NONLINEAR WAVELET COMPRESSION METHODS FOR ION ANALYSES AND DYNAMIC MODELING OF COMPLEX SYSTEMS

A dissertation presented to
the faculty of
the College of Arts and Sciences of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Libo Cao
November 2004
This dissertation entitled
NONLINEAR WAVELET COMPRESSION METHODS FOR ION ANALYSES AND
DYNAMIC MODELING OF COMPLEX SYSTEMS

BY
LIBO CAO

has been approved for
the Department of Chemistry and Biochemistry
and the College of Arts and Sciences by

Peter de B. Harrington
Professor of Chemistry and Biochemistry

Leslie A. Flemming
Dean, College of Arts and Sciences
Nonlinear wavelet compression (NLWC) was developed and applied to ion mobility spectra from IMS instruments mounted in an unmanned aerial vehicle. To obtain a higher compression ratio, a two-dimensional nonlinear wavelet compression (2D-NLWC) was created and applied to the IMS data of chemical warfare simulants. However, modeling the data during the compression procedure itself is important. The algorithm of SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) and alternating least squares (ALS) combined with 2D-NLWC, was designed to analyze the IMS data of chemical warfare simulants.

Wavelet compression is beneficial when applied to mass spectra. Mass spectral profiles retain information regarding ion peak shapes. This information can be used to
differentiate spectra from isomers or other structurally similar compounds. During the traditional process of integrating raw mass spectra into histograms, some useful information regarding ion peak shapes may be lost. However, without such integration, it is difficult for analytical instruments to process the large quantity of raw data. Wavelet compression was used to reduce the large size of the mass profiles while at the same time retaining the useful information of ion peak shapes. This dissertation concluded that the mass profiles after wavelet compression provided better projected difference resolutions (PDRs) among xylene isomers than the histogram.

Approved:

Peter de B. Harrington
Professor of Chemistry and Biochemistry
Acknowledgments

First of all, I would like to express my deepest gratitude to my research adviser, Dr. Peter de B. Harrington, for his invaluable support and guidance during my stay at Ohio University. This dissertation could not be accomplished without his help and I would never have the chance to make my dream come true. I would also like to thank my dissertation committee members, Drs. Gary W. Small, Howard D. Dewald, Greg Van Patten, Glen P. Jackson and Larry Snyder, for their great help in my academic progress and research pursuits.

I would like to thank the Department of Chemistry and Biochemistry at Ohio University for offering me the opportunity to conduct my doctoral research. The Center for Intelligent Chemical Instrumentation at Ohio University is thanked for supporting the conference trips. The US Army ECBC and GeoCenters are thanked for the partial support of the research. The Research Corporation is thanked for the Research Opportunity Award. GE Ion Track Instruments, Inc. is thanked for their support. Smiths Detection Ltd. is thanked for their UAV data collection and software package for reading the .gcv files.

I would also thank the members in Dr. Harrington’s research group: Mariela Ochoa, Preshious Rearden, Matt Rainsberg, George Bota and Ping Chen for their comments, suggestions, support and friendship. Special thanks are given to Dr. Guoxiang Chen for his consistent help over the years and life-long friendship.

Metara Inc. is thanked for providing me the first great job in Silicon Valley I have dreamed of since I was a little girl. I thank for her trust and encouragement when she
hired me before my graduation. I thank for her generosity of paying all the expense of my defense trip. I also thank everyone in Metara Inc. for the friendship they have given me.

Lastly, I would like to thank my dad Yuanchang Cao, my mom Chengming Zhang and my brother Hanzhi Cao who provided the most caring and love in my life. Without their support and encouragement, I would never have made this far.
Table of Contents

Abstract ...............................................................................................................................3
Acknowledgments ...............................................................................................................5
List of Tables ....................................................................................................................10
List of Figures ................................................................................................................... 12
List of Abbreviations ........................................................................................................ 20

Chapter 1 Introduction............................................................................................... 23

1.1 Background Information ........................................................................... 23

1.2 Ion Mobility Spectrometry (IMS)........................................................................ 24

1.3 Theoretical Principles ............................................................................... 27

1.3.1 SIMPLe-to-use Interactive Self-modeling Mixture Analysis
(SIMPLISMA) .......................................................................................................... 27

1.3.2 Alternating Least Squares (ALS) ...................................................... 29

1.3.3 Wavelet Transform (WT) and Wavelet Compression (WC) ............ 30

1.4 The Research Objectives ........................................................................... 33

Chapter 2 Nonlinear Wavelet Transform of Ion Mobility Spectra from Ion Mobility Spectrometers Mounted in an Unmanned Aerial Vehicle .................................... 35

2.1 Introduction ........................................................................................................ 35

2.2 Theory ............................................................................................................... 37
2.3 Experimental Section ................................................................................ 38
2.4 Results and Discussion ............................................................................. 40
2.5 Conclusions ............................................................................................... 55

Chapter 3 Two-Dimensional Nonlinear Wavelet Compressed Ion Mobility Data ... 59
3.1 Introduction ............................................................................................... 59
3.2 Theory ....................................................................................................... 61
  3.2.1 LWC .................................................................................................. 61
  3.2.2 Two-dimensional NLWC .................................................................. 62
  3.2.3 Reconstruction Error ......................................................................... 64
3.3 Experimental Section ................................................................................ 65
3.4 Results and Discussion ............................................................................. 69
3.5 Conclusions ............................................................................................... 96

Chapter 4 SIMPLISMA and ALS Applied to Two-Dimensional Nonlinear Wavelet
Compressed Ion Mobility Data ............................................................................. 97
4.1 Introduction ............................................................................................... 97
4.2 Theory ..................................................................................................... 101
  4.2.1 ASIMPLISMA ................................................................................ 101
  4.2.2 Two Dimensional-NLWC-ALS ...................................................... 102
  4.2.3 Evaluation Reconstruction Efficiency and Efficacy ....................... 105
4.3 Experimental Section ................................................................. 107
4.4 Results and Discussion .............................................................. 110
4.5 Conclusions ............................................................................. 135

Chapter 5 Pattern Recognition of Unintegrated Mass Spectra .............. 138
5.1 Introduction ............................................................................. 138
5.2 Theory ................................................................................... 140
5.3 Experimental Section ............................................................ 142
5.4 Results and Discussion ........................................................... 144
5.5 Conclusions ............................................................................. 164

Chapter 6 Summary and Future Work .................................................. 165

Bibliography .................................................................................. 168

Appendix A: Publications ................................................................. 178

Appendix B: Presentations ............................................................... 179

Appendix C: List of Common Chemical Warfare Simulants ................... 180
List of Tables

Table                                                  Page
Table 2.1 Root mean square reconstruction error for nonlinear compression of data set 2 52
with respect to wavelet filter.* .................................................................
Table 3.1 Data set collection information for the three CAM data sets and three 67
ITEMISER data sets used. .............................................................................
Table 3.2 Signal-to-noise ratio for a two-dimensional synthetic centered Gaussian with 71
intensity of 1.0 and standard deviation of 10 added to random normal deviates.*... Table 3.3 Compression factor, minimum SE$_{RMS}$, RE$_{RMS}$, and the wavelet types for 80
minimum SE$_{RMS}$ in drift Time dimension and data acquisition dimension.*........
Table 3.4 Minimum E$_{RMS}$ and RE$_{RMS}$ for the reconstructed three ITEMISER data sets for 83
several two-dimensional compression factors with all 484 wavelet filter combination
applied to nonlinear compression.* ............................................................... Table 3.5 The comparison of the reduced mobility (K$_0$) for all the peaks in CAM and 92
ITEMISER data sets using ALS spectra.* ...................................................... Table 3.6 ANOVA tables for linear wavelet compression and nonlinear wavelet 94
compression of CAM.* ............................................................................. Table 3.7 ANOVA tables for linear wavelet compression and nonlinear wavelet 95
compression of ITEMISER.* .................................................................... Table 4.1 Data set collection information for the three CAM data sets and three 111
ITEMISTER data sets. ..................................................................................
Table 4.2  ANOVA tables for 2D-NLWC-ALS of three CAM data sets with 15 wavelet filter combinations for each data set. * ................................................................. 121

Table 4.3  ANOVA tables for 2D-NLWC-ALS of three ITEMISER data sets with 15 different wavelet filter combinations for each data set. * ........................................ 122

Table 4.4  Comparison of the drift time of each peak of all the components in CAM and ITEMISER data set using ALS spectra. * ................................................................. 134

Table 4.5  The comparison of the reduced mobility ($K_0$) for all the peaks in CAM and ITEMISER data sets using ALS spectra. * ................................................................. 136

Table 5.1  Comparison of three types of mass spectra. ................................................................. 141

Table 5.2  Comparison of PDR between histogram, binned spectra at mass resolution 1.00, mass profiles, wavelet compression of mass profiles at level 1, 2 and 3, respectively. * ........................................................................................................... 149

Table 5.3  ANOVA table for the comparison of PDRs between mass profile and histogram................................................................. 155

Table 5.4  Evaluation of $E_{RMS}$ at three wavelet compression levels using different type of wavelet filters................................................................. 156

Table 5.5  ANOVA table for the comparison of PDRs between wavelet compressed mass profile at level 3 and histogram. ................................................................. 158
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>The schematic diagram of an ion mobility spectrometer.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Plot (a): Ion mobility spectra from data set 1 collected when the UAV flew through the DMMP plume with 1000 spectra and 1024 points in each spectrum. Plot (b): One ion mobility spectrum from data set 1 collected when the UAV flew through the DMMP plume and graphed from drift time 0.5 ms.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Plot (a) Ion mobility spectra from data set 2 collected when the UAV flew through the DMMP plume with 1000 spectra and 1024 points in each spectrum. Plot (b): One ion mobility spectrum from data set 2 collected when the UAV flew through the DMMP plume and graphed from drift time 2 ms.</td>
<td>42</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>ALS spectral profile for data set 1 collected when the UAV flew through the DMMP plume.</td>
<td>44</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Intensity profiles from data set 1 collected in positive ion mode when the UAV flew through the DMMP plume.</td>
<td>45</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>ALS spectral profile from data set 2 collected in positive ion mode when the UAV flew through the DMMP plume.</td>
<td>46</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Intensity profiles from data set 2 collected in positive ion mode when the UAV flew through the DMMP plume.</td>
<td>47</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>One ion mobility spectrum from data set 1 collected when the UAV flew through the DMMP plume. The graph was plot from drift time 2-12 ms. The uncompressed spectrum (Uncompressed); the reconstructed spectrum from 128...</td>
<td></td>
</tr>
</tbody>
</table>
linear wavelet coefficients compressed at level 3 using daublet 8 filter (Linear); and
the reconstructed spectrum from nonlinear wavelet compression at level 3 using
daublet 8 filter with 15% threshold (Nonlinear). The total number of wavelet
coefficients and position points are 93 points for the nonlinear compressed spectrum.

Figure 2.8  Evaluation of the root mean square errors with 22 types of wavelet filters on
data set 2 collected when the UAV flew through the DMMP plume at different
compression levels with linear wavelet compression and non-linear wavelet
compression. The full-scale signal for this data set was 8.0 V. The $E_{RMS}$ of linear
wavelet compression at different compression levels (Linear); and the $E_{RMS}$ of
Nonlinear wavelet compression with different thresholds (Nonlinear)................. 51

Figure 2.9  The number of points after the nonlinear compression with filter daublet 8 as
a function of the threshold using one spectrum from data set 2 collected when the
UAV flew through the plume of DMMP.............................................................. 54

Figure 2.10  Wavelet coefficient values at different percentage level as a function of
number of points using daublet 8 filter on data set 1........................................... 56

Figure 2.11  Reconstructed IMS spectrum from different compression efficiencies using
the daublet 8 filter, the spectrum used is one spectrum from data set 1. The graph
was plotted from 2 ms to 10 ms to give better vision ........................................... 57

Figure 3.1  Schematic diagram of the implementation principle of two dimensional
NLWC to compress a data matrix into a compressed wavelet coefficients and a
coefficient position control matrix........................................................................ 63
Figure 3.2  TEPO, DMMP, and DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM. The gating pulse has been removed........................................................................................................................................ 73

Figure 3.3  ALS spectra profile for the TEPO-DMMP-DPM data set collected by CAM. .............................................................................................................................................................................. 74

Figure 3.4  ALS concentration profile for the TEPO-DMMP-DPM data set collected by CAM. .............................................................................................................................................................................. 75

Figure 3.5  DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed....................................................................................................................................................... 76

Figure 3.6  ALS spectra profile for DPM-DMMP data set collected by ITEMISER....... 78

Figure 3.7  ALS concentration profile for DPM-DMMP data set collected by ITEMISER. ............................................................................................................................................................ 79

Figure 3.8  Minimum $SE_{RMS}$ value from 484 combination wavelet filters for row compression and column compression at each compression factor as a function of compression factor for three TEPO-DMMP-DPM CAM data sets. ......................... 82

Figure 3.9  Minimum $E_{RMS}$ value from 484 combination wavelet filters for row compression and column compression at each compression factor as a function of compression factor for three DMMP-DPM ITEMISER data sets. ......................... 84

Figure 3.10  Reconstructed TEPO-DMMP-DPM CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension. ............... 85
Figure 3.11  Reconstructed DPM-DMMP data set from nonlinear wavelet compression with compression factor of 1/128 the original ITEMISER data set. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension. ................................................................. 86

Figure 3.12  ALS spectra profile for reconstructed CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension. ............... 88

Figure 3.13  ALS concentration profile for the reconstructed CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension. .............. 89

Figure 3.14  ALS spectra profile for the reconstructed ITEMISER data set with compression factor of 1/128. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension;....................... 90

Figure 3.15  ALS concentration profile for the reconstructed ITEMISER data set with compression factor of 1/128. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension;....................... 91

Figure 4.1  Schematic diagram of the implementation principle of the 2D-NLWC-ALS algorithm. ................................................................. 103

Figure 4.2  TEPO-DMMP-DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode using CAM. The gating pulse has been removed........................................................................................................ 113
Figure 4.3  ALS spectra for TEPO-DMMP-DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM. .......................... 114

Figure 4.4  ALS concentration profiles for TEPO-DMMP-DPM CAM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM. ........................................................................................................ 115

Figure 4.5  DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed. Smaller data set comprising 300 spectra with 1,024 points per spectrum was used to plot this graph by cutting the ending points. ................................................. 117

Figure 4.6  ALS spectra for DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed. ........................................................................................................ 118

Figure 4.7  ALS concentration profiles for DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed.................................................... 119

Figure 4.8  Minimum $E_{RMS}$ evaluation from all 484 wavelet filter combinations applied to time dimension and spectra dimension using 2D-NLWC-ALS method. The time dimension was not compressed while the minimum $E_{RMS}$ was picked for different compress ratio in the dimension of spectra compression. Three data sets were from CAM on TEPO-DMMP-DPM and three were from ITEMISER on DMP-DMMP. ........................................................................................................ 123
Figure 4.9  Minimum $SE_{RMS}$ evaluation from all 484 wavelet filter combination applied to
time dimension and spectra dimension using 2D-NLWC-ALS algorithm. The time
dimension was not compressed while the minimum $E_{RMS}$ was picked for different
compress ratio in the dimension of spectra compression. Three data sets were
collected from CAM on TEPO-DMMP-DPM and three were ITEMISER data sets
on DMP-DMMP. ................................................................. 124

Figure 4.10  Reconstructed spectra of TEPO-DMMP-DPM data sets using 2D-NLWC-
ALS algorithm at compression ratio of 640 ppm. Wavelet filter of daublet 8 was
used at both the drift time dimension and the data acquisition dimension. The
compressed wavelet coefficients keep 100 points in time dimension and 300 points
in spectra dimension. ................................................................. 126

Figure 4.11  ALS spectra for TEPO-DMMP-DPM CAM data set reconstructed using 2D-
NLWC-ALS algorithm at compression ratio of 640 ppm. Wavelet filter of daublet 8
was used at both the drift time dimension and the data acquisition dimension...... 128

Figure 4.12  ALS concentration profiles for reconstructed TEPO-DMMP-DPM CAM
data set using 2D-NLWC-ALS algorithm at compression ratio of 640 ppm. Wavelet
filter of daublet 8 was used at both the drift time dimension and the data acquisition
dimension................................................................. 129

Figure 4.13  Reconstructed spectra of DMP-DMMP ITEMISER data set from 2D-
NLWC-ALS algorithm at compression ratio of 62 ppm. Wavelet filter of daublet 14
was used at both the drift time dimension and the data acquisition dimension. The
saved compressed spectra $S_c$ is a matrix of $200 \times 4$. The saved compressed concentration profiles $C_c$ is a matrix of $150 \times 4$. .......................................................... 130

Figure 4.14 ALS spectra for ITEMISER data set reconstructed using 2D-NLWC-ALS algorithm at compression ratio of 62 ppm. Wavelet filter of daublet 14 was used at both the drift time dimension and the data acquisition dimension. .......................... 132

Figure 4.15 ALS concentration profiles for reconstructed ITEMISER data set. Wavelet filter of daublet 14 was used at both the drift time dimension and the data acquisition dimension. ................................................................................................................ 133

Figure 5.1 $m$-xylene dataset comprised of 160 mass spectra in a TIC profile with 70 points per spectrum collected by GC/MS. The 100 spectra took 0.25 s of scan time. ........................................................................................................................................... 145

Figure 5.2 The representative of a single mass spectrum obtained from the three preprocessing approaches: (a) histogram; (b) binned at 1.00 $m/z$ mass resolution; (c) mass profile ........................................................................................................................................ 147

Figure 5.3 PCA cluster analysis of the histogram spectra. $m$: $m$-xylene; $p$: $p$-xylene; $o$: $o$-xylene ................................................................................................................ 148

Figure 5.4 PCA cluster analysis of the binned spectra at mass resolution of 1.00 $m/z$. $m$: $m$-xylene; $p$: $p$-xylene; $o$: $o$-xylene. ................................................................. 150

Figure 5.5 PCA cluster analysis of the mass profile. $m$: $m$-xylene; $p$: $p$-xylene; $o$: $o$-xylene ................................................................................................................. 151
Figure 5.6 Comparison of three PCA component cluster analysis between histogram and binned spectra with mass resolution 1.00 m/z. m: m-xylene; p: p-xylene; o: o-xylene................................................................. 153

Figure 5.7 Comparison of three PCA cluster analysis of mass profile and of binned spectra with mass resolution 1.00 m/z. m: m-xylene; p: p-xylene; o: o-xylene.... 154

Figure 5.8 Comparison of PCA cluster results between spectra wavelet compressed from mass profile at level 3 and binned spectra at mass resolution 1.00 m/z. m: m-xylene; p: p-xylene; o: o-xylene................................................................. 159

Figure 5.9 Comparison of PCA cluster analysis between spectra wavelet compressed from mass profile at level 3 and histogram spectra. m: m-xylene; p: p-xylene; o: o-xylene........................................................................................................ 160

Figure 5.10 PCA cluster analysis of average histogram spectra of different runs for a m-xylene spectrum from the ith TIC data set. pi (i = 1 to 3): a p-xylene spectrum from the ith TIC data set; oi (i=1 to 3): a o-xylene spectrum from the ith TIC data set.. 161

Figure 5.11 PCA cluster analysis of average binned mass spectra at mass resolution 1.00 m/z of different runs. pi (i = 1 to 3): a p-xylene spectrum from the ith TIC data set; oi (i = 1 to 3): a o-xylene spectrum from the ith TIC data set. .......................... 162

Figure 5.12 PCA cluster analysis of average mass profiles from different runs. pi (i = 1 to 3): a p-xylene spectrum from the ith TIC data set; oi (i = 1 to 3): a o-xylene spectrum from the ith TIC data set. ......................................................... 163
List of Abbreviations

2D.................................................................................. two-dimensional
3D.................................................................................. three-dimensional
ALS............................................................................. alternating least squares
ANOVA ........................................................................... analysis of variance
APCI ................................................................. atmospheric-pressure chemical ionization
CAM ................................................................. chemical agent monitor
CF ............................................................................. compression factor
CompE............................................................................. compression efficiency
CW ........................................................................... chemical weapon
CWT.............................................................................. continuous wavelet transform
DMMP .............................................................. dimethyl methylphosphonate
DPM........................................................................... 1 (or 2)-(2-Methoxymethyleneoxy) - propanol
DWT ................................................................. discrete wavelet transform
$E_{rr}$ ........................................................................... relative residual error
$E_{RMS}$ ........................................................................ root mean square reconstruction error
FC................................................................................. Fourier compression
FT................................................................................... Fourier transform
IMS .............................................................................. ion mobility spectrometry
ITMS.................................................................ion trap mobility spectrometer

$K_0$.................................................................reduced mobility

LabVIEW...........................................................laboratory virtual instrument engineering workbench

LCD.................................................................lightweight chemical detector

LWC.................................................................linear wavelet compression

MCR.................................................................multivariate curve resolution

MRA .................................................................multiresolution analysis

NIR.................................................................near infrared

NLWC............................................................nonlinear wavelet compression

PCA...............................................................principal component analysis

PDR .................................................................projected difference resolution

QMF...............................................................quadrature mirror filters

RIP .................................................................reactant ion peak

$RE_{RMS}$............................................................relative root-mean-square error

$SE_{RMS}$............................................................spectral reconstruction error

SFA.................................................................subwindow factor analysis

SIMPLISMA......................................................SIMPLE-to-Use Interactive Self-
modeling Mixture Analysis

ASIMPLISMA..................................................automated number of components

SIMPLISMA......................................................self-modeling curve resolution
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>signal-to-noise ratio</td>
</tr>
<tr>
<td>SPIDER</td>
<td>spectrometric point ionizing detector</td>
</tr>
<tr>
<td></td>
<td>expendable/recoverable</td>
</tr>
<tr>
<td>TEPO</td>
<td>triethyl phosphate</td>
</tr>
<tr>
<td>UAV</td>
<td>unmanned aerial vehicle</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>VI</td>
<td>virtual instrument</td>
</tr>
<tr>
<td>WC</td>
<td>wavelet compression</td>
</tr>
<tr>
<td>WFA</td>
<td>window factor analysis</td>
</tr>
<tr>
<td>WT</td>
<td>wavelet transform</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

1.1  Background Information

Chemometrics is the technique and operations associated with the mathematical manipulation and interpretation of chemical data defined since 1970’s. Chemometrics has been recognized as a powerful tool for identification, classification and interpretation of data in modern laboratories.\(^1\)–\(^7\) As a complement to laboratory automation, chemometrics seeks better solutions for data handling with mathematical and statistical operations. The basic chemometrics techniques used in processing and interpreting analytical data include several more important fields: experimental design,\(^8\) multivariate data analysis,\(^9\)–\(^11\) signal processing,\(^12\)–\(^15\) principle components and factor analysis,\(^16\) pattern recognition,\(^17\) and decision making.\(^18\)

Modern waveform digitizer cards have been developed to the point where automated data collection is so fast that large amounts of data can be generated in a few minutes. Many scientists are thus always looking for more efficient methods to properly save or process the generated large amount of data in a real-time way. Ion mobility spectrometry (IMS) is a technique that has broad applications for detecting chemical weapons (CW),\(^19\) explosives,\(^20\) bacteria,\(^21\) environmental pollutants,\(^22\) and drugs.\(^23\) When IMS sensors are used for monitoring gas-phase samples, they may acquire spectra at a rate of 40 Hz. Several hours of collection may result in millions of data points. Therefore, compression methods that facilitate the storage and handling of large collections of data are beneficial and necessary.
The purpose of data reduction is the replacement of large amount of measurements by a few characteristic numbers in which all relevant significant information has been preserved.\textsuperscript{24} Many different methods are available for data reduction depending on the type of data that is measured and the type of information wanted. This dissertation presents nonlinear wavelet compression (NLWC), two-dimensional nonlinear wavelet compression (2D-NLWC), and alternating least squares combined with two-dimensional nonlinear wavelet compression (2D-NLWC-ALS) with applications on IMS data of chemical warfare agent simulants.

Wavelet compression can also be used to reduce the larger mass profiles.\textsuperscript{25} The mass profiles after wavelet compression with smaller data size than the histograms still provide better projected difference resolutions (PDRs) among some isomers, for example xylenes. In addition, wavelet compression improved the PDRs among the isomers compared to the direct analysis of the uncompressed mass spectral profiles or histograms.

To understand the reaction mechanisms for simple and complex unimolecular and bimolecular chemical reactions, it is necessary and important to solve the derivation of analytical rate expressions and build a kinetic model based on experimental data. An automated method that can accurately solve the complicated analytical rate expressions and represent the result of the model at the same time is needed to study simple and complex chemical reaction schemas in a faster and easier manner.

1.2 Ion Mobility Spectrometry (IMS)

IMS has been widely used in screening explosives at airports,\textsuperscript{26-28} detecting
chemical agents for the military,\textsuperscript{29} and monitoring stack gas emissions in industry.\textsuperscript{30}

Figure 1.1 gives the schematic of a common ion mobility spectrometer. In practice, a vapor sample is introduced into the reaction region of a drift tube where neutral molecules of the vapor are ionized and the resultant ions are injected into the drift region for mobility analysis. This relationship of the variables is given in the following equation (1.1).

\[ V_d = K(T)E \]  

(1.1)

for which \( K(T) \) is the mobility that is determined from the drift velocity \( (V_d) \) attained by ions in a weak electric field \( (E) \) of the drift region at atmospheric pressure. Usually, electric fields of 200 V/cm and mobilities of 1 to 2 cm\(^2\)V\(^{-1}\)s\(^{-1}\) result in ion speeds of 200 to 1000 cm/s or drift times of 5 to 20 ms in 4 to 20-cm-long drift regions at ambient pressure.

The coefficient \( K(T) \) is a function of temperature and can be measured at a given temperature from physical measurements as given in equation (1.2).

\[ K(T) = \frac{V_d}{E} = \frac{L/t_d}{V/L} = \frac{L^2}{t_d V} \]  

(1.2)

With a potential difference \( V \) across a drift region of length \( L \), an ion with an arrival time \( t_d \) has the mobility of \( K(T) \). However, \( K(T) \) is usually converted to a reduced mobility, or \( K_0 \).

\[ K_0(T) = K(T) \cdot \frac{P}{760} \cdot \frac{273.16}{T} \cdot \frac{L^2}{t_d V} \cdot \frac{P}{760} \cdot \frac{273.16}{T} \]  

(1.3)

for which \( p \) is the drift gas pressure in Torr and \( T \) is the temperature in \( K \). From equation (1.3), at any given temperature and pressure, the reduced mobility \( K_0 \) of an unknown ion
Figure 1.1 The schematic diagram of an ion mobility spectrometer.
is obtained from the ratio of its \( t_d \) value to that of an agreed upon standard ion obtained under identical experimental conditions. Equation (1.4) gives the method to calculate the reduced mobility of an unknown ion by knowing its drift time using a standard ion.

\[
\frac{K_d(unknown)}{K_d(standard)} = \frac{t_d(standard)}{t_d(unknown)}
\]  

(1.4)

IMS spectra usually have a reactant ion peak (RIP) as a background component. The atmospheric pressure chemical ionization (APCI) charge transfer reactions are performed between the RIP and analytes. Reviews and tutorials can be found elsewhere.\(^{31-33}\)

1.3  Theoretical Principles

1.3.1  SIMPLE-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA)

SIMPLISMA is a pure variable selection method based on the principle that the intensity at a pure variable gives an estimate of the concentration of its associated components.\(^{34-40}\) A pure variable represents the concentration of one of the components in the data matrix. For example in equation (1.5)

\[
D = CS^T + E
\]  

(1.5)

for which \( D \) represents the original data matrix with the spectra in rows and \( C \) represents the concentration matrix, obtained from the columns of \( D \) that represent pure variables. \( S^T \) represents the transpose of the matrix with the spectra of the pure components in its columns. \( E \) represents the residual error. \( D \) is usually the known dataset collected from the experiment and \( C \) and \( S \) are unknown. SIMPLISMA first calculate the purity of each spectrum in the data set using equation (1.6).
\[ p_{ij} = w_{ij} \times \frac{\sigma_j}{\mu_j + \alpha} \]  

(1.6)

for which \( p_{ij} \) is the purity spectrum \( j \) of variable \( i \), \( w_{ij} \) is the weight that characterizes the linear independence of the \( j \)th candidate variable with respect to the previously selected pure \( i-1 \) variables. The weight is a determinant-based weight function that corrects for previously selected pure variables. The algorithm to determine pure variables in SIMPLISMA is described in detail elsewhere.\(^{41,42} \) \( \mu_j \) and \( \sigma_j \) are the mean and the standard deviation of the \( j \)th candidate variable, respectively. The number of components in a data matrix can be pre-defined as \( nc \). SIMPLISMA then estimates \( C \) with the pure variables of the largest \( nc \) values. After the \( C \) is known, the estimation of the pure spectra \( \hat{S} \) can be calculated by equation (1.7):

\[ \hat{S} = D^T C (C^T C)^{-1} \]  

(1.7)

In the next step, the concentration matrix is calculated again from \( \hat{S} \) as shown in equation (1.8):

\[ \hat{C} = D \hat{S} (\hat{S}^T \hat{S})^{-1} \]  

(1.8)

for which \( \hat{C} \) represents the projected \( C \). The dataset can be reconstructed as follows:

\[ D_{\text{reconstructed}} = \hat{C} \hat{S}^T \]  

(1.9)

If the proper number of components has been determined, the reconstructed dataset \( D_{\text{reconstructed}} \) and the original dataset \( D \) should be very similar. One iteration to predict \( C \) and \( S \) is used in SIMPLISMA to simplify the algorithm. For a more precise result, ALS can be used to predict \( C \) and \( S \).
1.3.2 Alternating Least Squares (ALS)

ALS has been widely used for multivariate curve resolution (MCR)\textsuperscript{43,44}. A data matrix $D$ ($m \times n$) contains $m$ rows of spectra and $n$ columns that correspond to the drift time measurements. A matrix $D$ is decomposed into the product of a matrix of concentration profiles ($C$) and a matrix of spectra ($S$) that will best fit the experimental data. This relationship is given by

$$D = CS^T + E$$ \hspace{1cm} (1.10)

An initial estimation of concentration profile ($C$) or pure spectra signals ($S$) is obtained from the data matrix by using SIMPLISMA. An iterative method is used for which the convergence criterion is the sum of the squared residual error. Each iteration comprises two steps.

The estimation of spectra is obtained by least squares

$$S^T = C^{-1}D$$ \hspace{1cm} (1.11)

for which $C^{-1}$ is the pseudo-inverse of matrix $C$, and $S^T$ is the transpose of the current estimate of the spectra. In this algorithm, two estimates are obtained. The first uses a classical least squares solution and then all the values in the spectra that are less than zero are set to zero. In the second solution, the spectra are obtained by a non-negativity constrained least squares fit that is implemented through the MATLAB function named “lsqnonneg”.

For the second step, a new estimation of the concentration profile is obtained by least squares
\[ C = D(S^T)^{-1} \quad (1.12) \]

for which \((S^T)^{-1}\) is the pseudo-inverse of the transpose of the pure spectra matrix \(S\), and \(C\) is the current estimation of the unit concentration. This step furnishes two estimates for \(C\) using the procedures described in step 1. The pair \((S\) and \(C)\) of estimates that furnishes the lowest residual error is used for the subsequent iteration.

The iteration above is repeated until the data matrix \(D\) the system converges with respect to the relative residual error \((E_{rr})\) that is defined as

\[ E_{rr} = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij}^2}} \quad (1.13) \]

for which \(d_{ij}\) is the \(i\)th row and \(j\)th column unit in the experimental data matrix, which is \(m \times n\) dimension, and \(\hat{d}_{ij}\) is the estimate obtained from the product of \(C\) and \(S\).

### 1.3.3 Wavelet Transform (WT) and Wavelet Compression (WC)

The WT has two forms that are the discrete wavelet transform (DWT) and the continuous wavelet transform (CWT). The DWT is advantageous over the CWT in that the DWT is considerably faster. A brief introduction of the DWT is provided. Many tutorials and a thorough presentation on WT can be found elsewhere. In addition, WC and nonlinear WC algorithms will be introduced.

For a finite signal \(X\) of length \(n\), the DWT can be implemented by passing it through two filters. The two orthogonal filters have the same set of coefficients but with alternating signs and in reverse order.
For spectral analysis, the forward DWT recursively bifurcates an object into smooth and detail components. This process occurs by applying pairs of low and high-pass filters with dyadic sampling, respectively. The two filters satisfy the condition of quadrature mirror filters (QMF) as given in equation (1.14):

$$g_m = (-1)^m h_{M-m-1}$$

for which $M$ is filter length in number of coefficients and $m$ is the index of the coefficients. The low-pass, smooth, or father filter is $g$ and $h$ is the high-pass, detail, or mother filter. Both of these filters are defined by basis wavelets that contain finite non-zero coefficients. The most commonly used wavelet filters are derived from the Daubechies family, coiflet family, and symmlet family. All of these filters are orthogonal. Different wavelet filters have different coefficients that determine the shape and length of the basis wavelet. Two operators $H$ and $G$, which are circulant matrices that contain the finite non-zero elements from $h$ and $g$, are obtained for the WT. The construction of operators $H$ and $G$ are described in the following example.

If a father wavelet $h$ is used with four coefficients $[c_1, c_2, c_3, c_4]$ as a row vector, the operation of $T^3(h-2)$ to $h$ constructs a matrix $H$ as given in equation (1.15).

$$H =
\begin{bmatrix}
c_1 & c_2 & c_3 & c_4 & 0 & 0 & 0 & 0 \\
0 & 0 & c_1 & c_2 & c_3 & c_4 & 0 & 0 \\
0 & 0 & 0 & 0 & c_1 & c_2 & c_3 & c_4 \\
c_3 & c_4 & 0 & 0 & 0 & 0 & c_1 & c_2 \\
\end{bmatrix}
$$

Similarly, we can construct $G$ by using a mother wavelet $g$ instead of $h$. If we have a spectrum $X$, and there are $n_x$ (a power of 2) points of data for each spectrum, we use $x^{(k)}$ and $d^{(k)}$ to denote the smooth and detail part of the spectrum at the $k$th level of.
compression, respectively. The uncompressed spectrum is denoted as \( x^{(0)} \). The method to calculate \( x^{(k+1)} \) and \( d^{(k+1)} \) are given in equations (1.16) and (1.17).

\[
x^{(k+1)} = Hx^{(k)} \quad \text{(1.16)}
\]

\[
d^{(k+1)} = Gx^{(k)} \quad \text{(1.17)}
\]

If the filter \( h \) is used with four coefficients as an example to process the spectrum \( x^{(0)} \), the operator \( H: x^{(k)} \rightarrow x^{(k+1)} \) is given in equation (1.18):

\[
H = \begin{bmatrix}
T^{(n/2^{k+1})}(h - 2)
\end{bmatrix}
\quad \text{(1.18)}
\]

Apparently, the length of \( x^{(k)} \) and \( d^{(k)} \) is \( n/2^k \). After each operation as given in equation (1.19), the length of \( x^{(k+1)} \) will be half of the length of \( x^{(k)} \). The \( x^{(k)} \) can also be reconstructed by equation (1.19).

\[
x^{(k)} = H^T x^{(k+1)} + G^T d^{(k+1)} \quad \text{(1.19)}
\]

A full DWT is thus performed by following the procedures above. After the WT is performed, a detailed coefficients \( d \) and a smooth coefficients \( x \) are saved. A WC can be obtained by discarding the detailed coefficients \( d \) and save the smooth coefficients \( x \) that are generated by a WT. The smooth coefficients \( x \) can be used as the input for next level of the WT, which will generate the next level smooth coefficients as the result of the next level WC. By recursion, the length of the spectrum will be reduced by a factor of two for each level. At the end of the full WT, there will be one smooth point and \( n-1 \) detail points.

A beneficial characteristic of WC is that the inverse wavelet transform is relatively simple. To reconstruct the spectrum at level \((k+1)\), equation (1.20) is used.
The estimation of the reconstructed spectrum is $\hat{x}^{(k)}$.

$$\hat{x}^{(k)} = H^T x^{(k+1)}$$  \hspace{1cm} (1.20)

The algorithm to decompose the raw spectrum $x^{(0)}$ using the above method is referred to as the forward DWT.

### 1.4 The Research Objectives

This dissertation involves three main research projects that include implementing SIMPLISMA and ALS combined with NLWC to IMS data and pattern recognition of unintegrated mass spectra.

A problem of linear WC applied to IMS data is that peak distortion occurs when the spectra are compressed. Nonlinear wavelet compression precisely preserves the peak location (i.e., drift time). The nonlinear wavelet compression (WC) was used to compress IMS data in the drift time dimensions on IMS data of chemical warfare simulants. Nonlinear WC was applied to achieve high compression rates and high fidelity in the reconstructed spectra. A method was also developed and evaluated for optimizing the compression. SIMPLISMA and ALS algorithms as chemometric modeling methods have been combined with the compressed data. Thus, the simpler models can efficiently reconstruct the entire measurement, while leaving the data set compressed. Another benefit is that nonlinear WC improves signal-to-noise ratios and thus the models built from compressed data are improved. Using this method, large volumes of data can be acquired and easily evaluated. The benefits of modeling compressed data will be presented as well as an approach for modeling nonlinearly compressed data.
An overlooked computational preprocessing step is the integration of mass spectral peaks and rendering a spectrum in the form of a histogram. With technological advances, mass spectral profiles can be processed thereby retaining information in the ion peak shapes. Although the size of the mass spectral profile can be quite larger than a histogram, nonlinear wavelet compression can reduce the spectra to a manageable size.

This dissertation includes six chapters. The general background is given in this chapter. Additional background and theory of each project will be presented in the corresponding chapters. The general theory of one-dimensional NLWC will be presented in Chapter 2 with an application on ion mobility spectra from ion mobility spectrometers mounted in an unmanned aerial vehicle. The application of two-dimensional NLWC of ion mobility spectra of chemical warfare agent simulants will be given in Chapter 3. A compression and modeling method that combines SIMPLISMA, ALS and two-dimensional NLWC with the ability to determine the number of components automatically for ion mobility spectra of chemical warfare agent simulants will be illustrated in Chapter 4. In Chapter 5, wavelet compressed high resolution mass profiles are proved to have better projected difference resolution on isomer data than traditional histogram data. Finally, the summary and future work will be given in the Chapter 6. The resulting papers and presentations will be included in Appendix A and B, respectively. Chemical structures for the chemical warfare simulants are given in Appendix C.
Chapter 2  Nonlinear Wavelet Transform of Ion Mobility Spectra from Ion Mobility Spectrometers Mounted in an Unmanned Aerial Vehicle

2.1  Introduction

IMS is one instrumental method for which data compression is important. IMS affords hand-held instruments that may detect parts-per-billion concentrations of volatile and semi-volatile compounds in air.\textsuperscript{26,27} When IMS sensors are used for monitoring gas-phase samples, they may acquire spectra at a rate of 40 Hz. Several hours of collection may result in millions of data points. Therefore, compression methods that facilitate the storage and handling of large collections of data are beneficial. IMS data are suitable for data compression because of the broad peaks that typify IMS data. Each IMS spectrum is a continuous function in the drift time dimension. The most important signal in an IMS spectrum is generally in the lower frequency components because noise occurs in the high frequency components. Recently, Fourier transform (FT) and wavelet transform (WT) have been used to compress the IMS data.\textsuperscript{51,52}

FT has been widely used for data compression.\textsuperscript{53,54} FT is useful for signal processing, such as convolution, correlation, and smoothing.\textsuperscript{55,56} A two-dimensional Fourier compression method was devised and applied to IMS data in both the drift time and sample acquisition dimensions.\textsuperscript{52}

The WT offers several advantages over FT including speed, simpler implementation, and compact support. Recently, WT methods have become a popular signal-processing tool for chemists.\textsuperscript{45,57} The WT technique has also been exploited for processing absorbance spectra,\textsuperscript{58,59} chromatograms, and electrochemical signals.\textsuperscript{60}
can been used with other chemometric approaches such as principal component analysis,\textsuperscript{59} and artificial neural network.\textsuperscript{61} Two-dimensional compression can be achieved with WT by transforming the signal into the wavelet domain and retaining a reduced number of coefficients.\textsuperscript{51} Reviews and tutorials about wavelet transform are available elsewhere.\textsuperscript{45, 62}

In this chapter, nonlinear wavelet compression (NLWC) was used to compress IMS data in the drift time dimension. A problem of linear WC applied to IMS data is that peak distortion occurs when the spectra are compressed.\textsuperscript{51} NLWC has been used in other applications such as temperature constrained cascade correlation neural networks,\textsuperscript{62} vector map data compression\textsuperscript{49} and optical tomography data.\textsuperscript{63} The advantage of NLWC is that it can achieve high compression rates and high fidelity in the reconstructed spectra. However, an optimum threshold value should be found to obtain the best compression efficiency and efficacy.

The nonlinear compression algorithm is dynamic in that it determines the degree of compression based on the complexity of the ion mobility spectrum. For nonlinear compression, two key components are furnished for each compressed spectrum. The first component is a reduced set of wavelet coefficients. The second component is a vector that stores the location of the wavelet coefficients. This component can be saved as a vector of single bytes so that it will not take as much space as the wavelet coefficients.

Comparisons were made between linear and nonlinear WC methods with in-flight data recorded from the airborne IMS instruments. Results showed that nonlinear WC was better for IMS data.
2.2 Theory

NLWC is implemented after a full forward DWT of the spectra. A garrotte threshold is applied in that all wavelet coefficients are removed whose absolute values are below a threshold. The threshold is defined as a percentage of the only smooth coefficient after the DWT (i.e., the lowest frequency coefficient). Coefficients whose magnitudes are larger than the threshold value are stored in the compressed data besides their locations. After nonlinear compression, two key components are furnished for each compressed spectrum. The first component is a reduced set of wavelet coefficients. The second component is a control vector that gives the location of the wavelet coefficients so that the data can be reconstructed. The number of consecutive saved wavelet coefficients and the number of the discarded wavelet coefficients are saved alternately in the control vector. This procedure allows single byte representation if the maximum number of retained or discarded coefficients in a sequence is less than 256, which is typical for IMS. Other strategies can be used to reduce the size of the control vector.

The reconstruction errors measure the compression efficacy and compression efficiency is measured as the extent of compression. The root mean square reconstruction error ($E_{RMS}$) is obtained from equation (2.1).

$$E_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}$$

(2.1)

for which $x_i$ is a point from the spectrum and $\hat{x}_i$ is the reconstructed data point. The number of points in the spectrum is $n$. 

The compression efficiency is the ratio obtained by dividing the compressed data size by the original data size as defined below

\[ CompE = \frac{n_{wc+ctrl}}{n_o} \]  

(2.2)

for which \( n_{wc+ctrl} \) is the size of the compressed spectrum plus the control vector file, \( n_o \) is the number of data points in the spectrum.

### 2.3 Experimental Section

The airborne IMS system consisted of two identical IMS chemical detectors operating in parallel. Each detector has two ion mobility spectrometers, one operating in the positive ion mode and the other operating in the negative ion mode. Because dimethyl methylphosphonate (DMMP) has a relatively high proton affinity and is detected as a positive ion, positive polarity was used for detection. The ion separation cells, sample inlets and signal generation electronics are identical to the Lightweight Chemical Detector (LCD) manufactured by Smiths Detection (Watford, U.K.).\textsuperscript{59} The drift length is 3.0 cm, and the drift gas flow is approximately 150 mL/min. Ammonia was used as the dopant agent and air was used as the drift gas. The IMS system is referred to as the SPIDER (spectrometric point ionizing detector expendable/recoverable) and is the payload of the unmanned aerial vehicle (UAV).

The data acquisition rate for each ion mobility spectrometer was two spectra per second. A single output spectrum comprised 1024 points that spanned 20 ms and was a co-addition of eight individual spectra. Data were collected as the UAV flew through DMMP plumes. The plumes were generated as three crop dusters disseminated
approximately 1000 pounds of DMMP over an area measuring 1000 feet by 1800 feet 800 feet above ground level. Three DMMP disseminations were accomplished – the UAV entered and exited the cloud four times during each of two of the disseminations and missed the plume altogether in the third. In each of the tests for which DMMP data were acquired, the data indicated that the UAV passed through the plume four times and ion mobility spectra were recorded for approximately four minutes. Four similar in-flight data sets were collected, and evaluated using the nonlinear wavelet compression algorithm, the result of two data sets is presented in this dissertation. Given the means of dispersion and measurement, different flights will have different concentrations. The wind turbulence, aircraft position, time of the release of the plume will cause a concentration gradient of the analyte.

The in-flight raw data were stored as binary files on onboard compact flash cards and were exported, post-flight, as ASCII text files from the Trimscan data acquisition program (version 0.11, Smiths Detection - Watford Limited, Watford, England). The spectrometers communicate with an embedded processor by an RS-232 interface.

MATLAB 6.5 R13 (The MathWorks, Inc, Natick, MA) was used for the computational evaluations and modeling of the wavelets. The exported text files were read by MATLAB and represented in matrix format with each row comprising an IMS spectrum. Because the UAV-mounted SPIDER system consisted of two onboard IMS instruments, four data sets were collected from two flights with each instrument. Two data sets are reported that were collected from different flights and instruments.

WaveLab 802 was used as a library of MATLAB routines for wavelet
Twenty-two of the most common wavelet filters were investigated from three different families of wavelets. The wavelet filters used in this dissertation comprised ten from the daublet families, five from the coiflet families, and seven from the symmlet families.

The nonlinear wavelet compression and reconstruction were written in the C language under the Microsoft Development Environment 2002 (version 7.0, Microsoft Corporation, Redmond, WA). MATLAB R13 programs were written to perform statistical calculations.

The IMS data matrix comprised 100 spectra of 1024 drift time measurements. The 100 spectra were selected from the region of the data that corresponds to when the SPIDER systems were flown through the chemical plume. ALS models were used to locate this region in the data set. Two data sets are presented that were acquired from the SPIDER payload in the UAV. The data in the drift time range of 0-0.5 ms were omitted to eliminate the gating pulse from the evaluations. Four data sets were used for study but only two are presented.

2.4 Results and Discussion

Figure 2.1 plot (a) and Figure 2.2 are two data sets that give 1000 ion mobility spectra that comprise 1024 points in each spectrum. For both data sets, the analyte is DMMP. These two data sets will be referred to as data set 1 and data set 2 in this chapter. In data set 1, the DMMP monomer peak was observed at drift time 5.27 ms (1.70 cm²V⁻¹s⁻¹) and dimer at 6.79 ms (1.31 cm²V⁻¹s⁻¹). For data set 2, the DMMP peak
Figure 2.1  Plot (a): Ion mobility spectra from data set 1 collected when the UAV flew through the DMMP plume with 1000 spectra and 1024 points in each spectrum. Plot (b): One ion mobility spectrum from data set 1 collected when the UAV flew through the DMMP plume and graphed from drift time 0.5 ms.
Figure 2.2  Plot (a) Ion mobility spectra from data set 2 collected when the UAV flew through the DMMP plume with 1000 spectra and 1024 points in each spectrum. Plot (b): One ion mobility spectrum from data set 2 collected when the UAV flew through the DMMP plume and graphed from drift time 2 ms.
locations were identical, 5.27 ms (1.70 cm$^2$V$^{-1}$s$^{-1}$) and 6.72 ms (1.31 cm$^2$V$^{-1}$s$^{-1}$) for monomer and dimer ions, respectively. The drift times and reduced mobility values were obtained using the Trimscan software. Figure 2.1 plot (b) gives one ion mobility spectrum from data set 1 collected when the UAV flew through the DMMP plume.

ALS has been widely used in multivariate curve resolution (MCR) of chemical data to simplify complicated large data set. ALS was used to decompose the large in-flight data set matrix into two small simpler matrices, a matrix of concentration profiles and a matrix of component spectra. Figure 2.3 gives the ALS spectra that have been normalized to unit vector length for data set 1. Figure 2.4 gives the integrated concentration profile as a function of time for data set 1. Figure 2.5 and Figure 2.6 give the spectra and concentration profiles for data set 2.

The compression ratios obtained from nonlinear compression are different from that of the linear compression. At each level of linear compression, the number of points is reduced by a factor of 2. The number of points in the compressed data is $1/2^n$ of that of the original spectrum at compression level $n$. The number of points in nonlinear compressed data depends on the data, the distribution of intensities (number of peaks in a spectrum) above the threshold and the threshold.

Figure 2.7 gives an uncompressed spectrum from data set 1, the reconstructed spectrum after compression to 128 points with the linear wavelet compression using filter daublet 8, and the reconstructed spectrum from a nonlinear spectrum compressed to 93
Figure 2.3 ALS spectral profile for data set 1 collected when the UAV flew through the DMMP plume.
Figure 2.4 Intensity profiles from data set 1 collected in positive ion mode when the UAV flew through the DMMP plume.
Figure 2.5 ALS spectral profile from data set 2 collected in positive ion mode when the UAV flew through the DMMP plume.
Figure 2.6  Intensity profiles from data set 2 collected in positive ion mode when the UAV flew through the DMMP plume.
Figure 2.7 One ion mobility spectrum from data set 1 collected when the UAV flew through the DMMP plume. The graph was plot from drift time 2-12 ms. The uncompressed spectrum (Uncompressed); the reconstructed spectrum from 128 linear wavelet coefficients compressed at level 3 using daublet 8 filter (Linear); and the reconstructed spectrum from nonlinear wavelet compression at level 3 using daublet 8 filter with 15% threshold (Nonlinear). The total number of wavelet coefficients and position points are 93 points for the nonlinear compressed spectrum. The inset graph is a enlarged graph cut from 3.1-4.4 ms.
points (including the compression coefficients and position control file) using the daublet 8 filter and a 15% threshold. There is no discernable difference between those three spectra. Reconstructed spectra from the linear and non-linear compression methods retained the dominant peaks in the original spectrum. Figure 2.7 also gives an enlarged window from 3.0 to 4.0 ms. Apparently, the reconstructed linear compression at level 3 creates artifacts that are related to the Gibbs effect (i.e., small ringing peak artifacts about sharp peaks). These artifacts are detrimental to peak selection routines and pattern recognition like neural network. The reconstructed nonlinear compression with a 3% threshold yields an undistorted reconstruction. The same is true for even higher thresholds up to 30% threshold. However, at further linear compressions, the ringing artifacts become severe. Another advantage of the nonlinear wavelet compression is that it precisely preserves the position and intensity of the maxima peaks. Reconstructed peaks from linear compressed data tend to have attenuated maxima with small shifts in position. The maximum intensities of the peaks at 3.82 ms and 8.94 ms in the original spectrum are 7.994 V and 0.964 V, respectively. For the linear compression at level 3 (128/1024 points), the maximum intensities of the peak at 3.84 ms and 8.96 ms are 7.935 V and 0.960 V, respectively. Compared to maximum intensities of the peak at 3.82 ms and 8.94 ms with 7.992 V and 0.965 V in the reconstructed spectrum of non-linear compression at 4% compression ratio (41/1024 points), linear compression at the same compression level causes peak shifts and intensity losses.

As always, there is a compromise between reconstruction fidelity and degree of compression. A comprehensive evaluation of the $E_{\text{RMS}}$ of reconstruction with 22 types of
wavelet filters on data set 2 at different compression levels with linear wavelet compression and non-linear wavelet compression was studied. The maximum intensity in the averaged unprocessed spectra was 7.8 V for this data set. Nonlinear compression with a 3% of the threshold (93/1024 points) for this data set had equivalent compression ratios with linear compression at level 3. The 8% and 30% thresholds were equivalent to level 4 (64/1024) and 5 (32/1024) linear compressions, respectively.

Figure 2.8 gives the comparison of $E_{\text{RMS}}$ of reconstruction between linear compression and nonlinear compression at different equivalent levels. The daublet 2 is equivalent to the Haar wavelet and due to its comparatively poor performance with IMS data it was excluded from the figures and tables.\textsuperscript{45} The best two filters for nonlinear compression are daublet 8 and symmlet 7. The $E_{\text{RMS}}$ for both filters have very low value at all three percentage of thresholds. Apparently, the change of threshold from 3% to 30% does not affect the reconstruction efficacy. Compared to nonlinear compression, the level 3 linear compression does not show significant variation in reconstruction efficacies for some filters. At the level of 4 and 5, there is an increase in the $E_{\text{RMS}}$ in all the filters. This trend results with linear wavelet compression, the removals of the wavelet coefficients at lower levels are relatively insignificant compared to higher compression levels.

Table 2.1 gives the $E_{\text{RMS}}$ for nonlinear compression with 22 different wavelet filters from data set 2. The compression ratios are decided by the percentage of the threshold. Different wavelet filters will not result in different compression ratios once the percentage of the threshold is defined. The percentages in the parenthesis after each $E_{\text{RMS}}$
Figure 2.8 Evaluation of the root mean square errors with 22 types of wavelet filters on data set 2 collected when the UAV flew through the DMMP plume at different compression levels with linear wavelet compression and non-linear wavelet compression. The full-scale signal for this data set was 8.0 V. The $E_{RMS}$ of linear wavelet compression at different compression levels (Linear); and the $E_{RMS}$ of Nonlinear wavelet compression with different thresholds (Nonlinear).
### Table 2.1 Root mean square reconstruction error for nonlinear compression of data set 2 with respect to wavelet filter.*

<table>
<thead>
<tr>
<th>Index</th>
<th>Filter Name</th>
<th>( E_{\text{RMS}} ) (mV) 0.5%</th>
<th>( E_{\text{RMS}} ) (mV) 3%</th>
<th>( E_{\text{RMS}} ) (mV) 8%</th>
<th>( E_{\text{RMS}} ) (mV) 30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daublet 2</td>
<td>40.67±1.40</td>
<td>52.05±1.80</td>
<td>101.18±3.67</td>
<td>211.77±7.73</td>
</tr>
<tr>
<td>2</td>
<td>Daublet 4</td>
<td>12.00±0.42</td>
<td>31.48±1.12</td>
<td>61.62±2.56</td>
<td>219.88±8.30</td>
</tr>
<tr>
<td>3</td>
<td>Daublet 6</td>
<td>6.22±0.24</td>
<td>20.37±0.74</td>
<td>31.48±1.29</td>
<td>107.44±4.37</td>
</tr>
<tr>
<td>4</td>
<td>Daublet 8</td>
<td>1.35±0.05</td>
<td>3.53±0.24</td>
<td>7.72±0.56</td>
<td>16.55±1.01</td>
</tr>
<tr>
<td>5</td>
<td>Daublet 10</td>
<td>1.81±0.06</td>
<td>7.38±0.27</td>
<td>13.38±0.45</td>
<td>32.11±1.41</td>
</tr>
<tr>
<td>6</td>
<td>Daublet 12</td>
<td>1.68±0.05</td>
<td>7.04±0.26</td>
<td>13.07±0.44</td>
<td>29.35±1.34</td>
</tr>
<tr>
<td>7</td>
<td>Daublet 14</td>
<td>6.41±0.22</td>
<td>19.31±0.66</td>
<td>39.11±1.40</td>
<td>143.71±5.56</td>
</tr>
<tr>
<td>8</td>
<td>Daublet 16</td>
<td>9.71±0.34</td>
<td>31.86±1.18</td>
<td>75.72±2.99</td>
<td>276.85±11.38</td>
</tr>
<tr>
<td>9</td>
<td>Daublet 18</td>
<td>10.77±0.37</td>
<td>41.68±1.68</td>
<td>86.02±3.53</td>
<td>353.76±13.07</td>
</tr>
<tr>
<td>10</td>
<td>Daublet 20</td>
<td>9.78±0.34</td>
<td>38.53±1.57</td>
<td>87.94±3.60</td>
<td>237.70±11.16</td>
</tr>
<tr>
<td>11</td>
<td>Coiflet 1</td>
<td>4.63±0.15</td>
<td>13.95±0.47</td>
<td>24.46±0.83</td>
<td>72.81±2.38</td>
</tr>
<tr>
<td>12</td>
<td>Coiflet 2</td>
<td>7.66±0.26</td>
<td>20.47±0.63</td>
<td>50.48±1.86</td>
<td>104.58±4.04</td>
</tr>
<tr>
<td>13</td>
<td>Coiflet 3</td>
<td>7.19±0.24</td>
<td>22.00±0.74</td>
<td>49.57±1.97</td>
<td>135.42±4.96</td>
</tr>
<tr>
<td>14</td>
<td>Coiflet 4</td>
<td>7.06±0.24</td>
<td>22.39±0.78</td>
<td>49.86±1.88</td>
<td>161.57±6.66</td>
</tr>
<tr>
<td>15</td>
<td>Coiflet 5</td>
<td>7.00±0.23</td>
<td>22.96±0.83</td>
<td>44.23±1.35</td>
<td>159.50±6.57</td>
</tr>
<tr>
<td>16</td>
<td>Symmlet 4</td>
<td>8.56±0.31</td>
<td>19.90±0.61</td>
<td>52.39±2.32</td>
<td>179.26±7.31</td>
</tr>
<tr>
<td>17</td>
<td>Symmlet 5</td>
<td>10.79±0.36</td>
<td>29.31±1.03</td>
<td>78.94±3.33</td>
<td>216.05±7.70</td>
</tr>
<tr>
<td>18</td>
<td>Symmlet 6</td>
<td>7.11±0.26</td>
<td>21.68±0.78</td>
<td>47.56±1.91</td>
<td>145.59±5.69</td>
</tr>
<tr>
<td>19</td>
<td>Symmlet 7</td>
<td>1.27±0.04</td>
<td>3.96±0.20</td>
<td>8.79±0.42</td>
<td>19.24±0.84</td>
</tr>
<tr>
<td>20</td>
<td>Symmlet 8</td>
<td>7.44±0.25</td>
<td>21.86±0.82</td>
<td>56.42±2.25</td>
<td>160.39±6.01</td>
</tr>
<tr>
<td>21</td>
<td>Symmlet 9</td>
<td>10.31±0.37</td>
<td>31.35±1.16</td>
<td>69.93±2.49</td>
<td>266.71±11.61</td>
</tr>
<tr>
<td>22</td>
<td>Symmlet 10</td>
<td>6.09±0.23</td>
<td>18.64±0.67</td>
<td>40.90±1.57</td>
<td>152.38±6.56</td>
</tr>
</tbody>
</table>

* The average and 95% confidence interval for 100 spectra from data set.

+ The percentages indicate the threshold with respect to the largest coefficient used for the nonlinear compression.

The full-scale signal for this data set was 7805 mV.
in the first row of the table represent the percentage of each threshold used for the nonlinear compression. Four thresholds were picked according to the different compression ratios that they had provided. At a low threshold of 0.5%, there is no difference of $E_{RMS}$ between filter daublet 8 and symmlet 7. At threshold of 30%, the compression ratio is 4%, and the $E_{RMS}$ was 0.2% of the maximum intensity of the uncompressed spectra. As indicated in the table, the $E_{RMS}$ of reconstruction at a thresholds of 30% is 0.25% of the maximum intensity of the data set, a nonlinear compression with the daublet 8 at a threshold of 30% can provide satisfactory compression ratios with marginal reconstruction error $E_{RMS}$. The threshold depends on the signal and noise attributes of the spectra. For nonlinear wavelet compression, the threshold should always be optimized when the signal and noise attributes of the data change.

Figure 2.9 gives a full range of number of points after the nonlinear compression with the daublet 8 filter with respect to the threshold using one spectrum from data set 2. For thresholds below 10%, the threshold will affect the number of compressed points, with lower thresholds increasing the number of points in the compressed data. As the threshold is increased from 10% to 30%, the number of compressed points will decrease but not as rapidly. The number of points does not decrease much when the threshold is raised above 30%. The control vectors do not change significantly with the different thresholds. The number of control vector points was included in the total number of compressed points. However, this approach is biased in that the coefficients are 4 byte single precision numbers, and in many cases the values of the control vector (i.e.,
Figure 2.9 The number of points after the nonlinear compression with filter daublet 8 as a function of the threshold using one spectrum from data set 2 collected when the UAV flew through the plume of DMMP. The inset graph is an enlarged one from percentage 1 to 10.
locations) can be single-byte unsigned integers or for complex data with high precision
two-byte integers.

Figure 2.10 gives the wavelet domain of data set 1. For nonlinear compression
coefficients, there is no significant difference of the magnitude of wavelet coefficients
between different threshold percentages. This feature is different from linear
compression. In linear compression, the coefficients from the first four levels of the
transform are insignificant when compared to the remaining coefficients at higher levels.
This trend explains why there is no significant change in reconstructed spectra when the
threshold is low enough that coefficients from these levels (i.e., 1-4) are used in the
compressed spectra.

Figure 2.11 gives several reconstructed spectra from data set 1 using different
nonlinear compression thresholds. The maximum intensity of the peak at 4.2 ms in the
original data is 7.9 V. For the reconstructed spectrum at all thresholds, the peak position
is preserved within 0.01 ms. The reconstructed spectrum at all thresholds has the same
maximal value of 7.9 V. Larger thresholds may cause insignificant decreases in peak
intensities. There is no significant visible peak position shifting or ringing effect created
for thresholds below 30%. When the threshold increases above 70%, at a drift time of
5.2 ms, a small artifact peak with an intensity of 0.5 V occurs at the right side of the
maximum peak. This small peak is 5.7% of the value of the maximum intensity of the
original spectrum.
Figure 2.10 Wavelet coefficient values at different percentage level as a function of number of points using daublet 8 filter on data set 1.
**Figure 2.11** Reconstructed IMS spectrum from different compression efficiencies using the daublet 8 filter, the spectrum used is one spectrum from data set 1. The graph was plotted from 2 ms to 10 ms to give better vision.
2.5 Conclusions

Linear and nonlinear wavelet compression of ion mobility measurements were compared and evaluated. IMS spectra can be compressed to ~ 4-5% of their original size using nonlinear wavelet compression without artifacts that are created when linear compression is applied. Nonlinear wavelet compression avoids peak shifts, peak maxima attenuations, and ringing artifacts that are frequently encountered when using linear compression. Smaller values of $E_{\text{RMS}}$ were obtained from nonlinear compression than from linear compression. A grid search was used to select the optimal filter parameters using four UAV IMS data sets. The results from two sets were reported, but all four data sets yielded consistent optimal parameters (i.e., filter types and thresholds).
Chapter 3 Two-Dimensional Nonlinear Wavelet Compressed Ion Mobility Data

3.1 Introduction

IMS instruments are widely used at airports to screen hand luggage for explosives because they furnish rapid scanning and low cost. They also have broad applications for chemical weapon (CW) and treaty verification, environmental pollutants, explosives, forensic, drugs of abuse, and many other research fields.\textsuperscript{20, 23, 26, 29, 30} Counter-terrorism and treaty verification require instrumentation that can be used for on-site identification of chemical warfare agents.\textsuperscript{27, 65} These instruments may collect spectra at a rate of 40 Hz and several hours of data collection may result in large quantities of data. Some IMS instruments are miniaturized hand-held devices and are furnished with onboard processors, which have limited storage capability. For some applications, the data may be transmitted from sensors using wireless communication to a computer.\textsuperscript{66} For wireless transmission, the bandwidth limits prohibit the rate at which data can be collected. Data compression can be useful when data are continually collected and analyzed in real-time via wireless transmission. The alternative approach of sampling at lower rates is inefficient because information available from the measurement is discarded.

Compression of analytical data was an active area of research for many years. Data compression seeks to reduce data size without losing important chemical information. The Fourier transform (FT) has been used for data compression.\textsuperscript{53, 67} A two-dimensional Fourier compression method was devised and applied to IMS data in both the drift time and sample acquisition dimensions.\textsuperscript{52} The wavelet transform (WT) offers several advantages over FT including speed, simpler implementation, and compact
support. Recently, WT methods have become a popular signal-processing tool for chemists. Many applications have applied de-noising and data compression on different types of data. Two-dimensional wavelet compression (WC) can be achieved with the WT by transforming the signal into the wavelet domain and retaining a reduced number of coefficients. Significant size reductions were achieved using this method. Reviews and tutorials of the WT are available, elsewhere.

The distinction between linear and nonlinear wavelet compression is important because linear wavelet compression and reconstruction can be accomplished solely from matrix operations. However, for nonlinear wavelet compression and reconstruction the data are compressed by a logical threshold applied to the coefficients magnitude so the compression cannot be accomplished by matrix operations. This distinction applies for both hard and soft threshold methods.

LWC uses the pyramid algorithm or retains the low frequency wavelet coefficients. Therefore, LWC performs as a low-pass digital filter for removing high frequency components. Because LWC has a fixed transfer function it is useful when all peaks in the spectral data set have the same shape. When this assumption is not met many of the problems of digital filters appear such as artifact peaks in the baseline or near peaks narrower than the filter band-pass. The spurious peaks can result in false positive errors. Peak detection algorithms are susceptible to artifact peaks that may arise from LWC. Peak-shifting can occur as well and may result in false negative errors if a peak shifts outside of a detection window.

NLWC retains wavelet coefficients that exceed an intensity threshold regardless
of the frequency or scale. Thus NLWC functions as a variable band-pass filter. NLWC gives the advantage that peak distortion and artifacts may be virtually eliminated in reconstructed spectra. The term de-noising is used specifically to describe the effect of the NLWC variable band-pass filter and should not be used with LWC. NLWC is especially useful for high compression rates, high-fidelity, and de-noised reconstruction. NLWC avoids the artifacts caused by LWC while achieving similar or improved compression.

3.2 Theory

3.2.1 LWC

The WT resembles the FT in that the FT decomposes the signal by projecting it onto sine and cosine waves of different frequencies. The WT breaks down a signal by projecting it onto scaled and shifted versions of a wavelet basis function.

For a finite signal \( D \) of length \( n \), the wavelet transform is implemented by passing through two filters. One filter is the scaling filter \( h \) that is called the father wavelet. The scaling filter is a low-pass filter and is implemented by a circulant matrix \( H \). The other filter called the mother wavelet \( g \), is a high-pass filter and is implemented by a circulant matrix \( G \). The low-frequency wavelet coefficients, which are called smooth, are generated by projecting the signal onto \( H \). The high-frequency wavelet coefficients, which are the detailed part, are generated by projecting the signal onto \( G \).

After each wavelet transform, LWC will keep the low-frequency wavelet coefficients for the next level of compression. After one level of wavelet transformation,
the data is bifurcated into two parts. The smooth part is retained and the detailed part is discarded. Thus, a 50% data compression is obtained at each level of transform. To perform the next level of WC, the smooth part of the last level compression will be used as the original data. After each level of compression, half of the original data will be saved and half will be discarded. At the $n$th level of compression, $1/2^n$ of the original size of data will be saved. The optimum data compression level depends on the data and the wavelet filter type that are used. An optimum wavelet filter should keep as much information as possible in the low-frequency region so that high compression level can be obtained.

3.2.2 Two-dimensional NLWC

To perform a NLWC, a full forward discrete WT is performed. The wavelet coefficients are sorted from largest to smallest based on magnitude (i.e., absolute value). A threshold criterion is defined as a percentage of the largest wavelet coefficient. Coefficients in the sorted array with magnitudes greater than the threshold and their indices are retained. As a result, after the NLWT, two components are furnished for each compressed spectrum. The first component is the reduced set of wavelet coefficients and the second is the position control vector that is used for reconstruction. The position vector is obtained from the sorted indices.

Figure 3.1 gives a flowchart for two-dimensional NLWC. In step 1, the WT is applied to each row of the data matrix, and in step 2 the WT is applied to each column. After these two transformations, the data set is transformed into a matrix of wavelet
Figure 3.1  Schematic diagram of the implementation principle of two dimensional NLWC to compress a data matrix into a compressed wavelet coefficients and a coefficient position control matrix.
coefficients. In step 3, the coefficient matrix is sorted from high to low magnitudes. A position control vector is generated from the indices of the sort to locate the coefficient’s position in the coefficient matrix. The position control vector is used to arrange the WT coefficients to their locations in the coefficient matrix in step 5. The coefficients that were not retained during compression are filled with zeros. An inverse WT is performed on the columns and the rows of the coefficient matrix consecutively to furnish the reconstructed data in step 6.

### 3.2.3 Reconstruction Error

Root mean square reconstruction error ($E_{\text{RMS}}$) is given by

$$E_{\text{RMS}} = \sqrt{\frac{\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} (d_{ij} - \hat{d}_{ij})^2}{n_i n_s}}$$

(3.1)

for which $d_{ij}$ is a point from the original data matrix for the data set and $\hat{d}_{ij}$ is the corresponding reconstructed data point. The number of points in each spectrum is $n_s$ and the number of spectra in each dataset is $n_s$. However, when the spectra have a low signal-to-noise ratio, this evaluation method will be biased because any noise that is removed may result in a larger reconstruction error. ALS spectra were compared between the reconstructed data and the original data. This comparison is less biased because the ALS spectra will have greater signal-to-noise ratios than the spectra in the data sets. Also, for IMS measurements, the noise is modeled in the concentration profiles and the ALS spectra have less noise. Therefore, the ALS spectra are compared from uncompressed and reconstructed datasets. The spectral reconstruction error ($SE_{\text{RMS}}$) is defined as
\[ SE_{\text{RMS}} = \sqrt{\frac{\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} (s_{ij} - \hat{s}_{ij})^2}{\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} s_{ij}^2}} \]

for which \( n_s \) is the number of points in each spectrum, \( n_c \) is the number of components in the model. The ALS spectra from the uncompressed data \( s_{ij} \) and from the reconstructed data set \( \hat{s}_{ij} \) are compared.

The compression efficiency is reported as compression factor (\( CF \)) that is a measure of the ratio of the number of points, which includes the number of coefficients and control values, retained in the compressed matrix \( (N_c) \) to the original data size \( (N_0) \).

\[ CF = \frac{N_c}{N_0} \]

The relative error of root mean square (\( RE_{\text{RMS}} \)) is also used as a ratio of the \( E_{\text{RMS}} \) value to the maximum intensity \( (I_{\text{max}}) \) in the dataset to evaluate the efficacy of the reconstruction.

\[ RE_{\text{RMS}} = \frac{E_{\text{RMS}}}{I_{\text{max}}} \]

### 3.3 Experimental Section

Two different ion mobility spectrometers were used. The first instrument used was a handheld chemical agent monitor (CAM, Type 482-301N, Graseby Ionics, Watford, Herts, UK). This instrument was modified by removal of the acetone dopant so that the ion chemistry was water based. The three chemical warfare simulants used were dimethyl methylphosphonate (DMMP) (97%, Lancaster Synthesis Inc., Windham, NH,
Lot No. P07087 212-052-3), triethyl phosphate (TEPO) (99.8%, Aldrich Chemical Company Inc., Milwaukee, WI, Lot No. 12502MI) and dipropylene glycol methyl ether (DPM) (99+%, Aldrich Chemical Company Inc., Milwaukee, WI, Lot No. 10613TA). The structures of the three chemical warfare simulants can be found in Appendix C. The instrument used was interfaced with a single processor PII 200 MHz/64MB RAM computer via a data acquisition board, type AT-MIO-16X (National Instruments, Austin, TX). The operating system was Microsoft Windows 98 Second Edition (Redmond, WA). Positive ion mode was used for data acquisition. Three data sets were collected from this instrument. Each data set contained TEPO, DMMP, and DPM. The IMS waveform was sampled at 80 kHz and a single spectrum comprised 1,500 points. For each data set, approximately 2,000 spectra were saved. Detailed information of data sets can be found in Table 3.1.

The second instrument was an ion trap mobility spectrometer (ITMS), ITEMISER contraband detection and identification system (GE Ion Track Instruments, Inc., Wilmington, MA). The ITMS system was interfaced to a laptop computer (Pentium III 850 MHz and 384 MB memory) via a DAQCard-AI-16XE-50 (National Instruments, Austin, TX). Positive mode was used for data acquisition. Microsoft Windows XP Pro with Service Pack 1 (Redmond, WA) was the operating system.

For both instruments, homemade virtual instrument (VI) programs were implemented in LabVIEW 6.1 (National Instruments, Austin, TX) to acquire data from the IMS instruments. The IMS waveform was sampled at 80 kHz and a single spectrum comprised 1,500 points. For each data set, 15,000 spectra were collected. Three data sets were used in this dissertation for data compression and analysis.
Table 3.1 Data set collection information for the three CAM data sets and three ITEMISER data sets used.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data set dimension (Scan number × Drift time)</th>
<th>Small data set dimension (Scan number × Drift time)</th>
<th>Collection time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM data set 1</td>
<td>2055 × 1500</td>
<td>205 × 1024</td>
<td>139.34</td>
</tr>
<tr>
<td>CAM data set 2</td>
<td>2073 × 1500</td>
<td>207 × 1024</td>
<td>142.20</td>
</tr>
<tr>
<td>CAM data set 3</td>
<td>2111 × 1500</td>
<td>211 × 1024</td>
<td>143.41</td>
</tr>
<tr>
<td>ITEMISER data set 1</td>
<td>15006 × 1500</td>
<td>300 × 1024</td>
<td>746.23</td>
</tr>
<tr>
<td>ITEMISER data set 2</td>
<td>15010 × 1500</td>
<td>300 × 1024</td>
<td>746.40</td>
</tr>
<tr>
<td>ITEMISER data set 3</td>
<td>15015 × 1500</td>
<td>300 × 1024</td>
<td>746.58</td>
</tr>
</tbody>
</table>
The spectrometer gating pulse was removed by removing the data points from the first 0.5 ms. Because the baseline region was located from 1.5 to 3 ms of the IMS spectrum, a region of the spectrum where usually no peaks occur, IMS signals were corrected by subtracting the average baseline signal from this region. Samples were not diluted before data collection. By placing a 5 mL open vial that contained 2 mL of pure sample at a distance of 1 cm from the inlet of the CAM for 1 to 3 seconds, the headspace was sampled at room temperature. Each data set was collected by placing the first sample near the instrument for 1 to 3 seconds, and then waiting until the instrumental response returned to the blank signal before the next sample was introduced.

The data processing was implemented on a personal computer with an AMD XP 2200+ processor and 512 MB RAM. The operating system was Window XP Pro SP1. MATLAB 6.5 R13 (The MathWorks, Inc, Natick, MA) was used to perform computational evaluations and modeling of the wavelets. The binary data collected from the VI were read by MATLAB and represented in matrix format with each row comprising an IMS spectrum. The 2D NLWC and reconstruction routine was written in MATLAB.

WaveLab 802 was used as a library of MATLAB routines for wavelet applications in this dissertation. Twenty-two of the most common wavelet filters were investigated from three different families of wavelets. The wavelet filters used in this dissertation comprised ten from the daubelet families, five from the coiflet families, and seven from the symmlet families.
The wavelet filters were optimized by using a grid search. The data sets were reduced to speed up the exhaustive evaluation. The data sets from CAM were decimated to smaller size by saving every 10th spectrum in the data acquisition dimension, and every 10th point in drift time dimension. The data sets collected from ITEMISER were decimated to smaller size by saving every 50th spectrum in the data acquisition dimension, and every 10th point in drift time dimension.

After the best wavelet filters were selected by using the small data set, the full-size data sets of both CAM and ITEMISER were used for compression and reconstruction evaluation using the optimal wavelet filters.

### 3.4 Results and Discussion

A synthetic data set was constructed in MATLAB for the proof of concept. A Gaussian peak with intensity of 1.0 and standard deviation of 10 was centered in a vector of 1024 points. The outer product of this vector was calculated to furnish a simple data set. To this data set normally (mean of zero and standard deviation of unity) distributed random deviates were added. The data set was reduced to furnish 3 other smaller data sets by sampling every sixteenth row, every sixteenth column, and sampling every sixteenth row and column. In addition, 2D-LWC and 2D-NLWC were applied to reduce the data down to the same size as the row and column sampled data set (i.e., $64 \times 64$), and then reconstructed. A daublet 10 was chosen for compressing both row and column dimensions.

The data sets were modeled using SIMPLISMA with a single component. The
data sets were reconstructed from the SIMPLISMA concentration profile and spectrum. The noise was calculated from the standard deviation of the sampling ten points around the perimeter of the smallest data set. For data sets that had dimensions, which were not sampled, the same corresponding points to the sampled data sets (i.e., every sixteenth) were used to calculate the noise. The signal was defined by the maximum in the reconstructed data set. This evaluation was run 1000 times with different random deviates.

The results are reported in Table 3.2. The first case corresponds to sampling the instrument and the IMS waveform as fast as possible. The second case corresponds to collecting spectra a slower rate. The third case corresponds to decreasing the resolution of the spectral measurement. The fourth case is a combination of the two reductions in sample rate. Both wavelet compressions compress the data to the same degree as the smallest data set, and demonstrate the SNR enhancement. The SNR is larger for the reconstructed spectra than the unprocessed data, because compression has removed the high frequency noise components. The NLWC was not optimized and could have been compressed further. As a consequence, it contained a greater noise contribution than the LWC data.

Two different sample suites were collected from the CAM and ITEMISER® instruments. Three chemical weapon simulants (TEPO, DMMP, and DPM) were used for the data sets collected from CAM. DMMP and DPM were used for the data sets collected from ITEMISER®. These data sets had different signal-to-noise ratios, different peak shapes, and transient instrumental responses. Three replicates of each sample suite
Table 3.2 Signal-to-noise ratio for a two-dimensional synthetic centered Gaussian with intensity of 1.0 and standard deviation of 10 added to random normal deviates.*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Full</th>
<th>Row Sampled</th>
<th>Column Sampled</th>
<th>Row and Column Sampled</th>
<th>LWC</th>
<th>NLWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>318.1±0.3</td>
<td>111.8±0.1</td>
<td>107.6±0.1</td>
<td>79.3±0.1</td>
<td>1396±1.4</td>
<td>1129.3±1.1</td>
</tr>
</tbody>
</table>

* Mean = 0; SD = 0.1. The evaluation comprised 1000 replicates that were used to compute the SNR averages and 95% confidence intervals.
were obtained for the statistical analyses.

Figure 3.2 gives a surface plot of the data set acquired from TEPO, DMMP, and DPM. This data set comprised 2,073 spectra with 1,500 points/spectrum that were collected in positive mode with the CAM. Because of the large number of spectra, the whole data set was represented by sampling every 10th spectrum from the original data set so that this figure could be rendered.

SIMPLISMA was used to obtain the initial concentration profile value for ALS using a damping factor, \(\alpha\), of 0.05.\(^{34,35,72}\)

Figure 3.3 gives ALS spectra for the CAM data sets. TEPO gave two peaks, a monomer at 7.96 ms (1.56 cm\(^2\)V\(^{-1}\)s\(^{-1}\)) and a dimer peak at 10.86 ms (1.14 cm\(^2\)V\(^{-1}\)s\(^{-1}\)). DMMP gave a monomer peak that was located at 6.85 ms (1.81 cm\(^2\)V\(^{-1}\)s\(^{-1}\)) and a dimer peak at 8.68 ms (1.43 cm\(^2\)V\(^{-1}\)s\(^{-1}\), as the standard). The peak occurred at 9.79 ms is the mixture of DMMP dimer and TEPO dimer peak. DPM peaks were obtained at 7.69 ms (1.61 cm\(^2\)V\(^{-1}\)s\(^{-1}\)) for the monomer and 10.00 ms (1.24 cm\(^2\)V\(^{-1}\)s\(^{-1}\)) for the dimer. Also it can be noted that the DPM dimer peak had two small shoulder peaks, which on the left at 9.46 ms is the peak for the mixture of DMMP dimer and DPM dimer, and on the right at 10.49 ms is the peak of the mixture of TEPO dimer and DPM dimer.

Figure 3.4 gives the concentration profiles for the TEPO-DMMP-DPM CAM data set. The compounds sampled by CAM were ordered as TEPO, DMMP, and DPM. Each compound was sampled once the instrumental response returned to baseline.

Figure 3.5 gives the ITEMISER DPM-DMMP data set comprising 15,000 spectra with 1,500 points per spectrum collected in positive mode. When comparing the data sets
Figure 3.2 TEPO, DMMP, and DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM. The gating pulse has been removed.
Figure 3.3 ALS spectra profile for the TEPO-DMMP-DPM data set collected by CAM.
Figure 3.4 ALS concentration profile for the TEPO-DMMP-DPM data set collected by CAM.
Figure 3.5  DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed.
obtained from CAM, the data sets collected from ITEMISER have less noise. Only a monomer was observed for DPM, however both monomer and dimer peaks were observed for DMMP.

Figure 3.6 gives the ALS spectra for the DPM-DMMP data set collected by the ITEMISER. The drift times from 3 ms to 11 ms were plotted. The DPM monomer peak occurred at drift time of 6.58 ms (1.60 cm$^2$V$^{-1}$s$^{-1}$). The DMMP monomer peak appears at a drift time of 5.64 ms (1.87 cm$^2$V$^{-1}$s$^{-1}$) and the dimer at 7.52 ms (1.40 cm$^2$V$^{-1}$s$^{-1}$). The concentration profiles of reactant ion peak (RIP), DPM, and DMMP can be found in Figure 3.7.

Selection of the optimum wavelet is also an important step for data compression because different wavelets may produce significantly different results in many applications. The reconstruction errors resulting from uniform compression of data sets with varying wavelets were compared to study the performance of different wavelets for compression efficacy. To evaluate the efficacy of the reconstruction, ALS was applied to both the uncompressed data and the reconstructed data, for which two sets of spectra were obtained. The $SE_{RMS}$ value was calculated from these two sets of spectra.

Table 3.3 gives the minimum $SE_{RMS}$, $RE_{RMS}$, and the wavelet types for minimum $SE_{RMS}$ in drift time and data acquisition dimensions with respect to different compression factors. The minimum $SE_{RMS}$ comes from 484 $SE_{RMS}$ values obtained from the combination of 22 wavelet filters applied to row compression and 22 wavelet filters applied to column compression. For each compression factor, there is a set of wavelet filters that can be applied to the NLWC to obtain a minimum $SE_{RMS}$. Note that at a
Figure 3.6 ALS spectra profile for DPM-DMMP data set collected by ITEMISER.
Figure 3.7  ALS concentration profile for DPM-DMMP data set data set collected by ITEMISER.
Table 3.3  Compression factor, minimum $SE_{RMS}$, $RE_{RMS}$, and the wavelet types for minimum $SE_{RMS}$ in drift time dimension and data acquisition dimension.*

<table>
<thead>
<tr>
<th>CF</th>
<th>Minimum $SE_{RMS}$</th>
<th>$RE_{RMS}$ (%)</th>
<th>Wavelet type for minimum $SE_{RMS}$ by dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Drift time</td>
</tr>
<tr>
<td>1/8</td>
<td>0.010±0.003</td>
<td>0.07±0.03</td>
<td>Daublet 20</td>
</tr>
<tr>
<td>1/16</td>
<td>0.010±0.005</td>
<td>0.07±0.04</td>
<td>Symmlet 4</td>
</tr>
<tr>
<td>1/32</td>
<td>0.010±0.007</td>
<td>0.07±0.05</td>
<td>Symmlet 4</td>
</tr>
<tr>
<td>1/64</td>
<td>0.014±0.007</td>
<td>0.10±0.06</td>
<td>Symmlet 10</td>
</tr>
<tr>
<td>1/128</td>
<td>0.012±0.007</td>
<td>0.09±0.07</td>
<td>Daublet 4</td>
</tr>
<tr>
<td>1/256</td>
<td>0.031±0.009</td>
<td>0.23±0.07</td>
<td>Symmlet 8</td>
</tr>
<tr>
<td>1/512</td>
<td>0.053±0.032</td>
<td>0.39±0.24</td>
<td>Daublet 12</td>
</tr>
<tr>
<td>1/1024</td>
<td>0.067±0.204</td>
<td>0.50±1.51</td>
<td>Coiflet 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data acquisition</td>
</tr>
<tr>
<td>1/8</td>
<td></td>
<td></td>
<td>Coiflet 1</td>
</tr>
<tr>
<td>1/16</td>
<td></td>
<td></td>
<td>Symmlet 8</td>
</tr>
<tr>
<td>1/32</td>
<td></td>
<td></td>
<td>Coiflet 5</td>
</tr>
<tr>
<td>1/64</td>
<td></td>
<td></td>
<td>Symmlet 4</td>
</tr>
<tr>
<td>1/128</td>
<td></td>
<td></td>
<td>Symmlet 9</td>
</tr>
<tr>
<td>1/256</td>
<td></td>
<td></td>
<td>Coiflet 2</td>
</tr>
<tr>
<td>1/512</td>
<td></td>
<td></td>
<td>Symmlet 8</td>
</tr>
<tr>
<td>1/1024</td>
<td></td>
<td></td>
<td>Daublet 6</td>
</tr>
</tbody>
</table>

* Data sets were collected from CAM in positive ion mode.
compression factor of 1/128, when applying the daublet 4 filter to the drift time
dimension and the symmlet 9 to the data acquisition dimension, the minimum $SE_{RMS}$ that
was obtained was 0.0124, and the $RE_{RMS}$ is 0.09%.

Figure 3.8 gives the plot of the minimum $SE_{RMS}$ value from 484 combination
wavelet filters for row compression and column compression as a function of
compression factor. The $SE_{RMS}$ increases significantly for compressions greater than
compression factor of 1/256.

For the ITEMISER data sets, the noise level is much lower than those for the
CAM. Thus, evaluation of $E_{RMS}$ is effective for the ITEMISER data sets. Table 3.4 gives
the minimum $E_{RMS}$ and $RE_{RMS}$ for the three reconstructed data sets for several two-
dimensional compression factors with all 484 wavelet filter combinations applied to
nonlinear compression. Figure 3.9 gives the minimum $E_{RMS}$ values from 484 combination
wavelet filters for row compression and column compression at each compression factor.
From compression factor 1/256, the value of $E_{RMS}$ rises. With the compression factor of
1/128, using wavelet filter of symmlet 10 in the drift time dimension and coiflet 2 in data
acquisition of the NLWC, the minimum $E_{RMS}$ can be obtained at 0.0055, and the $RE_{RMS}$ is
0.25%.

To illustrate the compression efficacy visually, the data sets presented in Figure
3.10 and Figure 3.11, respectively, are the CAM and ITEMISER data set after two-
dimensional compression and reconstruction. For the figures, a 1/128 compression factor
was achieved for each data set, which indicates that the data sets were reduced to 0.78%
of their original size. For each reconstructed data set, ALS was applied to evaluate the
Figure 3.8 Minimum $SE_{RMS}$ value from 484 combination wavelet filters for row compression and column compression at each compression factor as a function of compression factor for three TEPO-DMMP-DPM CAM data sets.
Table 3.4 Minimum $E_{RMS}$ and $RE_{RMS}$ for the reconstructed three ITEMISER data sets for several two-dimensional compression factors with all 484 wavelet filter combination applied to nonlinear compression.*

<table>
<thead>
<tr>
<th>CF</th>
<th>Minimum $E_{RMS}$ (x $10^3$)</th>
<th>RE$_{RMS}$ (%)</th>
<th>Wavelet type for minimum $E_{RMS}$</th>
<th>Drift time dimension</th>
<th>Data acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/16</td>
<td>4.8±0.2</td>
<td>0.22±0.01</td>
<td>Symmlet 8</td>
<td>Daublet 18</td>
<td></td>
</tr>
<tr>
<td>1/32</td>
<td>4.9±0.2</td>
<td>0.22±0.01</td>
<td>Daublet 16</td>
<td>Coiflet 5</td>
<td></td>
</tr>
<tr>
<td>1/64</td>
<td>5.1±0.2</td>
<td>0.23±0.01</td>
<td>Coiflet 4</td>
<td>Coiflet 5</td>
<td></td>
</tr>
<tr>
<td>1/128</td>
<td>5.5±0.3</td>
<td>0.25±0.01</td>
<td>Symmlet 10</td>
<td>Coiflet 2</td>
<td></td>
</tr>
<tr>
<td>1/256</td>
<td>6.2±1.5</td>
<td>0.28±0.07</td>
<td>Symmlet 7</td>
<td>Coiflet 2</td>
<td></td>
</tr>
<tr>
<td>1/512</td>
<td>7.9±4.9</td>
<td>0.36±0.22</td>
<td>Symmlet 5</td>
<td>Coiflet 2</td>
<td></td>
</tr>
<tr>
<td>1/1024</td>
<td>12.0±8.8</td>
<td>0.52±0.40</td>
<td>Symmlet 7</td>
<td>Coiflet 2</td>
<td></td>
</tr>
</tbody>
</table>

*The maximum intensities of the three ITEMISER data sets used are 2.2079 V, 2.2078 V, and 2.2099 V. Precision presents 95% confidence interval.
Figure 3.9 Minimum $E_{RMS}$ value from 484 combination wavelet filters for row compression and column compression at each compression factor as a function of compression factor for three DMMP-DPM ITEMISER data sets.
Figure 3.10  Reconstructed TEPO-DMMP-DPM CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension.
Figure 3.11 Reconstructed DPM-DMMP data set from nonlinear wavelet compression with compression factor of 1/128 the original ITEMISER data set. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension.
validation of the compression, as given in Figure 3.12 and Figure 3.13 for the CAM data sets, and Figure 3.14 and Figure 3.15 for the ITEMISER data sets. For both data sets, spectra and concentration profiles gave satisfactory representation of the original data. Note that in the ALS concentration profile for the original CAM data, at time of 12.69 s, there is a significant decrease in the RIP intensity. This problem was caused by the vibration of the data cable so that at that time point, a zero-signal spectrum was obtained. However, the corresponding concentration profile for the reconstructed CAM data set had this artifact removed which demonstrated a benefit of 2D-NLWC.

Table 3.5 gives the reduced mobility ($K_0$) of all the compounds in each of the experiments. The standard reduced mobility value of DMMP dimer for CAM is 1.43 cm$^2$V$^{-1}$s$^{-1}$ and for ITEMISER is 1.40 cm$^2$V$^{-1}$s$^{-1}$. The reconstruction errors of reduced mobility came from the ITEMISER data sets. Three different data sets for CAM and ITEMISER were used. The $K_0$ values were reported as the mean values for three data sets. The maximum 95% confidence level was 0.0057 cm$^2$V$^{-1}$s$^{-1}$ for all the $K_0$ averages. The $K_0$ of the DMMP monomer between 1.87 to 1.86 cm$^2$V$^{-1}$s$^{-1}$. Because the standard value for DMMP dimer reported had only three significant figures, our values were constrained. This variance in the monomer peak position was 0.01 ms between the unprocessed and reconstructed spectra.

To compare the compression efficacy for linear wavelet compression and nonlinear compression, ANOVA was applied to the CAM and ITEMISER reconstruction errors. Fifteen combinations of wavelet filters were randomly chosen from 22 wavelet filters for each dimension. Two-dimensional linear wavelet compression was applied to
Figure 3.12 ALS spectra profile for reconstructed CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension.
Figure 3.13 ALS concentration profile for the reconstructed CAM data set from a compression factor of 1/128 using wavelet filter symmlet 8 for compression at drift time dimension and coiflet 1 for compression at data acquisition dimension.
**Figure 3.14** ALS spectra profile for the reconstructed ITEMISER data set with compression factor of 1/128. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension.
Figure 3.15  ALS concentration profile for the reconstructed ITEMISER data set with compression factor of 1/128. Wavelet filter of symmlet 5 was used at the drift time dimension and coiflet 2 was used at the data acquisition dimension.
Table 3.5 The comparison of the reduced mobility ($K_0$) for all the peaks in CAM and ITEMISER data sets using ALS spectra.*

<table>
<thead>
<tr>
<th></th>
<th>Uncompressed CAM data (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Reconstructed CAM data (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Uncompressed ITEMISER (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Reconstructed ITEMISER (cm$^2$V$^{-1}$s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.56</td>
<td>1.56</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.14</td>
<td>1.14</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.81</td>
<td>1.81</td>
<td>1.87</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.43 (standard)</td>
<td>1.43 (standard)</td>
<td>1.40 (standard)</td>
<td>1.40 (standard)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.61</td>
<td>1.61</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.24</td>
<td>1.24</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* The maximum 95% confidence interval is 0.0057 cm$^2$V$^{-1}$s$^{-1}$. 
both CAM and ITEMISER data with compression levels of 4 for the drift time dimension and of 3 for the sample acquisition dimension using 15 combinations of wavelet filters drawn from a population of 484. The data was compressed to 1/128 of the original data size. The same combinations of wavelet filters were used for the two treatments (i.e., linear and nonlinear compression). For the CAM and ITEMISER data, the $E_{RMS}$ values for the 15 reconstructed data sets from linear wavelet compression were used as the first treatment of data, and the $E_{RMS}$ values for the 15 reconstructed data sets from nonlinear wavelet compression were used as the second treatment of data.

Two-way ANOVA was applied with the effects of wavelet filter given as rows and compression method as columns. The ANOVA results for the CAM data set are given in Table 3.6. The table indicates that the lower reconstruction errors are significant for the nonlinear wavelet compression method. At the same time, the ANOVA indicates that the differences among wavelet filters are insignificant. The ANOVA results for the ITEMISER data set are given in Table 3.7. The results indicate the same trend that was observed for the CAM data. The lower errors obtained from the nonlinear compression are significant and the differences among the wavelet filters are insignificant.

Two-dimensional wavelet compression is inefficient in that all spectra or samples and all drift times or channels are uniformly compressed. The nonlinear wavelet compression reported in this work uses an intensity threshold to guide the compression. Because IMS measurements usually contain empty channels, intensity threshold will characterize large spectral features with high fidelity. Baseline noise and small peaks below the threshold will be neglected.
Table 3.6 ANOVA tables for linear wavelet compression and nonlinear wavelet compression of CAM.*

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear vs.</td>
<td>0.00687</td>
<td>1</td>
<td>0.00687</td>
<td>53.3</td>
<td>0</td>
</tr>
<tr>
<td>nonlinear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelet filter</td>
<td>0.00172</td>
<td>14</td>
<td>0.00012</td>
<td>0.95</td>
<td>0.5357</td>
</tr>
<tr>
<td>Error</td>
<td>0.00180</td>
<td>14</td>
<td>0.00013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.01039</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The 15 randomly choose wavelet filter combinations used for the linear wavelet compression and nonlinear wavelet compression are:

<table>
<thead>
<tr>
<th>No.</th>
<th>Row Filter</th>
<th>Column Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>daublet 10</td>
<td>symmlet 8</td>
</tr>
<tr>
<td>2</td>
<td>coiflet 1</td>
<td>coiflet 3</td>
</tr>
<tr>
<td>3</td>
<td>symmlet 5</td>
<td>symmlet 7</td>
</tr>
<tr>
<td>4</td>
<td>daublet 20</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>5</td>
<td>symmlet 5</td>
<td>coiflet 3</td>
</tr>
<tr>
<td>6</td>
<td>symmlet 4</td>
<td>symmlet 8</td>
</tr>
<tr>
<td>7</td>
<td>daublet 18</td>
<td>daublet 8</td>
</tr>
<tr>
<td>8</td>
<td>symmlet 8</td>
<td>symmlet 8</td>
</tr>
<tr>
<td>9</td>
<td>symmlet 7</td>
<td>daublet 18</td>
</tr>
<tr>
<td>10</td>
<td>daublet 16</td>
<td>daublet 4</td>
</tr>
<tr>
<td>11</td>
<td>daublet 2</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>12</td>
<td>daublet 10</td>
<td>daublet 6</td>
</tr>
<tr>
<td>13</td>
<td>coiflet 3</td>
<td>daublet 10</td>
</tr>
<tr>
<td>14</td>
<td>daublet 10</td>
<td>daublet 12</td>
</tr>
<tr>
<td>15</td>
<td>symmlet 8</td>
<td>daublet 20</td>
</tr>
</tbody>
</table>
Table 3.7 ANOVA tables for linear wavelet compression and nonlinear wavelet compression of ITEMISER.*

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear vs. nonlinear</td>
<td>0.00014</td>
<td>1</td>
<td>0.00014</td>
<td>130.62</td>
<td>0</td>
</tr>
<tr>
<td>Wavelet filters</td>
<td>0.00001</td>
<td>14</td>
<td>0</td>
<td>0.94</td>
<td>0.5453</td>
</tr>
<tr>
<td>Error</td>
<td>0.00001</td>
<td>14</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.00016</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The 15 randomly choose wavelet filter combinations used for the linear wavelet compression and nonlinear wavelet compression are:

<table>
<thead>
<tr>
<th>No.</th>
<th>Row Filter</th>
<th>Column Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>daublet 18</td>
<td>daublet 20</td>
</tr>
<tr>
<td>2</td>
<td>coiflet 2</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>3</td>
<td>coiflet 5</td>
<td>daublet 10</td>
</tr>
<tr>
<td>4</td>
<td>daublet 16</td>
<td>coiflet 5</td>
</tr>
<tr>
<td>5</td>
<td>coiflet 1</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>6</td>
<td>daublet 20</td>
<td>coiflet 5</td>
</tr>
<tr>
<td>7</td>
<td>daublet 8</td>
<td>daublet 14</td>
</tr>
<tr>
<td>8</td>
<td>coiflet 5</td>
<td>daublet 10</td>
</tr>
<tr>
<td>9</td>
<td>coiflet 2</td>
<td>daublet 14</td>
</tr>
<tr>
<td>10</td>
<td>coiflet 5</td>
<td>daublet 8</td>
</tr>
<tr>
<td>11</td>
<td>symmlet 7</td>
<td>daublet 16</td>
</tr>
<tr>
<td>12</td>
<td>coiflet 3</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>13</td>
<td>symmlet 7</td>
<td>coiflet 1</td>
</tr>
<tr>
<td>14</td>
<td>coiflet 4</td>
<td>symmlet 6</td>
</tr>
<tr>
<td>15</td>
<td>daublet 14</td>
<td>daublet 16</td>
</tr>
</tbody>
</table>
When instrumental responses are infrequent (IMS waveform sampled less than 40 kHz), the same benefit of intensity threshold compression of the sample acquisition dimension will be beneficial. Because the threshold is applied to the two-dimensionally compressed matrix, a synergistic benefit is obtained that arises from compression characterizing regions (channels and spectra) with large features.

3.5 Conclusions

A two-dimensional nonlinear wavelet compression method was developed and applied to IMS data. A compression factor of 1/128 with relative error of root mean square ($RE_{RMS}$) of 0.25% was achieved for ITEMISER data. With same compression factor, $RE_{RMS}$ of 0.09% was achieved for CAM data, which had larger noise levels. The ALS models from reconstructed and unprocessed data sets gave indistinguishable spectra and concentration profiles. The wavelet filters for both drift time and acquisition time dimensions were optimized, although optimization may not be necessary to obtain good compression.

NLWC avoids peak distortion and eliminates artifacts caused by linear wavelet compression. Peak shifting was reduced to less than 0.01 ms for reconstructed spectra obtained from two different IMS instruments. This attribute of high fidelity compression is obtained because high frequency components that are important to characteristic features are retained during the compression, while low frequency components that do not characterize larger variations are omitted. By applying the threshold to 2D wavelet coefficients a synergistic benefit is obtained.
Chapter 4  SIMPLISMA and ALS Applied to Two-Dimensional Nonlinear Wavelet Compressed Ion Mobility Data

4.1  Introduction

Ion mobility spectrometry (IMS) has been broadly applied for the detection of trace levels of explosives,\textsuperscript{20, 26} bacteria,\textsuperscript{21, 73} pesticides,\textsuperscript{74} and chemical warfare agents.\textsuperscript{29, 65} The IMS signal arises from ion formation and characterization in a drift tube maintained at ambient pressure. IMS instrumentation can be miniaturized and affords portable devices for on-site chemical measurement. Hand-held IMS instruments may detect parts-per-billion concentrations of volatile compounds in air. The other advantages of IMS are its fast response and high sensitivity. Reviews on IMS can be found elsewhere.\textsuperscript{27, 75, 76}

The IMS signal arises from ion formation and characterization in a drift tube maintained at ambient pressure. IMS instrumentation can be miniaturized and affords portable devices for on-site chemical measurement. Hand-held IMS instruments may detect parts-per-billion concentrations of volatile compounds in air. The other advantages of IMS are its fast response and high sensitivity.

Multivariate curve resolution (MCR) methods model the transient responses of data.\textsuperscript{44, 67} Most methods decompose the data set into spectra and concentration profiles.\textsuperscript{43} Large data sets place demands on MCR calculations that may not be sufficed by the computational system. Data compression prior to MCR can dramatically reduce the storage requirements, concomitantly improve signal-to-noise ratios, and model the spectra and concentration profiles.
Data compression on IMS is able to reduce data size without losing important chemical information. Over the past decades, a variety of linear transforms have been developed for image compression. Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) are among the most popular ones used for chemical data. A two dimensional DFT-based method was devised and applied to IMS data in both the drift time and sample acquisition dimensions. The DCT can be regarded as a discrete-time version of the Fourier-Cosine series. DCT is currently the standard encoding scheme for JEPT still image compression, which has found tremendous applications in various areas of entertainment, engineering and science. Despite all the popularity and advantages of DFT- and DCT- based compression schemes, there are inherently drawbacks associated with these two approaches.

The DFT is not commonly used in image compression algorithms since the energy concentration properties are poor compared to other transforms, another degrading effect is that undesired high frequency components appear at the boundaries of the transformed block. In DCT-based image compression, the input image needs to be blocked to reduce the computational complexity. This results in the consequence that correlation across the block boundaries is not eliminated, which not only causes the loss of compression, but also the noticeable and annoying blocking artifacts. Another drawback for FT is that it is bad for non-periodic functions because of the non-locality of the basis functions.

Wavelets transformation (WT) has emerged as a better solution in overcoming the above shortcomings. Compared to the FT, the WT offers several advantages over FT
including faster, simpler implementation, and compact support that extends to infinity. Wavelet compression (WC) can be achieved with the WT by transforming the signal into the wavelet domain and retaining a reduced number of coefficients, significant size reductions were achieved using this method. Reviews and tutorials about WT are available.\textsuperscript{48, 70}

WC applied to IMS is useful when all peaks have the same shape. However, peak distortion in reconstructed ion mobility spectra from WC is problematic in that artifact peaks can result in false positive alarms. These artifacts can be detrimental when peak detection algorithms are used. Peak shifting also can result in false negative alarms when a peak shifts out of a detection window. To solve this problem, a nonlinear wavelet compression (NLWC) algorithm was devised to avoid peak distortion and eliminate artifacts in the reconstructed spectra. For NLWC, two key components are furnished for each spectrum. The first component is a reduced set of wavelet coefficients. The second component is a vector that stores the location of the wavelet coefficients. The algorithm works by storing the locations and coefficients of the wavelet coefficients that exceed an intensity threshold. NLWC is especially useful for high compression rates, high-fidelity reconstruction, and denoising in the reconstructed spectra. NLWC allows the removal of the artifacts caused by LWC while achieving similar or improved compression with respect to linear wavelet methods. The models built from compressed data are improved because NLWC improves signal-to-noise ratios. Using this method, large volumes of data can be acquired and easily evaluated.\textsuperscript{66, 79}

SIMPLe-to-use interactive mixture analysis (SIMPLISMA) is a self-modeling
curve resolution method that has been demonstrated as an effective tool for enhancing IMS measurements. Alternating least squares (ALS) is also a chemometric modeling method that can be used on the NLWC compressed data. The alternating procedure of regression applied to NLWC is with constraints, such as concentrations and spectra should not be negative. The general objective of SIMPLISMA and ALS is to decompose a large data matrix into two simpler matrices, a matrix of concentration profiles, and a matrix of component spectra. The simpler models can efficiently reconstruct the entire measurement, while leaving the data set compressed. Modeling as a form of compression can decrease the data size to 60 ppm to 6 ppm of the original size.

It is useful to automatically determine the size and complexity of the model. Therefore, large volumes of data can easily be acquired and evaluated. In addition, compression improves signal-to-noise ratios so that models built from compressed data are improved.

ALS applied to 2D-NLWC proved to be beneficial in the application of CAM and ITEMISER, which are two different IMS instruments with different level of signal to noise ratio. NLWC is advantageous compared to LWC, because it is resistant to shifts of the peak maxima (i.e., less than 0.02 ms). Smaller values of error of root mean square ($E_{\text{RMS}}$) were obtained from NLWC than from LWC. The procedure of building ALS models from 2D-NLWC compressed data will be called 2D-NLWC-ALS. The 2D-NLWC-ALS procedure simplifies the data and extracts the main principal variances with respect to of the spectra and their change with concentration.
Efficiency and efficacy are two evaluating criteria for compression. Compression efficiency is determined by the ratio of the size of compressed data to the uncompressed data. Compression efficacy is a measurement of the similarity of the reconstructed spectra compared to the unprocessed spectra. There is a tradeoff between these two criteria. For example, if the data are over compressed, high efficiency is obtained but relevant information may be lost, which represents low efficacy. In this paper, both compression efficiency and efficacy were evaluated.

4.2 Theory

4.2.1 ASIMPLISMA

For ALS to be used as a compression method, an automated approach to determining the number of components is required. For compression overestimating the number of components is preferred because to avoid losing important signals. SIMPLISMA was modified to automatically determine the correct number of components and evaluated with 6 sets of IMS data. ASIMPLISMA was modified to determine the number of components above a noise threshold.

The procedure uses a region of the IMS spectrum that is usually free of signal. This region may be at drift times early or late into the ion mobility measurement. The same region (e.g., 1 to 4 ms) for measuring the noise is often used to estimate and correct for baseline region of the ion mobility spectrum. The noise is the root-mean-square of the baseline corrected intensities for all spectra in the signal free drift time region. The noise is calculated by
\[ noise = \sqrt{\frac{\sum_{i=1}^{m_i} \sum_{j=1}^{n_1} d_{ij}^2}{m_in_1}} \] (4.1)

for which \( d_{ij} \) is a baseline corrected element in a region devoid of signal. The \( m_I \) and \( n_I \) are the dimension of the matrix to calculate the noise. The value for noise is used as the damping threshold \( \alpha \) in ASIMPLISMA.

After each component is calculated in ASIMPLISMA as \( c_i \), the signal value is calculated as

\[ sig_j = \sqrt{\frac{\sum_{i=1}^{m} c_{ij}^2}{mn}} \] (4.2)

The \( m \) and \( n \) are the time and drift time dimension for the original data set. The signal is divided by the number of drift time measurements \( n \), because the concentration profiles are integrated intensities across the spectra. When the signal is greater than twice the measured noise the component is retained as significant.

### 4.2.2 Two Dimensional-NLWC-ALS

ASIMPLISMA is applied to the original data to determinate the number of components in the data set before the implementation of 2D-NLWC-ALS algorithm. Figure 4.1 gives the schematic diagram of the implementation 2D-NLWC-ALS algorithm. The NLWC algorithm was described somewhere else.\(^{66}\) There are 8 steps involved in this algorithm.

1) A wavelet transformation is applied to the rows of the data set \( D (m \times n) \) to furnish wavelet coefficients \( W_c (m \times n) \).
Figure 4.1 Schematic diagram of the implementation principle of the 2D-NLWC-ALS algorithm.
2) A wavelet transformation is applied to the columns of the resulting matrix $W_e$ ($m \times n$) obtained from step 1. The data matrix is then transformed to a 2D-wavelet coefficient matrix $W_c \times W_c$.

3) In this step, all the coefficients in the matrix are sorted from highest to lowest magnitude. This sorted array is used to define a threshold value. The number of coefficients ($n_{coeff}$) that will be saved from the original data set is decided by the product of the expected compression ratio and number of points ($m \times n$) from the original data set. The threshold value is defined as the coefficient value in the list with index number equal to $n_{coeff}$. Values in the coefficient matrix that are greater than the threshold value are saved as a vector $W_{cc}$, and the remainders are discarded. The positions of the saved coefficients are saved in the position control vector so that the data may be reconstructed.

4) Using the compressed coefficients vector $W_{cc}$ and position control vector, a partially reconstructed wavelet coefficient matrix $D_{pr} (m_c \times n_c)$ is generated in reduced size.

5) The SIMPLISMA algorithm is applied to the partially reconstructed coefficient matrix $D_{pr} (m_c \times n_c)$. When the estimated total number of component is represented as $nc$, the coefficients matrix is decomposed into $C_0 (m_c \times nc)$ and $S_0 (n_c \times nc)$. The $C_0 (m_c \times nc)$ is used as the input of the initial value of concentration profiles to complete the calculation of ALS. $C_c (m_c \times nc)$ and $S_c (n_c \times nc)$ are output of ALS to represent the low resolution concentration profiles and the spectra. Note that the $C_c$ and $S_c$ are saved as the compressed data.
6) The WT is applied to both $C_c (m_c \times nc)$ and $S_c (n_c \times nc)$. The resulting wavelet coefficients are saved as $WC_c (m_c \times nc)$ and $WS_c (n_c \times nc)$. Then zeros matrices are used to pad the ends of $WC_c$ and $WS_c$ so that two full size wavelet coefficient matrix $WC (m \times nc)$ and $WS (n \times nc)$ are constructed.

7) The inverse WT is applied to both the $WC (m \times nc)$ and $WS (n \times nc)$ to reconstruct the $C (m \times nc)$ and $S (n \times nc)$ as high resolution concentration profiles and spectra. All the negative value in $C$ and $S$ are set to zero.

8) The data set can be reconstructed from the products of either the high- or low-resolution concentration profiles and spectra as in equation (1.9).

The combination values of $m_c$ and $n_c$ used in the algorithm are determined by author through a series of evaluation of the $E_{RMS}$. The optimal values of $m_c$ and $n_c$ should provide the best $E_{RMS}$ value after the reconstruction and reasonable compression ratio.

### 4.2.3 Evaluation Reconstruction Efficiency and Efficacy

The compression efficiency is evaluated with the compression ratio ($CR$).

$$CR = \frac{m_c \times nc + nc \times n_c}{m \times n} \times 100\%$$

(4.3)

The $CR$ is a percentage of the total number of elements in the low resolution concentration profiles $C_c$ and spectra $S_c$. The matrix $C_c$ and $S_c$ were obtained in step 5 in the 2D-NLWC-ALS algorithm.

To study the efficacy of the reconstruction errors associated with the data compression, the root mean square error ($E_{RMS}$) of the reconstructed data set with respect to the original data set was used as defined in equation (4.4).
\[ E_{RMS} = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (d_{ij} - c_i \times s_j^T)^2}{mn}} \]  

(4.4)

for which \( d_{ij} \) is the \( i \)th row and \( j \)th column unit in the experimental data matrix, which is \( m \times n \) dimension. The \( c_i \) is \( i \)th row in concentration profiles \( C \) and \( s_j^T \) is the \( j \)th column in the spectra matrix \( S \). The number of spectra in each data set is \( m \) and the number of points in each spectrum is \( n \).

The relative error of root mean square (\( RE_{RMS} \)) is also used as a ratio of the \( E_{RMS} \) value to the maximum intensity (\( I_{\text{max}} \)) in the data set to evaluate the efficacy of the reconstruction.

\[ RE_{RMS} = \frac{E_{RMS}}{I_{\text{max}}} \]  

(4.5)

Another error metric was used for spectra that have high signal-to-noise ratios. Because NLWC denoises spectra during compression, part of the reconstruction error can be attributed to noise removal. The spectra in obtained from ALS models were compared between the uncompressed and compressed data sets, because the magnitude of noise in these spectra was attenuated. The relative root mean square error between the ALS spectra (\( SE_{RMS} \)) from uncompressed and reconstructed data sets is defined as

\[ SE_{RMS} = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (s_{ij} - \hat{s}_{ij})^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij}^2}} \]  

(4.6)

for which \( n \) is the number of points in each spectrum, \( m \) is the number of spectra extracted from the data set after ALS is applied to it. The value of the \( j \)th point in the \( i \)th spectrum in the spectra extracted from the uncompressed data set is \( s_{ij} \), and \( \hat{s}_{ij} \) is value of
the $j$th point in the $i$th spectrum in the high resolution spectra that is obtained by applying 2D-NLWC-ALS to the original data set.

When the concentration profile is relatively smooth, the reconstruction error of the concentration profiles can be evaluated by the relative root mean square error between the ALS concentration profiles ($CE_{RMS}$) from the compressed and uncompressed data by

$$CE_{RMS} = \sqrt{\frac{\sum_{j=1}^{n} \sum_{i=1}^{m} (c_{ij} - \hat{c}_{ij})^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} c_{ij}^2}}$$

for which $m$ is the number of points in each concentration profile $c_i$, $n$ is the number of components in the ALS model. The value of the $j$th point in the $i$th spectrum in the concentration profile extracted from the uncompressed data set is $c_{ij}$, and $\hat{c}_{ij}$ is value of the $j$th point in the $i$th spectrum in the high resolution concentration profiles that is obtained by applying 2D-NLWC-ALS to the original data set.

### 4.3 Experimental Section

Two different ion mobility spectrometers were used to collect two types of data sets that have different noise levels. The first instrument was a handheld chemical agent monitor (CAM, Type 482-301N, Graseby Ionics, Watford, Herts, UK). This instrument was operated in positive ion mode. The acetone dopant was removed from the instrument, so that ionization was based on water chemistry.

The instrument was interfaced with a single processor PII 200 MHz/64MB RAM computer through a data acquisition (DAQ) board, type AT-MIO-16X (National
Instruments, Austin, TX). The operating system was Microsoft Windows 98 Second Edition (Redmond, WA). Three chemical warfare (CW) simulants were used to collect CAM data sets. The CW simulants used were dimethyl methylphosphonate (DMMP) (97%, Lancaster Synthesis Inc., Windham, NH, Lot No. P07087 212-052-3), triethyl phosphate (TEPO) (99.8%, Aldrich Chemical Company Inc., Milwaukee, WI, Lot No. 12502MI) and dipropylene glycol methyl ether (DPM) (99+%?, Aldrich Chemical Company Inc., Milwaukee, WI, Lot No. 10613TA). The data sets were acquired with a homebuilt VI (National Instruments, Austin, TX) program. Three data sets were collected from this instrument. Each data set contained TEPO, DMMP, and DPM. The IMS waveform was sampled at 80 kHz and a single spectrum comprised 1,500 points. For each data set, approximately 2,000 spectra were saved.

The second instrument was an ion trap mobility spectrometer (ITMS), ITEMISER contraband detection and identification system (GE Ion Track Instruments, Inc., Wilmington, MA). Positive mode was used with an ammonia reagent ion. The ITMS system was interfaced to a laptop computer (pentium III 850 MHz and 384 MB memory) via a card (Type DAQCard-Al-16XE-50). The operating system is Microsoft Windows XP Pro with Service Pack 1 (Redmond, WA). The IMS waveform was sampled at 80 kHz. DMMP and DPM were sampled for each data set. Three data sets were used in this paper for data compression and analysis. A single spectrum comprised exactly 1,500 points. For each data set, 15,000 spectra were collected.

For both instruments, homemade virtual instrument (VI) programs were implemented in LabVIEW 6.1 (National Instruments, Austin, TX) to acquire data from
the IMS instruments. Samples were not diluted before data collection. Each data set was collected by placing a 5 mL open vial that contained 2 mL of pure sample at a distance of 1 cm from the inlet of the CAM for 1 to 3 seconds, and the next sample was not introduced until the instrumental response returned to the blank signal. The data in the drift time range of 0 to 0.5 ms was omitted to eliminate the gating pulse from the evaluations.

The data processing was implemented on PC with an AMD XP 2200+ processor and 512 MB RAM. The operating system was Window XP Professional (service pack 1).

MATLAB 6.5 R13 (The MathWorks, Inc, Natick, MA) was used to perform computational evaluations and modeling of the wavelets. The binary data collected from the VI were read by MATLAB and represented in matrix format with each row comprising an IMS spectrum. The 2D-NLWC-ALS algorithm and reconstruction routine was written in MATLAB. The WaveLab 802 toolbox was used for wavelet applications in this paper. Twenty-two wavelet filters were investigated from three different families of wavelets, ten from the daublet families, five from the coiflet families, and seven from the symmlet families.

The data sets were reduced to speed up the exhaustive evaluation. The data sets with reduced data size, which will be referred as smaller data sets, are used to find number of components in each data set and evaluate the effect of wavelet filter type during the compression. The data sets from CAM were decimated to smaller size by saving every 10th spectrum in the data acquisition dimension, and every 10th point in
the drift time dimension. The data sets collected from ITEMISER were decimated to smaller size by saving every 50th spectrum in the data acquisition dimension, and every 10th point in drift time dimension. The wavelet filters were optimized by using a grid search.

The reduced ITEMISER data sets were also used for reconstruction of the original data sets. The original size CAM data sets were used for reconstruction. Table 4.1 gives all the dimension information of CAM and ITEMISER for both original and smaller data sets.

After the wavelet filters were selected by using the small data sets, the full-size data sets of both CAM and ITEMISER were used for compression and reconstruction evaluation using the optimal wavelet filters.

### 4.4 Results and Discussion

Two ion mobility spectrometers, the CAM and ITEMISER were used to collect two types of data sets with different noise levels. Three chemical weapon simulants (TEPO, DMMP, and DPM) were used for the data sets collected from CAM. To calculate the reduced mobility for each of the compounds, the DMMP dimer peak was used. This peak has a relatively constant reduced mobility value under different temperature environments.\(^8\) For the data sets collected by CAM, because the drift tube was at room temperature, the room temperature value of 1.43 cm\(^2\)V\(^{-1}\)s\(^{-1}\) was used as the reference mobility. For the data sets collected with the ITEMISER, the DMMP dimer reduced mobility of 1.40 cm\(^2\)V\(^{-1}\)s\(^{-1}\) was used because it corresponded with the drift tube
Table 4.1 Data set collection information for the three CAM data sets and three ITEMISTER data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data set dimension (Scan number × Drift time)</th>
<th>Small data set dimension (Scan number × Drift time)</th>
<th>Collection time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM data set 1</td>
<td>2055 × 1500</td>
<td>205 × 1024</td>
<td>139.34</td>
</tr>
<tr>
<td>CAM data set 2</td>
<td>2073 × 1500</td>
<td>207 × 1024</td>
<td>142.20</td>
</tr>
<tr>
<td>CAM data set 3</td>
<td>2111 × 1500</td>
<td>211 × 1024</td>
<td>143.41</td>
</tr>
<tr>
<td>ITEMISER data set 1</td>
<td>15006 × 1500</td>
<td>300 × 1024</td>
<td>746.23</td>
</tr>
<tr>
<td>ITEMISER data set 2</td>
<td>15010 × 1500</td>
<td>300 × 1024</td>
<td>746.40</td>
</tr>
<tr>
<td>ITEMISER data set 3</td>
<td>15015 × 1500</td>
<td>300 × 1024</td>
<td>746.58</td>
</tr>
</tbody>
</table>
temperature of 200 °C. The reduced mobilities for the ions from the other compounds were calculated using these reference values for high and low temperature DMMP dimer reduced mobilities.

Figure 4.2 gives a surface plot on data set of TEPO-DMMP-DPM comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM. This figure shows that TEPO was sampled by the CAM first, followed by DMMP, and then DPM. The RIP had a significant decrease in intensity when there was a new analyte introduced to the system.

ALS was used to model the TEPO-DMMP-DPM CAM data set. The ASIMPLISMA algorithm was used to obtain the initial concentration profiles for ALS. The baseline and noise values are calculated using the drift time from 1 ms to 4 ms. Maximum number of components is set to 25. For the three CAM data sets, the calculated estimate number of components is 4, which includes RIP, TEPO, DMMP, and DPM. For the three ITEMISER data sets, the calculated estimate number of components is 3, which includes RIP, DMMP, and DPM.

Figure 4.3 gives the spectra for the CAM data set. TEPO gave two peaks, a monomer at 7.96 ms ($K_0 = 1.56 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) and a dimer peak at 10.86 ms ($K_0 = 1.14 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$). DMMP gave a monomer peak at 6.85 ms ($K_0 = 1.81 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) and a dimer peak at 8.68 ms ($K_0 = 1.43 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$). DPM peaks were found at 7.69 ms ($K_0 = 1.61 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) for the monomer and 10.00 ms ($K_0 = 1.24 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) for the dimer.

Figure 4.4 gives the concentration profiles for the TEPO-DMMP-DPM data set. The compounds sampled by the CAM were introduced as TEPO, DMMP, and DPM.
Figure 4.2 TEPO-DMMP-DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode using CAM. The gating pulse has been removed.
Figure 4.3 ALS spectra for TEPO-DMMP-DPM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM.
Figure 4.4 ALS concentration profiles for TEPO-DMMP-DPM CAM data set comprising 2,073 spectra with 1,500 points per spectrum collected in positive mode with the CAM.
Each compound was sampled once the instrument returned to baseline response from the previous compound. Note that at time of 12.69 min, there is a significant decrease in the RIP. This problem was caused by the vibration of the cable so that at that time point, one zero-signal spectrum was obtained. This spurious signal was easily detected using ALS.

The data sets collected from the ITEMISER compared to that of the CAM have lower noise. TEPO did not give any signal in the ITEMISER, because of the competitive charge suppression by the ammonia dopant. DMMP and DPM were used for the data sets collected from ITEMISER. Figure 4.5 gives the DPM-DMMP data set comprising 15,000 spectra with 1,500 points per spectrum collected in positive mode using ITEMISER. The reduced mobility values for both DPM and DMMP were greater than for the CAM because of the higher temperature drift tube in ITEMISER. DMMP was introduced to the system before DPM.

ALS was applied to the original ITEMISER data set. Figure 4.6 gives the spectra for this data set in a drift time window of 3 ms to 11 ms. The DPM monomer peak occurred at drift time of 6.58 ms ($K_0 = 1.60 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$). The DMMP monomer peak appeared at a drift time of 5.64 ms ($K_0 = 1.87 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) and a dimer at 7.52 ms ($K_0 = 1.40 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$).

Figure 4.7 gives the concentration profiles for the DPM-DMMP ITEMISER data set. The smooth appearance of the integrated intensity for each of the compounds shows the lower noise levels of the ITEMISER instrument. From the concentration profiles, DMMP was introduced to the system before the introduction of DPM. The DPM concentration began to increase after 590 s.
Figure 4.5 DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed. Smaller data set comprising 300 spectra with 1,024 points per spectrum was used to plot this graph by cutting the ending points.
Figure 4.6 ALS spectra for DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed.
Figure 4.7 ALS concentration profiles for DPM-DMMP data set comprising 15,006 spectra with 1,500 points per spectrum collected in positive mode with the ITEMISER. The gating pulse has been removed.
Two way analysis of variance (ANOVA) was applied to both CAM and ITEMISER data sets to evaluate the change of error in reconstruction using different wavelet filters. For Table 4.2, fifteen combinations of wavelet filters were randomly chosen from 22 wavelet filters for each dimension and each wavelet combination was used for reconstruction. The $E_{RMS}$ values from 15 reconstructed data sets using 2D-NLWC-ALS with randomly selected wavelet filter combinations comprised the rows and 3 independent CAM measurements comprised the columns of a data set that was subjected to ANOVA. ANOVA indicates that there are no significant differences among the reconstruction errors from the CAM data sets. ANOVA indicates that the differences among wavelet filters are insignificant.

ANOVA was applied to the data in Table 4.3 for reconstruction errors from the ITEMISER data. The results are similar in that no significant differences among the 3 ITEMISER data sets or the 15 wavelet filters are obtained.

The reconstruction errors resulting from uniform compression of data sets with varying wavelets were compared to study the performance of different wavelets for compression efficacy. During this evaluation procedure, the reduced data sets were used to decrease computation time. Twenty two wavelet filters were applied to both row and column compression using all possible combinations, which resulted in 484 wavelet filters. Several compression levels in the spectral dimension were evaluated. As a result, 484 values of both $E_{RMS}$ and $SE_{RMS}$ were obtained for each level of compression level. The minimum values were selected for different compression levels and the results are given in Figure 4.8 and Figure 4.9.
Table 4.2 ANOVA tables for 2D-NLWC-ALS of three CAM data sets with 15 wavelet
filter combinations for each data set.*

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelet Filters</td>
<td>0.00030</td>
<td>14</td>
<td>2.13×10^{-5}</td>
<td>0.66</td>
<td>0.7876</td>
</tr>
<tr>
<td>Data Sets</td>
<td>0.00009</td>
<td>2</td>
<td>4.43×10^{-5}</td>
<td>1.38</td>
<td>0.2676</td>
</tr>
<tr>
<td>Error</td>
<td>0.00090</td>
<td>28</td>
<td>3.21×10^{-5}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.00129</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Each column is the $E_{RMS}$ value generated after reconstruction using the 2D-NLWC-ALS algorithm using a random filter combination applied to row compression and column compression. Each row is a different CAM data set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Row Filter</th>
<th>Column Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>symmlet 6</td>
<td>coiflet 2</td>
</tr>
<tr>
<td>2</td>
<td>daublet 16</td>
<td>coiflet 5</td>
</tr>
<tr>
<td>3</td>
<td>coiflet 2</td>
<td>daublet 20</td>
</tr>
<tr>
<td>4</td>
<td>coiflet 5</td>
<td>coiflet 4</td>
</tr>
<tr>
<td>5</td>
<td>symmlet 5</td>
<td>symmlet 9</td>
</tr>
<tr>
<td>6</td>
<td>coiflet 1</td>
<td>symmlet 7</td>
</tr>
<tr>
<td>7</td>
<td>daublet 8</td>
<td>symmlet 9</td>
</tr>
<tr>
<td>8</td>
<td>daublet 12</td>
<td>daublet 12</td>
</tr>
<tr>
<td>9</td>
<td>symmlet 7</td>
<td>symmlet 4</td>
</tr>
<tr>
<td>10</td>
<td>symmlet 7</td>
<td>daublet 10</td>
</tr>
<tr>
<td>11</td>
<td>daublet 14</td>
<td>coiflet 4</td>
</tr>
<tr>
<td>12</td>
<td>daublet 12</td>
<td>daublet 20</td>
</tr>
<tr>
<td>13</td>
<td>daublet 4</td>
<td>symmlet 9</td>
</tr>
<tr>
<td>14</td>
<td>coiflet 3</td>
<td>daublet 18</td>
</tr>
<tr>
<td>15</td>
<td>coiflet 1</td>
<td>daublet 16</td>
</tr>
</tbody>
</table>
Table 4.3 ANOVA tables for 2D-NLWC-ALS of three ITEMISER data sets with 15 different wavelet filter combinations for each data set. *

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelet Filters</td>
<td>0.16240</td>
<td>14</td>
<td>0.01160</td>
<td>1.04</td>
<td>0.4484</td>
</tr>
<tr>
<td>Data Sets</td>
<td>0.00293</td>
<td>2</td>
<td>0.00147</td>
<td>0.13</td>
<td>0.8776</td>
</tr>
<tr>
<td>Error</td>
<td>0.31317</td>
<td>28</td>
<td>0.01118</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.47850</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Each column is the $E_{RMS}$ value generated after reconstruction using the 2D-NLWC-ALS algorithm using a random filter combination applied to row compression and column compression. Each row is a different ITEMISER data set.

<table>
<thead>
<tr>
<th>No.</th>
<th>Row Filter</th>
<th>Column Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>daublet 6</td>
<td>daublet 20</td>
</tr>
<tr>
<td>2</td>
<td>symmlet 4</td>
<td>symmlet 7</td>
</tr>
<tr>
<td>3</td>
<td>daublet 12</td>
<td>daublet 12</td>
</tr>
<tr>
<td>4</td>
<td>symmlet 7</td>
<td>daublet 10</td>
</tr>
<tr>
<td>5</td>
<td>symmlet 5</td>
<td>symmlet 8</td>
</tr>
<tr>
<td>6</td>
<td>daublet 10</td>
<td>daublet 12</td>
</tr>
<tr>
<td>7</td>
<td>daublet 4</td>
<td>daublet 4</td>
</tr>
<tr>
<td>8</td>
<td>coiflet 4</td>
<td>daublet 10</td>
</tr>
<tr>
<td>9</td>
<td>symmlet 6</td>
<td>daublet 8</td>
</tr>
<tr>
<td>10</td>
<td>daublet 8</td>
<td>symmlet 9</td>
</tr>
<tr>
<td>11</td>
<td>daublet 20</td>
<td>daublet 16</td>
</tr>
<tr>
<td>12</td>
<td>daublet 14</td>
<td>daublet 16</td>
</tr>
<tr>
<td>13</td>
<td>daublet 18</td>
<td>coiflet 3</td>
</tr>
<tr>
<td>14</td>
<td>daublet 20</td>
<td>symmlet 7</td>
</tr>
<tr>
<td>15</td>
<td>symmlet 8</td>
<td>daublet 12</td>
</tr>
</tbody>
</table>
Figure 4.8 Minimum $E_{RMS}$ evaluation from all 484 wavelet filter combinations applied to time dimension and spectra dimension using 2D-NLWC-ALS method. The time dimension was not compressed while the minimum $E_{RMS}$ was picked for different compress ratio in the dimension of spectra compression. Three data sets were from CAM on TEPO-DMMP-DPM and three were from ITEMISER on DMP-DMMP.
Figure 4.9 Minimum $SE_{RMS}$ evaluation from all 484 wavelet filter combination applied to time dimension and spectra dimension using 2D-NLWC-ALS algorithm. The time dimension was not compressed while the minimum $E_{RMS}$ was picked for different compress ratio in the dimension of spectra compression. Three data sets were collected from CAM on TEPO-DMMP-DPM and three were ITEMISER data sets on DMP-DMMP.
After applying the 2D-NLWC-ALS algorithm to a data set, the saved compressed data is the low-resolution concentration profiles \( C_c \) and the low-resolution spectra \( S_c \) in step 5. The number of points saved in \( C_c \) and \( S_c \) that efficiently reflect the original data depends on the number of wavelet coefficients \( W_{cc} \) saved in step 3. To decide the number of points that should be saved in \( C_c \) and \( S_c \), the minimum \( E_{RMS} \) values are plotted as a function of the number of compressed points per spectrum for three CAM data sets and three ITEMISER data sets. For the three CAM data sets, the \( E_{RMS} \) increased significantly after compressing the spectra below 250 points. The ALS spectra from this level of compression yielded small artifacts. When the compression was greater than 300 points per spectrum, no artifacts were observed in the ALS spectra, thus the optimum number of points for each spectrum should be 300 for CAM data. For the CAM data sets, each spectrum was compressed to 300 points, the daublet 8 was the optimum wavelet filter for both row and column compression. For the ITEMISER data sets, 200 points per spectrum were found to be the optimum for compression and daublet 14 was the best filter for both row and column compression.

Figure 4.10 gives the reconstructed spectra of TEPO-DMMP-DPM from the CAM after application of the 2D-NLWC-ALS algorithm. The daublet 8 filter was applied to both row and column wavelet compression. The original size of the data set was 2,073 by 1,500. The low resolution spectra \( S_c \) saved is a matrix of \( 4 \times 300 \). The four columns are the ALS spectra for RIP, TEPO, DMMP, and DPM, respectively. Each spectrum with 1,500 points in the original data set was compressed to 300 points. The low resolution concentration \( C_c \) is a matrix of \( 4 \times 100 \). The original 2,073 spectra in the
Figure 4.10 Reconstructed spectra of TEPO-DMMP-DPM data sets using 2D-NLWC-ALS algorithm at compression ratio of 640 ppm. Wavelet filter of daublet 8 was used at both the drift time dimension and the data acquisition dimension. The compressed wavelet coefficients keep 100 points in time dimension and 300 points in spectra dimension.
data acquisition time dimension were compressed to 100 spectra. The compression ratio was 510 ppm. The $E_{RMS}$ is 0.0063 and the $RE_{RMS}$ is 1.62%.

Figure 4.11 gives the spectra obtained from the reconstructed CAM data set using ALS. No visible ringing effects can be observed for any of the peaks. The drift time of DMMP dimer was at 7.52 ms, the reduced mobility of $1.43 \text{ cm}^2 \text{V}^{-1} \text{s}^{-1}$ was used as the standard to calculate the other $K_0$ values of the other compounds. The drift times of RIP (5.97 ms), TEPO dimer (10.85 ms), DMMP dimer (8.68 ms), and DPM monomer (7.69 ms) were kept at the same value after the reconstruction. The drift time of TEPO monomer and DMMP monomer both were shifted for 0.01 ms. However, this insignificant drift time shift in the reconstruction is so small that all the compounds remained at the same reduced mobility after reconstruction.

One of the benefits of NLWC is that it has the ability to de-noise the spectra. Note that, the concentration profiles from the original CAM data set are very noisy. At the time of 12 second, there is a significant decrease in RIP concentration. This problem was caused by the vibration of the cable so that at that time point, zero-signal spectrum was obtained. However, the corresponding concentration profiles for the reconstructed CAM data set had this artifact removed which demonstrated the de-noising feature of NLWC. The concentration profiles in Figure 4.12 for the reconstructed ITEMISER data set were also smoother than those from the original data set.

Figure 4.13 gives the reconstructed spectra of DMMP-DPM from ITEMISER after applied 2D-NLWC-ALS algorithm. The daublet 14 filter was applied to both row and column for wavelet compression. The original size of the data set is $15,000 \times 1,500$. 
Figure 4.11  ALS spectra for TEPO-DMMP-DPM CAM data set reconstructed using 2D-NLWC-ALS algorithm at compression ratio of 640 ppm. Wavelet filter of daublet 8 was used at both the drift time dimension and the data acquisition dimension.
Figure 4.12  ALS concentration profiles for reconstructed TEPO-DMMP-DPM CAM data set using 2D-NLWC-ALS algorithm at compression ratio of 640 ppm. Wavelet filter of daubelet 8 was used at both the drift time dimension and the data acquisition dimension.
Figure 4.13 Reconstructed spectra of DMP-DMMP ITEMISER data set from 2D-NLWC-ALS algorithm at compression ratio of 62 ppm. Wavelet filter of daublet 14 was used at both the drift time dimension and the data acquisition dimension. The saved compressed spectra $S_c$ is a matrix of $200 \times 4$. The saved compressed concentration profiles $C_c$ is a matrix of $150 \times 4$. 
The saved low resolution spectra $S_c$ is a matrix of 3 by 200. The three columns are the ALS spectra for RIP, DMMP, and DPM respectively. Each spectrum with 1,500 points in the original data set was compressed to 200 points. The low resolution concentration $C_c$ is a matrix of 150 by 3. The original 15,000 spectra collected in the data acquisition time dimension were compressed to 100 spectra. The compression ratio is 46 ppm. The $E_{RMS}$ for the reconstruction of this data set is 0.0092 and the $RE_{RMS}$ is 0.42%.

To evaluate the efficacy of the reconstructed data set, ALS was applied. Figure 4.14 gives the spectra of the reconstructed ITEMISER data set. There is no peak shifting occurred after the reconstruction. The drift time of DMMP dimer was at 7.52 ms, the reduced mobility of $1.40 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$ was used as the reference value to calculate the other reduced mobilities of the other compounds. The drift times of DMMP monomer and DPM monomer maintained the same values after reconstruction. There were insignificant peak intensity attenuations less than 0.09% for the reconstruction. All characteristic ions retained the same reduced mobility after reconstruction with respect to the uncompressed spectra.

The ALS concentration profiles obtained from the reconstructed ITEMISER data set were plotted as Figure 4.15. Compared to the ALS concentration profile generated from original data set, the ALS concentration curves calculated from the reconstructed data set is smoother (less noise level) than the one from the original data set. The $CE_{RMS}$ is calculated as 0.0088.

Table 4.4 gives the comparison of the drift time for all the components in CAM and ITEMISER data set using ALS spectra. The reconstructed data sets were the product
Figure 4.14  ALS spectra for ITEMISER data set reconstructed using 2D-NLWC-ALS algorithm at compression ratio of 62 ppm. Wavelet filter of daublet 14 was used at both the drift time dimension and the data acquisition dimension.
Figure 4.15  ALS concentration profiles for reconstructed ITEMISER data set. Wavelet filter of daublet 14 was used at both the drift time dimension and the data acquisition dimension.
Table 4.4 Comparison of the drift time of each peak of all the components in CAM and ITEMISER data set using ALS spectra.*

<table>
<thead>
<tr>
<th>Component</th>
<th>Uncompressed CAM data (ms)</th>
<th>Reconstructed CAM data (ms)</th>
<th>Uncompressed ITEMISER data (ms)</th>
<th>Reconstructed ITEMISER data (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIP</td>
<td>5.97</td>
<td>5.97</td>
<td>4.33</td>
<td>4.33</td>
</tr>
<tr>
<td>TEPO monomer</td>
<td>7.96</td>
<td>7.95</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>TEPO dimer</td>
<td>10.85</td>
<td>10.85</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DMMP monomer</td>
<td>6.85</td>
<td>6.84</td>
<td>5.64</td>
<td>5.64</td>
</tr>
<tr>
<td>DMMP dimer</td>
<td>8.68</td>
<td>8.68</td>
<td>7.52</td>
<td>7.52</td>
</tr>
<tr>
<td>DPM monomer</td>
<td>7.69</td>
<td>7.69</td>
<td>6.58</td>
<td>6.58</td>
</tr>
<tr>
<td>DPM dimer</td>
<td>10.00</td>
<td>9.98</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* The values reported are (drift time/maximum peak intensity). Uncompressed CAM data set used as shown in Figure 4.2 and uncompressed ITEMISER data set is as in Figure 4.5. The reconstructed CAM data set is as Figure 4.10 and Figure 4.13 shows the reconstructed ITEMISER data set. The CAM data set was reconstructed from a matrix of spectra with size 300 × 4 and a matrix of concentration profiles with size of 100 × 4. The CAM data set was reconstructed from a matrix of spectra with size 200 × 3 and a matrix of concentration profiles with size of 150 × 3.
of the reconstructed high resolution spectra and concentration profiles. For the CAM data sets, the peak position of RIP, TEPO dimer, DMMP dimer, DPM monomer were kept the same between the compressed and original data. The peak positions of the TEPO monomer and the DMMP monomer were decreased by 0.01 ms. For ITEMISER data set, there is no shift in drift time after reconstruction.

Reduced mobility is used instead of drift time because it is independent of many experimental factors. The reduced mobilities were calculated based on standards. Table 4.5 gives the reduced mobility for both the uncompressed spectra and the reconstructed spectra using the 2D-NLWC-ALS algorithm. The DMMP dimer reduced mobility value of 1.43 cm$^2$V$^{-1}$s$^{-1}$ under room temperature was used as standard for CAM data sets. At high temperatures around 200 °C, the value of 1.40 cm$^2$V$^{-1}$s$^{-1}$ was used for the DMMP dimer as a standard to calibrate the other compounds for the ITEMISER data sets.

4.5 Conclusions

A two-dimensional nonlinear wavelet compression method combined with ALS (2D-NLWC-ALS) algorithm was developed and applied to IMS data. ALS served in this algorithm not only as a modeling method, but also a compression method. The smaller size ALS model is saved as the compressed data, and higher compression ratio are obtained along with the advantage of modeling the original data set. A compression ratio of 510 ppm, error of root mean square ($E_{RMS}$) of 6.3 mV, and relative root mean square ($RE_{RMS}$) of 1.62% were achieved for data sets collected by CAM. A compression ratio of 46 ppm, error of root mean square ($E_{RMS}$) of 9.2 mV, and relative of root mean square
Table 4.5 The comparison of the reduced mobility ($K_0$) for all the peaks in CAM and ITEMISER data sets using ALS spectra.*

<table>
<thead>
<tr>
<th></th>
<th>Uncompressed CAM data (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Reconstructed CAM data (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Uncompressed ITEMISER (cm$^2$V$^{-1}$s$^{-1}$)</th>
<th>Reconstructed ITEMISER (cm$^2$V$^{-1}$s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.56</td>
<td>1.56</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>TEPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.14</td>
<td>1.14</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DMMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.81</td>
<td>1.81</td>
<td>1.87</td>
<td>1.87</td>
</tr>
<tr>
<td>DMMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.43 (standard)</td>
<td>1.43 (standard)</td>
<td>1.40 (standard)</td>
<td>1.40 (standard)</td>
</tr>
<tr>
<td>DPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monomer</td>
<td>1.61</td>
<td>1.61</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td>DPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimer</td>
<td>1.24</td>
<td>1.24</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* Uncompressed CAM data set used as shown in Figure 4.2 and uncompressed ITEMISER data set is as in Figure 4.5. The reconstructed CAM data set is as Figure 4.10 and Figure 4.13 shows the reconstructed ITEMISER data set. The $K_0$ values were calculated from the drift times provided by Table 4.4.
(RE_{RMS}) of 0.42% were achieved for data sets collected by ITEMISER. NLWC avoids peak distortion and eliminates artifacts caused by linear wavelet compression while achieving similar or improved. The ALS models of the reconstructed spectra and the original spectra represent satisfactory spectra and concentration profiles with indistinguishable differences. The reduced mobilities for the characteristic ions from each compound in both instruments were unaltered by the compression and reconstruction.
Chapter 5  Pattern Recognition of Unintegrated Mass Spectra

5.1  Introduction

Histograms of peak areas are perhaps the most popular method for representing mass spectra. Chemical information manifested in peak shapes is discarded during this process. If peaks are not adequately resolved variability may be introduced that obfuscates the distinguishing characteristics of similar spectra.\textsuperscript{81, 82}

It is reported in 1947 that the mass spectra of the dissociation patterns of $o$-, $m$-, and $p$-diethylbenzene and $o$-, $m$-, and $p$-xylene were so much alike that no resolution of the isomers in either group could be made.\textsuperscript{83} It is well-known that the isomeric xylenes are practically indistinguishable from one another by using a conventional 70 eV electron impact mass spectrometer.\textsuperscript{84} A mass spectrometric method was developed to distinguish the molecular ions of the three isomeric xylenes via measurements of kinetic energy release accompanying photoexpulsion of a methyl radical.\textsuperscript{85} Metastable ion spectra and deuterium labeling have been used to investigate isomeric xylene ions.\textsuperscript{86} A Fourier-transform mass spectrometer, energetic (6.4 eV) multiphoton ionization (MPI) or photodissociation can differentiate isomers that yield similar electron ionization spectra.\textsuperscript{87} The high ionization efficiency makes possible high resolution MPI spectra.

However, all these methods that can differentiate the xylene isomers require specialized instrumentation. This chapter shows that GC/MS instruments have the capability of distinguishing among these isomers by omitting the histogram-preprocessing step and exploiting peak shape information contained in the mass spectral profiles, which are the unintegrated mass spectra.
The mass spectral profiles are larger than the histogram of integrated peak areas. Wavelet compression can be applied to reduce the size of the mass profiles to manageable sizes and at the same time achieve a signal-to-noise ratio improvement by removal of the high frequency noise components from the spectra.

The Fourier transform (FT) is widely used for signal processing, such as convolution, correlation, and smoothing. The FT methods are useful for data compression. A two dimensional compression method was devised and applied to ion mobility spectrometry data in both the drift time and sample acquisition dimensions. Compared to the FT, the wavelet transform (WT) offers several advantages over the FT including fast and simple implementation.

The WT methods have diverse application areas in the information sciences. Recently, WT methods have become a popular signal-processing tool for chemists. The WT technique has been exploited for processing absorbance spectra, chromatograms, and electrochemical signals. The wavelet transform can be used with other chemometric approaches, such as principal component analysis, partial least squares analysis, and artificial neural networks.

Data compression reduces the size of a data set while minimizing the loss of important chemical information. Noise can be preferentially loosed in some implementations thus improving the quality of compressed data. Two-dimensional compression can be achieved with the WT by transforming the signal into the wavelet domain and retaining a reduced number of coefficients. Reviews and tutorials about wavelet transform are available elsewhere.
5.2 Theory

The preprocessing methods for the three types of preprocessing methods are given in Table 5.1. The mass spectral profile is obtained with a 0.16 \( m/z \) point spacing between mass resolution elements and will hence be referred to as a profile. In the second preprocessing method that generates binned spectrum, the intensities are integrated across 1.00 \( m/z \) wide bins that are uniformly distributed across the mass range. The third method detects mass peaks and integrates the peak areas. The integrated peak areas were then integrated with 1.00 \( m/z \) wide bins to yield a histogram with an integer abscissa.

Wavelet compression was applied to the mass profiles to reduce the spectra to manageable sizes as well as digitally removing high frequency noise components. Comparisons between spectra acquired from different preprocessing methods and wavelet compressions were evaluated. During the procedures of WT implementation, two digital filters are used. One filter is a low-pass filter and the other is a high-pass filter. The wavelet filter pair is applied recursively to the data set by convolving the wavelet filter points with the data. Each recursive filter operation is referred to as a level. For linear compression, the detailed wavelet coefficients (i.e., the high-pass filtered data) are discarded. The numbers of low-pass coefficients that comprise the compressed data are halved at each level.

PCA is used for visualization of data by retaining variance in a reduce set of coordinates. Each isomer data set obtained from GC/MS can be expressed as a data matrix \( X \), in which each row vector represents a spectrum and the column dimension indicates the time of sample acquisition.
Table 5.1  Comparison of three types of mass spectra.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Main Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass profile</td>
<td>The mass spectrum collected from the GC-MS without any preprocessing. The instrument yielded a 0.16 m/z mass resolution element.</td>
</tr>
<tr>
<td>Binned Spectrum</td>
<td>The mass spectral intensities were integrated in uniformly distributed 1.00 m/z mass resolution wide bins.</td>
</tr>
<tr>
<td>Histogram</td>
<td>The spectrum obtained from the integration of the peak area from the mass profile and then were binned at 1.00 m/z mass resolution. This preprocessing is different from binning the spectrum at 1.00 m/z mass resolution.</td>
</tr>
</tbody>
</table>
Projected difference resolution (PDR) is used as the criteria to evaluate the discrimination two isomer data sets and defined as below

$$PDR(1, 2) = \frac{|\bar{t}_1 - \bar{t}_2|}{2(s_1 + s_2)}$$  \hspace{1cm} (5.1)$$

for which PDR(1,2) is the PDR between isomer 1 and 2, measured by the absolute difference between the averages of the projections ($\bar{t}_i$) divided by two times the sum of the standard deviations $s$ of the isomers scores about their averages. For each spectrum, the method to calculate the projection is given by equation (5.2)

$$t_i = x_i (\bar{x}_1 - \bar{x}_2)^T$$  \hspace{1cm} (5.2)$$

for which $t_i$ is the score obtained from mass spectrum $x_i$ by projecting onto the difference between the two isomer averages $\bar{x}_1$ and $\bar{x}_2$.

To evaluate the loss of important information, reconstruction errors were used to evaluate compression efficacy. The root mean square error ($E_{RMS}$) of the reconstructed dataset with respect to the original dataset is defined in equation (5.3).

$$E_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}$$  \hspace{1cm} (5.3)$$

for which $x_i$ is a point from a spectrum of the data set and $\hat{x}_i$ corresponds to the reconstructed point (i.e., application of the inverse WT to the zero-filled compressed spectrum), respectively. The number of points in the spectrum is $n$.

5.3 Experimental Section
Mass spectra were collected with a Hewlett Packard 5988 Gas Chromatograph-Mass Spectrometer. A RTX-65TG fused silica column (Restek Corp., Bellefonte, PA) was used with cross-bonded 65% diphenyl and 35% dimethyl polysiloxane, 30 m length, 0.32 mm i.d., 0.1 µm film thickness. Perfluorotributylamine (Scientific Instrument Services Inc., Ringoes, NJ) was used for mass spectrometer calibration.

The samples were, \( p \)-xylene (99.6%, Fisher Scientific, Fair Lawn, NJ, Lot No. 952739), \( o \)-xylene (95%, Spectrum, Gardena, CA, Lot No. IF151), and \( m \)-xylene (purified grade, Fisher Scientific, Fair Lawn, NJ, Lot No. 782730). For each xylene, the neat sample was used to rinse a 10 µL syringe three times. The sample vapor retained in the syringe was diluted by pumping 20 times with air. Ten µL of diluted sample vapor was injected into the GC. The GC temperature program was 80 °C for 3 min, 10 °C /min to 120 °C, 20 °C/min to 150 °C, and 150 °C for 10 min. For each xylene, three data sets were collected.

Data were acquired with a Dell 400 MHz personal computer with 128 MB RAM and the OS/2 Warp 4.0 operating system. The mass spectrometer was controlled using Prolab Vector/Two GC-LC/MS software, version 3.02.00 (Prolab Resources Inc., Madison, WI). Virtual instruments (VIs) were created using LabVIEW (National Instruments, Austin, TX) 6.1. The GC-MS data were processed using a VI that performed histogram calculations, binned spectrum calculations, and generated mass profiles.

A raw file that contains the spectral profiles (mass, intensity) was generated from the Vector/Two software (version 3.02.00). An interactive LabVIEW (National
Instruments, Austin, TX, version 6.1) VI was written that allowed the spectra to be selected from the run and preprocessed using different modes. The operating system was MS Windows XP SP1. The preprocessed spectra were saved as text files so that they could be imported into MATLAB 6.5 R13 (The MathWorks, Inc. Natick, MA) for processing and evaluation. The data was processed on a desktop PC (Gateway, Inc.) with a 1.2 GHz AMD (Athlon) processor and 512 MB RAM. The same raw spectral profiles were used for evaluation of the different preprocessing methods.

Wavelet compression routines were written in MATLAB R13. The wavelets evaluated in this paper were daublet 2-10, coiflet 2-10, and symmlet 4-10 for which the numbers indicate the vanishing moments of the filter. Thirty spectra in each xylene data set were compressed and reconstructed. The $E_{\text{RMS}}$ for each spectrum was calculated and totaled. The coiflet 4 filter gave the lowest reconstruction error at compression level 1 and 2, and was used in further wavelet compression studies with the three xylenes.

5.4 Results and Discussion

For $m$, $o$, and $p$-xylene respectively, 30 mass spectra were obtained from each GC run. The three types of mass spectra using different preprocessing methods are histogram, 1.00 $m/z$ binned spectra and the mass spectral profiles. They were evaluated with PCA. The first 3 components that accounted for $\sim 98$-99 % of the cumulative variance were used to plot the graph to show the separation of three isomers.

Figure 5.1 gives one of the $m$-xylene data sets that comprised 160 mass spectra with 70 points per mass profile. All scans began at $m/z$ 50 and ended at $m/z$ 120 for each
Figure 5.1  *m*-xylene dataset comprised of 160 mass spectra in a TIC profile with 70 points per spectrum collected by GC/MS. The 100 spectra took 0.25 s of scan time.
spectrum. The \( m/z \) at 51, 65, 77, 91, 105 and 106 are apparent in the graph.

The example of a histogram, a binned spectrum, and a mass profile based on single mass spectrum are given in Figure 5.2. The profiles from the three xylene isomers were reproducible for all sets of 30 mass spectra. Figure 5.2 (a) is a histogram for \( m \)-xylene. Figure 5.2 (b) is a binned spectrum of peak integrated areas at mass resolution of 1.00 \( m/z \). This binned spectrum was produced by integrating the range of [\(-0.5, 0.5\)] \( m/z \) for each \( m/z \). Figure 5.2 (c) increased the mass resolution to 0.16 u, so that the integration happen every other 0.16 \( m/z \) to generate the mass spectral profile. From the third plot, some important information, such as abundance at a specific \( m/z \) point and peak shape, were saved.

PCA is a method to visualize multivariate data. The data sets were mean-centered before the PCA. Figure 5.3 is a PCA score plot obtained from the histogram spectra. The scores, in which each point represents a spectrum, are overlapped. The PDR of the histogram spectra among 90 spectra of three xylenes were calculated in Table 5.2. The \( m \)-xylene and \( p \)-xylene were barely separated. In this table, the precisions in the table are standard deviation. Three data sets for each xylene isomers from the different runs were used. Confidence intervals were calculated from the 9 independent data sets about the mean PDRs. The histogram had the lowest PDRs for the \( m \)-xylene and \( p \)-xylene spectra. \( o \)-xylene and \( p \)-xylene spectral clusters were not resolved. The most compressed profiles yielded better PDRs than the histogram spectra.

The binned spectra had better PDR than histograms. Figure 5.4 gives the scores on the first 3 components obtained for the 1.00 \( m/z \) binned spectra. Figure 5.5 gives the
Figure 5.2 The representative of a single mass spectrum obtained from the three preprocessing approaches: (a) histogram; (b) binned at 1.00 m/z mass resolution; (c) mass profile.
Figure 5.3  PCA cluster analysis of the histogram spectra.  m: m-xylene; p: p-xylene; o: o-xylene.
Table 5.2  Comparison of PDR between histogram, binned spectra at mass resolution 1.00, mass profiles, wavelet compression of mass profiles at level 1, 2 and 3, respectively.*

<table>
<thead>
<tr>
<th></th>
<th>Histograms</th>
<th>Binned spectra (Resolution 1.00)</th>
<th>Mass profiles</th>
<th>Wavelet compressed at level 1</th>
<th>Wavelet compressed at level 2</th>
<th>Wavelet compressed at level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$ and $o$-xylene</td>
<td>1.89 ±0.05</td>
<td>2.20±0.06</td>
<td>2.52±0.07</td>
<td>2.38±0.07</td>
<td>2.25±0.07</td>
<td>2.14±0.06</td>
</tr>
<tr>
<td>$m$ and $p$-xylene</td>
<td>1.01 ±0.06</td>
<td>3.29±0.06</td>
<td>3.51±0.09</td>
<td>3.42±0.09</td>
<td>3.47±0.09</td>
<td>3.60±0.08</td>
</tr>
<tr>
<td>$o$ and $p$-xylene</td>
<td>1.54 ±0.04</td>
<td>1.72±0.03</td>
<td>1.80±0.03</td>
<td>1.72±0.03</td>
<td>1.64±0.03</td>
<td>1.49±0.03</td>
</tr>
</tbody>
</table>

*Precisions used are the standard deviation.
Figure 5.4 PCA cluster analysis of the binned spectra at mass resolution of 1.00 m/z.

m: m-xylene; p: p-xylene; o: o-xylene.
**Figure 5.5** PCA cluster analysis of the mass profile. m: \textit{m}-xylene; p: \textit{p}-xylene; o: \textit{o}-xylene.
scores on the first 3 principal components obtained for the mass spectral profiles. Clearly, all the three isomers are well separated and there is no overlap between any of the clusters of isomers. Figure 5.6 and Figure 5.7 give a clear perspective of the effects of preprocessing on the resolution of the isomers. Greater resolution may have been obtainable if the mass spectral waveform could have been sampled faster.

Two-way ANOVA was applied to the PDRs generated from mass profiles and histograms. Three different data sets were used for each xylene. Three PDRs between each pair of xylene are used as rows. One column is generated from mass profiles and the other is from histograms. Table 5.3 indicates that the PDRs generated from mass profiles are significant from the ones from histograms. At the same time, the ANOVA indicates that the PDRs between each of xylene pairs or interactions between the treatments are insignificant.

Wavelet compression can speed up data processing, but if the data are overly compressed some important information may be lost. The better PDRs were obtained when the spectra were compressed by one level. Ninety mass profiles, 30 from each type of xylene were used for the $E_{RMS}$ evaluation. Each mass profile was compressed at different compression level using 21 types of wavelet filters and then was reconstructed for evaluation. The $E_{RMS}$ value of each mass profile was calculated. The mean value of $E_{RMS}$ and the standard deviation of the 90 spectra were obtained and given in Table 5.4. At compression levels of 1 and 2, the coiflet 4 served as the wavelet filter with lowest $E_{RMS}$ value. However, at a level 3 compression, the Daublet 3 gave the lowest reconstruction error. ANOVA was applied to the PDRs with mass profile and wavelet
Figure 5.6 Comparison of three PCA component cluster analysis between histogram and binned spectra with mass resolution 1.00 m/z. m: m-xylene; p: p-xylene; o: o-xylene.
Figure 5.7  Comparison of three PCA cluster analysis of mass profile and of binned spectra with mass resolution 1.00 \textit{m/z}.  \textit{m}: \textit{m}-xylene; \textit{p}: \textit{p}-xylene; \textit{o}: \textit{o}-xylene.
Table 5.3 ANOVA table for the comparison of PDRs between mass profile and histogram.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>4.37745</td>
<td>1</td>
<td>4.37745</td>
<td>32.54</td>
<td>0.0001</td>
</tr>
<tr>
<td>Replicates</td>
<td>0.16161</td>
<td>2</td>
<td>0.08080</td>
<td>0.60</td>
<td>0.5642</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.46535</td>
<td>2</td>
<td>0.23268</td>
<td>1.73</td>
<td>0.2188</td>
</tr>
<tr>
<td>Error</td>
<td>1.61451</td>
<td>12</td>
<td>0.13454</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6.61892</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.4 Evaluation of $E_{RMS}$ at three wavelet compression levels using different type of wavelet filters.

<table>
<thead>
<tr>
<th>Filter Type</th>
<th>$E_{RMS}$ (level 1)</th>
<th>$E_{RMS}$ (level 2)</th>
<th>$E_{RMS}$ (level 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daublet 2</td>
<td>0.0190±0.0001</td>
<td>0.0293±0.0001</td>
<td>0.0338±0.0001</td>
</tr>
<tr>
<td>Daublet 3</td>
<td>0.0180±0.0001</td>
<td>0.0249±0.0002</td>
<td>0.0296±0.0001</td>
</tr>
<tr>
<td>Daublet 4</td>
<td>0.0159±0.0001</td>
<td>0.0214±0.0002</td>
<td>0.0301±0.0001</td>
</tr>
<tr>
<td>Daublet 5</td>
<td>0.0139±0.0001</td>
<td>0.0231±0.0001</td>
<td>0.0332±0.0001</td>
</tr>
<tr>
<td>Daublet 6</td>
<td>0.0133±0.0001</td>
<td>0.0263±0.0001</td>
<td>0.0338±0.0001</td>
</tr>
<tr>
<td>Daublet 7</td>
<td>0.0141±0.0001</td>
<td>0.0270±0.0001</td>
<td>0.0319±0.0001</td>
</tr>
<tr>
<td>Daublet 8</td>
<td>0.0153±0.0001</td>
<td>0.0249±0.0001</td>
<td>0.0305±0.0001</td>
</tr>
<tr>
<td>Daublet 9</td>
<td>0.0162±0.0001</td>
<td>0.0227±0.0001</td>
<td>0.0316±0.0001</td>
</tr>
<tr>
<td>Daublet 10</td>
<td>0.0162±0.0001</td>
<td>0.0230±0.0001</td>
<td>0.0332±0.0001</td>
</tr>
<tr>
<td>Coiflet 2</td>
<td>0.0156±0.0001</td>
<td>0.0252±0.0002</td>
<td>0.0333±0.0001</td>
</tr>
<tr>
<td>Coiflet 4</td>
<td>0.0128±0.0001</td>
<td>0.0207±0.0002</td>
<td>0.0347±0.0001</td>
</tr>
<tr>
<td>Coiflet 6</td>
<td>0.0132±0.0001</td>
<td>0.0272±0.0001</td>
<td>0.0322±0.0001</td>
</tr>
<tr>
<td>Coiflet 8</td>
<td>0.0134±0.0001</td>
<td>0.0221±0.0001</td>
<td>0.0303±0.0001</td>
</tr>
<tr>
<td>Coiflet 10</td>
<td>0.0136±0.0001</td>
<td>0.0266±0.0001</td>
<td>0.0322±0.0001</td>
</tr>
<tr>
<td>Symmlet 4</td>
<td>0.0173±0.0001</td>
<td>0.0220±0.0001</td>
<td>0.0345±0.0001</td>
</tr>
<tr>
<td>Symmlet 5</td>
<td>0.0145±0.0001</td>
<td>0.0238±0.0001</td>
<td>0.0323±0.0001</td>
</tr>
<tr>
<td>Symmlet 6</td>
<td>0.0166±0.0001</td>
<td>0.0268±0.0001</td>
<td>0.0327±0.0001</td>
</tr>
<tr>
<td>Symmlet 7</td>
<td>0.0138±0.0001</td>
<td>0.0240±0.0002</td>
<td>0.0322±0.0001</td>
</tr>
<tr>
<td>Symmlet 8</td>
<td>0.0134±0.0001</td>
<td>0.0220±0.0001</td>
<td>0.0303±0.0001</td>
</tr>
<tr>
<td>Symmlet 9</td>
<td>0.0150±0.0001</td>
<td>0.0266±0.0001</td>
<td>0.0336±0.0001</td>
</tr>
<tr>
<td>Symmlet 10</td>
<td>0.0161±0.0001</td>
<td>0.0259±0.0001</td>
<td>0.0315±0.0001</td>
</tr>
</tbody>
</table>
compressed mass profile at level 3. The result indicates that there is no significant
difference on the PDRs between the compressed and original mass profiles. Table 5.5 is
the ANOVA table for the comparison of PDRs between wavelet compressed mass profile
at level 3 and histogram. The columns represent the wavelet compressed mass profiles
and the histogram. The difference of the PDR between the compressed mass profile and
the histogram are significant. The rows are the PDRs of each pair of xylene from
different data sets with three replicates. The table indicates that there is no significant
difference between the PDRs among different pairs of xylenes. Also the interaction
between treatments is insignificant.

Figure 5.8 gives principal component scores of level 3 compressed profiles at
level 3 and 1.00 m/z binned spectra. The reason to compare these two types of spectra is
that they have same data size. This means that after wavelet compress the mass profile at
level 3, it has the same data size as binned spectra with mass resolution 1.0 u, however,
was able to obtain much higher PDR. A histogram spectrum also has the same data size
as the binned spectra with mass resolution 1.0 m/z. Figure 5.9 gives principal component
scores of level 3 compressed profiles and histograms. Even if at a high compression
level, some of the detailed peak information in mass profile was lost, the peak shape was
still kept to provide higher PDR for isomers than the histogram spectra. The mass
profiles were compressed to 1/8th of their original size and furnished better PDR than the
other preprocessing methods.

In Figure 5.10, Figure 5.11, and Figure 5.12, for the mass spectral profiles the
isomers were completely resolved for three independent sets of the spectra. The 1.00 m/z
Table 5.5 ANOVA table for the comparison of PDRs between wavelet compressed mass profile at level 3 and histogram.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>5.9028</td>
<td>1</td>
<td>5.90276</td>
<td>20.65</td>
<td>0.0007</td>
</tr>
<tr>
<td>Replicates</td>
<td>0.1479</td>
<td>2</td>
<td>0.07395</td>
<td>0.26</td>
<td>0.7762</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.7155</td>
<td>2</td>
<td>0.35777</td>
<td>1.25</td>
<td>0.3208</td>
</tr>
<tr>
<td>Error</td>
<td>3.4302</td>
<td>12</td>
<td>0.28585</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10.1964</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 5.8** Comparison of PCA cluster results between spectra wavelet compressed from mass profile at level 3 and binned spectra at mass resolution 1.00 m/z.  m: m-xylene; p: p-xylene; o: o-xylene.
Figure 5.9  Comparison of PCA cluster analysis between spectra wavelet compressed from mass profile at level 3 and histogram spectra. m: m-xylene; p: p-xylene; o: o-xylene.
Figure 5.10  PCA cluster analysis of average histogram spectra of different runs for a
*m*-xylene spectrum from the *i*th TIC data set.  $p_i$ ($i = 1$ to $3$): a *p*-xylene spectrum from
the *i*th TIC data set; $o_i$ ($i=1$ to $3$): a *o*-xylene spectrum from the *i*th TIC data set.
Figure 5.11  PCA cluster analysis of average binned mass spectra at mass resolution 1.00 m/z of different runs. pi (i = 1 to 3): a p-xylene spectrum from the ith TIC data set; oi (i = 1 to 3): a o-xylene spectrum from the ith TIC data set.
Figure 5.12 PCA cluster analysis of average mass profiles from different runs. $p_i$ ($i = 1$ to 3): a $p$-xylene spectrum from the $i$th TIC data set; $o_i$ ($i = 1$ to 3): a $o$-xylene spectrum from the $i$th TIC data set.
binned scores were not resolved for the \textit{o}-xylene and \textit{p}-xylene, and for histogram mass spectra, none of the xylenes were resolved.

5.5 Conclusions

Preprocessing mass spectra may discard useful information regarding ion peak shapes or introduce errors. In many commercial packages, little is known regarding the integration algorithm. Mass spectral profiles retain information regarding ion peak shapes. This information may be exploited to differentiate spectra from isomers or other structurally similar compounds. Projected difference resolution (PDR) is a useful metric for measuring pair-wise clustering in multidimensional spaces. Mass profiles gave better PDR among the three types of xylenes than the histograms and the binned mass spectra at 1.00 \textit{m/z} mass resolution.

Wavelet compression was used to reduce the larger size of the mass profiles. The mass profiles after wavelet compression with same data size as the histograms still provides better PDRs among xylene isomers. In addition, wavelet compression improved the PDRs among the xylene isomers compared to the direct analysis of the uncompressed mass spectral profiles.

This approach is also important to other methods of mass measurements. Especially for high-resolution measurements and sampling methods for which baseline resolution between the mass peaks is unobtainable.
Chapter 6  Summary and Future Work

In Chapter 1, a brief overview of IMS, SIMPLISMA, ALS and wavelet compression was given, and the motivations of the research projects were stated.

In Chapter 2, an airborne UAV system that consisted of two identical IMS chemical detectors operating in parallel was used to collect the IMS data. Data were collected as the UAV flew through DMMP plumes. Continuous collection of data may quickly exceed the storage capacity of the instrument. In order to solve this problem, a nonlinear wavelet compression algorithm was applied to the IMS data. The data were compressed to 4-5% of their original size without artifacts that are typically created when linear compression is applied. Nonlinear wavelet compression avoids peak shifts, peak maxima attenuations, and ringing artifacts that are frequently encountered when using linear compression. Smaller values of $E_{RMS}$ are obtained from nonlinear compression than from linear compression. A grid search was used to select the optimal filter parameters using four representative UAV IMS data sets. The results from two sets were reported, but all four data sets yielded consistent optimal parameters.

In Chapter 3, a two-dimensional nonlinear wavelet compression method was developed and applied to IMS data. Two different types of data sets collected from CAM and ITEMISER were used. Three chemical weapon simulants (TEPO, DMMP, and DPM) were used for the data sets collected from CAM. DMMP and DPM were used for the data sets collected from ITEMISER. These data sets had different signal to noise ratios, different peak shapes, and different transient instrumental response. A compression factor of $1/128$ with relative error of root mean square ($RE_{RMS}$) of 0.25%
was achieved for ITEMISER data. With the same compression factor, $RE_{RMS}$ of 0.09% was achieved for CAM data, which had larger noise levels. The ALS models from reconstructed and unprocessed data sets gave indistinguishable spectra and concentration profiles. The wavelet filters for both drift time and acquisition time dimensions were optimized, although optimization may not be necessary to obtain good compression.

In Chapter 4, a 2D-NLWC-ALS algorithm was created by applying ALS to partially reconstructed wavelet coefficients generated from 2D NLWC. The number of components in a data set can be determined automatically before the procedure of the modeling compression. The smaller ALS models were saved as the final compressed data and can be used to reconstruct the entire data set efficiently without maintaining the compressed wavelet coefficient matrix of the original data set. The 2D-NLWC-ALS algorithm provided greater compression ratios compared to regular wavelet compression and had the added advantage of building an ALS model. Using this method, large volumes of data can be acquired and easily evaluated through a simple compressed model. The 2D-NLWC-ALS algorithm was applied to IMS data of chemical warfare agent simulants collected by CAM and ITEMISER. This algorithm proved to be an efficient compression method that provided the added benefit of a simple ALS model.

Chemical information manifested in peak shapes is traditionally discarded during the process of integrating mass spectra to histograms. This information can be used to differentiate spectra from isomers or compounds with similar molecular structures. In Chapter 5, wavelet compression was used to reduce the larger size of mass profiles, while retaining information regarding ion peak shapes. PCA was used to visualize data by
retaining variance of a reduced set of coordinates. The mass profiles, with the same data size as the histograms, still provided better PDRs among xylene isomers even after wavelet compression. This approach was also important to other methods of mass measurements, especially for high-resolution measurements and sampling methods for which baseline resolution between the mass peaks was unobtainable.

The future work for the UAV project, which is presented in Chapter 2, is to apply 2D-NLWC to the DMMP data sets. One dimensional NLWC can compress the data to 4 – 5 % of the original data size. Two-dimensional NLWC should provide a compression ratio of 0.4 – 0.5% of the original data size. An optimized number of spectra should be calculated. During the time of collecting the data, once the optimized number of spectra is reached, the collected data can be packed as a package and the 2D-NLWC can be applied to the whole data package.

For the 2D-NLWC-ALS algorithm project presented in Chapter 4, a real-time algorithm can be created so that the compression/modeling algorithm can be applied to raw data during the data collection procedure. Since the number of components can be calculated before the implementation of the 2D-NLWC-ALS algorithm, experiments can be performed to find the optimum number of spectra for calculation of the number of components.

For Chapter 5, the future work includes building up a complete mass spectrum pattern recognition library using wavelet compressed mass profile. PCA should be applied to the wavelet compressed mass profiles of compounds with similar chemical structures. Isomers should be differentiated using this pattern recognition library.
Bibliography


(31) Eiceman, G. A. Quiet service from a field measurement technology; Ion mobility spectrometry is used worldwide for on-site analyses and current advances suggest a role in field measurements in the future. Field Anal. Chem. Technol. 2000, 4, 217-218.


Appendix A: Publications


Appendix B: Presentations


### Appendix C: List of Common Chemical Warfare Simulants

<table>
<thead>
<tr>
<th>CAS Name</th>
<th>Chemical Structure</th>
<th>Formula/MW/Composition(^{97,98})</th>
</tr>
</thead>
</table>
| Methylphosphonic acid dimethyl ester\(^97\) (DMMP) | ![Chemical Structure](image) | Molecular Formula: \(\text{C}_3\text{H}_9\text{O}_3\text{P}\)  
Molecular Weight: 124.08.  
Percent Composition:  
C 29.04%, H 7.31%, O 38.68%, P 24.96% |
| 1 (or 2)-(2-Methoxymethylethoxy) propanol\(^98\) (DPM) | ![Chemical Structure](image) | Molecular Formula: \(\text{C}_7\text{H}_{16}\text{O}_3\)  
Molecular Weight: 148.20.  
Percent Composition:  
C 56.73%, H 10.88%, O 32.39% |
| Triethyl Phosphate\(^98\) (TEPO)                   | ![Chemical Structure](image) | Molecular Formula: \(\text{C}_6\text{H}_{15}\text{O}_4\)  
Molecular Weight: 182.16  
Percent Composition:  
C 39.56%, H 8.30%, O 35.13%, P 17.00% |