(6) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformations. In *Chemometrics Research Studies Series*; Bawden, D., Ed.; Research Studies Press (Wiley): New York, 1988.
(7) Smellie, A. *CONSTRICTOR*. Oxford Molecular Ltd.: Terrapin House, University Science Area, South Parks Road, Oxford, England, 1989.
(8) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650.
(9) Silipo, C.; Hansch, C. *J. Am. Chem. Soc.* **1975**, *97*, 6849.
(10) Dietrich, S.; Smith, R. N.; Fukunaga, M.; Olney, M.; Hansch, C. *Arch. Biochem. Biophys.* **1979**, *194*, 600.
(11) Bron, C.; Kerbosch, J. *Commun. ACM* **1973**, *16* (9), 575-577.

Computer-Assisted Infrared Identification of Vapor-Phase Mixture Components

BARRY WYTHOFF,*[,†,‡] XIAO HONG-KUI,[§] STEVEN P. LEVINE,[§] and STERLING A. TOMELLINI[†]

Department of Chemistry, University of New Hampshire, Durham, New Hampshire 03824, and School of Public Health, The University of Michigan, Ann Arbor, Michigan 48109

Received January 3, 1991

The IRBASE/MIXIR system was originally tested on interpretation of infrared spectra of condensed-phase mixtures. The system has now been adapted to allow interpretation of vapor-phase mixture spectra. The dynamic interpretation capabilities of the system have been expanded to allow runtime manipulation of complete peak lists, allowing generation of the optimum spectral description for the interpretation at hand. The modifications to the system are described, along with the results of testing on actual mixtures of varying complexity.

INTRODUCTION

Infrared analysis of organic vapors has many potential applications, including on-site measurement of toxic compounds at hazardous waste sites and in the workplace, and analysis of unresolved effluents from GC-FTIR experiments. Efforts at computer-assisted identification of components of mixtures have been largely directed at condensed-phase systems. The analysis may be performed in either of two ways: through a hyphenated technique involving a separation prior to identification, e.g., GC-FTIR, or by analysis of the intact mixture.

Using a separation method prior to identification has the advantage that the data analysis is much simpler and ideally involves a spectral library search using some similarity metric for each of the separated components. Important characteristics include robustness in the presence of noise and experimental variation, and analysis times, particularly when real-time analysis of capillary GC-FTIR spectra is required. One-at-a-time library comparison methods using linear neural networks,[1] boolean logic,[2,3] cluster analysis,[4] parallel processing,[5] and information theory[6] have been demonstrated. Principal components analysis has been used to select library subsets for more detailed analysis,[7] and orthonormalized spectral libraries have been evaluated.[8] Interferometric search has been performed using only 100 data points from the interferograms.[9] The use of a second, coupled spectrometer subsequent to GC separation (GC-FTIR MS) has been demonstrated on a 30-component mixture.[10] The combination of independent sources of spectral information was found to aid the analysis.

A drawback of most library search methods is that they perform poorly on mixtures. There is no guarantee that separation will be complete when using a hyphenated separation-identification scheme. In addition, many compounds can undergo thermal degradation during gas chromatographic analysis or may adsorb irreversibly to liquid chromatographic column packings. Finally, the cost and complexity of instrumentation obviously increase with hyphenated techniques.

Infrared spectra of intact mixtures have been studied mathematically by principal components analysis to determine the number of components[11] and, in the Fourier domain by factor analysis, to quantitate mixtures where the component identities were known.[12] A comparison of four multivariate methods was performed on quantitative analysis of mixtures with known component identities.[13]

Quantitative analysis of vapor-phase mixtures where the component identities are not known has been explored by using least-squares fitting (LSF) techniques.[14,15] Qualitative identification of vapor-phase mixture components has been recently reported by using iterative least-squares fitting techniques (ILSF).[16] 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[17] and a knowledge-based system to manipulate these rules.[18] 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.[19] 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 challenge. 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 that 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 SECTION

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

* Corresponding author.
† University of New Hampshire.
‡ Present address: Division of Inorganic Analysis, Center for Analytical Chemistry, National Institute of Standards and Technology, Gaithersburg, MD 20899.
§ The University of Michigan.

0095-2338/91/1631-0392$02.50/0

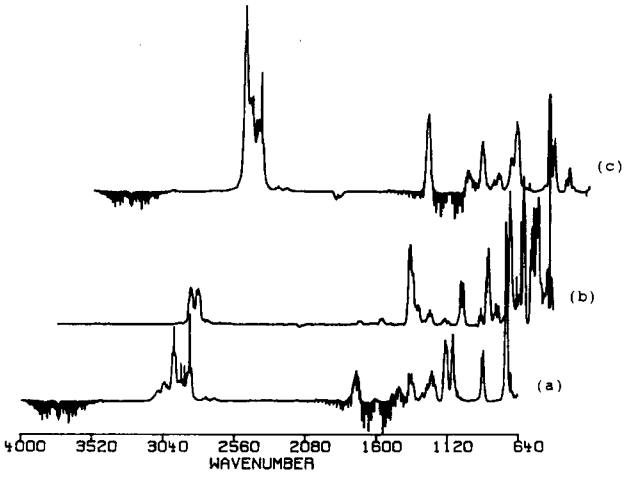

Figure 1. "Stacked plot" of the 50-ppm mixture spectra: (a) TANK A, (b) TANK B, and (c) TANK C.

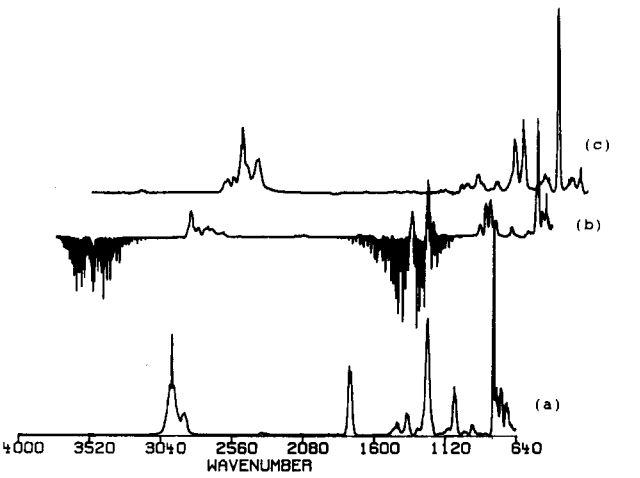

Figure 2. "Stacked plot" of the 50-ppm mixture spectra: (a) TANK D, (b) TANK E, and (c) TANK F.

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[15] and qualitative ILSF[16] 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 U.S. 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.[15]

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 Preprocessing. The 50-ppm vapor phase reference spectra were subjected to a 5-point Savitsky–Golay smoothing algorithm.[20] 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.[21] 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" (Figures 1–4). 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

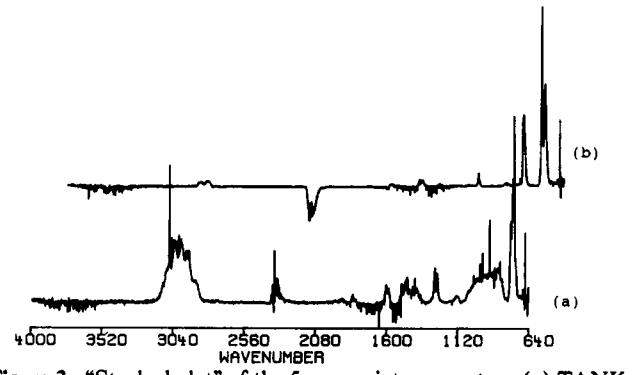

Figure 3. "Stacked plot" of the 5-ppm mixture spectra: (a) TANK 1 and (b) TANK 3.

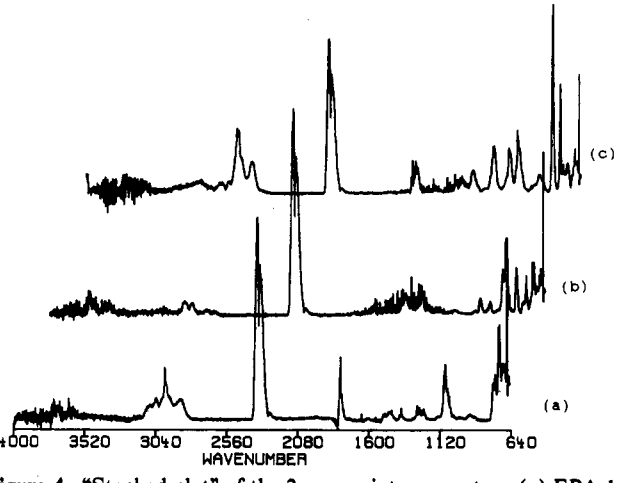

Figure 4. "Stacked plot" of the 2-ppm mixture spectra: (a) EPA 1, (b) EPA 2, and (c) EPA 3.

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 carbon dioxide features dominating the 2350-cm^{-1} region of the 2-ppm spectra (Figure 4), will interfere with sample peak signal detection. Even negative background contributions, as in the (inverted) water features in the 1300–1900-cm^{-1} regions of TANK A, TANK C (Figure 1), and TANK E (Figure 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 empirically set at 0–15, 15–35, 35–75, and greater than 75 cm^{-1}, corresponding to very sharp, sharp, average, and broad bandwidths, respectively. The actual width values in wavenumbers, were determined by the instrument peak-picking 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 specify interactively up to 50 points on the spectrum base line by manipulating a spectral cursor. A linear interpolation was performed between these base-line points and the interpolated base line subtracted from the spectrum. The adjusted spectrum allows a lower peak-picking threshold to be set, thereby providing access to the smaller spectral features. This program was used on any of the reference and mixture spectra that originally had ill-behaved base lines.

Program Description. The IRBASE and MIXIR systems have been adequately described elsewhere,[17,18] and so will be only briefly summarized here. IRBASE is a knowledge-based system that 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 that 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 interpret adaptively an unknown sample, using information gained from the user and determined by MIXIR throughout the analysis. For example, MIXIR can adaptively compensate for changed 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 that 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.[18] 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 1200–2100 cm^{-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 that provides a dynamic selection of bands for queries, based on the results from a spectral prescan. 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 20 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 that can be matched to these queries is then determined (I_{max}), and the minimum reference intensity desired is taken as

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

The MIXIR integral intensities are normalized to range from 0 to 20. Equation 1 therefore allows MIXIR to discard

reference peaks that 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 with 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 set limits adaptively.

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_{av}). 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:

$$I_{HI} = (R_{av}I_0) + 4 \quad (2)$$

or 20, whichever is larger, and

$$I_{LO} = (R_{av}I_0) - 4 \quad (3)$$

or 1, whichever is smaller.

This procedure allows MIXIR to set intensity windows to levels that 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 that 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 that cannot be attributed to compounds with scores greater than or equal to 0.20. The compound with the highest score that 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 by 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 data set consisted of four 5-component mixtures, one 2-component mixture, and one 6-component mixture (Table I).

The entire set of mixture spectra was interpreted by 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 expectation thresholds ranging from 0.40 to −0.50 was tabulated. (b) The average score of the compounds that were present in the mixtures and absent from the mixtures was tabulated.

Table I. 50-ppm Vapor-Phase Mixture Constituents and Concentrations

mixture	components	concentration[a] (ppm)
TANK A	toluene	46.8
	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	2-propanol	48.5
	ethyl ether	48.7
	3-chloropropane	49.1
	styrene	55.1
	ethylbenzene	50.8
	freon-11	51.1

[a] Analyzed by GC by Scott Specialty Gases.

An early version of the vapor-phase MIXIR system was first obtained by creating a program that 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[18] 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 data set might therefore be −0.10, where only three 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 that could be used as a prefilter for quantitative least-squares analysis.

Table II. Reference Compounds in the Vapor-Phase MIXIR Knowledge Base

1,3-butadiene	ethoxy ethanol
acetonitrile	ethyl ether
acetone	ethylene oxide
acetylaldehyde	freon-11
acrylonitrile	freon-114
butyl acetate	freon-12
benzene	freon-13
chlorobenzene	*n*-hexane
bis(2-chloroethyl) ether	2-propanol
2-butanone	methylene chloride
chloroform	4-methyl-2-pentanone
3-chloropropene	perchloroethylene
o-chlorotoluene	propylene oxide
cyclopentane	pyridine
carbon tetrachloride	styrene
1,2-dibromoethane	1,1,2-trichloroethane
1,2-dichloroethane	1,1,1-trichloroethane
1,1-dichloroethane	trichloroethylene
1,1-dichloroethene	tetrahydrofuran
dimethyl disulfide	toluene
1,4-dioxane	vinyl chloride
ethyl acetate	*o*-xylene
ethylbenzene	

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 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 peak-picked 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 II).

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 ben 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 III). 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 IV), in some cases it depressed the score of a compound actually 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

Table III. Summary Results on 50-ppm Mixtures, Using Null Intensity Windows; with and without Autocycling Option

level	false positives	false negatives	correct decisions
	With Autocycling[a]		
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[b]		
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

[a] Average score for the compounds which are present: 0.624. Average score for the compounds which are absent: −0.843. [b] Average score for the compounds which are present: 0.229. Average score for the compounds which are absent: −0.777.

Table IV. Abbreviated Score Reports for the 50-ppm Mixture TANK B, Obtained Using Null Intensity Windows; with and without Autocycling[a]

compound	score	peaks matched	peaks sought
	With Autocycling		
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-trichlorothane	0.240	3	12
vinyl chloride*	0.127	4	12
bis(2-chloroethyl) ether	0.113	3	12
o-chlorotoluene	−0.280	3	12
chlorobenzene	−0.679	4	12
(35 remaining compounds)			
	Without Autocycling		
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(2-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)			

[a] Asterisks mark actual mixture components.

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 that were matched had low uniqueness weightings, due to other mixture components that had coincident information, such as 4-methyl-2-pentanone. A spec-

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

compound	score	peaks matched	peaks sought
With Dynamic Query Selection			
freon-11*	0.999	2	2
styrene*	0.999	3	5
chloroform	0.272	1	3
bis(2-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
2-propanol*	−0.144	4	12
acetonitrile	−0.368	1	5
toluene	−0.371	2	5
(remaining compounds)			
Without Dynamic Query Selection			
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(2-chloroethyl) ether	0.028	3	12
ethylbenzene*	−0.079	5	12
ethoxy ethanol	−0.127	5	12
tetrahydrofuran	−0.134	3	11
2-propanol*	−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)			

[a] Asterisks mark actual mixture components.

troscopist 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; however, the distinction between the two interpretations is competition of compounds for the unknown spectral features. It is assumed therefore that a (fortuitous) screening by the dynamic intensity windows of several compounds with 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 that 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.

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

level	false positives	false negatives	correct decisions
With Auto-Peak Justification[a]			
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[b]			
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

[a] Average score for the compounds which are present: 0.271. Average score for the compounds which are absent: −0.789. [b] Average score for the compounds which are present: 0.206. Average score for the compounds which are absent: −0.792.

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 V presents abbreviated score reports for TANK F, using default intensity 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) benefited 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(2-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 VI. 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 that were improved 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 results. This indicates that peak intensity information in mixtures cannot generally be considered significant, 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 that were absent from the mixtures. Close examination of the results indicated that the conclusions which MIXIR reached, even when incorrect, were reasonable based

Table VII. 5-ppm Vapor-Phase Mixture Constituents and Concentrations

mixture	components	concentration[a] (ppm)
TANK 1	acrylonitrile	5.8
	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 VIII. Summary Results on the 5-ppm Mixtures Obtained Using Null Intensity Windows[a]

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

[a] Average score for the compounds which are present: 0.423. Average score for the compounds which are absent: −0.852.

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 VII). 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 VIII. 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 (Figure 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 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 Figures 1 and 2 with Figure 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.

Table IX. 2-ppm Vapor-Phase Mixture Constituents and Concentrations

mixture	components	concentration[a] (ppm)
EPA 1	tetrahydrofuran	2.3
	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
	2-propanol	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.

2-ppm Results. The mixture constituents for the 2-ppm mixtures are presented in Table IX. There were three such samples: one 6-component mixture (EPA 1), one 5-component mixture (EPA 2), and one 11-component mixture (EPA 3). The signal to noise ratio of these spectra was poor, as can be seen in Figure 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 11 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 (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 peak-picker. This situation cannot be cured by any interpretation logic—if no corresponding bands are detected, then the compound cannot be

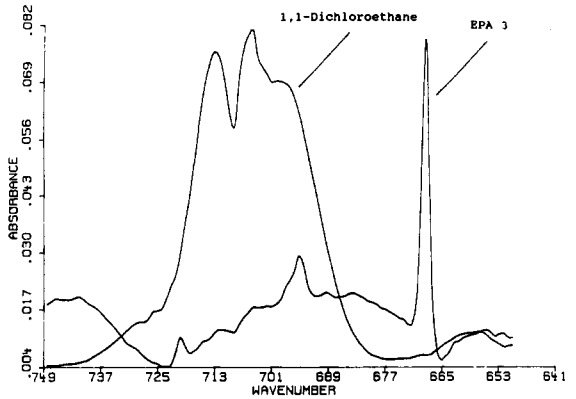

Figure 5. Portion of the C–Cl stretching region for the 1,1-dichloroethane reference spectrum and the 2-ppm mixture spectrum EPA 3.

Table X. Summary Results on the 2-ppm Mixtures Obtained Using Default Intensity Windows[a]

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

[a] Average score for the compounds which are present: 0.190. Average score for the compounds which are absent: −0.769.

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 X). 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 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.

ACKNOWLEDGMENT

This work was supported by Grant 1-R01-OH02404-01 from the National Institute for Occupational Safety and Health, Centers for Disease Control, and by a Dissertation Fellowship from the University of New Hampshire.

REFERENCES AND NOTES

(1) Liddell, R. W.; Jurs, P. C. *Anal. Chem.* **1974**, *46*, 2126.
(2) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1983**, *55*, 1288.
(3) Delaney, M. F.; Hallowell, J. R., Jr.; Warren, F. V., Jr. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 27.
(4) Zupan, J. E.; Munk, M. E. *Anal. Chem.* **1986**, *58*, 3219.
(5) Fields, R. E., III; White, R. L. *Anal. Chem.* **1987**, *59*, 2709.
(6) Dupuis, P. F.; Cleij, P.; Van't Klooster, H. A.; Dijkstra, A. *Anal. Chim. Acta* **1979**, *112*, 83.
(7) Bjerga, J. M.; Small, G. W. *Anal. Chem.* **1990**, *62*, 226.
(8) Nyden, M. R.; Pallister, J. E.; Sparks, D. T.; Salari, A. *App. Spectrosc.* **1987**, *41*, 63.
(9) de Haseth, J. A.; Azarraga, L. V. *Anal. Chem.* **1981**, *53*, 2292.
(10) Cooper, J. R.; Wilkens, C. L. *Anal. Chem.* **1989**, *61*, 1571.
(11) Rasmussen, G. T.; Isenhour, T. L.; Lowry, S. R.; Ritter, G. L. *Anal. Chim. Acta* **1978**, *103*, 213.
(12) Donahue, S. M.; Brown, C. W.; Obremski, R. J. *Appl. Spectrosc.* **1988**, *42*, 353.
(13) Donahue, S. M.; Brown, C. W.; Caputo, B.; Modell, M. D. *Anal. Chem.* **1988**, *60*, 1873.
(14) Ying, L. S.; Levine, S. P. *Anal. Chem.* **1989**, *61*, 677.
(15) Strang, C. R.; Levine, S. P. *Am. Ind. Hyg. Assoc. J.* **1989**, *50*, 78.
(16) Xiao, H. K.; Levine, S. P.; D'Arcy, J. B. *Anal. Chem.* **1989**, *61*, 2708.
(17) Wythoff, B. J.; Tomellini, S. A. *Anal. Chim. Acta* **1989**, *227*, 343.
(18) Wythoff, B. J.; Tomellini, S. A. *Anal. Chim. Acta* **1989**, *227*, 359.
(19) Ying, L. S.; Levine, S. P.; Tomellini, S. A. *Computer Enhanced Analytical Spectroscopy*; Plenum Press: New York, 1990; Vol. e.
(20) Savitsky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627.
(21) Enke, C. G.; Nieman, T. A. *Anal. Chem.* **1976**, *48*, 705A.