Int. J. Ion Mobil. Spec. (2008) 11:43–49
DOI 10.1007/s12127-008-0005-6

ORIGINAL RESEARCH

Preprocessing of ion mobility spectra by lognormal detailing and wavelet transform

Sabine Bader · Wolfgang Urfer · Jörg Ingo Baumbach

Received: 8 January 2008 / Revised: 18 April 2008 / Accepted: 19 April 2008 / Published online: 6 June 2008

Abstract This study has developed an efficient preprocessing strategy for ion mobility spectrometry (IMS) data allowing for improved peak clarity and comparability of different measurements. Using the discrete wavelet transform for data compression and denoising, and fitting a lognormal function to the strong tailing of the reactant ion peak (RIP), enables a data reduction to 25% or less, a significant increase of the signal-to-noise ratio, and the successful elimination of the RIP tailing. The preprocessing of breath measurements obtained by coupling an IMS to a gaschromatographic column, has resulted in the desired outcome of smooth peaks lying on a common base level. These results are transferable to other applications of one- and two-dimensional separations with IMS or instrumentations generating a similar data structure.

Keywords Ion mobility spectra · Detailing · Wavelet transformation · Data pre-processing · Tailing · Ion mobility spectrometry

Introduction

Ion mobility spectrometry (IMS) is a rapid, highly sensitive analytical method for the characterisation of gaseous samples with low detection limits [1, 2]. Characteristic peaks in the signal intensity of the resulting spectra data indicate specific analytes, however, these can be masked by high levels of noise. Additionally, the fast response is often counteracted by complex, time-consuming data analysis due to high dimensionality and redundancy of the data. This can especially be observed in two-dimensional separations e.g. IMS coupled to a gaschromatographic column (GC) [3]. Furthermore, the high variability in signal intensity and position of the peaks make data reduction to a predefined set of data points impossible [4].

Strong tailing after the high reactant ion peak (RIP) causes variable increases of peak height in different spectral areas. The amount of tailing can also vary between different measurements, thus complicating quantification and comparability of peak intensity within and between the measurements. In addition, the effect of the RIP tailing can disrupt peak detection and visualisation due to difficulties in distinguishing between noise and peaks by a single threshold. A strategy allowing reduction, simplification, and clarification of the IMS data is therefore required.

Recently, preprocessing via wavelet transform (WT) was shown to outperform the established compression methods of principal component analysis [5, 6] and Fourier transform [7, 8]. Moreover, its application was enhanced to two-dimensional compression [9] and real-time implementation [10]. Besides the linear WT, the nonlinear WT has also been used for spectra compression of one- and two-dimensional data [11–13], and furthermore for efficient data denoising [14].

Considering the beneficial effect of wavelet compression and denoising, this study has linked the application of both methods, thus combining the advantages of

S. Bader (✉) · J. I. Baumbach
ISAS - Institute for Analytical Sciences,
Bunsen-Kirchhoff-Str. 11, 44139 Dortmund, Germany
e-mail: Bader@isas.de

S. Bader · W. Urfer
Department of Statistics, University of Dortmund,
44221 Dortmund, Germany

reduced dimensionality and a higher signal-to-noise ratio. This, however, does not solve the problem of the RIP tailing affecting the height of other peaks, resulting in a need for clarification of IMS measurements. Accordingly, a lognormal function was adjusted to the spectra by optimisation via a newly developed penalty term and subtracted prior to the usage of the introduced wavelet procedure.

The proposed strategy of combining RIP detailing, wavelet smoothing and denoising, applicable for both one- and two-dimensional separations, has shown to be beneficial in breath measurements by GC-IMS. Data reduction to 25% or less with little information loss, a significant increase in the signal-to-noise ratio and better clarity of peaks can result in simplified further data processing with artificial neural networks, multivariate curve resolution or peak detection procedures.

Theory

Discrete WT The WT [15] simultaneously localises signal features in frequency and time, producing a decomposition allowing the reconstruction of the original signal without an error.

The application of discrete wavelet transform (DWT) with orthogonal wavelets results in a representation containing no redundancy, enabling the implementation of fast algorithms. The decomposition coefficients are calculated by determining the inner products of the signal with the functions of a wavelet base, which are formed by scaling and translating a mother wavelet function. The height and position of coefficients in the decomposition (Fig. 1b) of a spectrum (Fig. 1a), allow conclusions concerning the contribution of the different frequency scales to particular signal locations to be drawn.

For investigation of three-dimensional data structures, a separable two-dimensional WT, applying a one-dimensional DWT first for all rows and subsequently to all columns, can be useful. Filtering by low-pass and high-pass filters in each consecutive combination leads to a decomposition into four matrices. Further iterations in the next levels are applied only to the low-pass part, analogue to the one-dimensional case.

Using the WT for data smoothing, the first wavelet scales are eliminated before back-transforming to remove high-frequency components of the signal regardless of their amplitude.

The WT can furthermore be used for the denoising of signals by removing small-amplitude components via thresholding of the wavelet coefficients regardless of frequency before back-transforming.

Lognormal detailing Variations in the ion velocity, due to random ion-molecule reactions occurring in the drift tube, cause a strong tailing of the RIP and following varying heights for peaks in different parts of the drift time axis. A method for adjusting the strong tailing of the RIP was therefore developed by fitting and subtracting a function describing this instrumental behaviour.

A representative spectrum ensuring comparability between spectra after detailing was defined, allowing a reduced computational cost due to the performance of only a singular fitting step. The median spectrum $\mathbf{s}^{med}$ was found to meet the requirements of characteristic feature conservation for the whole spectra series, the low variance around zero in noise areas, and high robustness very well.

The detailing function itself, fitted to the tailing of the median spectrum, was chosen as a modified lognormal function of the form

$$L(\mathbf{x}) = \frac{a}{(\mathbf{x}-\theta)\,\sigma\,\sqrt{2\pi}} \cdot \exp\left[-\frac{\left[\log\left(\frac{\mathbf{x}-\theta}{m}\right)\right]^2}{2\sigma^2}\right],$$

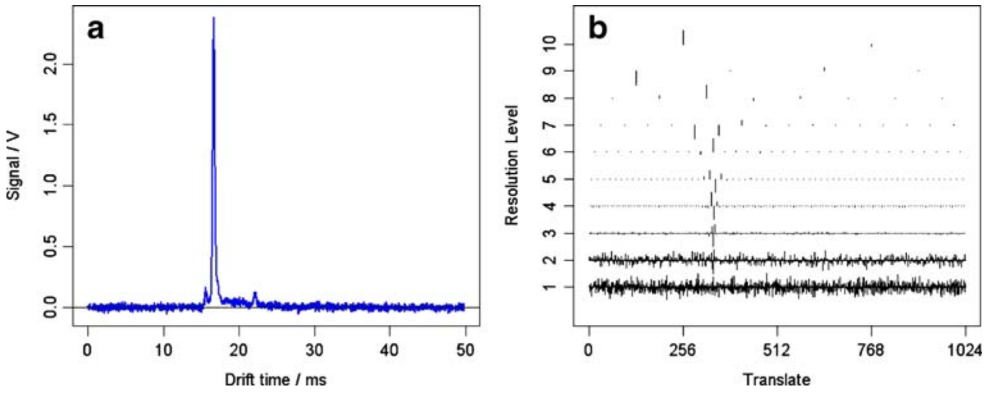

Fig. 1 Plot of **a** a raw ion mobility spectrum and **b** the corresponding wavelet decomposition

where **x** denotes the vector of drift times. The lognormal function met well the general assumption of Gaussian peaks and the right-screwed shape of the RIP tailing, and is often used to describe physical processes that are limited in one direction. The variables θ, σ, and m are parameters of location, shape, and scale, respectively, and the factor a allows for shrinkage to the actual intensity magnitude.

As a reasonable side condition it was claimed that the maximum positions $x_{\max}^{med}$ of the representative spectrum and $x_{\max}^{log}$ of the detailing function $L(\mathbf{x})$ should coincide. Since the maximum position of a lognormal function is known to be $x_{\max}^{log} = \theta + \frac{m}{\exp(\sigma^2)}$, the position parameter θ could be determined by

$$x_{\max} \overset{!}{=} \theta + \frac{m}{\exp(\sigma^2)} \quad \Leftrightarrow \quad \theta \overset{!}{=} x_{\max} - \frac{m}{\exp(\sigma^2)}.$$

Accordingly, only the three parameters of shape σ, scale m, and shrinkage a had to be optimised, achieved by the minimisation of the herefore developed penalty term P,

$$P = \sum_{i=1}^{n_D} \left[I_{[\Delta_i < -r_p]} p_{abs} + I_{[\Delta_i > r_p]} \min(\Delta_i, b_p) \right],$$

with $\Delta_i = s_i^{med} - L(x_i)$ constituting the difference between the ith value of the median spectrum s^{med} and the lognormal function $L(x)$, n_D giving the number of data points per spectrum, and $I_{[\,]}$ defining the indicator function, which takes the value 1 if the condition in its subscript is fulfilled, or 0 if this is not the case. Whilst the scalar r_p, dependent on the standard deviation in noise areas, allows for little variation of the adjusted function around the representative spectrum, the constant p_{abs} assigns an absolute penalty for parts of the detailing function lying over the spectrum **s**, to yield the detailing function nestling to the data from below, irrespective of occurring peaks. Downwards deviation was penalised by the actual deviation, constrained by

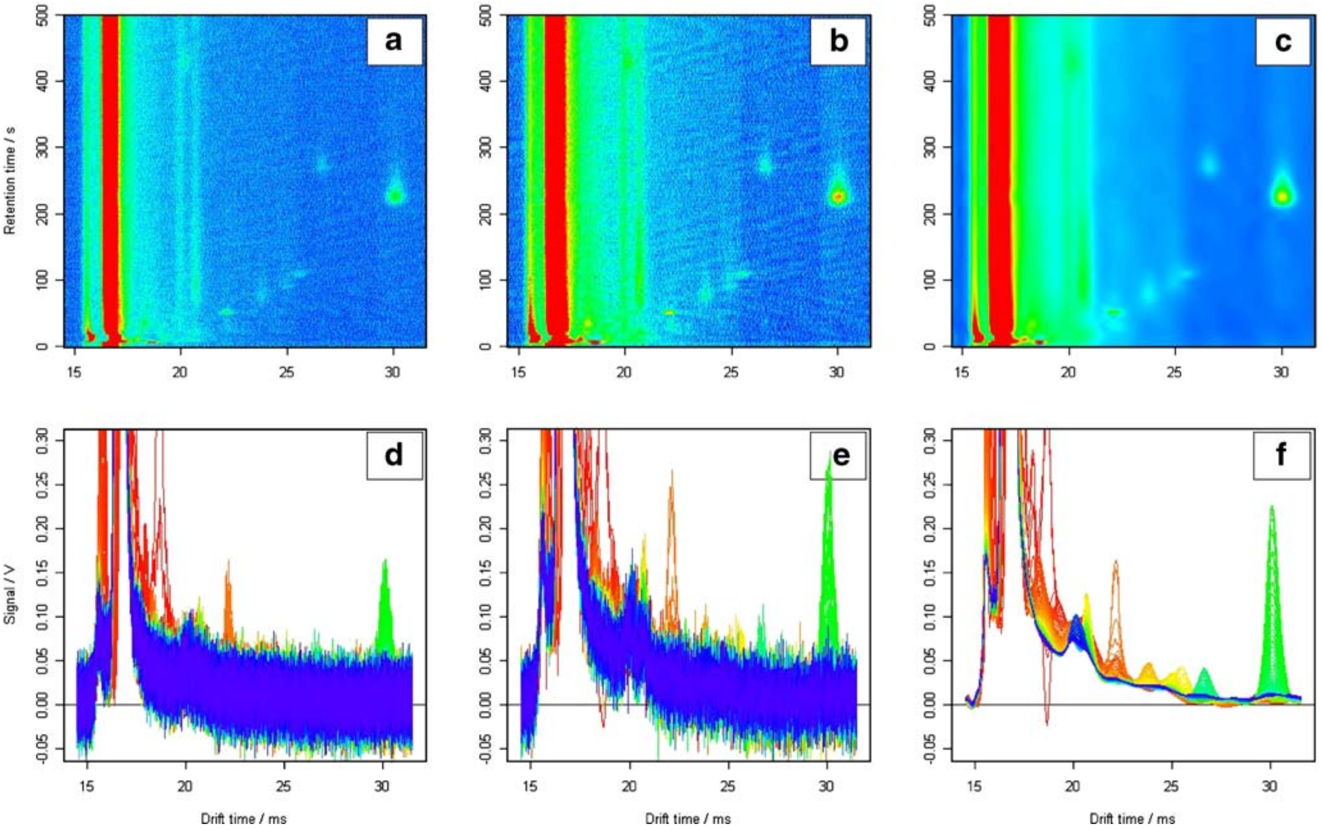

Fig. 2 Heatmaps of **a** raw, **b** smoothed, and **c** denoised data show the achieved information conservation concurrently with the reduction to a quarter of data points via smoothing and the reduction of noise by denoising, while the spectra series in a sideview of **d** raw, **e** smoothed, and **f** denoised data illustrate an improved signal-to-noise ratio, resulting from increasing peak heights after smoothing and the reduced amount of noise achieved through denoising, but concurrently also a not desired amplification of the RIP tailing

Table 1 Quantification of height and signal-to-noise ratio (SNR) after different processing steps for three exemplary peaks with the drift, retention time position pairs of (20.175 ms, 429 s) for peak A, (22.15 ms, 49 s) for peak B, and (30.05 ms, 225 s) for peak C

	Height	SNR
Peak A		
Axes transformed	0.062	59
& Detailed	0.040	38
& Smoothed	0.128	82
& Denoised	0.059	193
Peak B		
Axes transformed	0.069	65
& Detailed	0.057	54
& Smoothed	0.245	157
& Denoised	0.142	665
Peak C		
Axes transformed	0.114	108
& Detailed	0.113	107
& Smoothed	0.246	157
& Denoised	0.224	733

the threshold b_p giving a cut-off point for diminishing the influence of the RIP height in so much as deviations greater than this value were punished only by b_p.

For minimisation of this penalty term, yielding the optimal parameter set, a limited-memory modification of a quasi-Newton method was used, allowing the choice of box constraints for each variable [16].

Experimental

The data regarded in this work were obtained by an ISAS custom-designed IMS with radioactive nickel (^{63}Ni) as the ionisation source and a drift tube of length 12 cm, connected to a multi-capillary column with length 20 cm [17]. The measurements, recorded in the positive mode with a field strength of 236 V/cm, consisted of 501 spectra, each containing 2000 data points distributed equidistantly over a drift time of 50 ms. A high RIP emerges in all spectra due to air being used as the carrier gas for human breath measurements (Fig. 1a).

Before data were subjected to the presented preprocessing procedure, a baseline correction was used ensuring intensity varies around zero in areas of noise. Here the median intensity value in a measurement part of pure noise was subtracted from the entire data matrix.

The two dimensions of drift time and signal intensity resulting from IMS, can be regarded in ion mobility spectra (Fig. 1a). Considering also the additional dimension of retention time, when analytes pass from

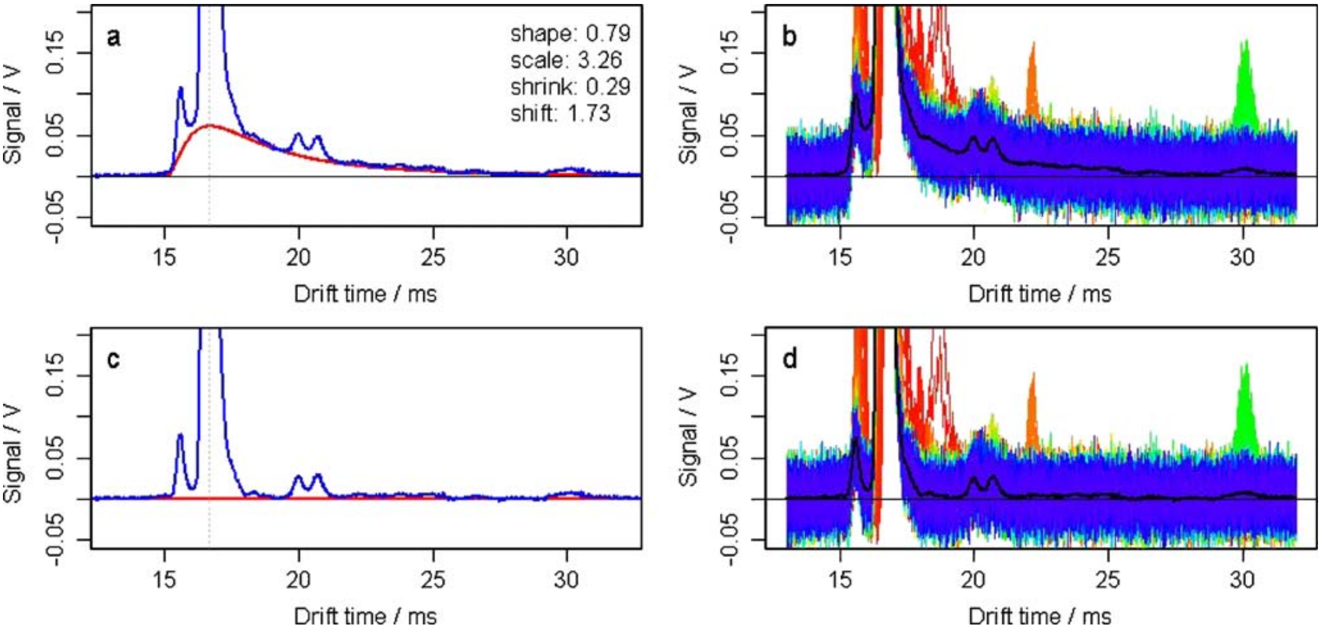

Fig. 3 Effect of the RIP detailing for **a** the characteristic median spectrum with lognormal function and optimised parameters before and **c** after subtracting of the detailing function, and **b** spectra series with median spectrum before and **d** after subtracting of the detailing function: After RIP detailing the peaks grow from a common base level for the characteristic median spectrum as well as for the whole spectra series

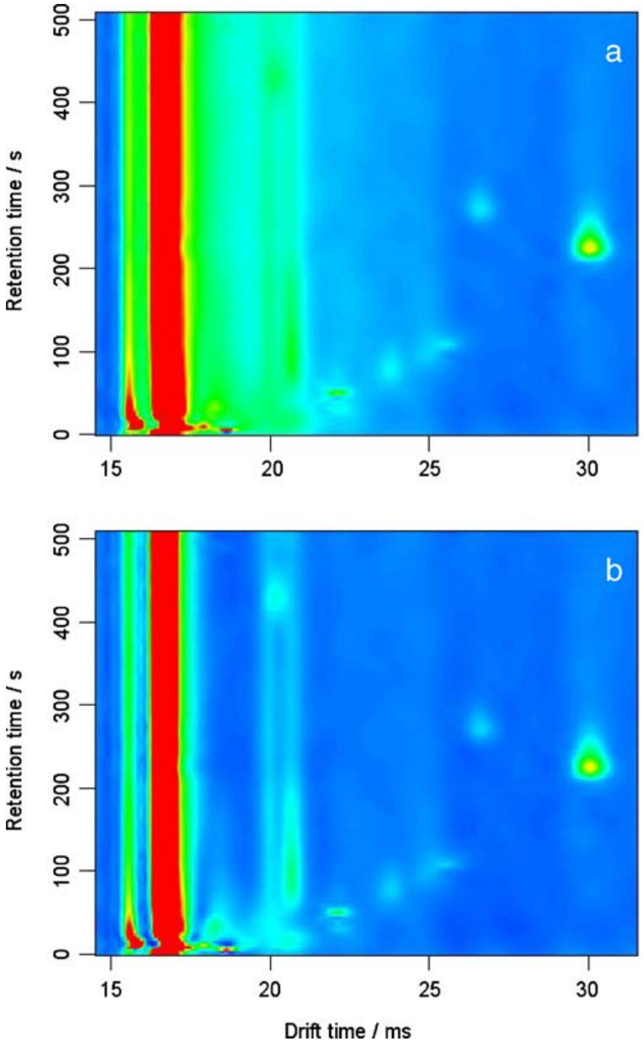

Fig. 4 Heatmaps of smoothed and denoised data **a** without and **b** with detailing: peaks in the spectra part after the RIP appeared more clear when the RIP detailing was applied, while the beneficial effects of the wavelet operations were retained

the column into the IMS, the measurement can be illustrated in topological heatmaps, encoding the signal intensity by a colour scheme (Fig. 2a).

Results and discussion

Smoothing and denoising The WTs in this study were executed with the programme package R [18] using the Daubechies extremal phase family function with length $N = 8$.

Applying one-dimensional wavelet smoothing with one compression level, first for all single spectra of a measurement, then for each time point across the spectra, a reduction to one quarter of the data points was achieved. The resulting data still contain the relevant information, but are more grainy and peaks are covered up to some degree by the amplified RIP tailing (Fig. 2a, b). At the same time, peak heights have increased, leading to an improved signal-to-noise ratio due to unchanged noise variation (Table 1, Fig. 2d, e).

Beyond, the signal-to-noise ratio could be further improved, achieving smooth and non-spiky graphs (Fig. 2c and f), by denoising via soft thresholding [14] of a two-dimensional WT using Donoho's universal threshold [19]. Peaks however tend to be relatively broad, thus hard thresholding, leading to sharper peak shapes, can also be indicated depending on the application. Unfortunately, a severe amplification of the RIP tailing appears concurrently with the beneficial effect of the increased signal-to-noise ratio, as it grows just like the peaks apparent in the data (Fig. 2c, f).

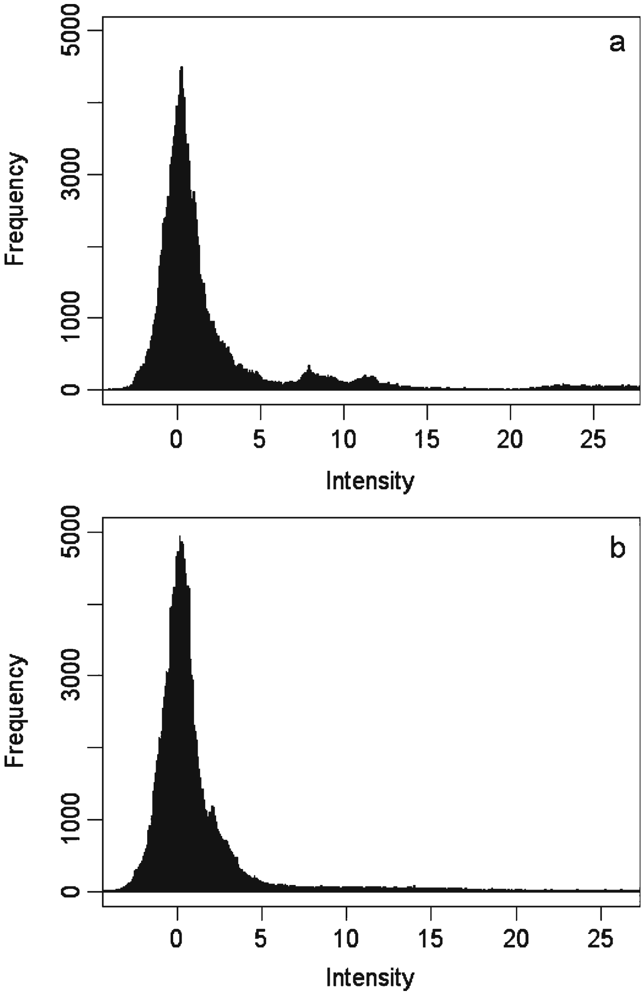

Fig. 5 Histogram of intensity values of the entire data matrix after wavelet smoothing and denoising **a** without and **b** with detailing: whilst without detailing the data contained several hills, indicating different intensity categories according to peaks lying on different levels of the RIP tailing, only a single hill was left with denoising, as all peaks were set to the same level

Adjustment of RIP tailing Using the developed penalty term for parameter optimisation, a lognormal detailing function was fitted to the median spectrum of the regarded measurement to describe and set down the RIP tailing by subtraction of the resulting curve. Fitting the tailing accurately, the adjusted lognormal function with the parameters given in Fig. 3a can be subtracted from the median spectrum smoothly (Fig. 3a, c). This beneficial effect is also transferred to the entirety of the spectra of a measurement (Fig. 3b, d).

Combined preprocessing Applying the RIP detailing prior to the introduced combination of smoothing and denoising by WT, peaks become more clear in the spectra parts after the RIP, while the advantageous effects of the wavelet operations are retained (Fig. 4). The varying impact of detailing on peaks in the different spectra parts can be quantified by consideration of their peak heights before and after this transformation, showing only little influence on peaks in the latter spectra parts (Table 1).

After RIP detailing all peaks share a common base level, which can be pointed out considering histograms of all intensities values in the entire data matrix. The histogram for smoothed and denoised raw data contains several hills, possibly indicating different intensity categories according to peaks lying on different levels of the RIP tailing (Fig. 5a). In the histogram of the intensity values after applying the detailing function and the combined wavelet procedure, only one hill is left aside from the big noise part around zero, as all peaks are on the same level now (cp. Fig. 5b).

Conclusions

The objective of this work was to develop an efficient preprocessing strategy for IMS data allowing for better processability and comparability of IMS measurements. This was necessary as characteristic IMS peaks can be masked by high levels of noise, variability in height, and strong tailing after the RIP.

To solve these problems, a linked application of smoothing and denoising by means of wavelets was combined with a new method for adjustment of the RIP tailing. Accordingly, a lognormal function was adjusted to the spectra by optimisation via a newly developed penalty term and subtracted prior to the introduced wavelet procedure.

The proposed strategy is applicable for both one- and two-dimensional separations, and was applied to breath measurements by GC-IMS. The wavelet procedure yielded a data reduction to 25% and a strongly improved signal-to-noise ratio, but also an increase of the RIP tailing. This effect was compensated by fitting a lognormal detailing function to the representative median spectrum and subtracting it from each spectra of a measurement. Applying the RIP detailing prior to smoothing and denoising by DWT, peaks were shown to share a common base level, which yielded a better comparability of peak heights, while the advantageous effects of the wavelet operations were retained.

Recapitulatory, combining the developed detailing function with the usage of WTs for smoothing and denoising, a powerful preprocessing could be reached, transferable to other applications of one- and two-dimensional separations with IMS or instrumentations generating a similar data structure.

Acknowledgements The financial support of the Bundesministerium für Bildung und Forschung and the Ministerium für Wissenschaft und Forschung des Landes Nordrhein-Westfalen is gratefully acknowledged. Furthermore, the valuable support of the lung hospital Hemer and the helpful discussions with Dr. Wolfgang Vautz and Dr. Darius Wilson are mentioned thankfully.

References

1. Baumbach JI, Eiceman GA (1999) Ion mobility spectrometry: arriving on site and moving beyond a low profile. Appl Spectrosc 53:338A–355A
2. Eiceman GA, Karpas Z (2005) Ion mobility spectrometry. CRC, Cleveland
3. Bader S, Urfer W, Baumbach JI (2005) Processing ion mobility spectrometry data to characterize group differences in a multiple clas comparison. Int J Ion Mobil Spectrom 8:1–4
4. Bader S, Urfer W, Baumbach JI (2006) Reduction of ion mobility spectrometry data by clustering characteristic peak structures. J Chemom 20:128–135
5. Davies AN, Baumbach JI (1999) Multidimensional data analysis—quantifying the hidden dimension. Spectrosc Eur 11(5):23–24
6. Mehay AW, Cai C, Harrington PB (2002) Regularized linear discriminant analysis of wavelet compressed ion mobility spectra. Appl Spectrosc 56:223–231
7. Harrington PB, Lijuan Hu (1998) Recovery of variable loadings and eigenvalues directly from fourier compressed ion mobility spectra. Appl Spectrosc 52:1328–1338
8. Harrington PB, Rauch PJ, Cai C (2001) Multivariate curve resolution of wavelet and fourier compressed spectra. Anal Chem 73(14):3247–3256
9. Urbas AA, Harrington PB (2001) Two-dimensional wavelet compression of ion mobility spectra. Anal Chim Acta 446:391–410
10. Chen G, Harrington PB (2003) Real-time two-dimensional wavelet compression and its application to real-time modeling of ion mobility data. Anal Chim Acta 490(1):59–69
11. Cai C, Harrington PB (1999) Wavelet transform preprocessing for temperature constrained cascade correlation neural networks. J Chem Inf Comput Sci 39(5):874–880

12. Cao L, Harrington PB, Harden CS, McHugh VM, Thomas MA (2004) Nonlinear wavelet compression of ion mobility spectra from ion mobility spectrometers mounted in an unmanned aerial vehicle. Anal Chem 76(4):1069–1077
13. Cao L, Harrington PB, Liu C (2004) Two-dimensional nonlinear wavelet compression of ion mobility spectra of chemical warfare agent simulants. Anal Chem 76(10): 2859–2868
14. Cai C, Harrington PB (1998) Different discrete wavelet transforms applied to denoising analytical data. J Chem Inf Comput Sci 38(6):1161–1170
15. Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
16. Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput 16(5):1190–1208
17. Ruzsanyi V, Baumbach JI, Sielemann S, Litterst P, Westhoff M, Freitag L (2005) Detection of human metabolites using multi-capillary columns coupled to ion mobility spectrometers. J Chromatogr A 1084:145–151
18. R Development Core Team (2007) R: a language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna
19. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90: 362–366