Multivariate Curve Resolution of Wavelet and Fourier Compressed Spectra

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 2001

Impact Factor: 5.64 · DOI: 10.1021/ac000956s · Source: PubMed

3 AUTHORS, INCLUDING:

Peter B Harrington
Ohio University
180 PUBLICATIONS **2,182** CITATIONS

SEE PROFILE

Anal. Chem. **2001,** *73,* 3247−3256

Multivariate Curve Resolution of Wavelet and Fourier Compressed Spectra

Peter de B. Harrington,*,† Paul J. Rauch,‡ and Chunsheng Cai§

Center for Intelligent Chemical Instrumentation, Chemistry Department, Ohio University, Athens, Ohio 45701-2979, Schaffner Manufacturing Company, Inc., 21 Herron Avenue, Pittsburgh, Pennsylvania 15202, and Aventis Pharmaceuticals, Mail Stop: C1-M0336 10236 Marion Park Drive, Kansas City, Missouri 64137.

The multivariate curve resolution method SIMPLe to use Interactive Self-Modeling Mixture Analysis (SIMPLISMA) was applied to Fourier and wavelet compressed ion-mobility spectra. The spectra obtained from the SIMPLISMA model were transformed back to their original representation, that is, uncompressed format. SIMPLISMA was able to model the same pure variables for the partial wavelet transform, although for the Fourier and complete wavelet transforms, satisfactory pure variables and models were not obtained. Data were acquired from two samples and two different ion mobility spectrometry (IMS) sensors. The first sample was thermally desorbed sodium γ-hydroxybutyrate (GHB), and the second sample was a liquid mixture of dicyclohexylamine (DCHA) and diethylmethylphosphonate (DEMP). The spectra were compressed to 6.3% of their original size. SIMPLISMA was applied to the compressed data in the Fourier and wavelet domains. An alternative method of normalizing SIMPLISMA spectra was devised that removes variation in scale between SIMPLISMA results obtained from uncompressed and compressed data. SIMPLISMA was able to accurately extract the spectral features and concentration profiles directly from daublet compressed IMS data at a compression ratio of 93.7% with root-mean-square errors of reconstruction <3%. The daublet wavelet filters were selected, because they worked well when compared to coiflet and symmlet. The effects of the daublet filter width and compression ratio were evaluated with respect to reconstruction errors of the data sets and SIMPLISMA spectra. For these experiments, the daublet 14 filter performed well for the two data sets.

Advances in analytical instrumentation have led to measurement systems that generate large quantities of data for single samples. Examples include hyphenated systems that include separation stages coupled to multichannel detectors, such as liquid chromatography−mass spectrometry (LC−MS) and LC-nuclear magnetic resonance spectrometry (LC−NMR). Some advances, such as the use of time-of-flight MS, furnish more resolution elements and can collect spectra at faster rates. Large volumes of data may be unwieldy for further online processing, especially if the processing is to be accomplished in real time.

Another area of advancement is the miniaturization of chemical sensors. As sensors decrease in size and cost, they will find widespread use outside of the controlled environment of the analytical laboratory. The analysis of complex samples in intricate environments will also force the replacement of the classical approach of signal averaging a stable sensor response with dynamic modeling methods. The advantage of modeling the dynamic sensor response is that temporal information may be exploited to resolve analytes in a mixture or to correct instrument drift resulting from changing ambient conditions. The use of dynamic modeling requires storing individual spectra or data objects. As sensors decrease in size, their storage and processing capabilities may become limited. For some applications, the data may be transmitted from sensors using wireless communication. Bandwidth limits may be encountered that may restrict the rate at which data can be conveyed, thereby necessitating the use of compression.

IMS sensors are amenable to miniaturization. These instruments are routinely used at airports for screening hand luggage for explosives, and have broad application for forensic, environmental, process, and industrial hygiene monitoring. Hand-held IMS sensors are commercially available that furnish detection limits below one part per million.

By modeling a data set using methods such as principal component analysis (PCA) and SIMPLISMA, the model can provide a clear and complete perspective of large and intricate sets of measurements. Multivariate curve resolution methods build mathematical models from variations in a data set that resolve spectra (i.e., curves). The models estimate the number of independent components (i.e., analytes). Overlapping peaks in the spectra may be modeled as separate components. Models furnish clear perspectives into complex trends in data.[1] Modeling methods should be used routinely for spectrometric and multichannel measurements.

Compression methods may be used to reduce the size of the spectra so that low capacity storage can be maintained and computational burdens alleviated. In many cases, compressing the data can also improve its quality by removing high-frequency noise components. In many compression methods, the data when

* Corresponding author.
† Ohio University.
‡ Shaffner Manufacturing Company, Inc.
§ Aventis Pharmaceuticals.

(1) Harrington, P. B.; Reese, E. S.; Rauch, P. J.; Hu, L.; Davis, D. M. *Appl. Spectrosc.* **1997**, *51*, 808−816.

10.1021/ac000956s CCC: $20.00
Published on Web 05/26/2001

compressed are rendered unintelligible. Once the data are compressed, the compressed data may be modeled, and the models can be transformed back to their original representation, that is, uncompressed format, while leaving the data compressed. The important consequence is that spectra can be compressed as they are acquired. The compressed data may be modeled by curve resolution methods so that the entire measurement can be evaluated with respect to the spectra and spectral changes with respect to time.

Multivariate curve resolution methods are useful for modeling multichannel data, specifically if the data objects have an implicit order, such as ordered with respect to time. Sensors used in process monitoring and chromatographic detectors have this attribute. SIMPLISMA,[2−4] is a multivariate curve-resolution method that has broad application. Perhaps the best application is the replacement of 3D surface plots of time-ordered multichannel data with two 2D figures[1]. SIMPLISMA has been implemented in near-real time and can be used for compression for which only the model components are stored.[5]

The Fourier transform (FT) was first suggested for compression of spectra in 1971,[6] and it has found broad application in analytical chemistry for signal processing.[7,8] The FT has been used to compress large volumes of data in the form of spectral libraries.[9−11] Two-dimensional Fourier compression (FC) was devised and applied to IMS data.[12] Singular value decomposition was applied to two-dimensionally Fourier compressed IMS data, and the eigenvectors were inverse-transformed back to the native domain.[13]

The wavelet transform (WT) is an old idea brought into use only recently. It is similar to the short-time or windowed FT and was developed to better utilize time and frequency information contained in the data being transformed.[14−18] The WT is considered an alternative to Fourier compression and processing.[19−21] FC has been used for IR[22,23] and UV−vis[24,25] spectra.

Another useful attribute for wavelet compression is that the transformed data retains both frequency and time information, unlike Fourier compression. This attribute is key to the success of SIMPLISMA applied directly to wavelet compressed data. PCA[26] has been applied to wavelet compressed data. Three studies have investigated partial least squares (PLS) regression applied to wavelet compressed data.[27−29] The wavelet transform has recently been implemented for IMS data as a compression and preprocessing method for artificial neural networks.[30,31]

The two IMS instruments that are used in this paper are handheld models that can be powered by 1.5 V batteries. GHB has recently been listed as a Schedule I drug and is abused as a date-rape and rave drug. IMS is a popular tool for the forensic analysis of drugs; however, the instruments used by law enforcement rely on the principle that drugs contain nitrogen and have relatively higher proton affinities than other organic materials.[32] GHB is interesting because it does not contain nitrogen. The Graseby chemical agent monitor (CAM), which is not the standard instrument for onsite drug detection, was used for this study.

The other sample was a liquid mixture of DCHA and DEMP. DEMP is a chemical warfare agent simulant. A Graseby MiniIMS (CT/AT) operated in positive ion mode was used to collect this chemical system, although this chemical system has been characterized using the CAM.[33] The MiniIMS differs from the CAM by having a pinhole inlet instead of a membrane interface and having a faster response to changes in analyte concentration. The mixture is interesting because DEMP completely suppresses the major peaks of DCHA through competitive charge inhibition with DEMP when detected with the CAM. This chemical system is extremely nonlinear in that a spectrum from a mixture of DCHA and DEMP resembles a DEMP spectrum. However, with dynamical modeling of these data with SIMPLISMA, DCHA can be detected in the presence of DEMP.

Theory. SIMPLISMA determines a set of pure variables in a data set with an implicit order.[2] SIMPLISMA will be briefly presented here, but it has been well-described in the literature.[1] The purity of a variable is a measure of selectivity for a particular analyte. The pure variable intensities are used to estimate the concentration profiles of the analytes. IMS data has features that vary nonlinearly with respect to concentration changes. These features may exhibit unique pure variables, so the term component is used instead of analyte. SIMPLISMA will model the data set by the changes in the spectral features and characterize these variances.

(2) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425−1432.
(3) Windig, W.; Heckler, C. E.; Agblevor, F. A.; Evans, R. J. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 195−207.
(4) Windig, W.; Stephenson, D. A. *Anal. Chem.* **1992**, *64*, 2735−2742.
(5) Rauch, P. J.; Harrington, P. B.; Davis, D. M. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 175−185.
(6) Wangen, L. E.; Frew, N. W.; Isenhour, T. L.; Jurs, P. C. *Appl. Spectrosc.* **1971**, *25*, 203−207.
(7) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, 1994.
(8) Horlick, G. *Anal. Chem.* **1972**, *44*, 943−947.
(9) Lam, R. B.; Foulk, S. J.; Isenhour, T. L. *Anal. Chem.* **1981**, *53*, 1679−1684.
(10) Azarraga, L. V.; Williams, R. R.; de Haseth, J. *Appl. Spectrosc.* **1981**, *53*, 466−469.
(11) Owens, P. M.; Isenhour, T. L. *Anal. Chem.* **1983**, *55*, 1548−1553.
(12) Cai, C.; Harrington, P. B.; Davis, D. M. *Anal. Chem.* **1997**, *69*, 4249−4255.
(13) Harrington, P. B.; Hu, L. *Appl. Spectrosc.* **1998**, *52*, 1328−1338.
(14) Massart, D. L.; Walczak, B. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 81−94.
(15) Alsberg, B. K.; Woodward, A. M.; Kell, D. B. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 215−239.
(16) Walczak, B.; Massart, D. L. *Trends Anal. Chem.* **1997**, *16*, 451−463.
(17) Depczynski, U.; Jetter, K.; Molt, K.; Niemooler, A. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 19−27.
(18) Leung, A. K. M.; Chau, F. T.; Gao, J. B. *Chemom. Intell. Lab Syst.* **1998**, *43*, 165−184.
(19) Barclay, V. J.; Bonner, R. F.; Hamilton, I. P. *Anal. Chem.* **1997**, *69*, 78−90.
(20) Walczak, B.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 81−94.
(21) Depczynski, U.; Jetter, K.; Molt, K.; Niemooler, A. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 151−161.
(22) Leung, A. K. M.; Chau, F. T.; Gao, J. B.; Shih, T. M. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 69−88.
(23) Chau, F. T.; Gao, J. B.; Shih, T. M.; Wang, J. *Appl. Spectrosc.* **1997**, *51*, 649−659.
(24) Ho, H. L.; Cham, W. K.; Chau, F. T.; Wu, J. Y. *Comput. Chem.* **1999**, *23*, 85−96.
(25) Chau, F. T.; Shih, T. M.; Gao, J. B.; Chan, C. K. *Appl. Spectrosc.* **1996**, *50*, 339−348.
(26) Walczak, B.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 39−50.
(27) Jouan-Rimbaud, D.; Walczak, B.; Poppi, R. J.; de Noord, O. E.; Massart, D. L. *Anal. Chem.* **1997**, *69*, 4317−4323.
(28) Trygg, J.; Wold, S. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 209−220.
(29) Alsberg, B. K.; Woodward, A. M.; Winson, M. K.; Jem, J. R.; Kell, D. B. *Anal. Chim. Acta* **1998**, *368*, 29−44.
(30) Cai, C.; Harrington, P. B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 874−880.
(31) Cai, C. Ph.D. dissertation, Ohio University, Athens, OH, 1999.
(32) Eiceman, G. A.; Karpas, Z. *Ion Mobility Spectrometry*, 1st ed.; CRC Press: Boca Raton, 1994.
(33) Harrington, P. B.; Wan, C.; Urbas, A. *Anal. Chem.* **2000**, *72*, 5004−5013.

The n rows of the data matrix (**D**) are spectra that are ordered with respect to time and the v columns are the resolution elements. The spectral scan number is reported instead of time, and each scan designates a row of **D**. The objective is to decompose **D** into two matrices

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{1}$$

for which the concentration profiles (**C**) comprises n rows and r columns; spectra (**S**) have v rows and r columns. The notation presents the components of the model as column vectors in **C** and **S**; thus, **S** is transposed to reconstruct the data matrix. The number of components in the model is r. The concentration profiles are the columns of **C**, and the spectra are the columns of **S**. The concentration profiles are not true concentrations, but profile changes in concentration, and have the same units as the intensities of the data. Thus, eq 1 can be applied to data acquired from semiquantitative or nonlinear measurements and still yield reliable models. However, the conversion of the concentration profiles to units of concentration would be problematic for nonlinear measurements.

The concentration profiles are initially modeled as the columns of **D** that yielded the r largest purity values. The spectra are calculated by regressing the data matrix onto the concentration profiles by

$$\mathbf{S} = \mathbf{D}^{\mathrm{T}}\mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1} \tag{2}$$

for which each column of **S** contains a spectrum that corresponds to each component. A new normalization method is proposed for SIMPLISMA as applied to compressed data. In previous work, the positive intensities of the spectra (i.e., columns of **S**) were summed, and this area was used to normalize the spectra.[1,34,35] This normalization was nonlinear in that only the positive intensities are used, and consequently, scale disparities were obtained for the SIMPLISMA concentration profiles and spectra. The SIMPLISMA spectra are normalized to unit Euclidean length by dividing each spectrum by the square root of its sum of the squared intensities. Because Fourier and wavelet compressions are linear and SIMPLISMA is linear, if the Euclidean length normalization is used, the concentration profiles and spectra from compressed data will correspond in scale with those obtained from SIMPLISMA of uncompressed data. The normalized spectra are used to generate new concentration profiles by

$$\mathbf{C} = \mathbf{D}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1} \tag{3}$$

In this work, the spectra are compressed, and SIMPLISMA is applied to the data in the Fourier and wavelet domains. The SIMPLISMA spectra are reconstructed to their original domain by use of the inverse transform.

Fourier compression can reduce the storage requirements per spectrum and yet retain the important IMS information in a format that is retrievable for feature extraction or other data analysis methods. The discrete Fourier transform was used and follows the procedure described in Cai et al.[12]

The forward and inverse FTs are scaled by the square root of the number of variables v.[13] This scaling causes the sines and cosines to form an orthogonal basis and preserves the scale of the transformed data. Both the forward and inverse transforms are scaled instead of following the convention of only scaling the inverse transform by v. The number of points in the uncompressed spectrum is v and is chosen as a power of 2. Once the data are in the frequency domain, the high-frequency data points may be removed to eliminate noise. The Fourier compression of IMS data acts as a low-pass filter by removing high-frequency noise. The data can then be converted back to the time domain for analysis by using the inverse FT.

The wavelet transform (WT) is similar in many respects to the FT and several tutorials exist.[15,16] The WT will be briefly described and will simply be presented as a pair of digital filters. One filter is a low-pass filter that yields smooth data, and the second filter is a high-pass filter that yields detail data analogous to the first derivative. The results of the filters produce smooth and detail points that characterize low and high-frequency information.

The special property of the wavelet filters is that the filter pairs form an orthogonal basis. The wavelet filter pair is applied recursively by convolving the wavelet filter points with the spectra, and each recursion is referred to as a level. For each level, a set of detail and smooth coefficients are obtained. The number of smooth points (i.e., output from the smooth wavelet function) is reduced by a factor of 2. In this work, the Daubechies wavelet family was used ,because this family performed well, as compared to the coiflet and symmlet families.

The data may be fully transformed using the wavelet filters, for which the number of levels is equal to the base 2 logarithm of the number of points in the spectrum (v). The fully transformed spectra will be composed of detail results and a single smooth data point. In the partial transform, the recursion stops once the number of smooth data points equals the desired number of compressed points. The number of levels is equal to the base 2 logarithm of the ratio of the initial to the compressed size of the data; therefore, the partial wavelet transform yields data composed of smooth filter results. Data from the complete wavelet transform contains the same information content as the data from the partial transform. The number of compressed points controls the information content. For example, compressing from a 1024-point spectrum to a 64-point spectrum yields a 94% compression. For a partial wavelet transform, only 4 levels of compression were computed instead of the 10 levels for a full wavelet transform.

Besides selecting the wavelet family, the number of points in the filter should also be defined. Analogous to the Savitsky-Golay filter,[36] the discrete wavelet filter can comprise various numbers of points (i.e., filter widths) that will be convolved with the spectra.

Both the compressed data and the SIMPLISMA spectra were reconstructed using the inverse FT and the inverse WT. For the FT reconstructions, it was beneficial to use a trapezoidal apodization function. The linear portion of the apodization was applied to the last 25% of the compressed data (e.g., for a 64-point compression, points 49−64 are linearly scaled to zero).

(34) Rauch, P. J.; Harrington, P. B.; Davis, D. M. *Anal. Chem.* **1998**, *70*, 716−723.

(35) Reese, E. S.; Harrington, P. B. *J. Forensic Sci.* **1999**, *44*, 68−76.

(36) Savitsky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−1639.

The root-mean-square error of reconstruction (RMSER) was determined for a data set and a set of spectra from a SIMPLISMA model as

$$RMSER = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{v}(d_{ij} - \hat{d}_{ij})^2}{nv}} \qquad (4)$$

for which n is the number of spectra, and v is the number of points in the spectrum. The ith reconstructed spectrum is given as $\hat{\mathbf{d}}_i$ and is obtained from the inverse transform. This measure is biased in that high-frequency noise components that may be removed during compression will contribute to the error. To overcome this bias, the average spectrum was calculated and compared with the average of the reconstructed spectra. The average spectra will have less indeterminate error.

A similar reconstruction error was calculated using eq 4 for the reconstructed SIMPLISMA spectra ($\hat{\mathbf{S}}$); however, the order of the SIMPLISMA components was observed to vary among the different compressions and filters. Therefore, the reconstructed spectra were sorted with respect to the component order of the uncompressed spectra to yield a minimum error. The sort routine minimized the Euclidean distance between $\mathbf{S}$ and $\hat{\mathbf{S}}$ with each $\hat{\mathbf{s}}_i$ used once and only once to compute the error. The SIMPLISMA spectra are relatively noise-free compared with the raw data spectra. If a reconstructed SIMPLISMA spectrum is missing artifacts (e.g., negative peaks) that are present in a SIMPLISMA spectrum from uncompressed data, a larger RMSER would be obtained; therefore, graphical examinations of the data are also important.

EXPERIMENTAL SECTION

IMS data was obtained in positive ion mode from a chemical agent monitor (CAM) type 482-301N or a Graseby MiniIMS CT/AT (v1.04) (Graseby Ionics, Ltd.; Watford, Herts, U.K.). These instruments were modified in that their dopant cartridges were removed so that the reagent ions were protonated water clusters.

The CAM uses a silicone membrane interface between the sample inlet and the reaction region. The data were acquired from the CAM using the Graseby WASP acquisition program and their analogue-to-digital converter (ADC). The data were acquired on an Intel 80486 computer operating at 25 MHz with 16 MB of RAM operating under Microsoft Windows 3.11. The computer used for data analysis was a 300 MHz Pentium II with 256 MB of RAM. The source code was developed using Borland C++ version 5.02.

Twenty-two blank spectra were collected, after which the IMS instrument was allowed to sample vapor from ~0.1 g of sodium γ-hydroxybutyrate (GHB) [Sigma Chemical Co.; St. Louis, MO (lot 106H5018)] as it was heated. The GHB was heated by placing the sample into a beaker that was set on its side in a grooved aluminum block. The aluminum block was placed on a hot plate with a thermometer in the aluminum block under where the sample was placed. When the hot plate was turned on, the block and the beaker would gradually heat; the temperature was read from the thermometer as the CAM data was collected. In this analysis, 500 spectra were collected over the 40-min period. For the CAM measurement, an 80 kHz data acquisition rate was used, and a spectrum was collected once every 5 s. The temperature increased from 22.3°C to 210 °C during the acquisition of the 500 spectra.

A solution of *N*,*N*-dicyclohexylamine (DCHA), 99%, [lot 02810MX, Aldrich Chemical Co.; WI] and diethyl methylphosphonate (DEMP), 98%, [batch 10010573; Lancaster, U.K.] was made by adding 1-mL volumes of the liquids into a 5-mL sample vial. The sample vial was placed 1 cm away from the inlet of the MiniIMS for several seconds and removed. This process was repeated several times during a 190-s data collection period to yield 1006 spectra.

The MiniIMS was interfaced to a National Instruments DAQ-Card 1200, and the data was collected using a LabVIEW virtual instrument on a Pentium 100 MHz notebook computer operating under Microsoft Windows 98 SE. The MiniIMS was operated in positive ion mode. Each spectrum was sampled at an 80 kHz data acquisition rate, and 5 spectra were acquired per second.

For each spectrum, the baseline was estimated using the average of the points in the range of 1.5 to 3.0 ms. This average was subtracted from all of the points in the spectrum to correct for baseline drift. The Fourier and wavelet compressions used 1024 points of each spectrum from 3 to 15.8 ms for the CAM and from 6.0 to 18.8 ms for the MiniIMS so that the total number of points in each spectrum was equal to a power of 2.

For the FT, the spectra were represented in wrap-around format[13] so that each spectrum was extended to 2048 points and represented as an even function. This manipulation of the data before the FT furnished a discrete cosine transform that yielded a transformed spectrum without complex numbers (i.e., real). A trapezoidal apodization function was used for the inverse FT of the SIMPLISMA components.

The wavelet transform required less preprocessing, because it does not include an imaginary component. The Daubechies family was selected because of the success of other comparative studies with IMS data.[30,31] When SIMPLISMA was implemented, three components were used; the α-value was 5% of the most intense peak of the mean spectrum.

RESULTS AND DISCUSSION

SIMPLISMA is an effective and efficient method to visualize data.[1] Figure 1 gives the 3 components that were obtained from SIMPLISMA on the GHB thermal desorption data set. The spectra are presented with both drift time and reduced mobility. The reduced mobilities were obtained from an internal standard of 2,4-lutidine (1.41 cm^2/Vs).[37] The reduced mobilities remove some of the instrumental variations caused by changes in the drift tube potential drop, temperature, humidity, and pressure. The orders of the SIMPLISMA components in the figure legends correspond to SIMPLISMA purity values. The fourth component characterized baseline variations, so three components were selected for the model. Including a fourth component had a marginal effect on the third component or peak 1. Using four components removed the baseline shift seen at high drift time (13 ms) from this peak.

Figure 2 gives concentration profiles with respect to scan number and temperature. For this experiment, each scan number also corresponds to 1 s of elapsed time. Again, the components

(37) Harden, C. S.; Schoff, D. B.; Davis, D. M.; Ewing, R. B. *Extended Abstract*; American Society for Mass Spectrometry: Palm Springs, CA, June 1997; 475.

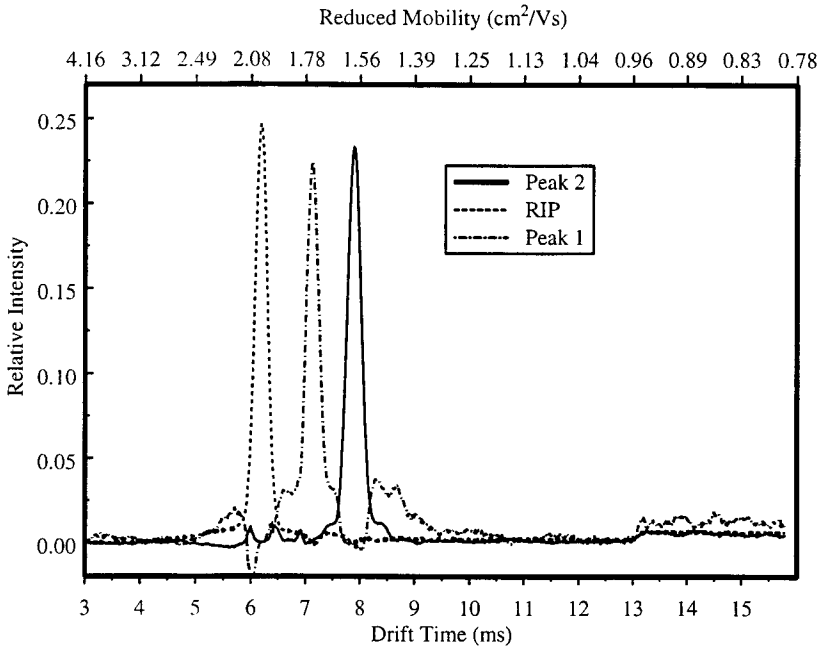

Figure 1. The SIMPLISMA spectra from positive ion mobility GHB data that was not compressed. The order in the legend corresponds to SIMPLISMA purity.

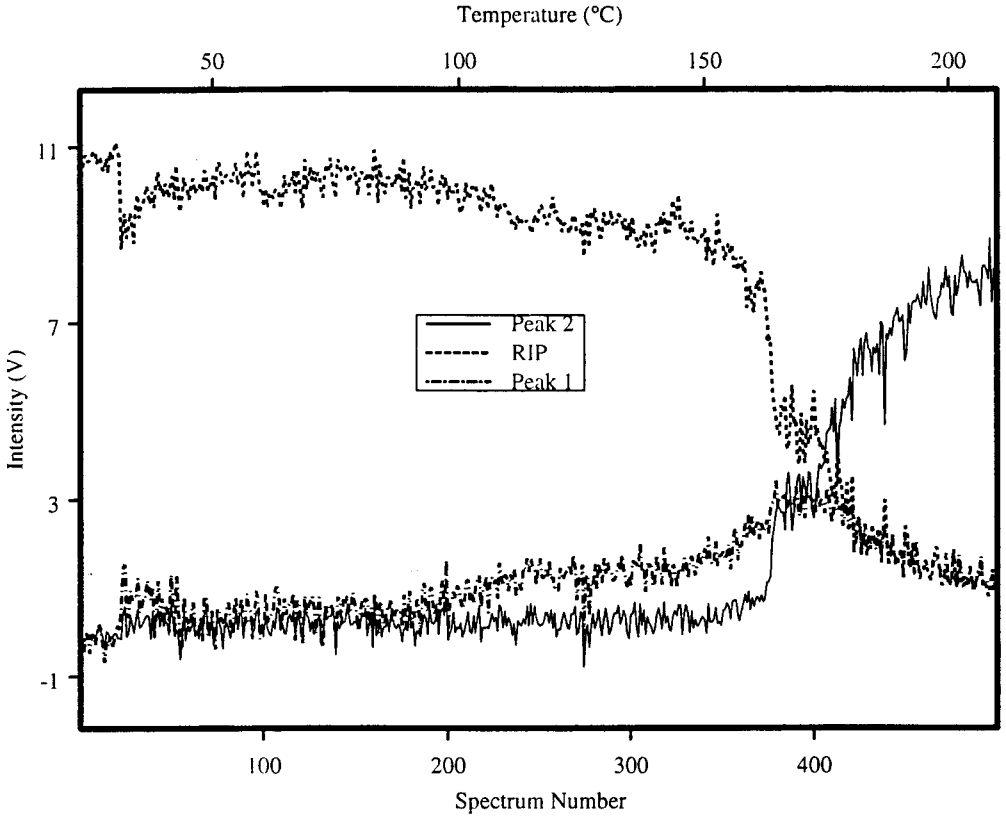

Figure 2. SIMPLISMA concentration profiles for GHB data that was not compressed.

are sorted by purity value in the legends. These two figures give a complete picture of the experiment. The reactant ions peak (RIP), which comprises protonated water cluster ions, is initially large for the first 20 spectra. When the experiment begins at scan 20, the water peak decreases when peak 1 appears at 7.2 ms. In a 3D surface plot, it would be very difficult to see this subtle trend. After 100 °C, the second peak at 7.9 ms begins to increase. Because there is competitive charge transfer (only a fixed amount of ions present during each measurement), peak 1 decreases as peak 2 increases.

The GHB data set collected using the CAM was used to compare SIMPLISMA models obtained from Fourier and wavelet compressed data. The two GHB features and the RIP are prominent in this figure. SIMPLISMA models these features as

Table 1. Comparison of Fourier and Wavelet Compressions[a]

data size	daublet filter points	RMSER Fourier compressed V	RMSER wavelet compressed V	RMSER FT SIMPLISMA spectra	RMSER full WT SIMPLISMA spectra	RMSER partial WT SIMPLISMA spectra
16	6	0.201	0.162	0.027	0.026	0.025
32	6	0.136	0.087	0.024	0.028	0.024
64	14	0.047	0.026	0.032	0.030	0.004
128	30	0.016	0.013	0.022	0.035	0.001

[a] A full-scale signal from the CAM was 2.7 V. The wavelet results are for the daublet optimal filter width. A full-scale SIMPLISMA peak is 0.25.

components in these data. This data set is well-characterized by 3 SIMPLISMA components. When four components were used, the fourth component characterized noise.

Both the Fourier and wavelet compressions compressed the data to a power of 2. Although Fourier compression can compress data to an arbitrary number of points, wavelet compression to a number of points other than a power of 2 would bias the results. Each level of the WT corresponds to a change in frequency or scale. The number of smooth points is reduced by a factor of 2 at each level. For wavelet compression, the resultant smooth information from each level is used for further compression. If points are removed from these levels, information is discarded pertaining to drift time. Starting with 1024 points after 4 levels of compression, the resultant data would be 64 points. For the partial wavelet transform, if the last 4 points were removed from the compressed data at this level to furnish a 60-point compression, then information at the longer drift times (i.e., 15−15.8 ms) would be removed. Removing these points may have little influence on the reconstruction error for the GHB data, because no IMS peaks occur in this range; however, peaks could occur in this range for other substances, and compressing by removing these data could bias the results in favor of wavelets. Therefore, only compressions to powers of 2 were evaluated.

This data set was optimized for 27 different wavelet filters, including the coiflet and symmlet filters and a compression range from 25 to 98.4%. Table 1 presents the results from the best daublet wavelet filter and results obtained from FC data. For each compression level, two figures of merit were generated. The first measured the RMSER between the spectra and the reconstructed spectra. Both wavelet and Fourier compressions remove high-frequency noise components. The RMSERs are biased, in that reconstructed spectra that have less high-frequency noise could have larger values. The second measures the RMSER between the set of SIMPLISMA spectra from the uncompressed IMS data and a sorted set of reconstructed SIMPLISMA spectra obtained from the compressed data.

The daublet, symmlet, and coiflet gave comparable results. The daublet was selected for further study, because they showed less variability and lower errors as compared to the other two wavelet families. The largest intensity in this data set is 2.5 V; thus, an RMSER of 0.05 V is tolerable and represents a 2% error, which is within the variation of the measurement. The results in Table 1 indicate that a 94% compression obtained by compressing the spectra from 1024 to 64 points furnished a good compression efficiency with relatively low errors. Although the RMSER results for the wavelet are less than those for the Fourier compression, it is important to note that wavelet filter width was optimized to yield the lowest RMSER. The Fourier results are all less than the filter widths (excluding daublet 2) that yielded the largest RMSER for the different levels of compression. Other work agrees, with the result that a 94% compression ratio is effective for IMS data collected using the CAM at an 80 kHz data acquisition rate.[30,38]

Although the RMSER results between the wavelet and Fourier compressions are comparable, the RMSER results for the SIMPLISMA spectra indicate that the partial wavelet compression is more effective than either the full-wavelet or Fourier compressions. The full-wavelet compression yielded results that were similar to the Fourier compression results. The IMS peaks have similar widths, so that they are convolved in the frequency domain, and in this domain, SIMPLISMA cannot obtain variables of sufficient purity to correctly model the components in the experiment. Impure variables result in negative and positive peaks in the SIMPLISMA spectra with corresponding positive and negative features in the concentration profiles. Note that the uncompressed IMS data do not contain any significant negative features. In the partial wavelet transform, the variables are expressed in drift time at a reduced resolution, so that pure variables may be obtained by SIMPLISMA.

Figure 1 gives the SIMPLISMA spectra obtained from the uncompressed IMS data. Figure 3 gives the reconstructed SIMPLISMA spectra obtained from the daublet 14 compressed data, which corresponds closely with the data obtained in Figure 1. Figure 4 gives the reconstructed SIMPLISMA spectra obtained directly from the Fourier compressed spectra. The negative peaks in the reconstructed spectra from the Fourier compressed data demonstrate that SIMPLISMA could not find pure variables in the frequency domain representation of the data. This result also corresponds to the data presented in Table 1. The SIMPLISMA RMSER exceeds the largest peak intensity in the uncompressed SIMPLISMA spectra, thus showing little correspondence between the reconstructed spectra and uncompressed spectra. Nearly identical results were obtained for the full-wavelet transform of the compressed data. The SIMPLISMA RMSER values do not increase with compression in Table 1 for the full-wavelet and Fourier compressed data, but are relatively constant.

The SIMPLISMA concentration profiles for the uncompressed data are given in Figure 2. The concentration profiles from the wavelet compressed data are given in Figure 5. These profiles correspond to pure variables and are virtually identical to those obtained from the uncompressed data. This correspondence is lost in the concentration profiles in Figure 6 that were obtained from the Fourier compressed data. It is not surprising to see

(38) Cai, C.; Harrington, P. B. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1161−1170.

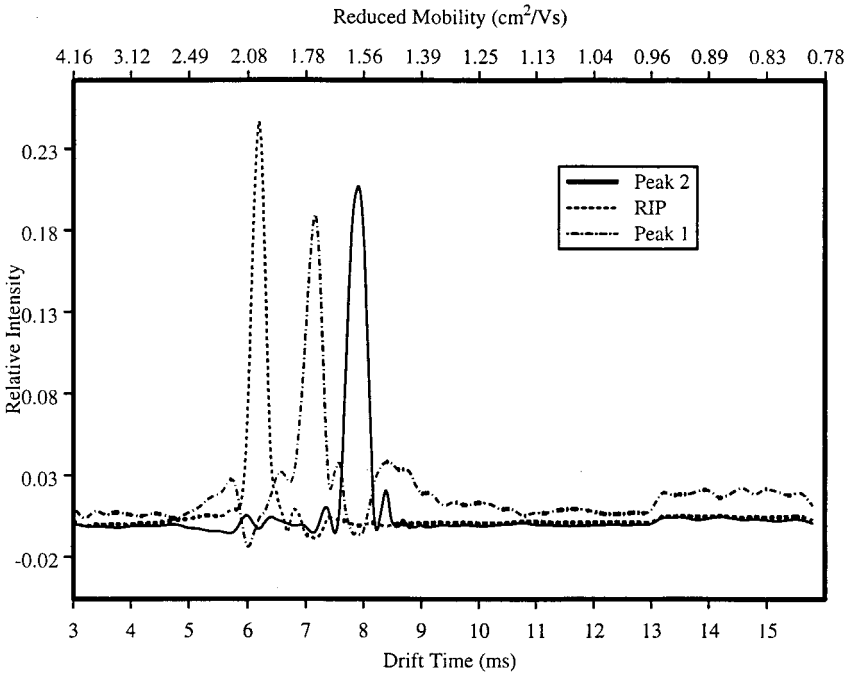

Figure 3. SIMPLISMA spectra for GHB data from 94% wavelet compressed data. The spectra were reconstructed using the inverse wavelet transform.

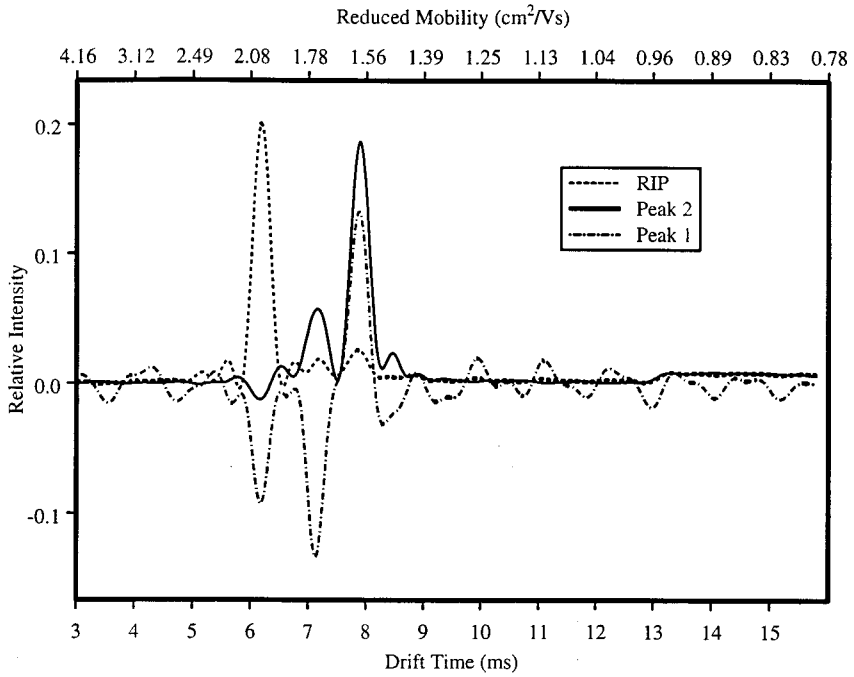

Figure 4. SIMPLISMA spectra for GHB data that has been Fourier compressed to 64 points (94%). The spectra were reconstructed using the inverse FT.

negative concentrations modeled from data in the Fourier domain, because the SIMPLISMA spectra also have negative peaks. Examination of the concentration profiles in Figure 6 and the spectra in Figure 4 suggests that alternating least squares equipped with nonnegative constraints might recover similar spectra and profiles as the uncompressed data. Alternating least squares with nonnegativity constraints helped to improve the SIMPLISMA spectra and concentration profiles, although the improvement was marginal.

Figure 7 compares SIMPLISMA spectra from 2 different compressions, 94% (i.e., 64 points) and 97% (i.e., 32 points), for the daublet 14 with the SIMPLISMA spectra from the uncompressed data. For display purposes, the SIMPLISMA spectra for the three components of the model were averaged. The 97% compression caused excessive distortion and peak shifts in the reconstructed spectrum. The 94% compression retained the key spectral features, including peak intensity, and reduced the noise content. Notice, however, that with the wavelet compression, the peak at 7.0 ms has shifted.

For IMS data collected at an 80 kHz sample rate, a level-4 compression that reduces a spectrum's size from 1024 to 64 points is efficient. The errors were computed for 500 GHB spectra

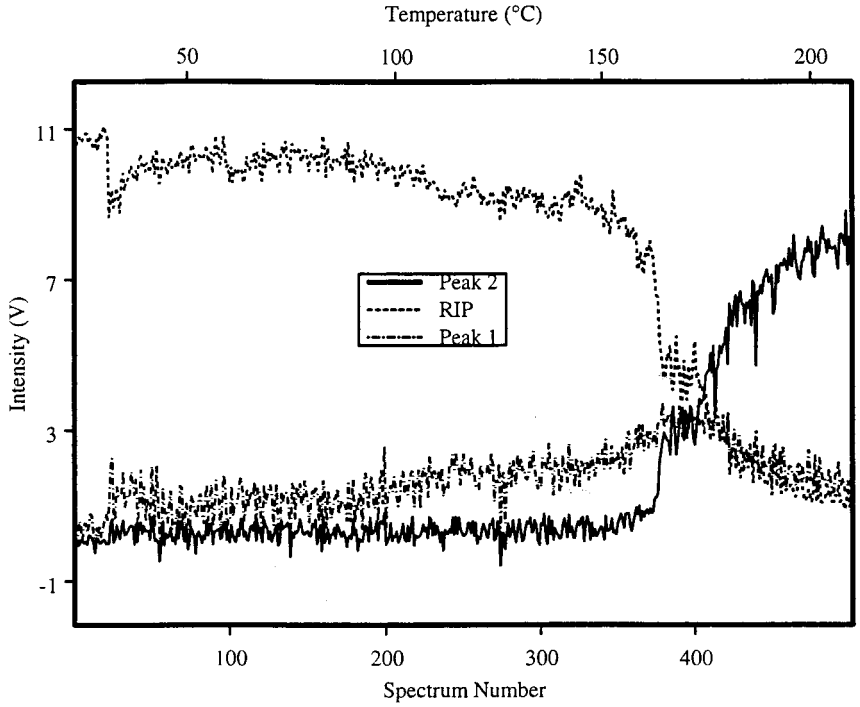

Figure 5. SIMPLISMA concentration profiles for the 94% compressed ion mobility GHB data using the daublet 14 filter. The concentration profiles are virtually identical to those obtained from the uncompressed data.

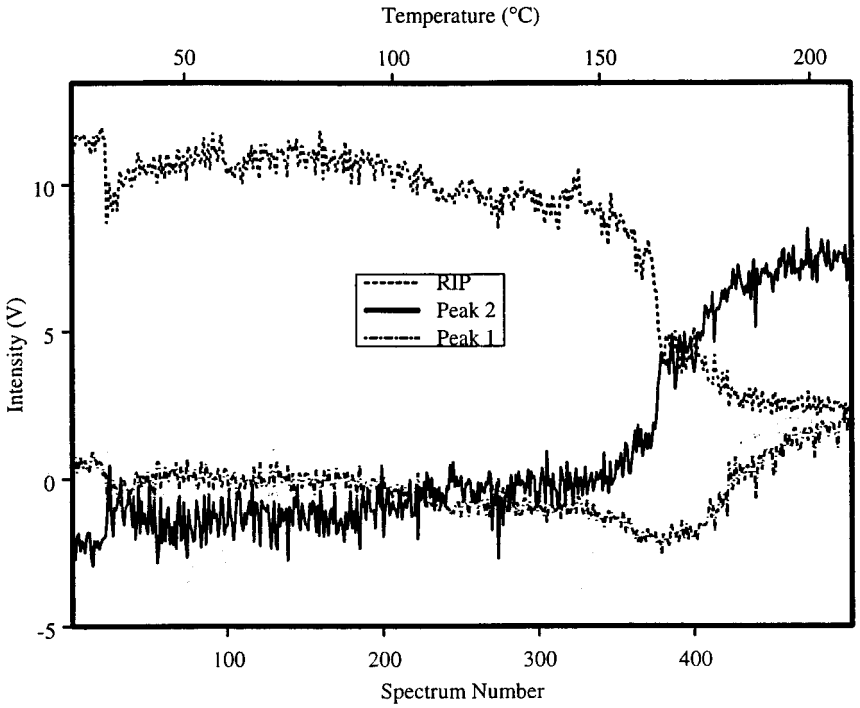

Figure 6. Concentration profiles from the 94% Fourier compressed ion mobility GHB data.

collected using the CAM and 1006 DCHA/DEMP spectra collected with the MiniIMS. The results for daublet filters of various widths at a level-4 compression are given in Table 2. The lowest RMSER values for the averaged spectra and the DCHA/DEMP data were obtained from the daublet 22 filter. For the averaged spectra, the daublet 14 filter RMSER of 0.007 V was slightly larger than the minimum RMSER of 0.005 V for the daublet 22 filter. The 0.007 V error is 3% of the full-scale signal.

An additional figure of merit was added that gives the average reconstructed errors. These results are tabulated with the GHB results in Table 2. The daublet 14 filter worked well with data collected on two different instruments and different chemical systems.

Figure 8 compares different reconstructed spectra obtained from the daublet 14 that yielded the minimum RMSER in Table 2 and the two nearest filter widths that yielded larger RMSER SIMPLISMA values (i.e., daublet 10 and daublet 24). The reconstructed SIMPLISMA spectra from the different daublet filter widths show that the daublet 14 reconstructed spectrum corresponds to the results obtained from the uncompressed spectra.

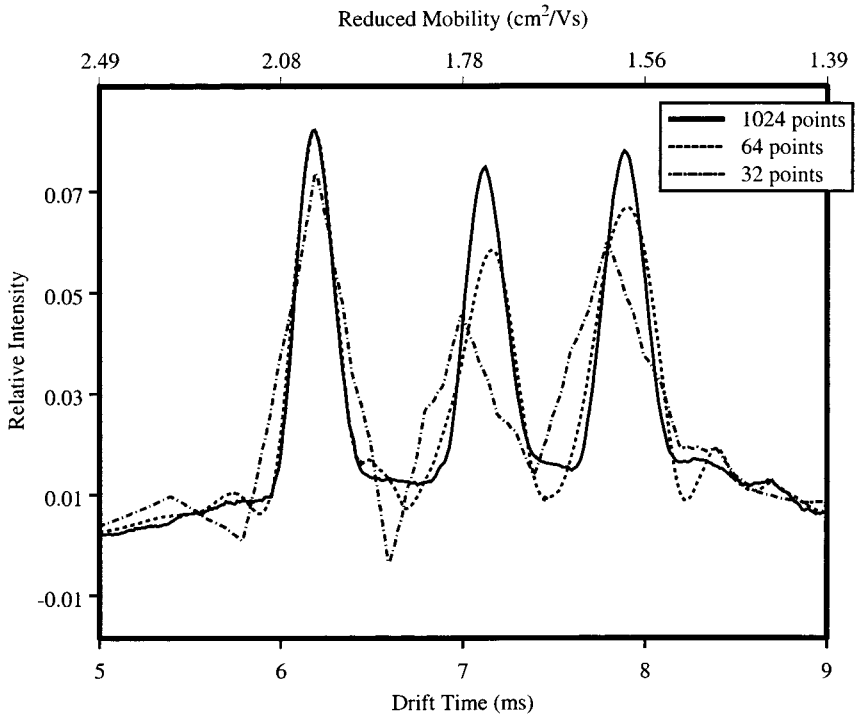

Figure 7. Wavelet reconstructions of SIMPLISMA spectra from different compressions. The solid line is the uncompressed GHB composite spectrum obtained by averaging the 3 SIMPLISMA spectra. The other spectra were generated from data compressed to 64 points (dashed line) and 32 points (dot−dashed line) using the optimal daublet filter widths of 14 and 6, respectively.

Table 2. Detail Set of Results from Daublet Compressed Data for the GHB Measurement with the CAM and a Mixture of DCHA/DEMP with the MiniIMS Instrument[a]

	RMSER					
	CAM GHB data			MiniIMS DCHA-DEMP data		
daublet filter points	wavelet compressed data, V	SIMPLISMA spectra	ave spectra V	wavelet compressed data, V	SIMPLISMA spectra	ave spectra V
2	0.101	0.010	0.089	0.053	0.006	0.051
4	0.057	0.006	0.047	0.027	0.003	0.026
6	0.038	0.006	0.028	0.019	0.002	0.018
8	0.056	0.006	0.049	0.020	0.002	0.019
10	0.061	0.019	0.054	0.018	0.002	0.017
12	0.044	0.004	0.037	0.012	0.002	0.011
14	**0.026**	**0.004**	**0.014**	0.008	0.001	0.007
16	0.041	0.005	0.033	0.013	0.002	0.012
18	0.054	0.005	0.047	0.015	0.002	0.014
20	0.049	0.004	0.042	0.012	0.001	0.011
22	0.031	0.004	0.023	**0.006**	**0.001**	**0.005**
24	0.030	0.017	0.020	0.008	0.001	0.007
26	0.045	0.011	0.038	0.013	0.002	0.012
28	0.050	0.018	0.044	0.013	0.015	0.012
30	0.040	0.018	0.033	0.009	0.011	0.007

[a] The full-scale signal on the CAM is 2.7 V; on the MiniIMS, is 1.6 V.

The daublet 14 corresponds well with the RIP, but the 2 GHB peaks are attenuated. The daublet 10 has an attenuated RIP, but corresponds with peak 2. All of the wavelet filters have difficulty with peak 1, which is also the smallest peak in the data set. The SIMPLISMA spectra are normalized, and the intensity information is given by the concentration profiles. For this peak, the reconstructed peaks also are shifted with respect to position, which could have important ramifications. Peak positions are important for identifying ions by their reduced mobilities; however, the peak positions for the daublet 14 were within 0.3%. The other filters' widths yielded a maximum deviation of 1.3% in peak positions, which are within the experimental error of the measurement. Interestingly, there were no shifts in peak positions for SIMPLISMA spectra obtained from the Fourier compressed data.

CONCLUSIONS

Wavelet compressed data may be modeled directly by multivariate curve resolution methods such as SIMPLISMA. The data were compressed using the partial wavelet transform. This method allows data to be compressed, as they are collected and facilitates real-time collection and the dynamic analysis of instrument response. The SIMPLISMA spectra may be reconstructed using

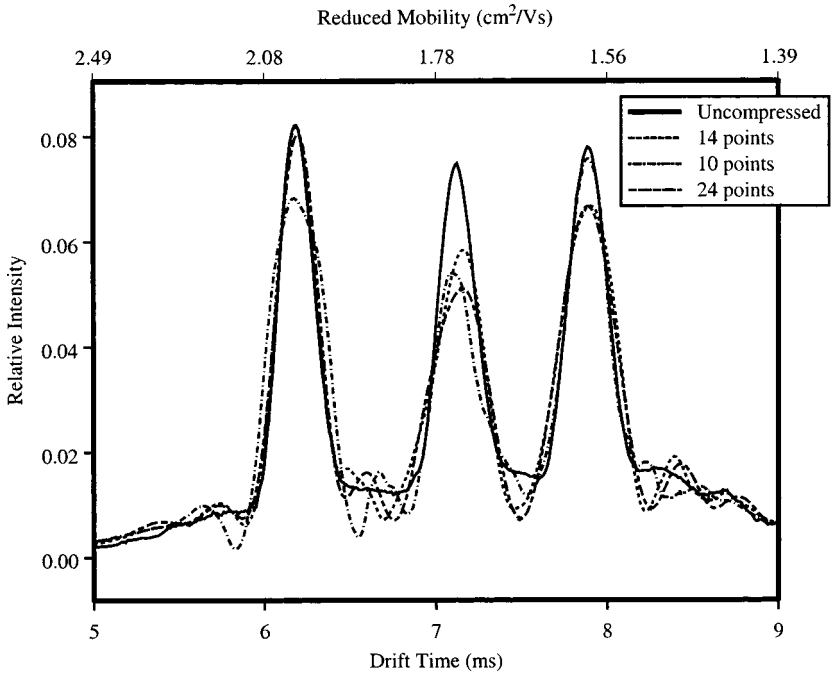

Figure 8. Effect of daublet filter widths on reconstructed spectra obtained from a 94%, 64-point compression of the GHB spectra. The solid line is the averaged SIMPLISMA spectra from uncompressed data; the dashed line, dot−dashed line, and dashed-dashed lines are designated reconstructed spectra from daublet 14, 10, and 24 filters, respectively.

the inverse wavelet transform so that they can be studied in their native domain. The results from SIMPLISMA applied to daublet 14 compressed data agree closely with the results obtained from the uncompressed data.

The same method applied to Fourier compressed data and fully transformed wavelet data did not work. Similar concentration profiles can be obtained, but the SIMPLISMA spectra were not representative of the components in the data. Pure variables in the frequency domain were not available to adequately resolve the features in the GHB or the DCHA/DEMP data sets. Because compressed data from a partial wavelet transform preserves some drift time information, similar pure variables and models were obtained. An added benefit is that the partial wavelet transform requires fewer calculations and is better-suited for real-time implementation than either the Fourier and full wavelet compression methods.

SIMPLISMA spectra acquired from wavelet compressed data may exhibit negligible errors in peak position and peak intensity. No peak shifts were observed in the SIMPLISMA spectra obtained from the Fourier compressed data; however, the models were poor and larger errors were obtained in the peak intensities.

Ion mobility spectra resemble data acquired from many other measurement methods. SIMPLISMA analysis of data in compressed formats allows the real-time analysis of high-resolution spectra or large collections of spectra that might otherwise be computationally prohibited. This approach may provide a method for modeling large collections of data, such as high-resolution spectral images or real-time screening of liquid chromatographic mass spectral or nuclear magnetic resonance data.

ACKNOWLEDGMENT

The U.S. Army ERDEC is thanked for the loan of the Graseby CAM and Mini CT/AT instruments. Aaron Mehay, Aaron Urbas, Tricia Buxton, Guoxiang Chen, and Libo Cao are thanked for their helpful discussions and comments.

Received for review August 11, 2000. Accepted May 2, 2001.

AC000956S