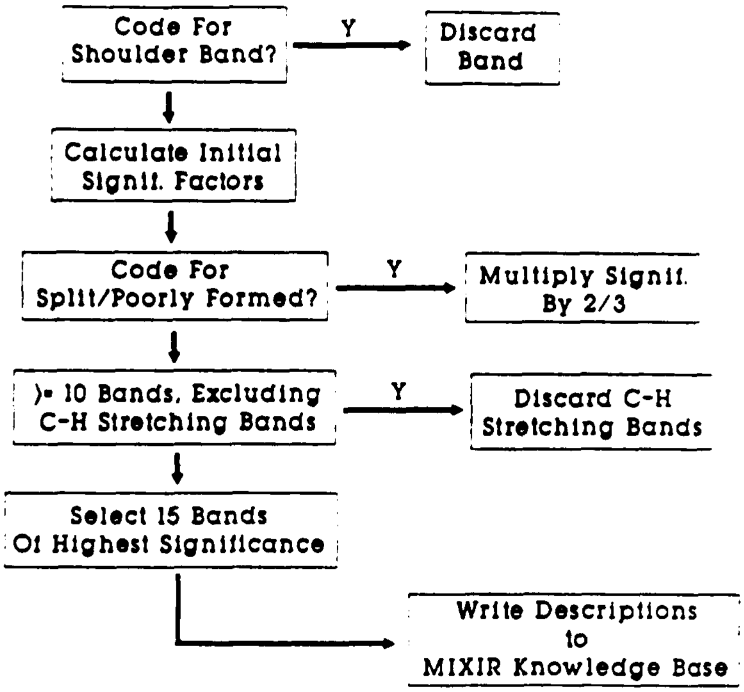
$$\mathrm{UNIQ}_i = \sum_n \mathrm{POS}_n|\mathrm{POS}_i \qquad (1)$$

where UNIQ_i represents the uniqueness of band i, and $\mathrm{POS}_n|\mathrm{POS}_i$ represents the fractional overlap of the position window of band n with that of band i.

The second factor, INT_i, is the intensity of a spectral band, relative to others for the compound. The more intense features of the spectrum for a given compound are more likely to be observed in the spectrum of a mixture containing that compound. The normalized values used for intensity, calculated during preprocessing, represent this factor.

The third factor is the intensity of a band relative to others found in the same region for the compounds in the data

Figure 2-4. The Scheme Used by IRBASE for Feature Selection.


Code For Shoulder Band?
Y
Discard Band
Calculate Initial Signif. Factors
Code For Split/Poorly Formed?
Y
Multiply Signif. By 2/3
>= 10 Bands, Excluding C-H Stretching Bands
Y
Discard C-H Stretching Bands
Select 15 Bands Of Highest Significance
Write Descriptions to MIXIR Knowledge Base

set. Even bands of moderate intensity are likely to be observed, if only small interfering bands due to other components are found in the same region of the spectrum. This factor is taken as:

$$\mathrm{IWT}_i = \frac{\mathrm{INT}_i}{\sum_n ((\mathrm{POS}_n | \mathrm{POS}_i) * \mathrm{INT}_n) \div \mathrm{UNIQ}_i} \qquad (2)$$

where IWT_i is the overlap intensity weight of band i.

Another way of viewing these coefficients is that the uniqueness factor reflects the likelihood of the presence of a compound, given a spectral feature, while the latter two factors reflect the likelihood of observing a spectral feature, given the presence of a compound. A number of compromises must be made in choosing both strategies and parameters in any spectral interpretation system. Frequently, a clear choice can be seen between strategies minimizing false positive results (i.e., the system indicates a compound is present in a mixture when it actually is absent), as opposed to false negative results (i.e., the system indicates a compound is absent in the mixture when it actually present). It was our desire to minimize false negative results, at the expense of false positive results. False negative errors were considered to be of greater concern, since the result in error is eliminated from the analyst's consideration, in contrast to false positive results. The two factors concerning the significance of band absence are important to minimizing false negatives, while the factor related to band presence is important for minimization of false positives. Therefore, only the latter two factors were used in peak selection for spectral descriptions of the compound.

These two factors are then normalized and combined to yield the overall band significance as follows:

$$SIG_i = (C1*IWT1_i + C2*INT_i) * ATTN \qquad (3)$$

where SIG_i is the significance of band i, and $IWT1_i$ and INT_i are the normalized intensity weights. C1 and C2 are weighting coefficients which have been empirically set to 0.67 and 1.00, respectively, and ATTN is an empirical attenuation factor. The ATTN factor is 1.0 for all bands except those which are marked as split or poorly formed. "Poorly formed" refers to bands which are ill-defined or non-Gaussian. It is difficult to determine peak maxima locations for such bands. Poorly formed bands will likely be missed by the peakpicking program for the spectrum of a mixture. Split bands may merge, due to peak shifting in a mixture. Therefore, bands so marked have an ATTN equal to 2/3 to account for the lower likelihood of assigning them in the mixture. The value of 2/3 for ATTN was chosen empirically following preliminary testing of the knowledge base. The resulting SIG values are then used to select the 15 most significant features for the compound.

Results and Discussion:

The creation of the spectral description for the compound ethyl acetate will be used to illustrate the spectral descriptions generated by IRBASE. The spectral peak table resulting from processing the IR spectrum of ethyl acetate is given in Table 2-3, along with the relative peak weighting factors determined by IRBASE. The peak table information, along with the functionality information Ester-Acetate, were entered as data into IRBASE. Fourteen of the 28 bands originally picked were rejected for use in the final spectral description due either to having insufficient intensity (intensity=1) or being classified as "shoulders". Spectral intervals were still derived for these bands, however, and included in the total peak file used for uniqueness derivation.

The spectral intervals derived for the remaining bands, as major components of the mixture, as they appear in an intermediate spectral data file, are given in Table 2-4. Using the correlation table, IRBASE determined the polar C=O absorption at 1742 cm^{-1}, and the C-O stretch at 1241 and assigned +5/-12 and +/-8 cm^{-1} position windows respectively, to them. The remaining bands were assigned an initial +/-4 cm^{-1} window. The range was increased by 4 cm^{-1} at each end for the bands coded as poorly formed. All bands were assigned an upper intensity limit of I_o + 4, as previously noted, and a lower intensity limit of 1. The width limits for all of these bands are 1 to 3, corresponding to no width discrimination. Both the C=O and C-O stretching bands were considered to be essential, and an additional essential peak

Table 2-3. Peak Table and Weighting Factors for Ethyl Acetate

Position (cm^{-1})	Relative Intensity	Width	Special Code	Relative Weight
463	2	2	2	*
608	5	1	-	0.576
634	4	2	-	0.394
786	3	2	-	0.228
847	4	2	-	0.555
918	3	2	2	*
938	4	2	-	0.420
1004	3	2	2	*
1048	15	1	-	0.851
1098	6	2	1	0.320
1160	2	2	2	*
1173	2	2	2	*
1241	20	2	-	1.000
1301	7	2	2	*
1374	14	1	-	0.802
1447	7	2	3	0.166
1465	6	2	3	0.137
1479	5	2	3	0.146
1555	1	2	-	*
1567	1	2	2	*
1742	17	2	-	0.858
1830	1	2	2	*
1889	1	2	3	*
2878	2	2	2	*
2908	4	2	2	*
2943	5	2	2	*
2985	9	2	-	*
3464	1	2	2	*

Notes: Intensities are normalized from 1 to 20.
Width: 1= Sharp, 2= Medium, 3= Broad
Code: 1= Split, 2= Shoulder, 3= Poorly Formed
* Denotes a band not included in the final description.

window was written for these, denoted by the "Essential" code.

The spectral windows derived for ethyl acetate as a minor component are given in Table 2-5. Only four bands, those with at least 50 percent of the intensity of the largest peak, were selected for this group. The position windows for these peaks are the same as for those included in the major component description. The upper intensity limit, however, is set to I_0-8. This different intensity limit is chosen since these bands should be observed at reduced intensity if the compound is a minor component.

The peak weighting factors derived for the bands chosen to be included in the spectral description are presented in Table 2-3. These weights are expected to reflect the likelihood of observing the band, given the presence of ethyl acetate in a mixture. The C-H stretching band at 2985 cm^{-1} was eliminated since at least 10 rule peaks remained, and the C-H stretch is expected to be of little diagnostic value. As one would expect, there is a general trend toward higher weighting factors for larger bands. The bands with the lowest weights, (i.e., those in the region from 1447 to 1479 cm^{-1}) are both poorly formed and also appear in a region where many intense bands, due to other compounds, are observed.

Once established, the spectral windows, special codes, alternative regions, etc., for the selected bands are written to a binary file. This process produces both the major and minor component tables. The resulting binary files are utilized by the MIXIR analysis program during an

Table 2-4. Spectral Windows for Ethyl Acetate as a Major Component of the Mixture.

Band Position (cm^{-1})	Intensity	Width	Special Code	Window Position (cm^{-1})	Intensity	Width
608	5	1	-	604 - 612	1 - 9	1 - 3
634	4	2	-	630 - 638	1 - 8	1 - 3
786	3	2	-	782 - 790	1 - 7	1 - 3
847	4	2	-	843 - 851	1 - 8	1 - 3
938	4	2	-	934 - 942	1 - 8	1 - 3
1048	15	1	-	1044-1052	1 - 19	1 - 3
1098	6	2	1	1094-1102	1 - 10	1 - 3
1241	20	2	-	1233-1249	1 - 20	1 - 3
				Essential 1191-1291	1 - 20	1 - 3
1374	14	1	-	1370-1378	1 - 18	1 - 3
1447	7	2	3	1439-1455	1 - 11	1 - 3
1465	6	2	3	1457-1473	1 - 10	1 - 3
1479	5	2	3	1471-1487	1 - 9	1 - 3
1742	17	2	-	1730-1747	1 - 20	1 - 3
				Essential 1692-1792	1 - 20	1 - 3
2985	9	2	-	2981-2989	1 - 10	1 - 3

Note: An "Essential" prefix denotes a corresponding essential peak query window, with same format as main window entries.

Table 2-5. Spectral Windows for Ethyl Acetate as a Minor Component of the Mixture.

Band Position (cm^{-1})	Intensity	Width	Special Code	Window Position (cm^{-1})	Intensity	Width
1048	15	1	-	1044-1052	1 - 7	1 - 3
1241	20	2	-	1233-1249	1 - 12	1 - 3
			Essential	1191-1291	1 - 20	1 - 3
1374	14	1	-	1370-1378	1 - 6	1 - 3
1742	17	2	-	1730-1747	1 - 9	1 - 3
			Essential	1692-1792	1 - 20	1 - 3

Note: An "Essential" prefix denotes a corresponding essential peak query window, with same format as main window entries.

interpretation session. Peak significance factors are continually recalculated by MIXIR for the rule peaks during the interpretation process. It should be noted that the mathematical manipulations used to calculate peak significance factors in MIXIR are different from those used in IRBASE, due to the different logical analysis possible. In IRBASE, we are looking forward and attempting to forecast the significance of spectral features, the goal being to select the most important features for rule queries. In MIXIR, we have a good deal of information available specific to the problem at hand. This information is updated as knowledge is gained during the interpretation process, and band significance may be calculated by looking back at what is known. The MIXIR system is described in detail in chapter 3.

The spectral windows used by IRBASE to generate the compound descriptions are so-called "hard" or Boolean windows. It may be advantageous to use a "fuzzy" approach (78), or a modified Boolean approach as was used in PAWMI. Alternatively, a "smart" interpreter would ideally use the likelihood of component presence in a mixture, determined in a first pass, to consider the likelihood of observing a given spectral shift. Thus, rather than simply assuming that a -8 cm^{-1} spectral shift is unlikely for C=O, the interpreter would check the likelihood of the presence of strong hydrogen bonding species in the mixture, to attempt to "rationalize" the observation. Simulating logical deduction may prove more powerful than application of statistics to such interpretation problems.

Conclusion:

IRBASE, provides rapid and consistent generation of compound-specific knowledge bases for a dynamic, knowledge based spectral interpretation system, MIXIR. The modular construction of IRBASE and the use of metareasoning allows rapid knowledge base development and practical program maintenance. The spectral descriptions produced by the system for each compound are highly specific. IRBASE presently includes heuristic rules describing the infrared characteristics of 39 polar functionalities. Infrared spectral descriptions have been created for 50 organic compounds using the system. Extending this approach represents an important step toward the creation of practical knowledge based systems for the rapid analysis of complex mixtures.

CHAPTER 3

DYNAMIC, COMPUTER-ASSISTED INTERPRETATION OF INFRARED SPECTRA OF CONDENSED-PHASE MIXTURES

Introduction:

Demand is increasing for the rapid analysis of complex mixtures for environmental and industrial applications. While such analyses can usually be accomplished using a combination of a separation and identification techniques (e.g. GC/MS), there are time, cost, and often scientific advantages to analyzing the intact mixture. Recent reductions in the cost of FTIR instrumentation have made this technique especially attractive for mixture analysis due to the volume of information which can be obtained rapidly. Interpretation of the complex spectral data is, however, often a problem. Due to the complexity and time required for such analyses, there have been numerous attempts to develop so-called expert or knowledge based systems for interpreting IR spectra (15-37).

Briefly, expert systems apply an "inference engine" to process input data through a "knowledge base", to arrive at a logical analysis of the data (2-4). There has been a tremendous surge of interest in this area in the past several years, as artificial intelligence (AI) applications have begun to move from the laboratory to the workplace. Such systems are still evolving, however, and many limitations must yet be overcome before AI technology can become commonplace in analytical chemistry.

One such limitation is frequently the knowledge base itself. Hard coding of rules for every possible event in the solution of complex problems is nearly impossible. What can be done, however, is to broadly classify the possible events, and to generalize the problem solving strategy for each event. Once this has been accomplished, rules may be derived according to the generalizations, as the need arises. This approach involves the use of "meta-reasoning" (3-4). The use of meta-reasoning was deemed necessary to accomplish a major goal of the MIXIR system, the dynamic derivation of rules during the interpretation process.

The problem of determining the components of condensed-phase mixtures by IR spectroscopy presents several interesting challenges. It is a spectral recognition problem where many features may be missing, or altered. Additionally, the spectral features of the mixture are the result of a combination of the spectral features for each of the unknown components. A set of bands which could be attributed to compound "X" may in fact be due to the simultaneous presence of compounds "Y" and "Z". Although "fuzzy" matching criteria have been defined (78), conventional AI search methods cannot be directly applied, since the question is not simply, "Which compound does this resemble most?", but "Which combination of compounds might give rise to this set of features?". Applying an AI searching algorithm to the various combinations possible from the database is possible, in principle, if allowances are made for changes in the spectrum of the components by matrix interactions. This is prohibited in practice, however, by the time required to search the various combinations for even a small data set. Due to practical constraints then, the fundamental question MIXIR seeks to answer is: "Could a subset of the spectral features arise from the presence of compound "X"?"

The goal of this work was to develop a user-interactive knowledge based system to assist chemists in determining the likely components of complex mixtures. The user decides on the logical paradigms used by the system to interpret the spectral data and, if so desired, may participate in the interpretation process. Thus, the system functions as a so-called "smart assistant" in the interpretation process.

Experimental:

A Nicolet 3600 FTIR instrument (Nicolet Analytical Instruments, Madison, WI) was used for all spectral acquisition. The spectral data were processed using Nicolet 1180 and 620 minicomputers. Spectra of the pure compounds were obtained from the high resolution Nicolet-Aldrich FTIR spectral library. Mixtures were prepared, by weight, using Aldrich (Milwaukee, WI) spectral grade solvents. Spectra of these mixtures were acquired as thin films pressed between KBr plates, to reproduce the sample preparation of the pure compounds contained in the spectral library.

Data Pre-Treatment:

Spectra were obtained at a nominal 2 cm^{-1} resolution by coaddition of 100 interferograms. Spectral post-processing was performed exactly as in the preparation of the knowledge base (36).

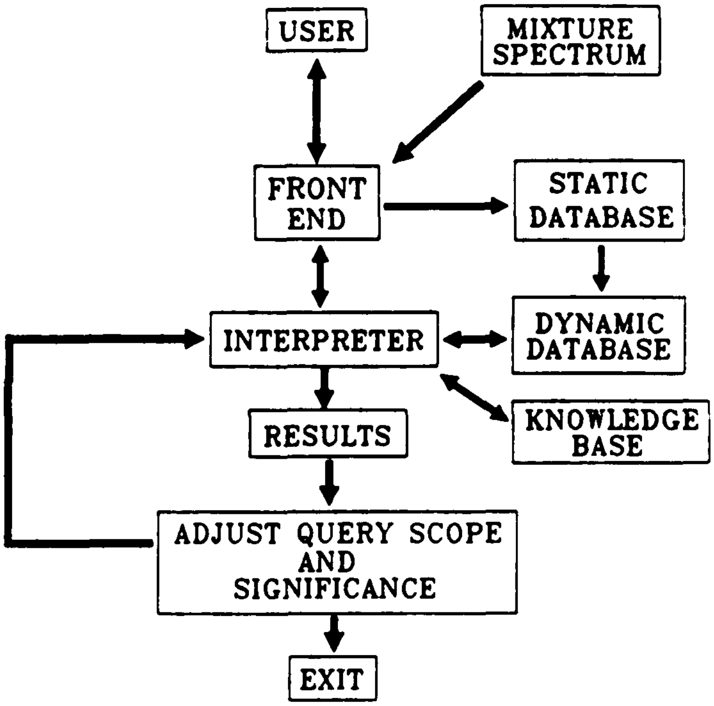
Program Description:

MIXIR was written in ANSI standard FORTRAN 77 and comprises over 4700 lines of code. A VAX 8650 superminicomputer (Digital Equipment Corp., Maynard, Mass.) was used for program development. The programs and associated data tables, once developed, were downloaded to an Dell System 310 (Dell Computer Corporation, Austin, Texas). The microcomputer version of MIXIR requires approximately 260 kilobytes of system memory for execution, and was compiled with the Microsoft FORTRAN compiler, version 4.1 (Microsoft Corporation, Redmond, Washington).

Execution times on the microcomputer, which was equipped with a 20 MHz Intel 80386/387 CPU and math coprocessor, were dependent on the options chosen, and the number of iterations required. A typical value is 30 seconds, using the "optimum" combination of options, detailed in the results and discussion section. No attempt was made to optimize the code for speed. Clarity of code was instead placed foremost during the software design process.

A block diagram of the basic interpretive flow of MIXIR is given in Figure 3-1. The system accesses a knowledge base

Figure 3-1. A Schematic Representation of the Overall MIXIR Program Structure.



previously compiled by the IRBASE program (36). This knowledge base consists of the spectral descriptions for all compounds in the data base. No IF-THEN-ELSE constructs are coded in the knowledge base. The separation of logic from data allows unlimited flexibility in the use of the information contained in the knowledge base and is a key feature of MIXIR. This freedom is critical to providing the dynamic interpretation capabilities sought in developing MIXIR. The system design also simplifies the revision of both the data and controlling logic (20).

Since MIXIR derives interpretation rules at runtime, using the spectral descriptions provided by IRBASE, the user is able to provide information concerning the sample being analyzed. The user can designate compounds which are known to be present or absent in the mixture. Compounds suspected of being present can also be so designated for later use by the interpreter. The information provided by the user is stored in a "static" database. During the subsequent interpretation, inquiries for compounds known to be present or absent are bypassed since their presence or absence has already been established by the user (A conservative approach would be to perform the interpretation with and without restricting the domain). The information contained in the static database is also used during the calculation of significance factors for the remaining substances.

An analogous "dynamic" database is maintained during the interpretation process. This dynamic data base is initialized to the values of the static database at the start of an interpretation and is updated after each pass of the interpreter. It indicates the current state of inquiry for each compound and is probed while determining peak significance factors. The allowed states are: "present", "absent", or "unknown". The designation "unknown" is accompanied by a value indicating the current state of belief for the presence of that compound. This value ranges from +0.999 to -0.999, indicating maximum belief in compound presence or absence, respectively.

The idea behind the dynamic data base is to achieve some of the evolutionary qualities of human problem solving techniques. Consider, for example, the following hypothetical spectroscopist's approach: Noting a strong band in the mixture spectrum at 1745 cm^{-1}, the spectroscopist may think, "A saturated ester appears to be present". Examining a group of likely components reveals five saturated esters. The 1745 cm^{-1} band is not significant evidence for any of these, on the first pass. By checking for the remaining features for these compounds, however, the scientist will revise the relative significance of the 1745 cm^{-1} band for each ester. In an analogous way, MIXIR probes the dynamic database during each interpretation cycle to reevaluate the relative significance of expected features, using the likelihood of competing explanations.

Evaluation of Band Significance:

When interpreting an IR spectrum of a mixture, there is significance in observing a band which corresponds to a band in the spectrum of a pure compound. It is also significant if a band in the spectrum of any given pure compound is not found in the spectrum of the mixture. These two possible events are evaluated independently during a MIXIR interpretation. This division is in many ways analogous at a conceptual level to the "Correlation factor" and "Uniqueness factor" reported by Palmer, et al. (41). for mass spectral interpretation rules.

The significance of finding a spectral feature in the mixture spectrum, which correlates with a band in the pure compound, is related to the confidence that the band is due to the presence of the compound in question. This factor, UWT_i, is given by:

$$UWT_i = (1/NQ) * (S_i/S_{max}) \qquad (1)$$

where NQ is the number of queries for the compound, S_i is the compound score from the previous pass, and S_{max} is the maximum of all compound scores which may explain the band, from the previous pass of the interpreter. The second term in the equation provides a competition for each band sought, between the compound in question, and the best explanation in the knowledge base for that band. The relative "quality" of a given explanation depends on the total evidence for a given compound. The value of the competition term is set to 1.00 during the first pass of the interpreter, since no information exists at that time to determine any preference of one explanation over another.

The significance of not finding a band in a given spectral region is related to the confidence that the band is actually missing, rather than masked by another peak or below the intensity threshold of the peakpicking program. To evaluate the significance of not finding a band, the raw intensity weight, $RIWT_i$ is first calculated:

$$RIWT_i = INT_i * ATTN1 * ATTN2 \qquad (2)$$

where INT_i is the intensity of the band for the pure compound, ATTN1 and ATTN2 are used to account for moderating factors. ATTN1 is set to 2/3 for peaks which were indicated as being poorly formed, or split in the pure compound. Split bands may merge, and poorly formed (i.e., ill-resolved or non-Gaussian) bands will likely be missed by the peakpicking program. ATTN2 reflects the likelihood that a band is masked by another nearby band. The value assigned to ATTN2 ranges from 0.20 to 1.00, corresponding to a high and low likelihood for the band being masked, respectively. The factor is determined using a heuristic procedure to be further described in a subsequent section.

The intensity weights are adjusted to sum to 1.00 using:

$$IWT_i = \frac{RIWT_i}{\sum_j RIWT_j} \qquad (3)$$

Interpretation Process:

An outline of the overall program scheme is presented in Figure 3-2, and a diagram of the basic interpretation process is given in Figure 3-3. The system was designed to continually evaluate the effective uniqueness of spectral features. Uniqueness is considered to be the number of compounds in the database which are potential explanations for a feature found in the spectrum of the mixture. The uniqueness is recalculated during each interpretation cycle using the information contained in the dynamic database which is revised after each cycle, using the interpretation results. This feedback process will ideally converge with a uniqueness of one for a given spectral feature for each compound present.

Only those compounds whose presence is deemed "unknown" are interpreted in each pass. The remaining compounds, however, influence band uniqueness calculations. Each of the "unknown" compounds is queried in turn, with the goal being to determine what evidence exists to support the presence of the compound in the mixture. Matching a spectral feature results in the addition of the significance determined for finding a band in the region queried (UWT), for the compound in question. A missing feature results in the subtraction of the determined significance (IWT) for a band being absent from this region. The significance factors related to presence and absence of bands are each scaled to sum to a value of one. Thus, the resulting compound expectations may range from -1.00 to 1.00, indicating minimum to maximum evidence for the presence of a compound.

Figure 3-2. A Procedural Flow Diagram of MIXIR.

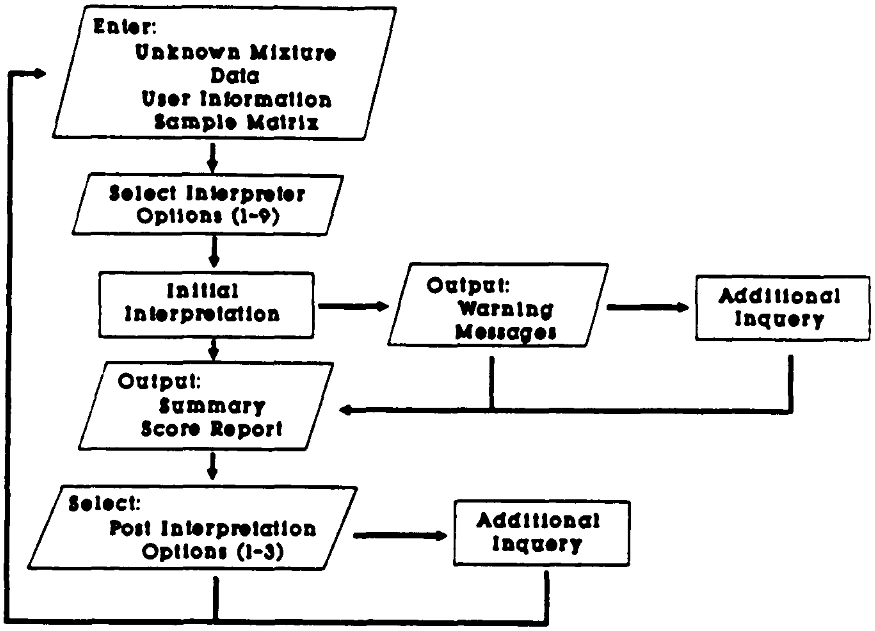
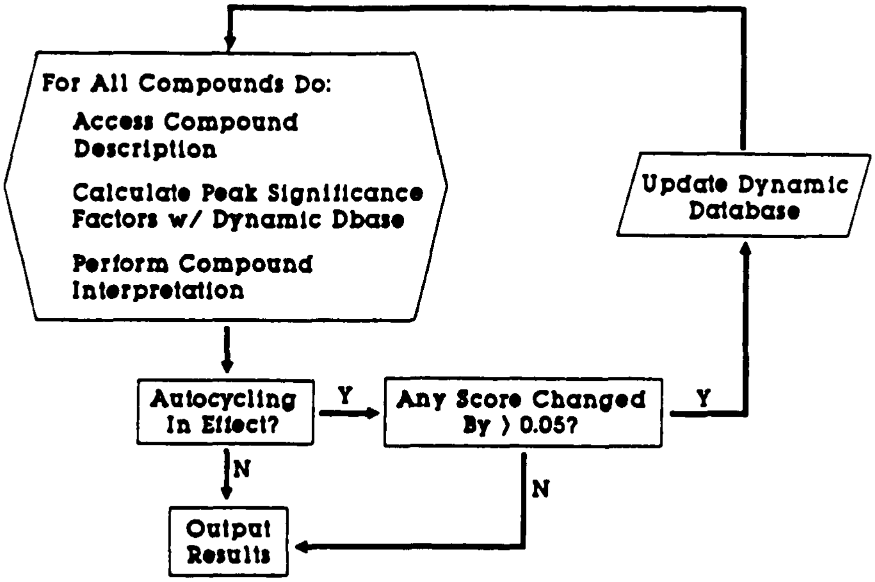


Figure 3-3. An Algorithmic Description of the Basic MIXIR Interpretation Process.


For All Compounds Do:
Access Compound Description
Calculate Peak Significance Factors w/ Dynamic Dbase
Perform Compound Interpretation
Update Dynamic Database
Autocycling In Effect?
Y
N
Any Score Changed By > 0.05?
Y
N
Output Results

User Selectable Interpretation Options and Paradigms:

A major goal in designing MIXIR was to incorporate as many problem solving strategies as possible, realizing the limitations imposed by the use of peak tables to represent the spectra. While these strategies provide potentially greater discriminatory capabilities and allow the user to participate in the interpretation process, they may also provide misleading results, particularly to the uninformed user. It was decided, therefore to include these paradigms as options, and provide a brief on-line explanation of each. Thus, the user is forced to make informed decisions concerning the specific problem solving paradigms employed. The chosen strategy may include any or all of the options available. The system includes both pre and post-interpretation options which are user selectable. The nine pre-interpretation options are:

1) **User Override-** This option provides a complete interpretation trace, and causes the interpreter to pause after presenting each query and answer. The user is permitted to override the decision, if so desired. Thus, the user is actively involved in the interpretation at the most detailed level. The ability to interactively participate in the interpretation process has proven to be very useful for other knowledge based spectral interpretation systems (35).

2) **Relative Peak Intensity Checking-** Absolute peak intensities are not significant when interpreting spectra of mixtures, since an unknown dilution factor renders

them meaningless. Relative peak intensities, however, may be significant. If two bands in a spectrum of a pure compound have a ratio of 3:1 then it is not expected that corresponding bands in the mixture will have a ratio of 1:4. Gross intensity mismatches are checked using this option. For each unknown peak matched for a given compound, the intensity relationship to the next peak matched is determined. The intensity relationship is assigned one of five values, ranging from "much greater than" to "much less than". The intensity relationships for the peaks in the spectrum of the pure compound are then compared to the corresponding intensity ratios in the spectrum of the mixture. If a corresponding or adjoining relationship is not found, the compound score is significantly reduced. This option was designed as a coarse filter, to reduce decisions leading to false positive results.

3) **Essential Peak Checking-** Discrimination by this filter corresponds to the spectroscopist's argument: "Without a band in the carbonyl region, I know that acetone cannot be present in my sample". This option makes use of any essential peak descriptions which may be encoded in the spectral description for each compound. Failure of an essential peak query results in the score for the compound being set to the minimum value, and further queries for the compound are ignored.

4) **Extended Scoring System-** This option considers the

cumulative importance of matching bands which can only be attributed to a few compounds. When invoked, the number of bands matched for a given compound which can only be attributed to 2 and 3 compounds in the knowledge base is noted. Additional credit is then given for the cumulative number of each of these, using an empirically derived exponential function:

$$Sig = Sig_0 * 1.20^x * 1.10^y$$

"Sig" is a value normalized from 0 to 1, Sig_0 is the compound score prior to the function application, also normalized from 0 to 1, while x and y represent the number of bands matched for the compound which could only be attributed to two and three compounds, respectively. The function output is then converted to the -1 to 1 scale normally used by MIXIR.

5) **Automatic Cycling-** Using this option, the interpreter will iteratively interpret the mixture spectrum, updating the dynamic database after each interpretation. The dynamic database is probed during each cycle to continually update the band significance factors, using the new information. This iterative process continues until no compound score changes by more than 0.05. This capability is extremely important, as it provides the interpreter with the ability to use knowledge gained during the interpretation process to reevaluate the significance of the presence or absence of spectral

features. The problem is attacked by a progression through a series of successive states.

6) **Interpreter Trace-** It is important that the user be able to determine how the program arrived at a given conclusion. This option provides a complete real time trace of the interpreter queries, answers, and actions. This option is a non-interactive user information utility.

7) **Peak Justification-** This option checks that all major bands in the mixture spectrum can be attributed to a compound which has been determined likely to be present at high significance. The interpreter pauses after the initial interpretation process to determine whether the compounds with scores greater than 0.20 can account for all of the major unknown peaks. If not, then all compounds which could explain this feature are examined, and those with the two highest scores are determined. The user is informed of the situation and can attempt to distinguish between the various compounds which may account for each unclaimed band using a post-interpretation "either/or" procedure.

8) **Reduced Peak Set Check-** If a compound is present at a low concentration, it is expected that only the most intense bands of that compound will be observed in the spectrum of the mixture. If one were to arrange the bands sought for that compound, in order of decreasing relative intensity in the pure compound, and mark those which are

observed in the mixture, then one would expect to obtain two distinct groups of peaks. The first group would contain the peaks which are above the threshold used by the peakpicking program while the second group would contain those peaks which are below it. Using reduced peak set checking causes the interpreter to look for situations where (a) Three bands are matched, and those matched are all larger than those which are missed, or (b) More than three bands are matched, and at most one anomaly exists in the expected pattern. Such anomalies can be explained if an intense band which we expect to be present may is still missed due to peak overlap, or a less intense band that we expect not to be observed may be correlated due to another component. If either of these conditions (a or b) are met, the compound score is increased by 25 percent.

9) **Peak Swamping Check-** This paradigm considers situations where a band from a component at low concentration may be masked by a large band due to another component. Using this option, the interpreter does not deduct for a missed spectral feature if it is reasonable to expect that the band may be present but masked. Decisions concerning the likelihood of band masking are made using a heuristic tree consisting of approximately twenty rules. These rules take into account the intensity of the band in the pure compound spectrum, the distance to the closest band in the unknown mixture, and the intensity and width of

the closest band. When selected, this procedure is used for all compounds. It can also be invoked for individual compounds which are suspected to be components of the mixture, as described previously.

Three user selectable post-interpretation paradigms are available to provide additional information. They are:

1) **Peak Assignment-** This corresponds to a spectroscopist saying "I know that these bands can be attributed to compounds X,Y..etc, what information is still left to be explained, and what compounds can account for it?". To answer this question MIXIR determines all bands in the spectrum of the mixture which can be attributed to compounds above a user specified score. The interpretation is then repeated for the remaining compounds, neglecting the assigned bands. This option may reduce false positives, but there is the danger that false negatives may result. Erroneous results are expected if coincident information exists for compounds present in the mixture at greatly disparate concentrations. The compound at high concentration would then be expected to have a high score, and have bands attributed to it eliminated. Any unresolved band due to the compound at lower concentration would then also be eliminated.
2) **Either/Or Procedure-** This procedure is very useful when structurally similar compounds are present in the

knowledge base. If one such compound is present in the unknown mixture, then any analogs will likely also be assigned high scores. This option should be used with two chemically similar compounds as the arguments. A set of interpretation rules is then derived for each compound which excludes any coincident spectral information. Thus, the results of a subsequent interpretation indicates the evidence for each compound, neglecting the coincident information.

3) **Interpretation Cycle-** This option allows the user to single step through interpretation cycles as in the auto mode. The user is thereby permitted to examine the developing solution by reviewing the numerical results after each pass of the interpreter.

Results and Discussion:

The MIXIR system was evaluated using a knowledge base produced by IRBASE (36). The knowledge base consists of spectral descriptions for 50 compounds as both major and minor components. The test compounds are listed in Table 3-1. The system was evaluated using a set of 20 mixtures (Table 3-2), which included 10 two-component mixtures and 10 three-component mixtures, with components ranging from 10 to 90%, by mass.

To evaluate the system, the entire set of mixture spectra was interpreted with and without using combinations of the interpreter options previously described. The results are

Table 3-1. Compounds Included in the MIXIR Knowledge Base.

acetone
anthracene
benzaldehyde
benzene
2-butanone
butyl-acetate
chlorobenzene
1-chlorodecane
chloroform
o-chlorotoluene
1-phenyl-1-propanol
o-cresol
p-cresol
cyclohexanone
dibutyl-phthalate
o-dichlorobenzene
m-dichlorobenzene
1,2-dichloroethylene
dichloromethane
2,4-dichlorophenol
1,2-dichloropropane
1,3-dichloropropane
1,3-dichloropropene
dicyclopentadiene
diethyl-phthalate
dioctyl-phthalate
2-ethoxyethyl-acetate
ethyl-acetate
ethyl-alcohol
ethylbenzene
2-ethylphenol
m-ethyltoluene
p-ethyltoluene
eugenoltoluene
hexachloro-1,3-butadiene
hexachlorocyclopentadiene
n-hexane
2-hexanone
methyl-alcohol
3-methylpentane
4-methyl-2-pentanone
pentachloroethane
phenol
n-propanol
n-propyl-acetate
2-propanol
styrene
1,2,3,4-tetrachlorobenzene
1,2,4,5-tetramethylbenzene

Table 3-2. Test Mixtures Used to Evaluate MIXIR.

	Components	Composition (w/w)
1.	Ethylbenzene/1,3-Dichloropropane	1:1
2.	Ethylbenzene/1,3-Dichloropropane	4:1
3.	Chloroform/Ethyl Acetate	3:1
4.	Chloroform/Ethyl Acetate	1:5
5.	n-Propanol/2-Butanone	3:1
6.	n-Propanol/2-Butanone	1:9
7.	Chlorobenzene/n-Hexane	2:1
8.	Chlorobenzene/n-Hexane	1:5
9.	Ethylbenzene/Chlorobenzene	9:1
10.	Ethylbenzene/Chlorobenzene	1:4
11.	Ethylbenzene/1,3-Dichloropropane/Ethyl Acetate	3:1:1
12.	Ethylbenzene/1,3-Dichloropropane/Ethyl Acetate	1:3:1
13.	Ethylbenzene/1,3-Dichloropropane/Ethyl Acetate	1:1:3
14.	n-Propanol/2-Butanone/Ethyl-Acetate	1:1:1
15.	n-Propanol/2-Butanone/Ethyl-Acetate	5:1:1
16.	n-Propanol/2-Butanone/Ethyl-Acetate	1:1:8
17.	Chlorobenzene/n-Hexane/Chloroform	2:2:1
18.	Chlorobenzene/n-Hexane/Chloroform	1:9:1
19.	Chlorobenzene/n-Hexane/Chloroform	8:1:1
20.	Ethylbenzene/Chlorobenzene/n-Hexane	1:1:6

reported as: 1.) the average scores for the compounds present and absent, and 2.) the number of false positive and false negative results at 7 different score cutoff values. The reason for reporting results at a number of different thresholds is to avoid interpreting a blanket increase or decrease in compound scores as an actual improvement or degradation of system performance.

Non-Iterative Interpretation: The interpretation results follow a predictable pattern when the automatic recycle option (i.e., iterative interpretation) is not used. The effect of using the other options can be determined by comparison of the interpretation results with those produced when no user-selectable options are specified (Table 3-3). The relative peak intensity checking and essential peak checking options are expected to reduce the compound scores when the spectra do not meet the criteria previously described. This is indeed the case, as can be seen from the results in Table 3-4. Several points should, however, be emphasized. Relative peak intensity checking reduces the scores of the compounds present, as well as those of compounds absent. Thus, although a higher number of correct results is obtained for most cutoff thresholds, there is some tradeoff of increased false negative results for decreased false positive results, using this option.

Using the essential peak checking option results in a slight decrease in the average scores of the compounds present in the mixture but also results in a marked decrease in the

Table 3-3. Baseline MIXIR Results (No User Selectable Options in Use) for the Interpretation of the Mixtures Given in Table 3-2.

Average score for the compounds which are present: 0.516
Average score for the compounds which are absent: -0.560

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	4	16	980
0.30	7	14	979
0.20	13	13	974
0.10	26	7	967
0.00	42	6	952
-0.10	65	5	930
-0.20	114	3	883

Table 3-4. Interpretation Results for the Non-Iterative Use of the Intensity Checking and Essential Peak Checking Options.

A. Intensity Checking Option

The average score for the compounds which are present: 0.451
The average score for the compounds which are absent: -0.590

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	2	18	980
0.30	4	17	979
0.20	8	15	977
0.10	18	11	971
0.00	30	9	961
-0.10	48	8	944
-0.20	88	5	907

B. Essential Peak Checking Option

The average score for compounds which are present: 0.516
The average score for compounds which are absent: -0.680

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	4	16	980
0.30	7	14	979
0.20	13	13	974
0.10	25	7	968
0.00	40	6	954
-0.10	60	5	935
-0.20	104	3	893

average scores for the compounds which are absent. This result confirms the desired selectivity for this logic filter. It should also be noted, however, that the number of false positive and false negative results is unaffected, compared to the baseline results, until a score cutoff threshold of 0.10 is reached. Thus, essential peak checking did not discriminate greatly against false positive results at high significance for the test mixtures. Since the strong false positive results tend to come from closely related compounds, it is not surprising that they contain major spectral features which are similar.

The extended scoring system, option 4, had no effect without performing iterative interpretations. This result simply indicates that no peaks of uniqueness 2 or 3 could be matched without reducing the problem bounds, and updating the effective uniqueness, i.e., performing an iterative interpretation.

The results obtained using the peak swamping check and reduced peak set check options are presented in Table 3-5. Reduced peak set checking (option 8) produces enhanced scores for compounds present, while leaving the scores for compounds which are absent largely unaffected. The result is reasonable since it is more likely that compounds which are present in the mixture will demonstrate the expected intensity pattern than compounds which are absent.

While reducing the number of false positive results, the consideration of peak swamping (option 9) also produces a

Table 3-5. Interpretation Results for the Non-Iterative Use of the Reduced Peak Set Checking and Peak Swamping Check Options.

A. Reduced Peak Set Checking Option

The average score for compounds which are present: 0.593
The average score for compounds which are absent: -0.558

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	4	14	982
0.30	7	12	981
0.20	14	11	975
0.10	27	6	967
0.00	43	5	952
-0.10	66	4	930
-0.20	114	3	883

B. Peak Swamping Option

The average score for the compounds which are present: 0.571
The average score for the compounds which are absent: -0.487

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	6	16	978
0.30	20	11	969
0.20	33	8	959
0.10	50	7	943
0.00	68	5	937
-0.10	117	4	879
-0.20	162	2	836

significant enhancement in the scores for compounds which are absent. In fact, the proportional gain for absent compounds is greater than for those compounds which are present. This observation can be explained by considering the following facts. The peak swamping check discriminates between compounds for which band masking is unlikely. If, however, band masking is expected, then it is equally likely that the judgement on masking will be made for a band of a compound which is present as for one which is absent. Since this option only has an effect when a band is not observed, and compounds which are not present will have more missing bands, then on average these compounds will have a greater enhancement. This option should, therefore, only be chosen if the user is willing to accept a significant increase in the number of false positive results, for the greater margin of safety afforded.

Iterative Interpretation: The success of interpretations using the autocycle option can be seen by comparing the results given in Table 3-6 with the baseline results of Table 3-3. Even more dramatic gains are made when the extended scoring system is used in conjunction with autocycling (Table 3-7). As noted above, the extended scoring system only has an effect when just two or three compounds can explain a given band. This high uniqueness is obtained through iterative interpretation.

Use of the essential peak checking option in conjunction with autocycling produced interpretation results which may

Table 3-6. Interpretation Results Using the Autocycle Option.

The average score for compounds which are present: 0.507
The average score for compounds which are absent: -0.709

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	2	18	980
0.30	4	15	981
0.20	6	13	981
0.10	6	9	985
0.00	14	8	978
-0.10	23	6	971
-0.20	36	5	959

Table 3-7. Interpretation Results Using the Extended Scoring System and Autocycle Options, Concurrently.

The average score for compounds which are present: 0.751
The average score for compounds which are absent: -0.713

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	3	8	989
0.30	4	8	988
0.20	6	8	986
0.10	9	8	983
0.00	12	7	981
-0.10	21	6	973
-0.20	35	5	960

Table 3-8. Interpretation Results Using the Essential Peak Checking and Autocycle Options, Concurrently.

The average score for compounds which are present: 0.574
The average score for compounds which are absent: -0.781

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	2	16	982
0.30	4	13	983
0.20	6	11	983
0.10	6	9	985
0.00	14	8	978
-0.10	23	6	971
-0.20	34	5	960

seem puzzling initially. It was expected that using the essential peak checking option would reduce the scores of those compounds which failed the essential peak check test, while leaving the scores of the other compounds unaffected. The results given in Table 3-8 show the expected decrease in the scores of absent compounds. What is not so obvious is that the scores of the compounds not directly affected (ideally, only compounds present), must then increase. This is readily explained by the competition term of equation 1. Autocycling, which was designed to be a feedback amplification mechanism, tends to magnify both the desirable and undesirable effects produced by the logical algorithms employed.

The "optimum" combination of user selectable options is defined as the condition which produces the greatest separation between scores for the compounds which are present and those which are absent for the test mixtures. Of the combinations of options tested, optimum results were obtained using the autocycle, essential peak checking, extended scoring system and reduced peak checking options, concurrently. The results achieved using this combination of options to interpret the mixture spectra are presented in Table 3-9.

A portion of the score report for the interpretation of a 1:1:6 w/w mixture of ethylbenzene, chlorobenzene, and n-hexane, using the optimum option settings, is given in Table 3-10. The likelihood of each compound as a major and as a minor component is reported along with the number of bands sought, and the number of bands matched. This latter

Table 3-9. Interpretation Results for the "Optimum" Option Combination Specified in the Text.

The average score for compounds which are present: 0.803
The average score for compounds which are absent: -0.780

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	6	7	987
0.30	6	6	988
0.20	6	5	989
0.10	10	5	985
0.00	17	5	978
-0.10	23	5	972
-0.20	33	5	962

Table 3-10. Partial Score Report for a 1:1:6 w/w Mixture of Ethylbenzene/Chlorobenzene/n-Hexane.

COMPOUND	MAJOR	Peaks Matched	Peaks Sought	MINOR	Peaks Matched	Peaks Sought
n-HEXANE	.999	6	8	-.999	0	4
ETHYLBENZENE	.999	12	15	-.603	1	5
CHLOROBENZENE	.999	10	12	-.331	3	8
TOLUENE	.795	9	12	-.710	1	6
DICHLOROMETHANE	-.333	2	7	-.999	0	2
:						
: (44 Intermediate Compound Scores)						
:						
1,2,3,4-TETRACHLOROBENZENE	-.999	0	15	-.999	0	8

information may be important for cases where the pure compound has only a few peaks such that finding a few corresponding bands in the mixture spectrum will generate a high score for that particular compound. The user should, therefore, suspect a false positive result. Similarly, a false negative result should be suspected when a number of bands in the mixture are attributed to a particular compound, but the score for that compound is low.

One of the difficulties encountered in designing a spectral interpreter for mixture analysis is determining how to distinguish between compounds which are close structural analogs and, thus, have a high degree of spectral similarity. It is often observed if one of the structural analogs is a component of the mixture, that interpretation of the spectral data results in a high likelihood for the other analogs being present as well. The MIXIR program was designed to allow the user to perform a two part interpretation of the data to resolve this problem. The first interpretation is used to answer the question, "What compounds are likely to be present in this mixture?". Subsequently, the user recognizing that structurally similar compounds are indicated as likely components of the mixture can perform a second interpretation, using the "either/or" post-interpretation option. This option is used to answer the question; "If only one of these two compounds is present, which is more likely, X or Y ?". Since the coincident spectral features for the two compounds provide no information which allows MIXIR to discriminate between the two compounds, the interpreter must evaluate the spectral features which are unique for each compound. The separation of logic from data in the design of the system, allows MIXIR to derive a set of rules to discriminate between two compounds. To perform an interpretation with the "either/or" option, the user simply indicates which two compounds to distinguish between. After determining which spectral bands can be attributed to both of these compounds, MIXIR discards the corresponding peak queries. The remaining peak queries are then used to derive a set of interpretation rules which correspond to the spectral features unique to each compound.

The score report for the interpretation of a 1:1:6 w/w mixture of ethylbenzene, chlorobenzene, and n-hexane (Table 3-10) will be used to illustrate an "either/or" interpretation. On examining this report, it is apparent that toluene is falsely reported to be a component of the mixture with a score of 0.795. The user should question these results since nine bands of the mixture spectrum were matched for toluene, while ten bands were matched for a close structural analog, ethylbenzene. Thus, much of the spectral evidence for toluene may be a subset of that for ethylbenzene. A second interpretation was performed using the "either/or" option to discriminate between these two compounds. The interpretation resulted in scores of 0.534 and -0.033, for ethylbenzene and toluene, respectively, which indicates that little spectral information is unique to toluene.

The 20 mixture spectra were interpreted using the previously described optimum combination of user selectable options. If compounds with scores above 0.20 are assumed to be present, then 6 false positives results are obtained. Only 3 false positives remained after using the post-interpretation "either/or" option to distinguish between obvious structural analogs which had scores above 0.20 from the first interpretation. The criterion used for discrimination between the two structural analogs was that their scores, after using the "either/or" option, differ by at least 0.500.

Users must be aware, however, of the limitations of the "either/or" procedure. For example, several mixtures were prepared containing structural analogs. Two of these also became subject to the either/or procedure, and in one case, a false negative result was produced. In that case, interpretation of the spectrum of a 9:1 w/w ethylbenzene/chlorobenzene mixture resulted in scores of 0.999 for both compounds. A second interpretation was performed using the either/or option to distinguish between these two compounds. The resulting scores were 0.999 for ethylbenzene and 0.291 for chlorobenzene. The difference, 0.707, was greater than the 0.500 discrimination criterion chosen previously, thus indicating, ethylbenzene is more likely to be present in the mixture than chlorobenzene. This result should not be surprising, since ethylbenzene was present in the mixture at nine times the mass fraction of chlorobenzene. Similarly, an interpretation of a 1:1:6 mixture of ethylbenzene/chlorobenzene/n-hexane again yielded initial results of 0.999 for chlorobenzene and ethylbenzene. Interpretation using the either/or procedure in this case yielded 0.576 for chlorobenzene and 0.758 for ethylbenzene. Since the two compounds were present at equal concentrations, neither was found as a preferred explanation by this procedure. The either/or procedure is invoked to perform a discriminatory function: to determine which of two compounds is the preferred explanation for spectral evidence. It should be realized that if structural analogs are actually present at widely different levels, then the one which is more concentrated will likely be preferred.

A discussion of system performance would not be complete without an easily digestible summary. Table 3-11 presents the results of testing the interpreter using the 20 mixture test set, as described. The "optimum" option combination was used to interpret the mixture spectra. The expectation cutoff for compound presence was set at 0.20. Defining system reliability as the number of correct decisions divided by the total number of decisions, as given in Table 3-11, resulted in a derived reliability of 0.99 for the test data. This value demonstrates the discrimination abilities of MIXIR.

The possible pitfalls associated with the various optional interpretation procedures have already been discussed. The major limitation associated with the present system, however, is that it is currently a peak-based system operating on hardware remote from the instrument computer. The advantages and disadvantages of peak-based interpretation methods were examined by Coates (32). Ideally, one would have a peak-based method for coarse searching, followed by the use of spectral curve-fitting to resolve ambiguities. Such a system would reside directly in the instrument computer, and interact with the existing instrumental software to gain maximum effectiveness. While not a scientific question, ease of use would be improved if a (hardware-dependent) graphical window-based environment were developed, as exists on most modern commercial

Table 3-11. MIXIR Performance Summary.

The Following Interpretation Options Were Used Concurrently: Essential Peak Checking, Extended Scoring System, Automatic Recycle, Peak Swamping Check

	TP	TN	FP	FN
Without Either/Or Procedure:	45	944	6	5
After Either/Or Procedure*:	44	947	3	6
Best Case Results:	50	950	0	0
Worst Case Results:	0	0	950	50

* Used to derive the following:

$$\% \text{ False Positives} = \frac{FP}{(FP + TP)} * 100 = 5.9\%$$

$$\% \text{ False Negatives} = \frac{FN}{(FN + TN)} * 100 = 0.6\%$$

$$\text{Reliability} = \frac{(TP + TN)}{(TP + TN + FP + FN)} = 0.99$$

software. Resolution of the above issues, along with incorporation of additional interpretation paradigms, should be the goal of future work.

The true usefulness of a knowledge based system is not simply a function of the distilled output (i.e., the compound scores), but of the user's ability to understand and interpret the results (35). The design of MIXIR provides a "smart assistant" for interpreting IR spectra of mixtures. The informed user is allowed to participate in the interpretation process, specify the logic to be used to interpret the spectral data, perform a dynamic interpretation of the spectral data and as a result is better able to utilize the interpretation results.

CHAPTER 4

SPECTRAL PEAK DETECTION WITH A MULTI-LAYERED PERCEPTRON

Introduction:

Artificial neural networks are mathematical models of biological neural systems. Although they are a gross simplification of actual physical cognitive processes, application of these models has indicated that artificial neural networks have strengths and weaknesses in the same areas as humans. For example, they excel at recognition of visual patterns, but are poorly suited to precise mathematical calculation. Despite the advances in traditional mathematical approaches to pattern recognition, humans are generally still considered to be the most effective pattern recognition system available for audible and visual patterns. Together, the superiority of humans at visual pattern recognition and the link between biological and artificial neural networks provided the inspiration for the research to be described. The object of this research was to investigate the ability of an artificial neural network to reproduce human judgements on the presence of peak-shaped signals in infrared spectral data.

Much of the interest in neural nets stems from their remarkable ability to robustly process information containing a degree of uncertainty. This uncertainty may correspond to a variable, noisy, or incomplete input. Moreover, no underlying statistical assumptions are made on the behavior of the data. As a result, any type of variation in the inputs may be accounted for by simply including example input patterns containing such variation during the training process.

The roots of adaptive artificial neural networks date back to ideas proposed in the late 1940's by D.O. Hebb (51). Hebb first described a system where a group of interconnected neurons "learns" by adjusting the strength of the synaptic connections between them. A great deal of work was done in the 1960's using "single-layered perceptrons", also known as "linear learning machines" (52), including some applications in infrared spectroscopy (60-62). Interest in the area waned, however, when it became apparent that the computational machinery available then was not sufficently powerful to support the volume of calculations involved in training a large network, which may involve billions of floating point calculations. A further blow was dealt by the theoretical work of Minsky and Papert (53), which proved that single-layered neural networks using linear transfer functions could not solve a non-linearly separable classification problem, e.g. the "exclusive-or" problem.

The 1980's brought the development of more complex network architectures, and more sophisticated activation and learning rules. Classification is now possible in pattern spaces of arbitrary complexity. Concomitant advances in computer hardware allow the practical implementation of neural network models on relatively inexpensive, readily available hardware. As a result of these advances, a renewed surge of interest has occurred in neural network theory and application. Munk and Robb (63) recently developed a system for recognition and identification of functional group patterns in infrared spectra. Donahue, Brown and Kumaresan have also reported a system for neural networks to identify functional groups from infrared spectra (64). Long and Gemperline have used a feed-forward network to perform quantitation of wheat samples using near-infrared reflectance spectroscopy (65).

Recent work on developing a peak-based infrared spectral interpretation program for mixture analysis (36-37) has shown that a fundamental limitation of such systems lies in the quality of the peak table. That is, the power of the system is limited not only by the power of the processing logic or mathematics, but by the quality and amount of the spectral input information provide to the system.

Many conventional signal detection algorithms are available (83-86). The infrared spectroscopist is however, often limited to those algorithms which are commonly available with commercial infrared instrumentation or employing visual peak detection using a "trained eye". Spectroscopists often find it necessary to perform the additional step of visual validation of peak lists produced by the instrument's software. The problem with this human intervention is that it is time consuming and requires a trained scientist to obtain reasonable results. Additionally, human judgements on signal presence cannot generally be reproduced. Studies have shown however, that the "trained eye" can provide equivalent or better results than other peak detection methods (86). It would, therefore, be highly desirable to have an automated method of peak validation with the proficiency of a trained human. The development of a peak validation system and stand alone peak recognition system, based on a neural network model termed the multi-layered perceptron (54-59), is described here.

Experimental:

The vapor phase infrared spectra used for this study were acquired at a nominal 0.3 cm^{-1} resolution, and were transformed to 2 cm^{-1} resolution representation for use in a knowledge based system designed to interpret infrared spectra of vapor phase mixtures. The 2 cm^{-1} resolution data was used for this study, since a major goal was to improve the peak table input to the knowledge based system. The spectral transformation and derivation of the training peak tables were carried out on a Nicolet 620 FT-IR workstation (Nicolet Analytical Instruments, Madison, WI). The spectral data and training peak table were subsequently downloaded to a Dell 310 microcomputer (Dell Computer Corp., Austin, TX) via a DEC 8820 superminicomputer (Digital Equipment Corp., Maynard, MA), using Nicolet VAXtran software. The Dell 310 was equipped with a 20 MHz Intel 80386 CPU and the companion 20 MHz 80387 math coprocessor, along with an Intel 82385 cache controller.

The neural network simulation and data visualization programs were written in their entirety by the author at the University of New Hampshire. The source code for these programs was written in Pascal, comprising approximately 3000 lines. The programs were compiled using the Turbo Pascal Compiler, version 5.0 (Borland International, Scotts Valley, CA). The graphic displays were created using a set of routines written by the author to drive the routines provided in the Borland Graphics Interface (BGI).

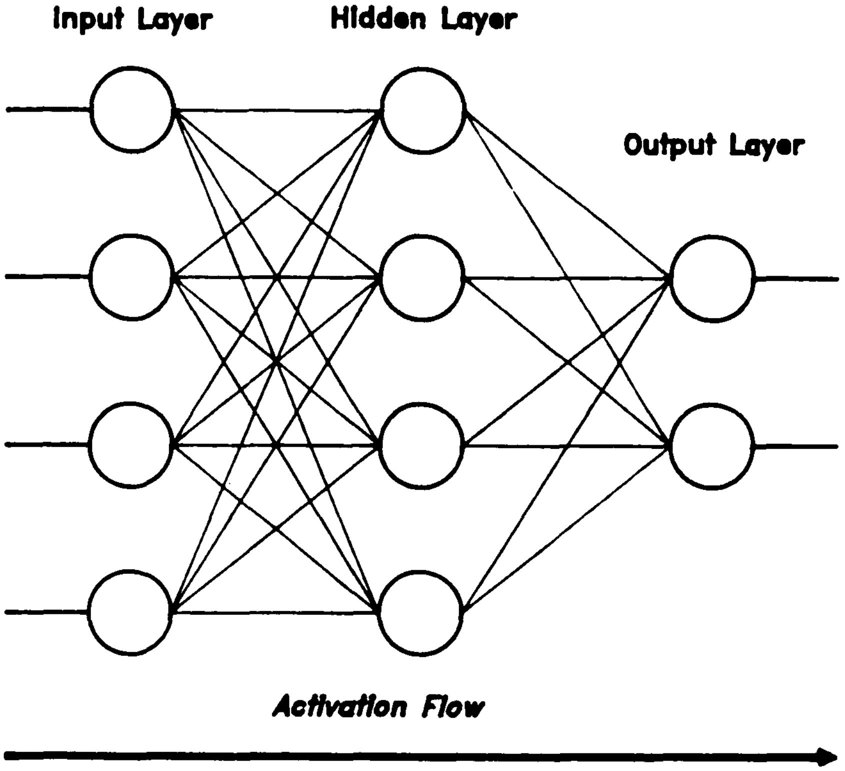
Program Description:

The system consists of three programs: (1) a visual peak confirmation program, (2) a network training and diagnostic program, and (3) a stand-alone peak-picking program. The peak confirmation program is used to provide the "correct" output values used in network training, and has been very useful in determining an appropriate form for the network input. Each input pattern is graphically displayed by this program, scaled analogously to the network input normalization algorithm, and the human "teacher" specifies the correct output for that input pattern. The system was designed such that it allows the teacher to obtain essentially the same view of the data as the network training program. This is important since the network is presumed to carry out the pattern classification in a manner analogous to a human pattern classifier. It was decided that any change in the form of the input pattern which made it easier to distinguish visually between peaks and noise would be incorporated into the network preprocessing calculations.

A three-layered back-propagation network was employed. The layers are termed the input layer, the hidden layer, and the output layer. A fully connected architecture was used, i.e. every node in a given layer was connected to every node in the layer below it (Fig. 4-1). The input nodes simply distribute the input signal to each of the hidden layer nodes. The input vector was composed of a set of absorption values taken from a digitized infrared spectrum. The input vector was first range-scaled (77) as a preprocessing measure to prevent decisions from being made on absolute absorption magnitude.

The network training and diagnostic program constructs the specified network structure in the computer and carries out the calculations and operations involved in training the network: forward propagation of neuron activation, and backward propagation of the error, with concomitant adjustments to the connection weights (54-55,58-59). The equations governing these processes are normally expressed from the viewpoint of a single node, and are understood to be carried out over all nodes in the network. The calculations which follow describe the hidden and output layers, not the input layer. As noted above, the input layer merely distributes the input vector to the first hidden layer inputs, and does not alter the values in doing so.

Figure 4-1. A Schematic Representation of an Arbitrary Fully Connected Neural Network Architecture.



Forward propagation begins with calculation of the input to a node, given by a simple weighted sum of the inputs:

$$I = (\Sigma\ w_i x_i) + w_T T \qquad (1)$$

where the x_i are the input signals to the node, and the w_i are the corresponding weights. The last term, $w_T T$, represents the threshold or bias to be applied to the node. The T-value is a constant, which was given a value of 1.0 here, as is the convention.

The input is then processed through a nonlinear transfer function to obtain the activation for the node according to the general perceptron "sigmoidal" activation equation:

$$A = 1/(1 + e^{-I}) \qquad (2)$$

This activation level "A" is then passed as the output from the node in question to the input of the nodes in the layer above it.

Back propagation of the error is carried out according to a modification of the "generalized Delta learning rule" (54,59). The delta value for an output node (i) is defined as:

$$E_i = y_i(1-y_i)(d_i-y_i) \qquad (3)$$

where d_i represents the desired node output for the input pattern presented, and y_i represents the actual output of the node. There is no known value for the correct output for a hidden layer node, and so the delta for a hidden node is recursively defined in terms of the errors of the nodes in the layer above it:

$$E_i = y_i(1-y_i)\ \Sigma E_j w_{ij} \qquad (4)$$

where the y_i is the output of the hidden node on the previous pass, and the error summation is taken over all the output nodes fed by that hidden node, weighted by the connection strengths (weights) between them.

The adjustment to the weights which accomplishes the so-called "learning" process is defined by a modification to the generalized Delta rule:

$$w_{i(t+1)} = w_{i(t)} + \beta E x_{i(t)} \qquad (5)$$

where the x and w are an input-weight pair, and β is the so-called "gain" constant, which controls the rate of learning. The gain constant was set at 0.60 for all the experimental results to be presented. The "(t)" and "(t+1)" subscripts refer to the old and new values of a variable for a training iteration step. In this work, a so-called "momentum" term was added, to produce:

$$w_{i(t+1)} = w_{i(t)} + \beta E x_i + \alpha(w_{i(t)}-w_{i(t-1)}) \qquad (6)$$

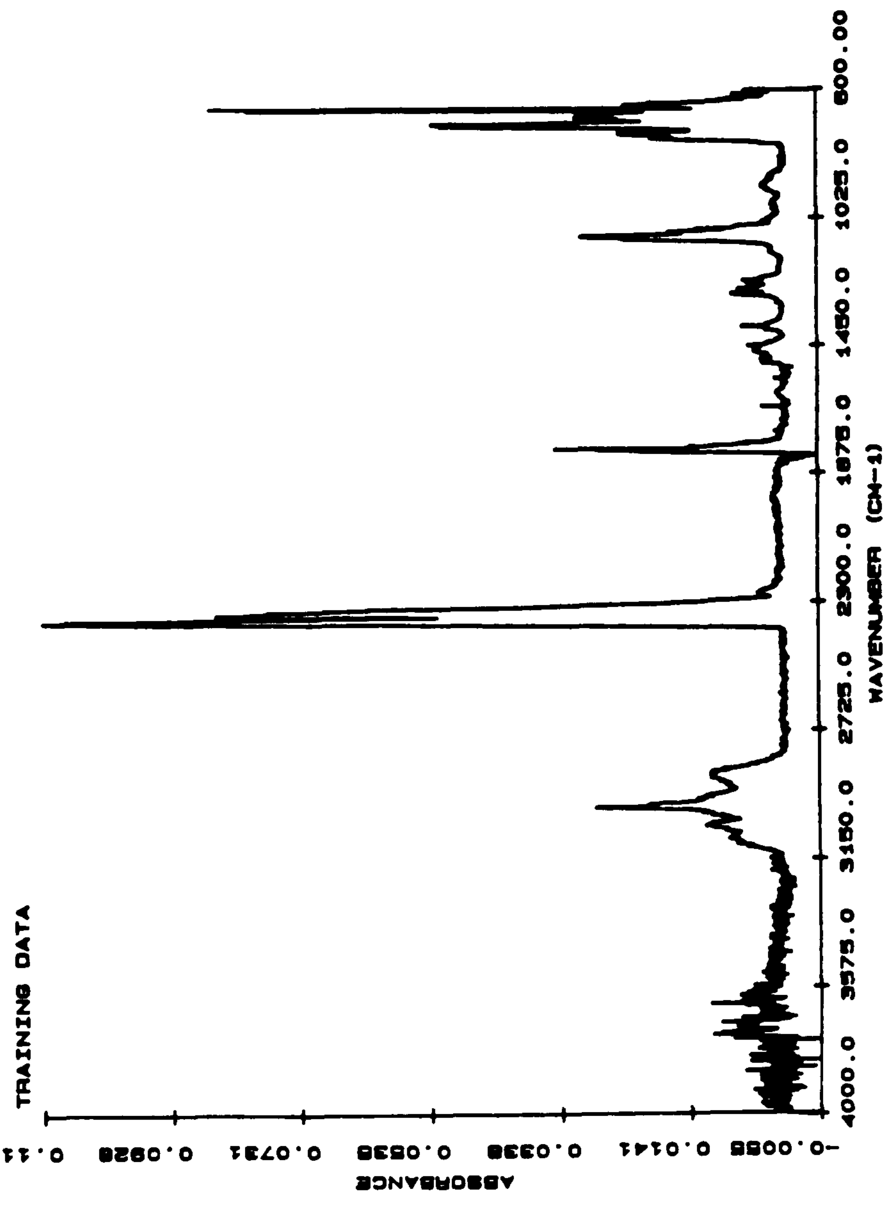
The momentum term tends to preserve the direction and magnitude of a trend in movement of the weight vector during optimization. The purpose of this added term is twofold: to attempt to carry the weight vector out of local minima during the optimization process, and to filter out high frequency variations in the error surface (59). Due to these two characteristics, the presence of the momentum term often speeds the convergence of the learning process, as well. The value of α selects the fraction of the previous step to be summed with the present step. The momentum constant was set at 0.80 for all the experimental results presented here.

Results and Discussion:

The infrared spectral data used in this study were 2 cm^{-1} resolution spectra of vapor phase species. These data presented an exceptional challenge, since the natural width of some of the features was on the order of the spectral resolution. As a result, spectral peaks were often represented by a single data point in this data. This made the task of distinguishing signal from noise very difficult, since no band shape information was available for the narrowest features.

The spectrum used in training the networks (Fig. 4-2) was obtained from a vapor phase mixture of tetrahydrofuran, 1,1-dichloroethane, benzene, ethylbenzene, methylene chloride and 1,1,1-trichloroethane, at concentrations of approximately 3 ppm each. A peak table composed of 132 peaks was produced

Figure 4-2. IR Spectrum of a Vapor Phase Mixture of Tetrahydrofuran, 1,1-Dichloroethane, Benzene, Ethylbenzene, Methylene Chloride and 1,1,1-Trichloroethane, at Approximately 3 ppm Each. Data from this Spectrum was Used for Training the Neural Network Systems.



from this spectrum, using the peak-picking algorithm from the instrument workstation. The peakpicking threshold (in absorption units) was deliberately set very low, to include a number of false and suspect peaks in the resulting peak table. The peak table, and the spectral absorption (Y) values in the (X) range from 400 to 4000 cm^{-1} were then transferred to the microcomputer.

Initially, the input vector was chosen as a group of 17 data points centered about a "peak maximum" from the peak list. After some experience, it was apparent from the view provided by the peak confirmation program that a modification of the input vector was necessary. It was often impossible for the teacher to decide if the input pattern corresponded to an actual or a noise peak, when using only a few data points surrounding a test pattern (Fig. 4-3a,b). Further analysis of the human signal detection process showed that a "visual" comparison of a signal pattern with the form and amplitude of the spectral noise was being performed.

A portion of noise data from the spectrum was then added to each input vector. This greatly facilitated judgements on peak presence (Fig. 4-3c,d). The noise data are chosen by the teacher using a spectral cursor which may be interactively manipulated on the display. A vertical line in the left portion of the display output serves to visually separate the signal pattern (left side) from the noise reference (right side), and the circled data point indicates the peak maximum. The X-values assigned to the noise points are meaningless, and

Figure 4-3. Two Test Patterns, (a,b) Without a Noise Reference, and (c,d) With the Noise Reference. The Circled Point Indicates the Peak Position Being Evaluated.

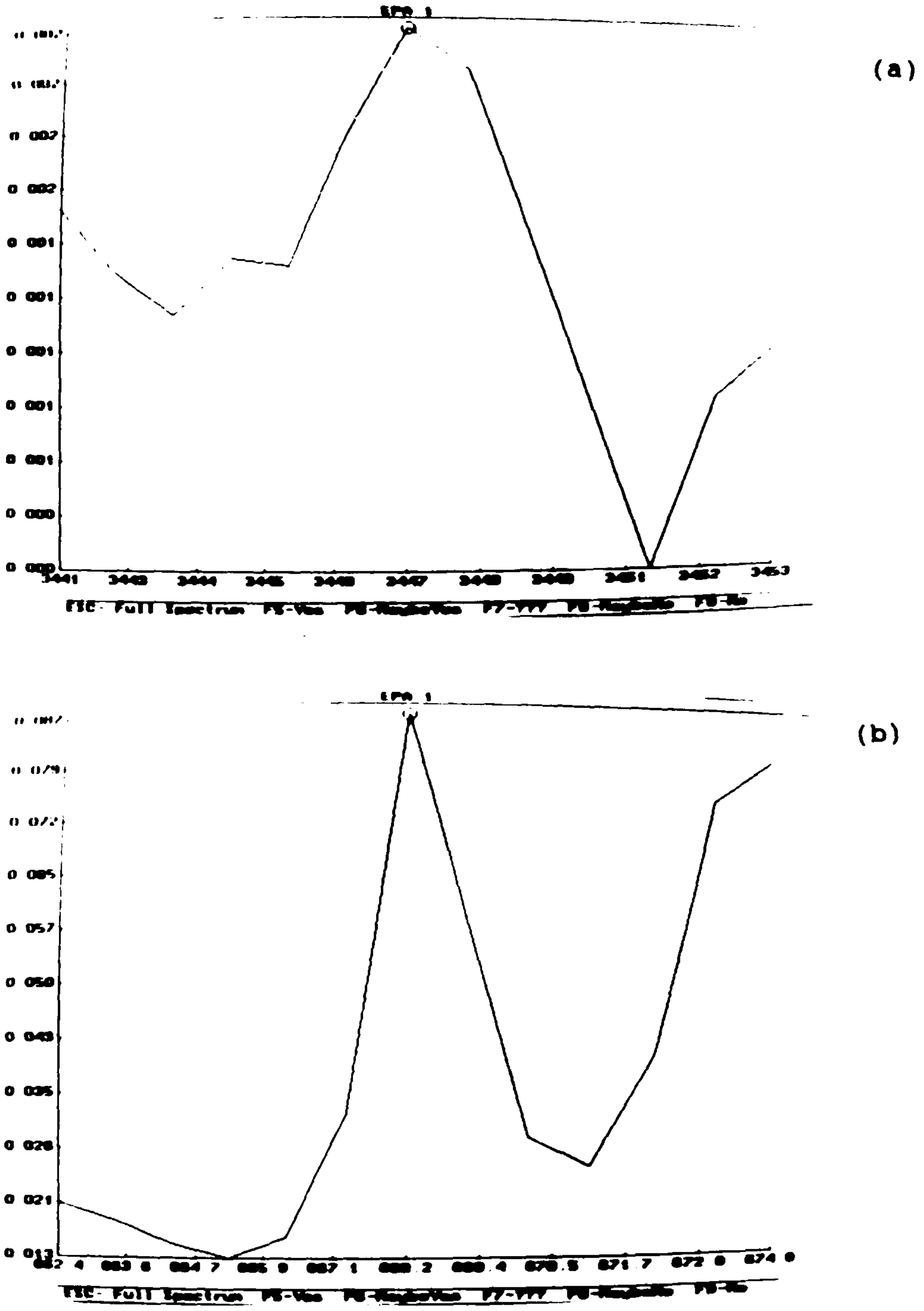
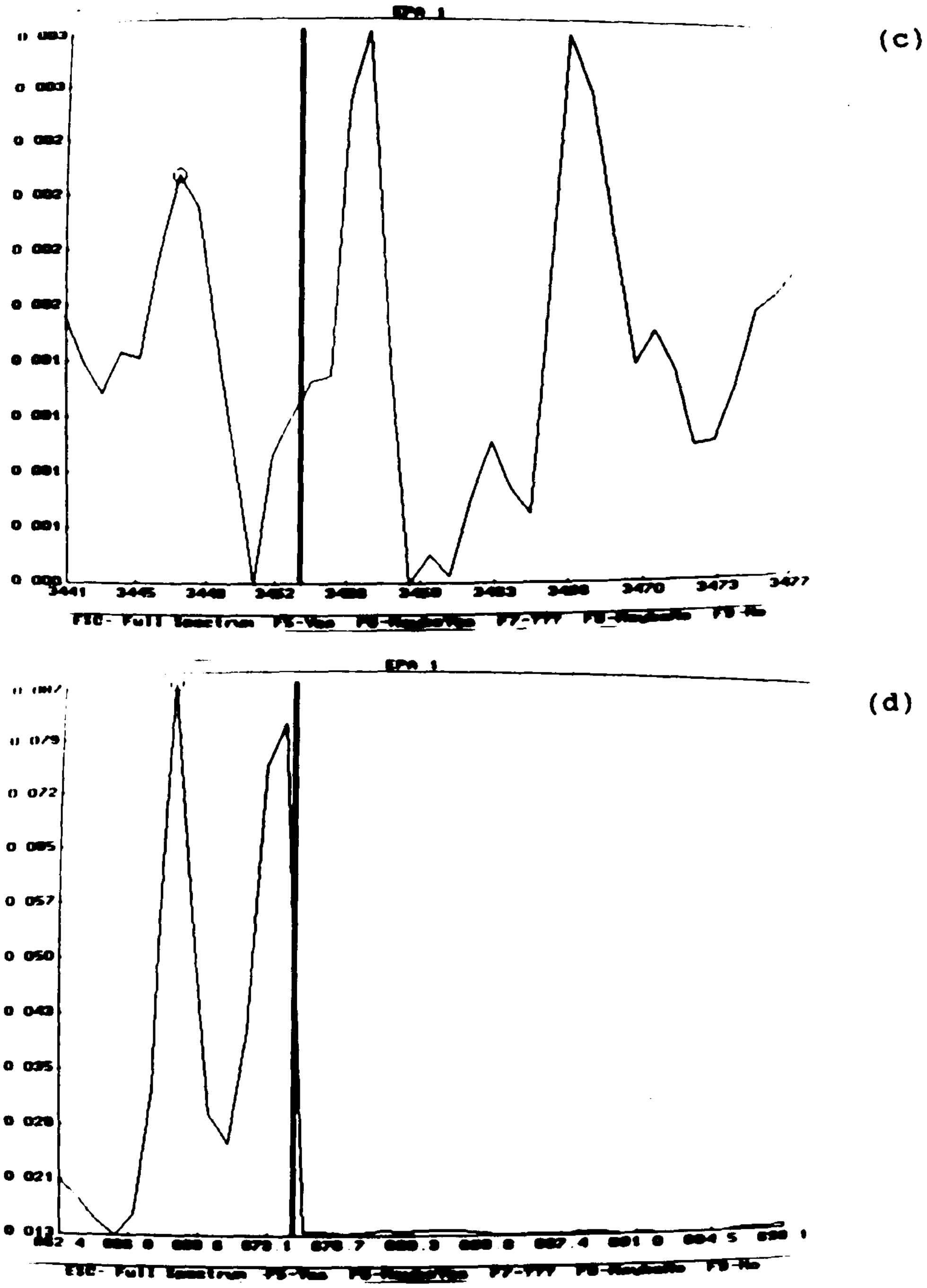


Figure 4-3. Two Test Patterns, (a,b) Without a Noise Reference, and (c,d) With the Noise Reference. The Noise Reference Provided for Comparison in Examples c and d are Presented to the Right of the Solid Line. The Circled Point Indicates the Peak Position Being Evaluated. (Continued from Previous Page).



are simply assigned in a manner to make them contiguous to the data pattern X-values, for plotting. Only the spectrum Y-values are presented to the network. The noise region was "wiggled" over a range of 20 data points at random, during training. This was done to prevent the network from attaching any significance to the form of a particular noise pattern.

Experiments with the noise-augmented input pattern did show dramatic improvements in learning rate, accuracy, and generalization, compared to the original pattern results. However, the stand-alone peakpicking program still produced some undesirable results. Two significant characteristics of the noise portion of the input pattern could be identified: the frequency and the amplitude. Consideration of the mathematics performed in the learning process showed that the network could not learn the frequency of a randomly varying signal (the noise appears at the frequency of the infrared sampling rate).

Since the characteristics learned during training cannot be directly controlled, it was decided to use the standard deviation of the noise data to augment the input pattern. This was done to force the network to learn the desired noise characteristic. A single value was then defined to represent the noise. The noise input value was taken as the minimum of the spectral pattern absorbance values plus 20 times the standard deviation of the noise:

$$X_{noise} = Y_{min} + 20\sigma_{noise} \qquad (7)$$

The factor of 20 was chosen empirically to increase the noise measure by a constant factor. The value of 20 was chosen because this placed the magnitude of the noise input at approximately the same magnitude as a typical signal. The purpose of multiplying by this factor is to insure that the network learning algorithm "recognizes" the importance of that value in determining the correct output (see equations 1 and 5 for justification). The backpropagation procedure corresponds to a gradient search of an error surface in weight space. "Noise" is observed on this surface which can misdirect the search procedure. Increasing the magnitude of an input results in a proportional increase in the gradient of the error surface along the corresponding weight dimension (59). It was presumed that this action would help to overcome the effect of local fluctuations in the error surface due to noise.

The next step was to test if the network shared the same improvement in reliability of signal detection as the human, with the addition of a noise reference to the signal. After some experimentation, it was decided that 13 data points can describe most of the bands in the test data, and that 25 noise points is sufficient to provide an adequate description of the noise. The network was trained twice using the same parameters: once using 13 data points surrounding the test patterns, and again using 13 data points plus the noise input derived according to equation 7 from the 25 noise points. These noise points are the same as those included in the visual pattern confirmation process. Two noise regions were included from each spectrum, to account for the variation in noise with wavelength. The region closest to the center of a given test pattern was chosen as the reference by the training program. The training patterns were presented repetitively and in random order to the network to minimize the possibility of cycling of the weight vector during training.

The values assigned for the correct output for each test pattern were assigned to one of five discrete values: 0.00, 0.25, 0.50, 0.75, and 1.00 with the peak confirmation program. These values are intended to reflect the teacher's belief that a given pattern actually represented a spectral peak, rather than noise.

While values of 0.00 and 1.00 only could have been used, this was not done, owing to the following: Even with the noise reference, it is often difficult to make a simple yes/no decision on peak presence in the test patterns. Keeping in mind that the network performs a mathematical mapping of the input vector to the output vector, it is difficult to reconcile assigning a 1.00 to a questionable "yes", and a 0.00 to a questionable "no". These types of patterns are normally more similar to each other than they are to the ideal yes and no cases. It was felt that providing a range of output values for intermediate cases would aid the network in providing a continuous mapping for the full range of input patterns. Judgements finer than five levels of belief, however, were deemed excessive.

A "training epoch" was defined as 1000 successive presentation/correction cycles. After each epoch, the complete set of 132 training patterns was presented to the network, each in turn, performing only the forward-feeding of activation. The actual output was compared to the desired output for each training pattern, and the mean absolute difference was calculated. The descent of the error function was not a smooth one (Fig. 4-4). This may be due to a combination of factors, including a complex, error surface, noise on the surface, and the inertia of the search dynamics caused by the momentum term (eqn. 6). Due to the spikes superimposed on the error descent during training, training was allowed to proceed for 500 epochs, and was then halted after the mean error decreased for three successive epochs.

After training was accomplished, two sets of test data were then presented to each network. The first was obtained from a vapor phase mixture of vinyl chloride, trichloroethane, toluene, tetrachloroethylene and chlorobenzene at concentrations of approximately 5ppm each. The second test spectrum was derived from a vapor phase methyl isobutyl ketone sample, also at approximately 5ppm. In all, 189 test patterns were presented to the networks. The distribution of "correct" scores that were defined with the peak confirmation program for the training and test patterns is shown in Figure 4-5.

A network with nine hidden layer nodes was trained with and without an added noise reference, and the mean absolute difference between the actual and desired network output was

Figure 4-4. Plot of the Mean Absolute Error Difference Between the "Correct" and the Actual Network Output vs the Training "Epoch" Number for a Network with 9 Hidden Nodes. A Training Epoch was Defined as a Period of 1000 Learning Cycles

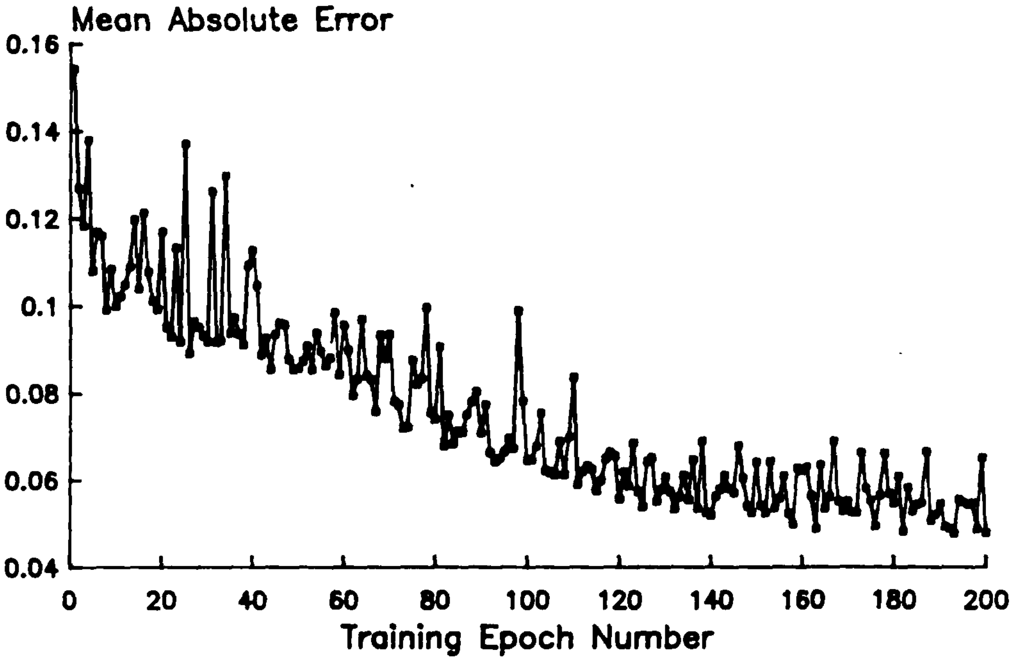
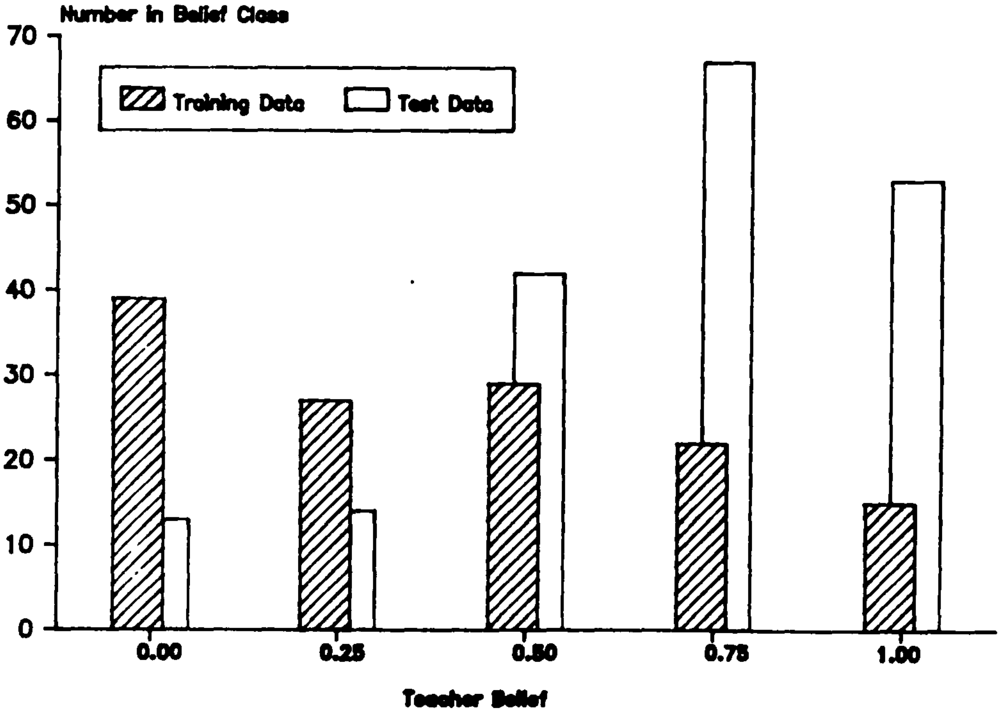


Figure 4-5. Histogram Showing the Distribution of Belief Classes for the Spectral Patterns Used to Train and Test the Networks. The Belief Classes Ranged from a Low of 0.0, Indicating the Pattern was Definitely not a Peak, to a High of 1.0, Indicating the Pattern Definitely Represented a Peak. The Patterns were Classified by the Author.



tabulated for the training and test patterns.

The results derived without the added noise reference will be considered first. On the training data, the mean absolute difference between actual and desired scores was 0.038. With nine hidden layer nodes, the network was able to map very closely to the training data. It may at first seem then, that a useful system has been derived. In fact, however, with a sufficient number of hidden nodes, any arbitrary mapping of training inputs to outputs may be performed. Whether or not any useful mapping has been learned, however, is determined by performance on the unknown, or test data.

On the test data, the mean absolute difference was 0.357, which is very close to the mean difference of 0.33 between two values chosen at random from the interval 0 to 1. This indicates that nothing generally useful was learned from the representation without a noise reference.

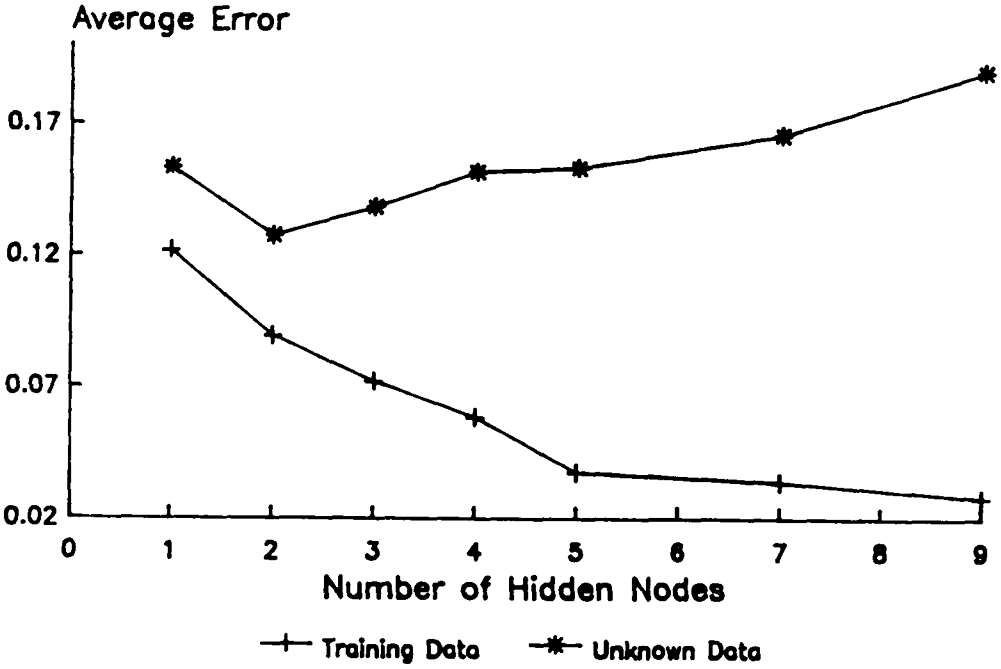
The results for the network trained with a noise reference were slightly better for the training data (0.028), and dramatically better for the test data (0.190). These results indicate that an effective mathematical mapping is possible when a noise reference is present in the input pattern, and seem to confirm the similarity between the human and network pattern recognition processes.

The next factor investigated was the influence of the number of hidden nodes on network performance. The same training and test data were used as in the study described above.

Networks with one to nine hidden layer nodes were evaluated. The solution to the error minimization derived during training is in general not unique. For this reason, each training procedure was repeated in triplicate, starting from a different set of random weights, and the results averaged. The mean absolute difference between the desired and actual network outputs was then plotted for both the training and test data (Fig. 4-6).

The error in mapping the training values decreased with the number of hidden nodes. The larger the number of surfaces we can position in the pattern space, the more closely we can trace the topology of the training pattern group shapes. The error in emulating the human judgements on the test patterns, however, passed through a minimum at 2 nodes (Fig. 4-6). This can be explained in the following way: the training set contains only a sample of the entire set of all possible patterns. Therefore, many topological features of a training set group in the pattern space will not be representative of the true shape of the group, but just the shape of the example group. As we increase the number of hidden nodes, we will more closely follow the topology of the training group clusters on optimization. This results in a greater "recall accuracy" for the training set. Above an optimum level, however, adding more nodes will result in tracing the training patterns too closely, thereby falsely excluding some unknown (test) patterns from the group during classification of

Figure 4-6. The Mean Absolute Difference Between the Desired and Actual Outputs for the Training and Test Data, Plotted as a Function of the Number of Hidden Nodes.



unknowns, or "generalization". Below the optimum level of nodes, we cannot accurately reproduce the true shape of the pattern distributions. This behavior is analogous to the familiar problem of overfitting a regression equation to a set of experimental data points. While increasing the order of the equation will continue to improve the correlation coefficient for the model, it will also inevitably exceed the true functional complexity of the data.

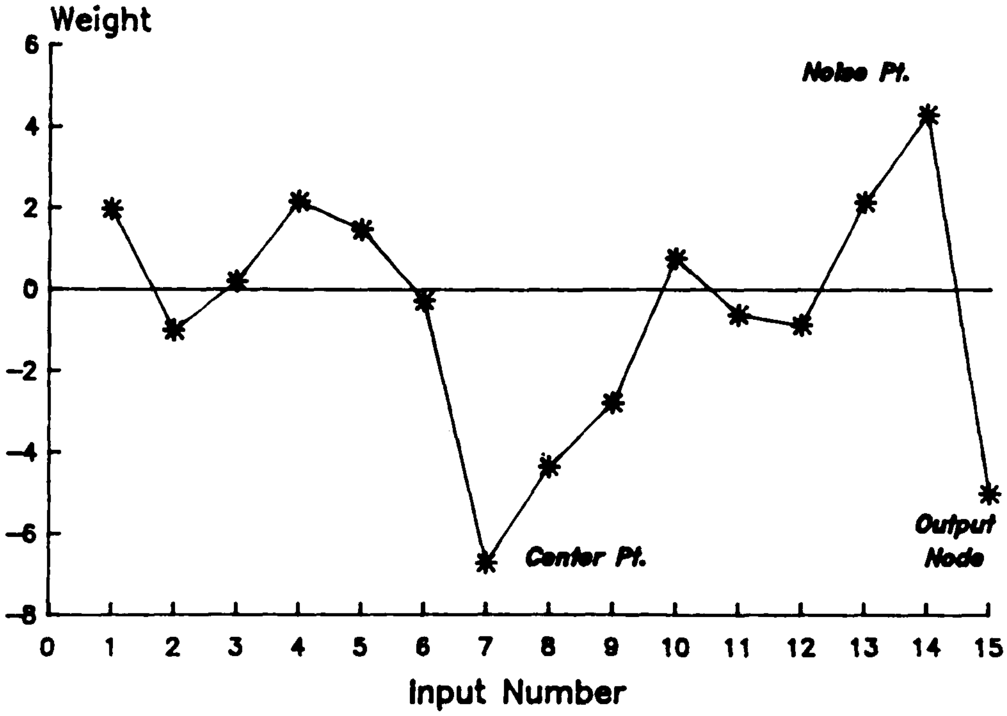
All of the accuracy values for the test data must be interpreted with the following in mind: the network is being trained to reproduce the judgement of a chemist in fuzzy decisions on signal presence. It was often difficult to decide whether a pattern should be a 0.00 or a 0.25, a 0.50 or a 0.75, and so on. Therefore, even if the network perfectly emulates the judgement of the chemist, it should be expected that "errors" on the order of 0.25 will occasionally occur, corresponding to the variation in judgement of the teacher.

An objective measure of peak presence, such as the output of a cross-correlation algorithm could alternatively be suggested to train the network. This would then provide a reproducible and continuously variable range of outputs. However, this would also produce no useful results. If a mathematical function were used to provide the teaching values, we would at best be providing a difficult means to emulate an existing mathematical function. The implicit assumption made here is that humans are better at recognizing signals in non-ideal data than mathematical algorithms, so a human judgement must therefore be used to train the network.

In order to learn more about the learned network mapping function, the weighting function of the simplest network was investigated. This network contains only one hidden node between the inputs and the output node. The weighting procedure (eqn. 1) is analogous to that applied during signal cross-correlation (87), an alternative signal detection procedure. Since the transfer function is continuous and increasing (eqn. 2), it was expected that the learned weighting function should be similar in appearence to the optimum cross-correlation function for the set of data inputs. The optimum cross-correlation function, while unknown, should be dominated by a peak-shaped component. The noise input was presumed to take on a negative weight, since neuron activity should be inhibited by an increased noise measure. The actual network input weights produced a mirror image of the function proposed above (Fig. 4-7).

While at first confusing, the inversion is readily explained by the last input weight shown, which corresponds to the weight applied to the connection between the hidden node and the output node. The negative value of this weight effectively inverts the mapping performed by the hidden node, hence the mirror image. The network has learned what we would expect, but without the familiar convention of image orientation. The analysis of the weights for the larger networks would of course be more complex. However, this example provides increased insight into and confidence in

Figure 4-7. Input Weights to the Hidden (1-14) and Output (15) Nodes Learned When Using Only One Hidden Node. "Center Pt." Refers to the Midpoint of the Test Pattern Window, "Noise Pt." to the Weight Given to the Noise Reference Input, and "Output Node" to the Connection Between the Hidden Node and the Output Node.



these networks.

On the basis of the results of these studies, a standalone peakpicking program was written to perform autonomous evaluation of a test spectrum. This program can use any network architecture and connection weights created with the training/evaluation program. A data window equal in width to the number of inputs for the network is incrementally moved down the spectrum data points, one at a time. A simple prefilter is employed to reduce the volume of data processed in the network: The data in the window is only passed through the network if the centerpoint of the window has a larger amplitude than both its neighbors. At the start of the program, the test spectrum is displayed, and the user is asked to position an arrow shaped cursor at the location of two noise regions to be used for the reference portion of the input patterns. The user is then asked to choose a threshold value for the output of the network mapping function. Only test patterns producing a network output greater than this value are then presented. The variable threshold allows the user to select the degree of conservatism to be employed in generating the peak list.

The peakpicking program may be operated in either an interactive mode, or a fast file dump mode. The interactive mode of operation causes the program to pause at each pattern with an output greater than the threshold. This pattern is then graphically displayed on the screen, along with the peak position, intensity, and network output value. The file dump mode generates an ASCII file containing the peak list. The peak list consists of a list of records, with each record containing peak position, intensity, and network output values.

The network calculations are efficient. When the system is prevented from pausing to display matching patterns, the region from 600 to 4000 cm^{-1} (ca. 3400 data points) is processed in about 1 second. The interactive mode produces some intriguing results. Five sample patterns obtained from the test spectrum of methyl isobutyl ketone are shown in Figure 8a-e, along with the network output values. These plots were photographed from the microcomputer display with the program running in the interactive mode. It is clear from these results that the network is performing a useful function. Since the network is an adaptive feedback learning system, further training will be performed in the future, using the results of this first generation system. This is expected to yield improved network performancc. The use of alternative network architectures, perhaps with an additional hidden layer, and/or a restricted connectivity, will also be explored.

Conclusion:

The feasibilty of exploiting neural network technology to recognize peak-shaped signals in analytical data has been evaluated. While the system described has been developed to interpret infrared spectral data, peak detection has

Figure 4-8. Sample Results Produced by the Stand-Alone Peakpicking Program on the Test Data, Using Three Hidden Nodes. The Network Output is Given in the Block in the Upper Right Hand Corner for each Evaluation. The Network Outputs for these Patterns Ranged from a Low of 0.208 for Figure 8a, Indicating that the Pattern is Probably not a Peak, to a High of 0.959 for Figure 8e, Indicating that the Pattern Probably is a Peak, on the Basis of the Data Used to Train the Network.

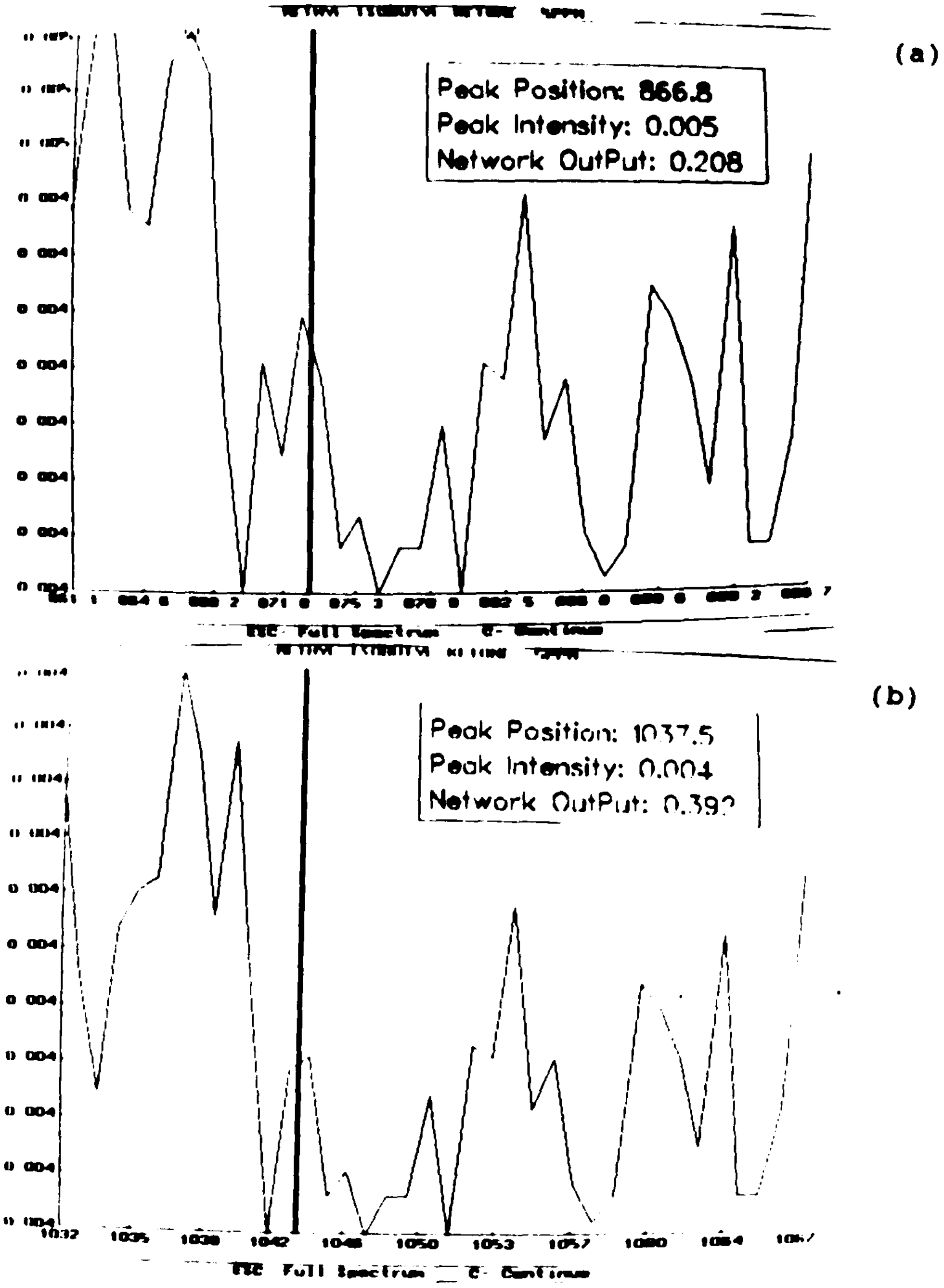


Figure 4-8. Sample Results Produced by the Stand-Alone Peakpicking Program on the Test Data, Using Three Hidden Nodes. (Continued from Previous Page).

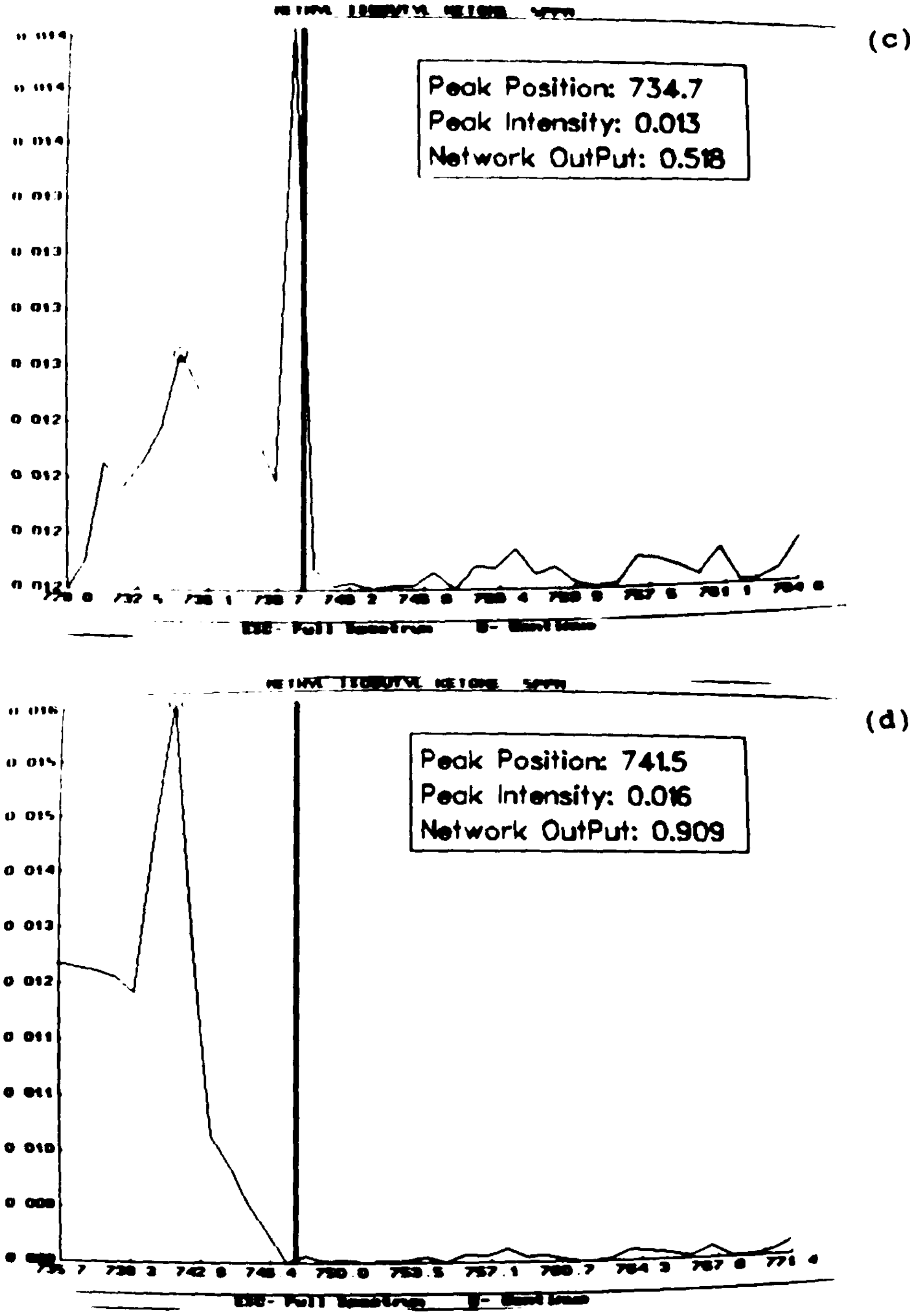
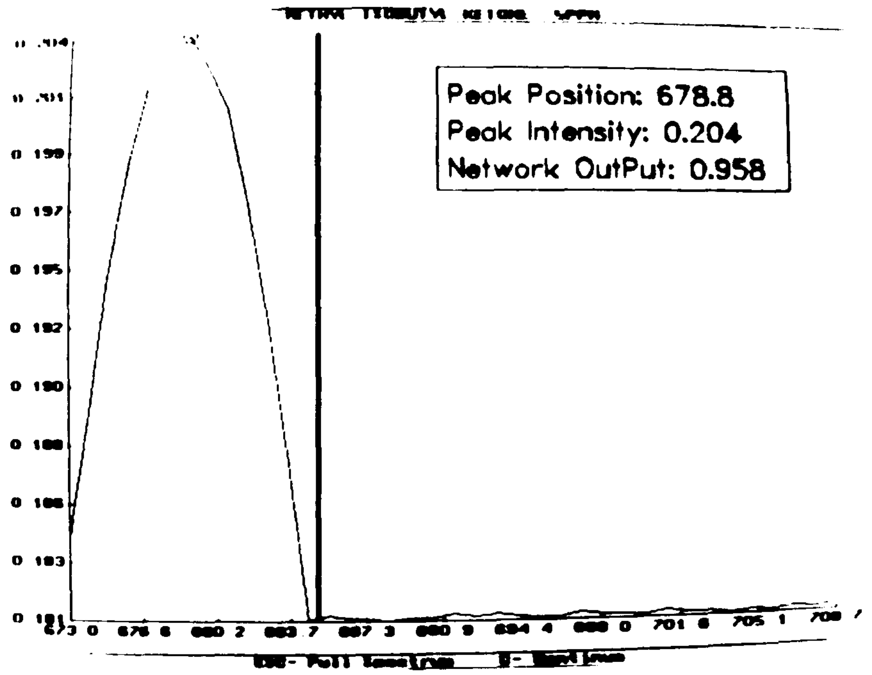


Figure 4-8. Sample Results Produced by the Stand-Alone Peakpicking Program on the Test Data, Using Three Hidden Nodes. (Continued from Previous Page).



implications in every chemical application where the recognition of peak-shaped signals in analytical data is important. Chemical applications could include virtually all spectroscopic and chromatographic methods, as well as flow injection analysis and the scanning electrochemical methods.

The idea that the human and artificial neural network would perform the signal detection task best under similar conditions was examined. The incorporation of a noise reference was found to aid both the human and the network signal detection processes. Other modifications of the input pattern and the network connectivity must be explored, however it is clear that there is a great deal of potential in applying artificial neural networks to perform signal recognition for chemical data.

CHAPTER 5

COMPUTER ASSISTED INFRARED IDENTIFICATION OF VAPOR-PHASE MIXTURE COMPONENTS

Introduction:

Infrared analysis of organic vapors has many potential applications, including on-site measurement of toxic compounds at hazardous waste sites, in the workplace, and analysis of unresolved effluents from GC-FTIR experiments. Efforts at computer-assisted interpretation of infrared spectra of mixtures has been largely directed at condensed phase analysis. Quantitative analysis of vapor phase mixtures has recently been explored in the Fourier domain using factor analysis (88), and in the spectral domain using least squares fitting (LSF) techniques (89,90). Qualitative identification of vapor phase mixture components has been reported using Iterative Least Squares Fitting Techniques (ILSF) (91). An intriguing possibility is the use of a knowledge based system to reduce the number of components fed into an LSF quantitation program. This should greatly reduce the workload required for the LSF calculations, and provide more accurate quantitative results, as well.

The IRBASE/MIXIR system is a knowledge based system developed to identify the likely components of mixtures from infrared spectral data. The original work concerned the development of a compound-specific automated rule generator (36), and a knowledge based system to manipulate these rules (37). The experimental test data were condensed phase mixtures; however, most of the interpretation algorithms and logic are applicable to vapor phase samples as well. A previous attempt at adapting a condensed phase expert system for vapor phase analysis has been made (92). It was concluded in that research that a peak based expert system was "not appropriate" to vapor phase analysis. It was recognized from the outset, therefore, that these data would present a difficult challlenge. The goal of this research was to attempt to define the limits of knowledge based systems for interpreting peak based information from infrared spectra of mixtures.

While adapting these programs for vapor phase analysis, many improvements have been made which can also be used for condensed phase analysis. This paper describes these modifications and enhancements, which represent another phase in the continuing evolution of the MIXIR/IRBASE system.

Experimental:

The vapor phase infrared spectra used for this study were acquired at a nominal 0.3 cm^{-1} resolution, and were transformed to a 2 cm^{-1} resolution representation for this work. The raw spectral data were obtained from Xiao Hong-kui at the University of Michigan. There were three groups of mixture spectra, consisting of mixtures with component concentrations of approximately 50, 5, and 2 parts per million (ppm). The 50 ppm mixtures, designated TANK A through TANK F, were obtained from undiluted 50 ppm reference standard gases, and have been the subject of previous quantitative LSF (90) and qualitative ILSF (91) studies. The 5 ppm mixtures, designated TANK 1 and TANK 3, and the 2 ppm mixtures, designated EPA 1 through EPA 3, were obtained by dilution with "zero air" containing very low levels of carbon dioxide and water vapor. These low concentration mixtures were prepared by the US EPA Atmospheric Research and Exposure Assessment Laboratory, Research Triangle Park, NC. The 2 ppm mixture data have been the subject of a previous quantitative ILSF study (90).

The spectral transformation and derivation of the peak tables were carried out on a Nicolet 620 FT-IR workstation (Nicolet Analytical Instruments, Madison, WI). The derived spectral data were subsequently uploaded to a DEC 8820 superminicomputer (Digital Equipment Corp., Maynard, MA), using Nicolet VAXtran software.

The IRBASE and MIXIR systems were both written entirely in standard FORTRAN 77 and therefore run both on VAX and IBM-PC hardware. The majority of the present program development and testing, however, were carried out on the Digital computer. Approximately 1500 lines of FORTRAN code were added or modified during this vapor phase work.

Data Pre-Processing:

The 50 ppm vapor phase reference spectra were subjected to a 5 point Savitsky-Golay smoothing algorithm (82). The mixture spectra received a 13 point smooth. The smoothing was performed to reduce noise-induced false peaks in the peak tables. The degree of smoothing is always a compromise between removing noise-induced false peaks and removing small, poorly resolved signal peaks. Another value of the smoothing window size may have produced better results on some spectra, however the 13 point window used in the mixtures was found to be a good compromise in most cases.

The mixture spectra were plotted in a "stacked format" (Figs. 5-1 to 5-4). Examination of these plots shows that the signal to noise ratio decreased with decreasing component concentration, as expected. In addition to instrumental noise, "chemical noise" presented difficulties.

Correction for background absorptions, primarily of water and carbon dioxide, is often incomplete. A variation of the amount of water and/or carbon dioxide vapor in the optical path between acquiring the background spectrum and the sample spectrum will cause an incomplete ratio correction for the background. In addition, variation in the experimental conditions, among them, temperature and pressure, can cause band shifting in the background spectrum, which further hampers efforts to correct for background absorptions. It is clear that positive background contributions, such as the the carbon dioxide features dominating the 2350 cm^{-1} region of the 2 ppm spectra (Fig. 5-4), will interfere with sample peak signal detection. Even negative background contributions, as

Figure 5-1. A "Stacked Plot" of the 50 ppm Mixture Spectra (a) TANK A, (b) TANK B, and (c) TANK C.

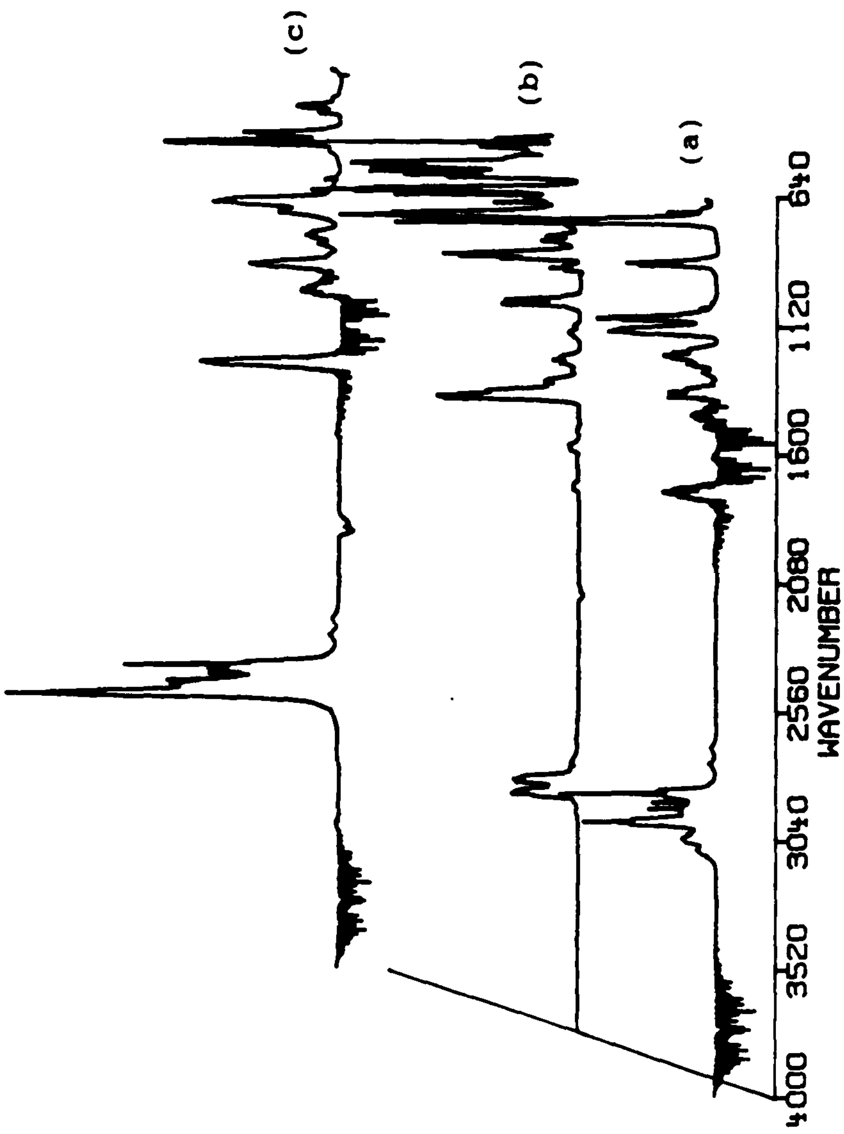


Figure 5-2. A "Stacked Plot" of the 50 ppm Mixture Spectra (a) TANK D, (b) TANK E, and (c) TANK F.

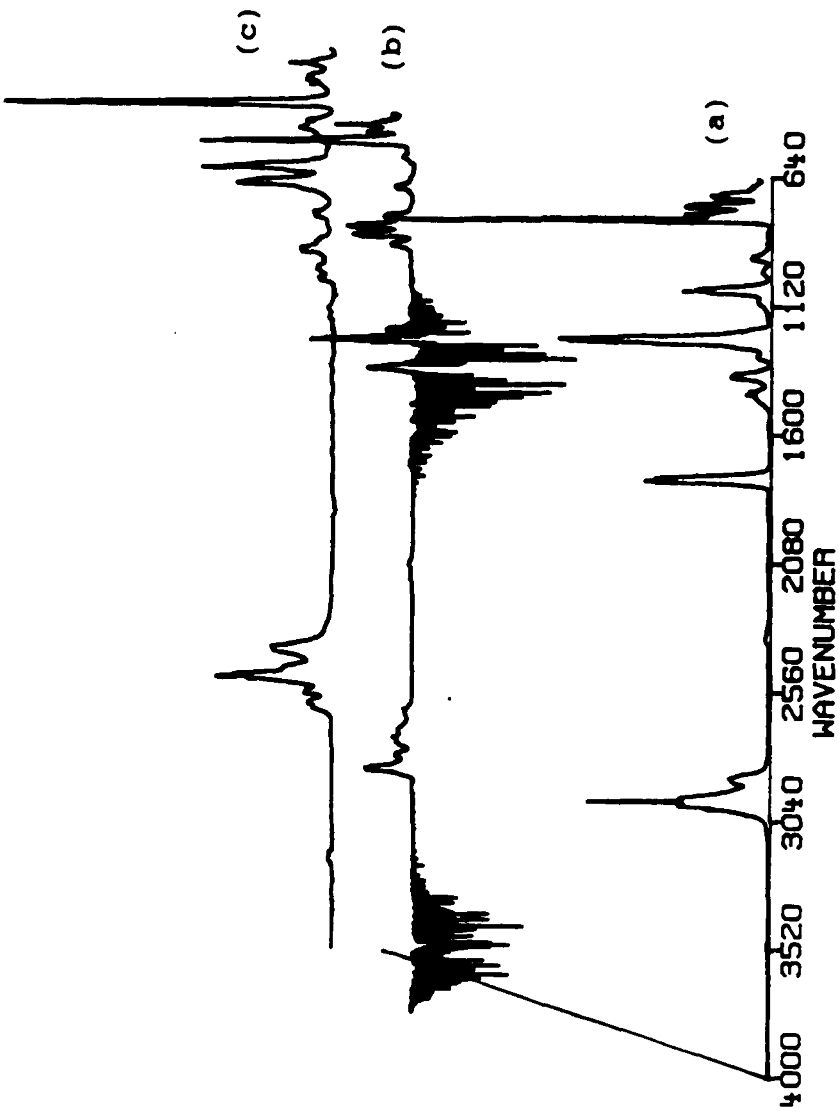


Figure 5-3. A "Stacked Plot" of the 5 ppm Mixture Spectra (a) TANK 1 and (b) TANK 3.

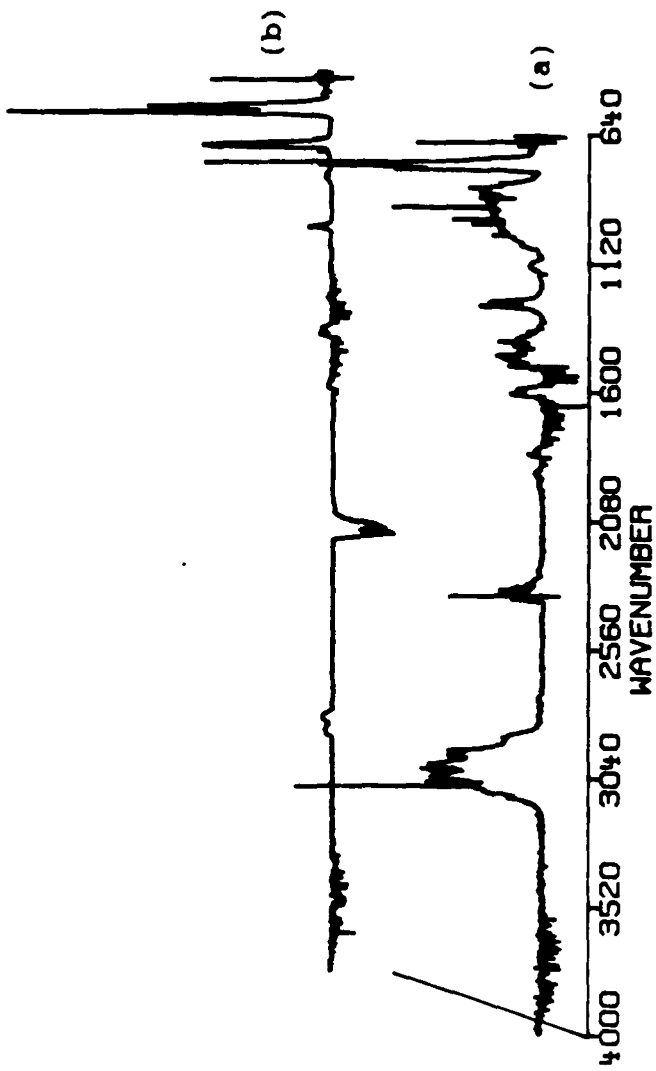
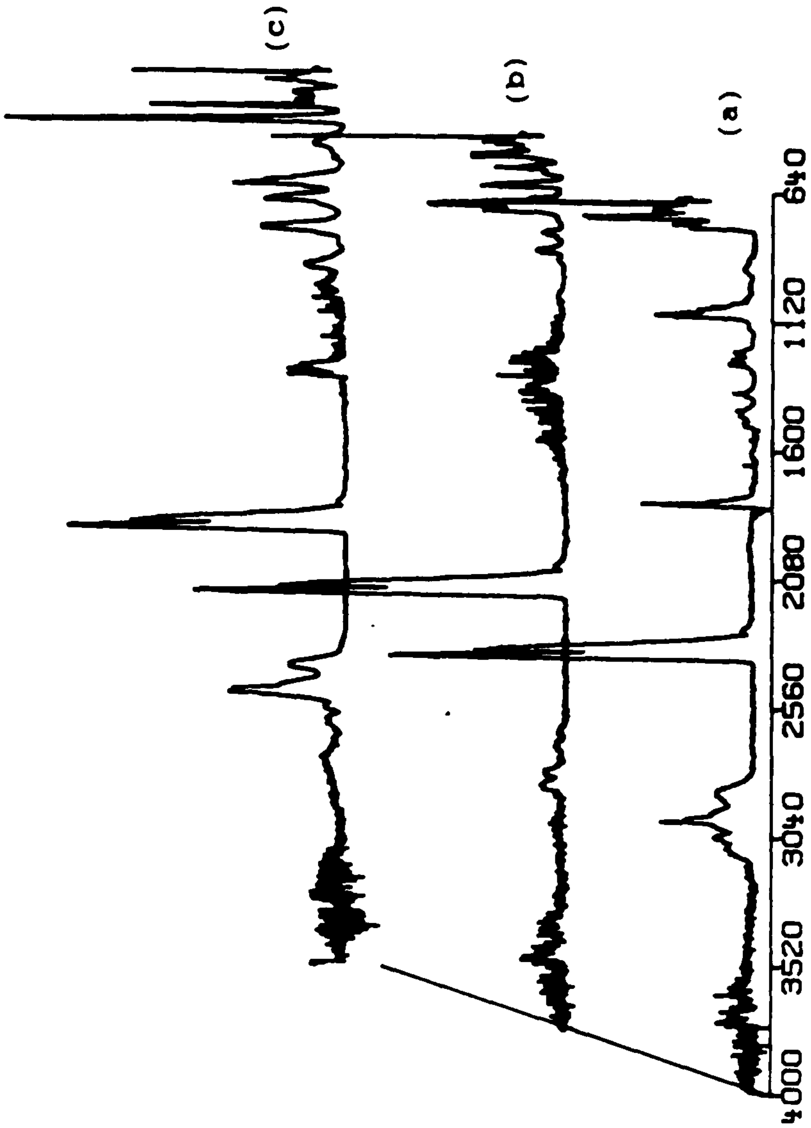


Figure 5-4. A "Stacked Plot" of the 2 ppm Mixture Spectra (a) EPA 1, (b) EPA 2, and (c) EPA 3.



in the (inverted) water features in the 1300-1900 cm^{-1} regions of TANK A, TANK C (Fig. 5-1) and TANK E (Fig. 5-2) present major problems for signal detection. The inverted water features superimposed on the sample features create false peak maxima in the spectrum at the valley points of the true water absorptions. Another form of chemical noise in the mixture spectra is the overlap of spectral features from different mixture components, which often render individual sample features undetectable.

Since isolated molecules should exhibit little change in molar absorptivities, and vapor phase spectra can be easily taken with reproducible optical pathlengths, real-valued absorption intensities were included in the knowledge base. Previously, these intensities were preprocessed to integral values ranging from 0 to 20, normalized against the most intense peak in the spectrum.

Peak widths were coarsely classified into ranges emperically set at 0 to 15, 15 to 35, 35 to 75, and greater than 75 cm^{-1}, corresponding to very sharp, sharp, average, and broad band widths, respectively. The actual width values, in wavenumbers, were determined by the instrument peakpicking algorithm, and integer codes corresponding to the four classes described above were then assigned by a utility program developed to create MIXIR format peak files.

It was determined from early testing that the quality of the results was greatly influenced by the ability to include small mixture bands in the MIXIR input. A program was written to allow the user to interactively specify up to 50 points on the spectrum baseline by manipulating a spectral cursor. A linear interpolation was performed between these baseline points, and the interpolated baseline subtracted from the spectrum. The adjusted spectrum allows a lower peakpicking threshold to be set, thereby providing access to the smaller spectral features. This program was used on any of the reference and mixture spectra which originally had ill-behaved baselines.

Program Description:

The IRBASE and MIXIR systems have been adequately described elsewhere (36,37), and so will be only briefly summarized here. IRBASE is a knowledge based system which creates a condensed-phase compound specific knowledge base for use by MIXIR. Information on functional groups present in a compound to be included in the knowledge base is used to predict the likely range of shifts in mixtures of the peak parameters: position, intensity and width. A subset of the resulting band descriptions is chosen for each compound, based on the program's judgement of the likelihood of observing the feature in a spectrum of a mixture containing that compound.

MIXIR is an adaptive knowledge based system which reports the likelihood of presence or absence of reference compounds in an unknown mixture. A flexible set of interpretation routines is available for the user to manipulate the data, providing various logical algorithms. The interpretation is approached in a dynamic fashion, making use of information gained during the interpretation process.

Modifications to the MIXIR Knowledge Base:

Since vapor phase spectra normally show features of isolated molecules, the tailoring of band position windows to functional group origin is unnecessary. Instead, band position windows were scaled to the width of the peaks in the pure compound spectra, to account for the greater uncertainty in determining the location of band maxima for broader bands. Initial values were arrived at empirically by considering the spectral resolution employed, (2 cm^{-1}) and the magnitude of likely errors in determining the positions of the various band maxima. Subsequent testing allowed refinement of these windows. The final position window settings were 2,3,10, and 25 cm^{-1} above and below the band position in the reference spectrum, for very sharp, sharp, average, and broad bands, respectively.

It has been determined that reduction of the original spectral features in the reference spectra prevents dilution of the significance of important spectral features during an interpretation, and improves execution speed. However, no reduction was performed at the stage of creating the knowledge base, as had been done in IRBASE. It was decided that the optimum description should be created dynamically at runtime, from the entire set of spectral features for a given compound, using what is known about the sample matrix. Delaying decision making as long as possible allows the MIXIR system the maximum ability to adaptively interpret an unknown sample, using information gained from the user, and determined by MIXIR throughout the analysis. For example, MIXIR can adaptively compensate for changes in background absorptions due to H_2O and CO_2, by ignoring unknown features in these regions (important for normalization of intensities), and eliminating queries which refer to these regions which may contain strong interfering absorptions.

MIXIR was modified to use an estimate of the relative strength of compound absorptions in the unknown matrix to determine which features from the reference set could be reasonably expected to appear in the particular mixture under study. Elimination of preliminary feature reduction allows MIXIR a larger set of features to select from when performing an "either/or" interpretation, used to discriminate between two structurally similar compounds (37). Although this type of approach is costly in terms of runtime processing requirements, it also provides the maximum use of the available information. This provides a higher degree of system "intelligence", and should therefore provide more accurate spectral analyses.

The knowledge base server routines now calculate the integral normalized intensity values from the real valued intensities at runtime. In the future, absolute intensities will allow the interpreter to make coarse quantitation estimates. These results will be useful to the interpreter as well as to the user. The interpreter may later use estimates of large component concentrations to determine what spectral regions should be avoided in making subsequent peak queries.

Separate descriptions are no longer written for the major and minor component mixture features. This workload has been shifted to the MIXIR program, and is described below.

Modifications to the MIXIR System:

The regions where the sample matrix may provide strong interfering absorptions have been noted in the program for vapor phase as follows:

H_2O: 1200-2100cm^{-1}, 3200-4000

CO_2: 2225-2400

The presence of these spectral interferences is determined by queries to the user. This information is then used by the knowledge base server routines. Spectral descriptions provided by these routines will not contain any band queries in an interfering region. Normalized band intensities will exclude reference to bands in this region, preventing, for example, normalization of sample bands against a strong CO_2 absorption, which would otherwise provide inappropriate intensity values.

As mentioned above, the full set of band descriptions is available to the knowledge base server routines at runtime. An option has been added which provides dynamic selection of bands for queries, based on the results from a spectral pre-scan. This procedure is as follows: the query server routines accept parameters which specify the minimum reference peak intensity to be accepted for a query set, the maximum number of queries requested, and the intensity window selection scheme to be used. To determine the minimum reference intensity desired, a band query set is requested with a maximum of twenty queries, and with a "null" intensity window scheme, i.e., all intensities are passed by the query. The null scheme is used here since no information exists at this stage to allow adaptive settings. The most intense unknown band which can be matched to these queries is then determined (I_{max}), and the minimum reference intensity desired is taken as

$$I_{min} = 20.0/I_{max} \qquad (1)$$

The MIXIR integral intensities are normalized to range from 0 to 20. Equation (1) therefore allows MIXIR to discard reference peaks which can be expected to correspond to bands with insufficient intensity in the mixture under study to be detected.

In addition to the null intensity window scheme mentioned above, two other intensity window schemes are provided. The first is identical to that produced in the condensed phase rule generation program, IRBASE, and will hereafter be termed the "default scheme". This method sets the upper intensity limit to a value equal to the normalized reference intensity plus 4, or 20, whichever is less. The lower intensity limit is set to a value of 1. This scheme, like the null scheme, is static- it does not use any information from the unknown mixture to adaptively set limits.

The third scheme is a dynamic approach. First, the average ratio of matched unknown peaks to reference peaks is determined as follows: The server routine is prompted for a query set with a maximum of 12 queries, a minimum reference intensity of 1, and the null intensity window scheme. A weighted average of the ratio of normalized unknown peak intensities to matching reference peak intensities is calculated (R_{avg}). The weighting factor used is the reference peak intensity. This factor was chosen because larger reference bands are more likely to be matched. Subsequent calls to the server routines requesting dynamic intensity windows produce the following intensity limits:

$$I_{HI} = (R_{avg} * I_o) + 4 \qquad (2)$$

or 20, whichever is larger, and

$$I_{LO} = (R_{avg} * I_o) - 4 \qquad (3)$$

or 1, whichever is smaller.

This procedure allows MIXIR to set intensity windows to levels which reflect the average intensity of bands which appear to match the queries for a particular compound. It is known that even for compounds which are present in an unknown mixture, spectral bands which are primarily due to other components in the mixture will often be "matched", causing the intensity ratios described above to vary across a query set. The underlying assumption, however, is that this variation will be smaller on the average than that observed for a compound which is absent from the mixture. It is expected that this behavior would provide more frequent band rejection for compounds which are absent from the mixture.

An automated peak justification option was added to prevent false negative results. Similar to the condensed phase MIXIR peak justification, this procedure checks for any mixture bands with an intensity of 4 or greater which cannot be attributed to compounds with scores greater than or equal to 0.20. The compound with the highest score which can explain this feature is determined, and its score is set to 0.20, since it can be assumed that one of the compounds with a score less than 0.20 must then be responsible.

Results and Discussion:

A knowledge base consisting of the spectral descriptions derived from the spectra of 40 vapor phase compounds of toxicological significance was generated using a utility program written for this purpose. Many of the compounds are very similar in structure, and so have similar spectral features. Three sets of vapor phase mixture data were then presented to MIXIR. The first set to be discussed was composed of six mixtures at approximately 50 ppm concentration for each component. The dataset consisted of four 5-component mixtures, one 2-component mixture, and one 6-component mixture (Table 5-1).

The entire set of mixture spectra was interpreted using MIXIR with different combinations of the optional procedures. The results were summarized in two ways: (a) The number of false positive results and false negative results at each of

Table 5-1. The 50 ppm Vapor Phase Mixture Constituents and Concentrations.

Mixture	Components	Concentration*
TANK A	Toluene	46.8 ppm
	1,1,1-Trichloroethane	47.5
	1,4-Dioxane	42.8
	Acetone	50.1
	1,2-Dichloroethane	47.7
TANK B	Vinyl Chloride	49.9
	Benzene	49.9
	Methylene Chloride	50.2
	1,1-Dichloroethene	46.8
	Trichloroethylene	53.2
TANK C	2-Butanone	46.1
	n-Hexane	58.9
	4-Methyl-2-Pentanone	48.7
	Perchloroethylene	53.1
	1,4-Dioxane	48.1
TANK D	Cyclopentane	48.6
	Ethyl Acetate	49.9
	1,1-Dichloroethane	49.0
	1,1,2-Trichloroethane	51.1
	Carbon Tetrachloride	50.2
TANK E	o-Chlorotoluene	26.1
	Chlorobenzene	24.4
TANK F	Isopropanol	48.5
	Ethyl Ether	48.7
	3-Chloropropene	49.1
	Styrene	55.1
	Ethylbenzene	50.8
	Freon-11	51.1

*Analyzed by GC by Scott Specialty Gases.

ten expectation thresholds ranging from 0.40 to -0.50 was tabulated and (b) the average score of the compounds which were present in the mixtures and absent from the mixtures was tabulated.

An early version of the vapor phase MIXIR system was first obtained by creating a program which produced IRBASE formatted descriptions, with position windows scaled to peak width as described above, and the "null" intensity windows. The position windows were set at 3,5,10, and 25 cm^{-1} about the position of the reference peaks. At this stage MIXIR itself was only modified to handle the larger peak tables of vapor phase spectra. The knowledge base used at this stage was a subset of the final knowledge base, containing 41 compounds. The interpreter options used were "Essential Peak Checking", "Reduced Peak Set Checking", and the "Extended Scoring System". These options have been previously described in detail elsewhere (37), and were used in producing all of the results discussed in this work. The highest percentage of correct decisions at a given threshold was 91.3%, at thresholds of 0.30 and 0.20. These results, however, were obtained with an unacceptable number of false negative results (13 and 9), representing false rejection of 46% and 32% of the actual mixture components, respectively.

False negative results are of much greater concern to us than false positives. It is envisioned that this system can be used as an aid to an analytical chemist, to reduce the number of possibilities under consideration to a manageable number. Incorrectly eliminating a compound from further consideration, is therefore potentially more harmful than incorrectly retaining a compound for consideration. A better threshold to use routinely for this dataset might therefore be -0.10, where only 3 false negative results were obtained, at the cost of 23 false positives. It should be noted that setting a binary decision threshold for compound presence/absence is a compromise between rejecting as many false positive results as possible, while sustaining as many false negative results as acceptable.

One goal of this work was to produce a system which could be used as a prefilter for quantitative least squares analysis. Fulfilling this goal, however, meant that no false negative results would be acceptable. Since 72 false positive results were produced at this level (-0.10), it was obvious that further work was necessary.

After producing the system modifications which were described above, and further tuning the position windows, the system was retested on the the same early knowledge base, using the same interpreter options. The results produced were clearly superior to those obtained with the early version of the system. A maximum of 94.2% correct decisions were obtained, and three false negatives were obtained at the expense of only 13 false positives. More significantly, zero false negative results were obtained with 28, instead of 72 false positives. This corresponds to eliminating 90.1% of the compounds absent from the mixture from further consideration, without eliminating any actual components of the mixture.

The reference spectra were then peakpicked again with less aggressive smoothing (a 5 point Savitsky-Golay smooth was performed), and five more compounds were added to the knowledge base, making a total of 45 reference compounds in the final knowledge base (Table 5-2).

It was determined that weighting the significance of not finding a queried spectral band by a factor proportional to the square of the reference intensity, instead of directly proportional to the reference intensity, as had been done previously, was beneficial. The average score of the compounds present in the mixture, Avg_P, increased by a significant amount (approximately 0.10, depending on the other options selected). The average score of the compounds absent from the mixture, Avg_A, increased by only a very small amount (approximately 0.005). This change was subsequently incorporated into the MIXIR significance evaluation scheme.

Comparison of the results summaries for evaluation with null intensity windows, and with and without autocycling, showed a large overall benefit from autocycling (Table 5-3). The addition of autocycling provides a much larger spread between the values of Avg_P and Avg_A. Despite this success, it is more instructive to examine the failures of interpretation strategies. While in most cases autocycling provided significant and selective score enhancement for compounds present in the mixtures (e.g. Tank B results, Table 5-4), in some cases, it depressed the score of a compound actually

Table 5-2. The Reference Compounds in the Vapor Phase MIXIR Knowledge Base.

1,3-Butadiene

Acetonitrile

Acetone

Acetylaldehyde

Acrylonitrile

Butyl Acetate

Benzene

Chlorobenzene

Bis-chloroethyl Ether

2-Butanone

Chloroform

3-Chloropropene

o-Chlorotoluene

Cyclopentane

Carbon Tetrachloride

1,2-Dibromoethane

1,2-Dichloroethane

1,1-Dichloroethane

1,1-Dichloroethene

Dimethyl Disulfide

1,4-Dioxane

Ethyl Acetate

Ethylbenzene

Ethoxy Ethanol

Ethyl Ether

Ethylene Oxide

Freon-11

Freon-114

Freon-12

Freon-13

n-Hexane

Isopropanol

Methylene Chloride

4-Methyl-2-Pentanone

Perchloroethylene

Propylene Oxide

Pyridine

Styrene

1,1,2-Trichloroethane

1,1,1-Trichloroethane

Trichloroethylene

Tetrahydrofuran

Toluene

Vinyl Chloride

o-Xylene

Table 5-3. Summary Results on 50ppm mixtures, using Null Intensity Windows; with and without the Autocycling Option.

With Autocycling:

Average score for the compounds which are present: 0.624
Average score for the compounds which are absent: -0.843

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	3	8	259
0.30	4	8	258
0.20	6	8	256
0.10	9	5	256
0.00	10	5	255
-0.10	12	5	253
-0.20	15	4	251
-0.30	18	3	249
-0.40	22	1	247
-0.50	24	0	246

Without Autocycling:

Average score for the compounds which are present: 0.229
Average score for the compounds which are absent: -0.777

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	2	20	248
0.30	3	20	247
0.20	4	17	249
0.10	7	12	251
0.00	13	6	251
-0.10	17	4	249
-0.20	22	2	246
-0.30	24	0	246
-0.40	28	0	242
-0.50	36	0	234

Table 5-4. Abbreviated Score Reports for the 50ppm Mixture TANK B, obtained using Null Intensity Windows; with and without Autocycling.

With Autocycling:

Compound	Score	Peaks Matched	Peaks Sought
TRICHLOROETHYLENE*	0.999	4	12
METHYLENE CHLORIDE*	0.999	7	9
1,1-DICHLOROETHENE*	0.999	6	12
BENZENE*	0.999	7	12
CARBON TETRACHLORIDE	0.337	1	4
1,1,2-TRICHLOROETHANE	0.240	3	12
VINYL CHLORIDE*	0.127	4	12
BIS-CHLOROETHYL ETHER	0.113	3	12
O-CHLOROTOLUENE	-0.280	3	12
CHLOROBENZENE	-0.679	4	12
:			
:			
(35 Remaining Compounds)			

Without Autocycling:

Compound	Score	Peaks Matched	Peaks Sought
METHYLENE CHLORIDE*	0.749	7	9
BENZENE*	0.534	7	12
CARBON TETRACHLORIDE	0.195	1	4
1,1,2-TRICHLOROETHANE	0.120	3	12
TRICHLOROETHYLENE*	0.041	4	12
BIS-CHLOROETHYL ETHER	0.030	3	12
1,1-DICHLOROETHENE*	0.019	6	12
VINYL CHLORIDE*	-0.114	4	12
O-CHLOROTOLUENE	-0.121	3	12
CHLOROBENZENE	-0.401	4	12
ETHYL ETHER	-0.484	3	12
:			
:			
(34 Remaining Compounds)			

Note: Asterisks mark actual mixture components.

present in a mixture.

Consider, for example, the results for TANK C: one component, 2-butanone, received scores of -0.466 and -0.286, when interpreted with and without autocycling, respectively. The score was lower in the presence of autocycling because MIXIR attached less significance to the three features matched for 2-butanone in TANK C. Autocycling introduces a competition between compounds to explain unknown spectral features. In this case, only a few features were matched for 2-butanone. Most of the major features were not detected in the mixture, due to spectral overlap with other component absorptions. The few reference features which were matched had low uniqueness weightings, due to other mixture components which had coincident information, such as 4-methyl-2-pentanone. A spectroscopist might have come to the same conclusion here. It is reasonable to believe that a particular compound is absent from a mixture when only a few matching features can be found for that compound, and these can also be explained by other compounds which appear more likely to be present.

The dynamic intensity window scheme described above enhanced the results slightly in the presence of autocycling, and degraded the results slightly in the absence of autocycling. The exact origin of these results was not explored further, but since the distinction between the two runs is competition of compounds for the unknown spectral features, it is assumed that a (fortuitous) occurrence of screening out several compounds which had features matched which had significant coincidence with those of actual mixture components occurred. It should be noted that even at the lowest threshold monitored of -0.50, there was still one false negative result, both with and without autocycling. These results suggest that intensity changes due to spectral overlap preclude general application of tight constraints on peak intensities, in mixtures.

It was also found that the dynamic query selection procedure had little overall effect on the average scores for this set of mixture data. Individual compound scores, however, often showed significant changes. TANK F was a troublesome spectrum to interpret due to the fact that one of the components, freon-11 ($CFCl_3$), has C-Cl stretching absorptions which are approximately four times as intense as the most intense absorptions due to the other components. As a result, in addition to missing features due to spectral overlap, many of the smaller features in the spectrum had an integral intensity of zero when normalized. Attempts to allow zero intensity mixture bands to satisfy band queries however, seriously degraded results overall, since many false peaks were then included.

It was expected that dynamic query selection would benefit this spectrum in particular, assuming that the larger features were matched, and the smaller ones missed, due to the normalization effect noted above. Table 5-5 presents abbreviated score reports for TANK F, using default intensity

Table 5-5. Abbreviated Score Reports for the 50ppm Mixture TANK F, Obtained with Default Intensity Windows; with and without Dynamic Query Selection.

With Dynamic Query Selection:

Compound	Score	Peaks Matched	Peaks Sought
FREON-11*	0.999	2	2
STYRENE*	0.999	3	5
CHLOROFORM	0.272	1	3
BIS-CHLOROETHYL ETHER	0.267	2	5
ETHYL ETHER*	0.185	5	12
3-CHLOROPROPENE*	0.182	5	9
ETHYLBENZENE*	0.176	3	5
TRICHLOROETHYLENE	0.053	4	12
1,1,2-TRICHLOROETHANE	0.046	1	5
1,3-BUTADIENE	-0.034	1	5
ETHOXY ETHANOL	-0.127	5	12
TETRAHYDROFURAN	-0.134	3	11
ISOPROPANOL*	-0.144	4	12
ACETONITRILE	-0.368	1	5
TOLUENE	-0.371	2	5
:			
:			
(Remaining Compounds)			

Without Dynamic Query Selection:

Compound	Score	Peaks Matched	Peaks Sought
FREON-11*	0.999	2	2
STYRENE*	0.278	6	12
CHLOROFORM	0.272	1	3
ETHYL ETHER*	0.185	5	12
3-CHLOROPROPENE*	0.112	6	12
TRICHLOROETHYLENE	0.053	4	12
BIS-CHLOROETHYL ETHER	0.028	3	12
ETHYLBENZENE*	-0.079	5	12
ETHOXY ETHANOL	-0.127	5	12
TETRAHYDROFURAN	-0.134	3	11
ISOPROPANOL*	-0.144	4	12
1,1,2-TRICHLOROETHANE	-0.166	1	12
1,3-BUTADIENE	-0.288	1	12
TOLUENE	-0.416	4	12
:			
:			
(Remaining Compounds)			

Note: Asterisks mark actual mixture components.

windows, both with and without dynamic query selection. The autocycling option was not used in producing these results. Three of the six mixture components: styrene, 3-chloropropene, and ethylbenzene benefitted from this approach, and none suffered from it. It should also be noted, however, that a compound which was not present in the mixture, bis-chloroethyl ether, also had its scored increased by dynamic query selection.

The automatic peak justification feature, when used without autocycling, had a significant effect on compounds having scores in the middle of the range, as shown in Table 5-6. The number of false negative results produced at the 0.20 threshold is cut by 61% when automated peak justification is employed. Further examination of this table shows that the scores of those mixture compounds which had the least evidence to indicate their presence (those scoring below 0.00) were not significantly increased by this procedure. There was little effect produced by automated peak justification in the presence of autocycling, since nearly all of the mixture components scores which were improveded by automated peak justification were already being boosted by autocycling.

Conclusions on 50 ppm Results:

Significant improvements were made in MIXIR by tuning the position windows to appropriate vapor phase limits, and using squared intensity weighting. Modifications which attempted to use band intensity information more fully met with mixed

Table 5-6. Summary Results on the 50ppm mixtures, Obtained with Default Intensity Windows; with and without the Auto-Peak Justification Option.

With Auto-Peak Justification:

Average score for the compounds which are present: 0.271
Average score for the compounds which are absent: -0.789

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	1	20	249
0.30	3	20	247
0.20	6	7	257
0.10	7	6	257
0.00	13	5	252
-0.10	15	3	252
-0.20	22	2	246
-0.30	25	0	245
-0.40	27	0	243
-0.50	34	0	236

Without Auto-Peak Justification:

Average score for the compounds which are present: 0.206
Average score for the compounds which are absent: -0.792

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	1	20	249
0.30	3	20	247
0.20	4	18	248
0.10	6	13	251
0.00	12	10	248
-0.10	14	4	252
-0.20	21	2	247
-0.30	24	0	246
-0.40	27	0	243
-0.50	34	0	236

results. This indicates that peak intensity information in mixtures, cannot generally be considered significent, due to spectral overlap. Even at the zero tolerance level for false negative results, MIXIR could, with several combinations of options, reject 90% of the compounds which were absent from the mixtures. Close examination of the results indicated that the conclusions which MIXIR reached, even when incorrect, were reasonable based on the information presented to the system. The major limitation at this time appears to be the reliability with which component peaks can be detected in mixtures.

5 ppm Results:

There were two spectra in this group, one with six components, and one with five components (Table 5-7). Overall, the results for these spectra with the option combinations investigated were about equal to the quality of results obtained with the 50 ppm mixtures. A sample results summary obtained with the default intensity scheme, is presented in Table 5-8. At the level of zero false negative results, 98% of the compounds which were absent from the mixture were eliminated from further consideration. These spectra still had good signal to noise ratios, and little interference from carbon dioxide and water absorptions (Fig. 5-3), hence, the results were quite good.

The false positive rejection was numerically better for the 5 ppm results just described than for the 50 ppm results

Table 5-7. The 5 ppm Vapor Phase Mixture Constituents and Concentrations.

Mixture	Components	Concentration[a]
TANK 1	Acrylonitrile	5.8 ppm
	1,3-Butadiene	5.4
	Ethylene Oxide	5.4
	Methylene Chloride	5.9
	Propylene Oxide	8.1
	o-Xylene	8.4
TANK 3	Carbon Tetrachloride	4.3
	Chloroform	3.2
	Perchloroethylene	6.8
	Benzene	3.3
	Vinyl Chloride	3.1

[a] Analyzed by GC.

Table 5-8. Summary Results on the 5 ppm Mixtures Obtained Using Null Intensity Windows.

Average score for the compounds which are present: 0.423
Average score for the compounds which are absent: -0.852

LEVEL	False Positives	False Negatives	Correct Decisions
0.40	0	5	85
0.30	0	5	85
0.20	0	3	87
0.10	0	1	89
0.00	2	0	88
-0.10	3	0	87
-0.20	3	0	87
-0.30	5	0	85
-0.40	9	0	81
-0.50	10	0	80

obtained with the identical interpreter options. We suspect that this difference might be explained by one or more of the following factors: (1) While the signal to noise ratio was better for the 50 ppm spectra, they also had greater interferences from background water (compare Figs. 5-1,5-2 with Fig. 5-3). Therefore, signal detection may have been more reliable in the 5 ppm spectra. (2) Mixtures containing different components will present different spectral patterns-it cannot easily be determined why one pattern is more difficult to analyze than another.

Dynamic query selection also performed better on the 5 ppm data than on the 50 ppm data. This too, indicates that more reliable signal detection was obtained in the 5 ppm data, since dynamic query selection requires fairly reliable information to be effective. The largest separation observed between Avg_P and Avg_A was in one of the tests on the 5 ppm data. In this run, default intensity windows, autocycling, and dynamic query selection were used. The average score of the actual mixture components was 0.902, and the average score of the compounds not present in the mixture was -0.863. The success achieved in this run can be attributed to the high quality of the input information, which allowed more complex inference procedures to be effective.

2 ppm Results

The mixture constituents for the 2 ppm mixtures are presented in Table 5-9. There were three such samples: one

Table 5-9. The 2 ppm Vapor Phase Mixture Constituents and Concentrations.

Mixture	Components	Concentration[a]
EPA 1	Tetrahydrofuran	2.3 ppm
	1,1-Dichloroethane	3.5
	Benzene	2.3
	Ethylbenzene	2.1
	Methylene Chloride	2.4
	1,1,1-Trichloroethane	2.5
EPA 2	Vinyl Chloride	2.4
	Trichloroethylene	3.7
	Perchloroethylene	2.0
	Toluene	2.5
	Chlorobenzene	2.1
EPA 3	Cyclopentane	1.3
	Ethyl Acetate	1.3
	1,1-Dichloroethane	1.2
	1,1,2-Trichloroethane	1.4
	Carbon Tetrachloride	1.3
	Isopropanol	2.8
	Ethyl Ether	2.5
	3-Chloropropene	2.6
	Styrene	1.6
	Ethylbenzene	2.4
	Freon-11	2.9

[a] Results of GC analysis.

six component mixture (EPA 1), one five component mixture (EPA 2) and one eleven component mixture (EPA 3). The signal to noise ratio of these spectra was poor, as can be seen in Figure 5-4. In addition, the presence of large (relative to the sample absorptions) carbon dioxide absorptions in all of these spectra necessitated the use of the matrix interference option for carbon dioxide described previously. The spectra for mixtures EPA 2 and EPA 3 had significant positive interference from background water, as well. These two were therefore treated with the water interference option. The use of these options prevented the normalization of the sample absorptions against the large carbon dioxide absorption which dominated them. Additionally, queries in the interfering regions were eliminated, which prevented false negative or positive judgements of component peak presence in these areas. Of course, eliminating these spectral regions, although necessary, also reduced the number of features which could be queried in the resulting interpretations.

In addition to the increased problems with instrumental noise and background absorptions, one of the 2 ppm mixtures, EPA 3, contained eleven components. Due to these difficulties, one would expect the results on the 2 ppm mixtures to be poorer than those previously described. This was indeed the case. The score thresholds examined were reduced by 0.30, since the components scored lower on the average in these mixtures. Even at -0.80, however, at least one false negative result still remained, regardless of the options used. This false negative was 1,1-dichloroethane in mixture EPA 3. No matching bands were found for this compound, regardless of the options used.

The spectrum of 1,1-dichloroethane is dominated by the C-Cl stretching absorptions, which appear at about 710 cm^{-1}. This region of the absorption spectrum of 1,1-dichlorethane is shown superimposed on the same spectral region of EPA 3 in Figure 5-5 (the absorption axis shown is that of the mixture). The two large absorptions of 1,1-dichloroethane have virtually disappeared into the absorption background of the mixture, and hence were not detected by the peakpicker. This situation cannot be cured by any interpretation logic- if no corresponding bands are detected, then the compound cannot be judged to be in the mixture in question. Only better signal recognition algorithms can help in such cases.

At a level of one false negative, the minimum number of false positive results observed with any option combination was 35, using the default intensity windows only (Table 5-10). This corresponds to 69% of the possible false positive compounds rejected, along with one actual mixture component.

Conclusions:

It has been demonstrated that effective vapor phase spectral descriptions can be generated from peak tables of a reference set. A modified form of the condensed phase spectral interpreter, MIXIR, was found to reject 90 and 98% of the possible false positive results for the 50 ppm and 5 ppm

Figure 5-5. A Portion of the C-Cl Stretching Region for the 1,1-Dichloroethane Reference Spectrum, and the 2 ppm Mixture Spectrum EPA 3.

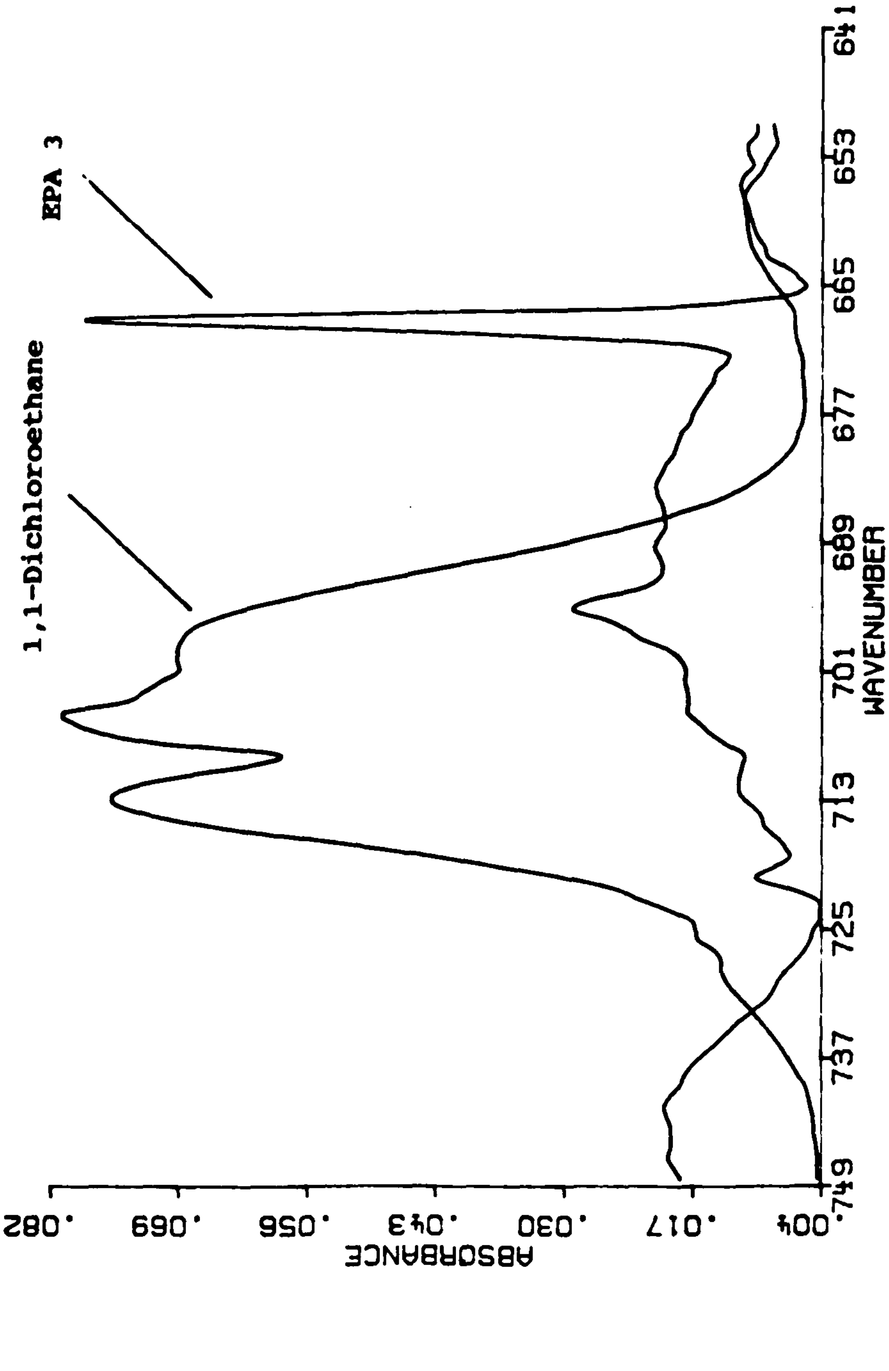


Table 5-10. Summary Results on the 2 ppm Mixtures Obtained Using Default Intensity Windows.

Average score for the compounds which are present: 0.190
Average score for the compounds which are absent: -0.769

LEVEL	False Positives	False Negatives	Correct Decisions
0.10	8	10	117
0.00	10	6	119
-0.10	15	5	115
-0.20	16	4	115
-0.30	16	3	116
-0.40	20	3	112
-0.50	23	2	110
-0.60	24	2	109
-0.70	27	2	106
-0.80	35	1	99

mixtures tested, respectively, without eliminating any actual mixture components. MIXIR could not, however, produce results at the level of zero tolerance for false negatives, for the 2 ppm mixtures tested.

Several new interpretation paradigms were developed and tested. These provided extended dynamic capabilities, based largely on peak intensity information. These paradigms were found to be useful, however, only so long as the unknown peak information was reliable, and extensive overlap between bands of widely differing intensity did not occur in the mixtures.

The major limitation currently facing MIXIR is the question of reliable signal detection. More sensitive and selective peak detection algorithms must be developed before the system can be advanced to work under more adverse conditions, i.e., those involving poor signal to noise ratios, and extensive component band overlaps. The issue of interfering background absorptions due to water and carbon dioxide might be solved by developing a library of water and carbon dioxide spectra taken under varying conditions. Computer selection of the best background match from this library to that observed in a sample spectrum would likely produce more effective correction for these spectral interferences. If effective, this would greatly extend the limits of infrared spectral detection in complex mixtures.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

Significant progress was made in developing adaptive knowledge based systems for spectral analysis. The major advances made with these systems stem from their dynamic approach to data interpretation. This approach is largely made possible due to the creation of computer procedural methods used to mimic the human interpretive processes which have been identified. It has been shown that the major components of condensed phase and vapor phase mixtures may be reliably determined using peak-based information only.

At this stage, several factors have been identified which limit the IRBASE/MIXIR system as it stands. From the chemist's perspective, the most important of these is that the system already makes nearly optimum use of the peak-based information which is used. As noted in the research chapters, attempts to use increasingly subtle relationships in the peak data often failed, due to the reliability of the information. Significant further advances will probably not be possible without increasing the information which is presented to the system.

One way to increase the information is to improve the signal detection process itself. Exploratory work into the use of neural networks for infrared peak detection has provided some promising results, however this work is still immature. It may be that the preprocessing performed on the network input should be eliminated, or modified. Another modification which might prove beneficial would be to alter the method of noise representation. The reason for presenting the noise input was that it was determined that visual comparison of signal height to noise magnitude was being performed in human peak detection. However, with the method used, the network is left to determine the proper relationship of the noise representation to the rest of the data. Since the human use of the information is a comparison of magnitudes (i.e. a ratio), the ratio of the signal magnitude to the noise magnitude would likely prove more effective. When a specific characteristic which we wish the network to learn can be identified, it makes sense to provide this characteristic directly. This places less reliance on the network learning process to extract meaningful information from a series of examples.

It is my opinion that any information extraction which can be performed through preprocessing, without actually eliminating other information as a side effect, is beneficial and should be performed. Providing this aid to the network may seem to conflict with the notion of the network as a general purpose "learning machine". Consider, however, that a neural network is useful when specific mathematical properties of data which are needed to solve a problem cannot be identified. It is hoped that the network will then be able to implicitly extract these properties from the data set through the learning process. This process may be misdirected, however, by the variation and noise present in the data which cause us to choose the network approach in the first place. By essentially "pointing out" those features which can be identified we can not only ensure that the network learns these characteristics, but can also improve the chances that the network will be able to correctly identify those features which we cannot.

Further improvements may result from modification of the network architecture. A fully connected network architecture was used, since this is the simplest and most common approach in what still constitutes art as well as science. However, again, when we can identify any manner in which information is processed in a human pattern recognition process, advantage should be taken of this knowledge. In the way that the author processes spectral data, the signal to noise ratio is one piece of information, and the shape of a prospective signal pattern is another. These two are evaluted separately, and the results of these two evaluations are then combined. This behavior can be forced in the network by allowing the hidden layer to perform the shape analysis, and providing the signal to noise ratio as an input to be combined at the output layer. In this way, the topology of our network would mimic the topology of the author's human information processing flow.

As stated in chapter 4, a comparison of the frequency of the noise with the width of a prospective signal pattern is also performed in the author's "signal processor". In order for the network to do this, however, it must first be given the frequency of a noise sample from the spectrum. This might be done by performing an FFT on the noise train, and using the largest peak of the resulting noise spectrum for the noise frequency input to the network. It still remains then for the network to learn how to extract the "width" of an arbitrary pattern such as will confront it in use. This is more daunting, since it is difficult to think of what might constitute the width of an arbitrary function which is truncated by our spectral window. If this can in some way be conceived, then at least we can have hope that the network will be able to learn it, and can begin to take steps to aid this process.

Returning to the question of providing additional information to a peak-based spectral interpreter, the use of least-squares fitting (LSF) techniques for vapor phase spectra, and in limited spectral regions, for condensed phase spectra, should be explored. The two methods could be combined into a large system which uses the peak based expert system to narrow down the possibilites. Traditional least squares quantitation of the reduced reference set to the unknown could then be performed much more quickly and reliably.

Alternatively, LSF could be performed locally in the spectrum to perform conflict resolution for the expert system.

When several compounds appear to offer a matching feature for an unknown feature, LSF could be used to determine which of these compounds are contributing to this feature in the unknown. Since LSF provides for simultaneous solutions to the system, the competition which can sometimes be a disadvantage for MIXIR could be avoided, when necessary.

From a computer science perspective, a major issue facing MIXIR and IRBASE is their complexity, and the implementation language. FORTRAN is not well suited to writing complex programs which require modelling abstract processes. The lack of structure type definition capabilities in particular severely handicaps the FORTRAN programmer. MIXIR and IRBASE are currently in a state where they are very difficult to understand mechanistically. They should both be recast in a modern language such as Pascal or C. Maximum advantage should be made of the chosen language capabilities to redesign, rather than simply translate the resulting programs. An object-oriented languages such as Object Pascal, C^{++}, or the like could probably be used to even better advantage to control complexity. Although these are more engineering than scientific considerations, the practical need for this cannot be overemphasized.

LIST OF REFERENCES

1. Smith, D.H.; Gray, N.A.B.; Nourse, J.G.; Crandell, C.W. Anal. Chim. Acta, 1981, 133, p.471.

2. Barr, A.; Feigenbaum, E.A. "The Handbook of Artificial Intelligence", William Kaufmann, Inc., Los Altos, Ca., 1982.

3. Winston, P.H. "Artificial Intelligence", Addison-Wesley: Menlo Park, Ca., 1984.

4. Nilsson, N.J. "Principles of Artificial Intelligence", Tioga, Palo Alto, Ca, 1980.

5. J. Zupan, "Computer Supported Spectroscopic Databases", John Wiley & Sons, New York, 1986.

6. Gray, N.B. "Computer-Assisted Structure Elucidation", John Wiley & Sons, New York, 1986.

7. Pierce, T.H.; Hohne, B.A. "Artificial Intelligence Applications in Chemistry", ACS Symposium Series, American Chemical Society, Washington D.C., 1986.

8. Small, G.W. Anal. Chem., 1987, p.535A.

9. Gray, N.A.B. Anal. Chim. Acta, 1988, 210, p.9.

10. Hohne, B.A.; Pierce, T.H. "Expert System Applications in Chemistry", ACS Symposium Series, American Chemical Society, Washington D.C., 1989.

11. Wythoff, B.J.; Tomellini, S.A. R & D Magazine, April 1989, p.52.

12. Tomellini, S.A.; Wythoff, B.J.; Woodruff, H.B. "Developing Knowledge Based Systems: A Learning Process". ACS Symposium Series- Expert System Applications in Chemistry, American Chemical Society, 1989, p.236.

13. Gray, N.A.B.; Crandell, C.W.; Nourse, J.G.; Smith, D.H.; Dageforde, M.L.; Djerassi, C. J. Org. Chem., 1981, 46, p.703.

14. Zupan, J.; Novic, M.; Bohanec,S,; Razinger, M.; Lah, L.; Tusar, M.; Kosir, I. Anal. Chim. Acta, 1987, 200, p.333.

15. Visser, T.; van der Maas, J.H. Anal. Chim. Acta, 1980, 122, p.357.

16. Visser, T.; van der Maas, J.H. Anal. Chim. Acta, 1980, 122, p.363.

17. Woodruff, H.B.; Smith, G.M. Anal. Chem., 1980, 52, p.2321.

18. Tomellini, S.A.; Saperstein, D.D.; Stevenson, J.M.; Smith, G.M.; Woodruff, H.B. Anal. Chem., 1981, 53, p.2367.

19. Visser, T.; Van der Maas, J.H. Anal. Chim. Acta, 1981, 133, p.451.

20. Trulson, M.O.; Munk, M.E. Anal. Chem., 1983, 55, p.2137.

21. Tomellini, S.A.; Hartwick, R.A.; Stevenson, J.M.; Woodruff, H.B. Anal. Chim. Acta, 1984, 162, p.227.

22. Smith, G.M.; Woodruff, H.B. J. Chem. Inf. Comp. Sci., 1984, 24, p.33.

23. Tomellini, S.A.; Hartwick, R.A.; Woodruff, H.B. Appl. Spectrosc., 1985, 39, p.331.

24. Puskar, M.A.; Levine, S.P.; Lowry, S.R. Anal. Chem., 1986, 58, p.1156.

25. Puskar, M.A.; Levine, S.P.; Lowry, S.R. Anal. Chem., 1986, 58, p.1981.

26. Saperstein, D.D. Appl. Spectrosc., 1986, 40, p.344.

27. MacDonald, R.S. Anal. Chem., 1986, 58, p.1906.

28. Blaffert, T. Anal. Chim. Acta, 1986, 191, p.161.

29. Hamoudi, N.A.A. J. Pet. Res., 1986, 5, p.121.

30. Ying, L.S.; Levine, S.P.; Tomellini, S.A.; Lowry, S.R Anal. Chem., 1987, 59, p.2197.

31. Luinge, H.J.; Kleywegt, G.J.; Van't Klooster, H.A.; Van der Maas, J.H. J. Chem. Inf. Comput. Sci., 1987, 27, p.95.

32. Coates, J.P. Spectroscopy, 1988, vol.3, no.2, p.18.

33. Ying, L.S.; Levine, S.P.; Tomellini, S.A.; Lowry, S.R Anal. Chim. Acta, 1988, 210, p.51.

34. Schlieper, W.A.; Isenhour, T.L.; Marshall, J.C. J. Chem. Inf. Comput. Sci., 1988, 28, p.159.

35. Wythoff, B.J.; Buck, C.F.; Tomellini, S.A. Anal. Chim.

Acta, 1989, 217, p.203.

36. Wythoff, B.J.; Tomellini, S.A. Anal. Chim. Acta, 1989, 227, p.343.

37. Wythoff, B.J.; Tomellini, S.A. Anal. Chim. Acta, 1989, 227, p.359.

38. Haraki, K.S.; Venkataraghavan, R.; Mclafferty, F.W. Anal. Chem., 1981, 53, p.386.

39. Enke, C.G.; Wade, A.P.; Palmer, P.T.; Hart, K.J. Anal. Chem., 1987, 59, p.1363A.

40. Scott, D.R. Anal. Chim. Acta, 1988, 211, p.11.

41. Palmer, P.T.; Hart, K.J.; Enke, C.G.; Wade, A.P. Talanta, 1989, 36, p.107.

42. Sasaki, S.; Fujiwara, I.; Abe, H.; Yamasaki, T. Anal. Chim. Acta, 1980, 122, p.87.

43. Shelley, C.A.; Munk, M.E. Anal. Chim. Acta, 1981, 133, p.507.

44. Debska, B.; Duliban, J.; Guzowska-Swider, B.; Hippe, Z. Anal. Chim. Acta, 1981, 133, p.303.

45. Moldoveanu, S.; Rapson, C.A. Anal. Chem., 1987, 59, p.1207.

46. Bach, R.; Karnicky, J.; Abbott, S. "Artificial Intelligence Applications in Chemistry", pp.278-296, American Chemical Society, Washington D.C., 1986.

47. Gunasingham, H.; Srinivasan, B.; Ananda, A.L. Anal. Chim. Acta, 1986, 182, p.193.

48. Ananda, A.L.; Foo, S.M.; Gunasingham, H. J. Chem. Inf. Comput. Sci., 1988, 28, p.82.

49. Settle, F.A.; Diamondstone, B.I.; Kingston, H.M.; Pleva, M.A. J. Chem. Inf. Comput. Sci., 1989, 29, p.11.

50. van Leeuwen, J.A.; Vandeginste, B.G.M.; Kateman, G.; Mulholland, M.; Cleland, A. Anal. Chim. Acta, 1990, 228, p.145.

51. Hebb, D.O. "The Organization of Behavior", Wiley, New York, 1949.

52. Nilsson, N.J. "Learning Machines", McGraw Hill, New York, 1965.

53. Minsky, M.; Papert, S. "Perceptrons", MIT Press, Cambridge, MA, 1969.

54. Rumelhart, D.E.; McClelland, J.L. "Parallel Distributed Processing", vol.1, MIT Press, Cambridge, MA, 1986.

55. Lippmann, R.P. IEEE ASSP Magazine, 1987, vol.4, no.2, p.4.

56. Caudill, M. AI/Expert, 1987, vol.2, no.12, p.46.

57. Caudill, M. AI/Expert, 1988, vol.3, no.2, p.55.

58. Caudill, M. AI/Expert, 1988, vol.3, no.6, p.53.

59. McClelland, J.L.; Rumelhart, D.E. "Explorations In Parallel Distributed Processing", MIT Press, Cambridge, MA, 1988.

60. Kowalski, B.R.; Jurs, P.C.; Isenhour, T.L.; Reilley, C.N. Anal. Chem., 1969, 41, p.1945.

61. Preuss, D.R.; Jurs, P.C. Anal. Chem., 1974, 46, p.520.

62. Liddell III R.W.; Jurs, P.C. Anal. Chem., 1974, 46, p.2126.

63. Robb, E.W.; Munk, M.E. Mikrochim. Acta., 1990, I, p.131.

64. Donahue, S.M.; Brown, C.W.; Kumaresan, R. "Neural Networks and the Interpretation of Infrared Spectra", 1990 Pittsburgh Conference On Analytical Chemistry and Applied Spectroscopy, N.Y., paper no. 1062.

65. Long, J.R.; Gregariou, V.G.; Gemperline, P.J. Anal. Chem., 1990, 62, p.1791.

66. Conley, R.T. "Infrared Spectroscopy", Allyn and Bacon, Boston, Ma., 1966.

67. Bellamy, L.J. "Infrared Spectra of Complex Molecules", Chapman and Hall, New York, 1975.

68. Nakanishi, K.; Solomon, P. "Infrared Absorption Spectroscopy", Holden-Day, Oakland, Ca., 1977.

69. Dolphin, D.; Wick, A. "Tabulation of Infrared Spectral Data", John Wiley & Sons, Inc, New York, 1977.

70. Bellamy, L.J. "Infrared Spectra of Complex Molecules, Vol.2 Advances in Group Frequencies", Chapman and Hall, New York, 1980.

71. Jurs, P.C.; Isenhour, T.L. "Chemical Applications of Pattern Recognition", John Wiley and Sons, New York, 1975.

72. Rasmussen, G.T.; Isenhour, T.L.; Lowry, S.R.; Ritter, G.L. Anal. Chim. Acta, 1978, 103, p.213.

73. Varmuza, K. "Pattern Recognition in Chemistry", Lecture Notes in Chemistry Series, no.2, Springer Verlag, New York, 1980.

74. Haraki, K.S.; Venkataraghavan, R.; McLafferty, F.W. Anal. Chem., 1981, 53, p.386.

75. Heite, F.H.; Dupuis, P.F.; van't Klooster, H.A.; Dijkstra, A. Anal. Chim. Acta, 1978, 103, p.313.

76. Dupuis, P.F.; Cleij, P.; van't Klooster, H.A.; Dijkstra, A. Anal. Chim. Acta, 1979, 112, p.83.

77. Massart, D.L.; Kaufman, L. "The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis", Chemical Analysis Series, vol.65, John Wiley and Sons, New York, 1983.

78. Blaffert, T. Anal. Chim. Acta, 1984, 161, p.135.

79. Seil, J.; Kohler, I; Leith, C.W.v.d.; Opferkuch, H.J. Anal. Chim. Acta, 1986, 188, p.219.

80. Sharaf, M.A.; Illman, D.L.; Kowalski, B.R. "Chemometrics", Chemical Analysis Series, vol.82, John Wiley and Sons, New York, 1986.

81. Matthews, M.H. AI/Expert, 1987, vol.2, no.9, p.43.

82. Savitsky, A.; Golay, M.J.E. Anal. Chem., 1964, 36, p.1627.

83. Grushka, E.; Monacelli, G.C. Anal. Chem., 1972, 44, p.484.

84. Reich, G. Anal. Chim. Acta, 1987, 201, p.171.

85. Bryant, W.F.; Trivedi, M.; Hinchman IV, B.; Sofranko, S.; Mitacek, P. Anal. Chem., 1980, 52, p.38.

86. Currie, L.A. J. Res. Natl. Bur. Stand., 1985, 90, 409.

87. Helstrom, C.W. "Statistical Theory of Signal Detection", Pergammon Press, New York, NY, 1968.

88. Donahue, S.M.; Brown, C.W.; Obremski, R.J. Appl.

Spectrosc., 1988, 42, p. 353.

89. Ying, L.S.; Levine, S.P. Anal. Chem., 1989, 61, p.677.

90. Strang, C.R.; Levine, S.P. Am. Ind. Hyg. Assoc. J., 1989, 50, p.78.

91. Xiao, H.K.; Levine, S.P.; D'Arcy, J.B. Anal. Chem., 1989, 61, p.2708.

92. Ying, L.S.; Levine, S.P.; Tomellini, S.A. "Computer Enhanced Analytical Spectroscopy", vol. 2, p.245, Plenum Press, NY, 1990.

APPENDIX A.

TRACE OF THE INTERPRETATION PROCESS FOR THE FUNCTIONALITY "ACID" INCLUDING EXPLANATORY COMMENTS.

```
FUNCTIONALITY ACID
PASSED                   INITIAL EMPIRICAL FORMULA TEST
* (Carboxylic Acids undergo extremely strong intermolecular
*  interactions, therefore, vapor phase spectra require separate
*  rules:)

QUERY--IS THE SOLVENT/SAMPLE STATE VAPOR?
     ANSWER------NO------

 * (Begin Pyridine-Like Acid queries. This functionality also
*  requires at least one nitrogen, in addition to the other
*  formula stipulations already queried:)

 FORMULA QUERY------ANY NITROGEN(S)     ?
     ANSWER------YES-----

* (Since Pyridines are aromatic, and this class has already been
*  determined, we check for a reasonable value here, before
*  proceeding any further:)

 EXPECTATION VALUE QUERY
     IS AROMATIC             GREATER THAN    0.20?
     ANSWER------YES-----
* (The acid group on the molecule is capable of protonating the
*  pyridyl nitrogen. For protonated pyridine-like bases, there is
*  evidence to indicate the formation of a (pyHpy)+ complex, with
*  with a double minimum proton potential, observed spectro-
*  scopically at about 2500 and 2000 cm-1. Such a species may
*  be involved here, as similar absorptions are observed in
*  pyridine-like acids, to which the following two queries
*  pertain:)

 PEAK QUERY
   ANY PEAK(S)  POSITION:  2251 -  2520     INTENSITY:   1 -   6
                WIDTH: BROAD
      ANSWER------YES-----
 ACTION------SET ACID-PYRIDINE-LIKE    TO  0.10
                                       CURRENT VALUE =  0.10
 PEAK QUERY
    ANY PEAK(S) POSITION:  1801 -  2000     INTENSITY:   1 -   6
                WIDTH: BROAD
      ANSWER------NO------

* (Begin "normal" Acid queries- here, Acid-Sat. Looking for the
*  extremely strongly H-bonded O-H stretch which characterizes
*  the Acid functionality. It is broadened to the degree that
*  such hydrogen bonding exists, due to the increased environ-
*  mental heterogeneity. The best possible peak match is 7-10,B,
*  3050-2900. The peak queries will attempt to converge on
*  these values:)

 PEAK QUERY
    ANY PEAK(S)  POSITION:  2780 -  3150     INTENSITY:   4 -  10
```

```
                WIDTH: BROAD
      ANSWER------YES-----
 ACTION------SET ACID TO  0.25     CURRENT VALUE =  0.25
 PEAK QUERY
    ANY PEAK(S)   POSITION:  2780 -  3150     INTENSITY:   7 - 10
                 WIDTH: BROAD
      ANSWER------NO------
 PEAK QUERY
    ANY PEAK(S)   POSITION:  2900 -  3050     INTENSITY:   4 - 6
                      WIDTH: BROAD
      ANSWER------YES-----
 ACTION---ADD  0.10 TO ACID         CURRENT VALUE =  0.35

* (Looking for the H-bonded C=O stretch of the Acid dimer,
*  which often is observed concurrently with the monomeric form,
*  but is the more prevalent, ordinarily, of the two:)

 PEAK QUERY
    ANY PEAK(S)   POSITION:  1661 -  1750     INTENSITY:   7 - 10
                      WIDTH: SHARP TO  BROAD
      ANSWER------YES-----
 ACTION---ADD  0.25 TO ACID           CURRENT VALUE =  0.60

 * (Looking for C-O stretch:)

 PEAK QUERY
    ANY PEAK(S)   POSITION:  1191 -  1300     INTENSITY:   4 - 10
                 WIDTH: SHARP TO  BROAD
      ANSWER------YES-----
 ACTION---ADD  0.10 TO ACID           CURRENT VALUE =  0.70

* (Narrowing peak-intensity "window" to converge on most likely
*  values:)

 PEAK QUERY
    ANY PEAK(S)   POSITION:  1191 -  1300     INTENSITY:   7 - 10
                      WIDTH: SHARP TO  BROAD
      ANSWER------YES-----
 ACTION---ADD  0.10 TO ACID             CURRENT VALUE =  0.80

* (Looking for O-H bending vibration, which typically is broad
*  also, due to hydrogen bonding. Here, we begin with the best
*  match to expected values, and broaden the query range
*  slightly, to include less typical cases, if necessary. Such an
*  approach is rarely taken in PAIRS, for although this has no
*  effect on the outcome, it is formally a less efficient way of
*  asking the questions=>think about this...:)

 PEAK QUERY
   ANY PEAK(S)   POSITION:   900 -   950   INTENSITY:   3 - 6
                    WIDTH: BROAD
      ANSWER------NO------
 PEAK QUERY
```

```
   ANY PEAK(S)   POSITION:   876 -   970   INTENSITY:   1 -  10
                     WIDTH: BROAD
     ANSWER------NO------
 PEAK QUERY
     ANY PEAK(S)  POSITION:   900 -   950   INTENSITY:   3 -   6
                    WIDTH: AVERAGE
     ANSWER------YES-----
 ACTION---ADD  0.10 TO ACID              CURRENT VALUE =  0.90
 * (Begin speciation (Acid subclasses) based on position match of
 *  C=O stretch for the various subclasses. Again, all queries
*  refer to the dimeric form predominant in condensed phases:)

 PEAK QUERY
    ANY PEAK(S)   POSITION:  1661 -  1750   INTENSITY:   7 -  10
                  WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
 PEAK QUERY
    1 PEAK(S)      POSITION:  1661 - 1750    INTENSITY:   7 - 10
                  WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
 * (Looking for C=O stretch for acid with an electron withdrawing
 *  group alpha to the carbonyl. Such substitution increases the
*  C=O stretching frequency, due to the increased importance of
*  the canonical form 2(R)C+--O-, here:)

 PEAK QUERY
    ANY PEAK(S)  POSITION:  1726 -  1750    INTENSITY:   7 -  10
                    WIDTH: SHARP TO  BROAD
     ANSWER------NO------
 * (The following peak position range corresponds to an area
*  where both Alpha-withdrawing Acids and Saturated Acids
*  appear. A "hit" here will cause both these subclasses to be
*  set to the class value here; then reduced by 20%, to indicate
*  the ambiguity of the subclass assignment:)

 PEAK QUERY
    ANY PEAK(S)  POSITION:  1721 -  1725    INTENSITY:   7 -  10
                WIDTH: SHARP TO  BROAD
     ANSWER------NO------
 PEAK QUERY
    ANY PEAK(S)     POSITION:  1708 -  1720   INTENSITY:   7 - 10
                   WIDTH: SHARP TO  BROAD
     ANSWER------NO------

 * (The following peak position range corresponds to an area
*  where both Saturated and Unsaturated Acids appear. A "hit"
*  here will cause both these subclasses to be set to the class
*  value here; then reduced by 20%, to indicate the ambiguity of
*  the subclass assignment:)

 PEAK QUERY
    ANY PEAK(S)   POSITION:  1700 -  1707   INTENSITY:   7 -  10
                  WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
ACTION------SET ACID-UNSATURATED TO  0.90   CURRENT VALUE = 0.90
ACTION---MULTIPLY ACID-UNSATURATED BY  0.80  CURRENT VALUE = 0.72
ACTION------SET ACID-SATURATED TO  0.90    CURRENT VALUE = 0.90
ACTION---MULTIPLY ACID-SATURATED BY  0.80  CURRENT VALUE =  0.72

 * (Compare Acid score to that for the Acid-Pyridine-Like
 *  subclass, and save the larger as the Acid class value:)

 PROBABILITY QUERY
     IS ACID                 LESS THAN      0.10?
     ANSWER------NO------
```

APPENDIX B.

TRACE OF THE NONINTERACTIVE INTERPRETATION FOR THE "ACETAL" FUNCTIONALITY, INCLUDING EXPLANATORY COMMENTS, FOR THE SPECTRAL DATA OF 1,1-DIETHOXYETHANE.

```
FUNCTIONALITY ACETAL
PASSED                       INITIAL EMPIRICAL FORMULA TEST
* (Looking for C-O-C-O-C symmetric and asymmetric stretches:
*  For Acetals, the C-O stretches are split into multiple
*  components, due to multiple coupling of adjacent C-O
*  vibrations, along with any molecular asymmetry about these
*  groups:)

PEAK QUERY
   AT LEAST 4 PEAK(S)  POSITION: 1035 -  1210  INTENSITY: 1 - 10
                       WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
ACTION------SET ACETAL TO  0.15     CURRENT VALUE =  0.15
* (Likely observe further splitting:)

PEAK QUERY
  AT LEAST 5 PEAK(S)  POSITION: 1035 -  1210  INTENSITY: 1 - 10
                      WIDTH: SHARP TO  BROAD
     ANSWER------NO------
* (Two of the components of the C-O-C-O-C multiplet should be
*  in this region:)

PEAK QUERY
   AT LEAST 2 PEAK(S)  POSITION: 1120 -  1195  INTENSITY:  1 -10
                      WIDTH: SHARP TO  BROAD
     ANSWER------NO------
* (Looking for a particular component of the C-O-C-O-C
*  multiplet:)

PEAK QUERY
   ANY PEAK(S)  POSITION:  1156 -  1190   INTENSITY:   1 -  10
                WIDTH: SHARP TO  BROAD
     ANSWER------NO------
* (Looking for a particular component of the C-O-C-O-C
*  multiplet:)

PEAK QUERY
   ANY PEAK(S)  POSITION:  1061 -  1100   INTENSITY:   1 -  10
                WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
ACTION---ADD 0.05 TO ACETAL          CURRENT VALUE =  0.20
* (Looking for C-H adjacent to C-O, this is the sole band
*  differentiating Acetals from the closely related Ketals:)

PEAK QUERY
   ANY PEAK(S) POSITION:  1101 -  1110    INTENSITY:   1 -  10
               WIDTH: SHARP TO  BROAD
     ANSWER------YES-----
ACTION---ADD 0.25 TO ACETAL            CURRENT VALUE =  0.45
* (The following functionality(s), already queried, have similar
*  absorptions, and can be determined with greater reliability.
*  If they appear to be absent then, we can increase our
*  confidence in assigning these peaks to Acetal. Implicit here
*  is an "either/or" approach..when might this type of reasoning
*  fail? :)

 EXPECTATION VALUE QUERY
      IS ACID GREATER THAN 0.50?
      ANSWER------NO------
 EXPECTATION VALUE QUERY
      IS ALCOHOL GREATER THAN 0.50?
      ANSWER------NO------
 EXPECTATION VALUE QUERY
      IS AMIDE GREATER THAN 0.50?
      ANSWER------NO------
 EXPECTATION VALUE QUERY
      IS AMINE GREATER THAN 0.50?
      ANSWER------NO------
 ACTION---MULTIPLY ACETAL BY 1.29     CURRENT VALUE =  0.58
```

APPENDIX C.

PARTIAL TRACE OF AN INTERACTIVE INTERPRETATION FOR 1,1-DIETHOXYETHANE

```
 FUNCTIONALITY ACETAL
 PASSED                  INITIAL EMPIRICAL FORMULA TEST
 * (Looking for C-O-C-O-C symmetric and asymmetric stretches:
 *  For Acetals, the C-O stretches are split into multiple
 *  components, due to multiple coupling of adjacent C-O
 *  vibrations, along with any molecular asymmetry about these
 *  groups:)

 PEAK QUERY
    AT LEAST 4 PEAK(S)  POSITION: 1035 -  1210  INTENSITY: 1 - 10
                        WIDTH: SHARP TO  BROAD
      ANSWER------YES-----

DESIRED ANSWER (YES,NO, RETURN): y
THE ANSWER IS: YES

 ACTION------SET ACETAL TO  0.15     CURRENT VALUE =  0.15

 * (Likely observe further splitting:)

 PEAK QUERY
   AT LEAST 5 PEAK(S)  POSITION: 1035 -  1210  INTENSITY: 1 - 10
                       WIDTH: SHARP TO  BROAD
      ANSWER------NO------

DESIRED ANSWER (YES,NO, RETURN): y
THE ANSWER IS: YES

 ACTION-----ADD 0.10 TO ACETAL       CURRENT VALUE = 0.25
      .
      .      [Deleted portion of trace ouput.]
      .
ACTION---MULTIPLY ACETAL BY 1.29     CURRENT VALUE =  0.71
```