Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1071 (2005) 239–246
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

A new mathematical procedure to evaluate peaks in complex chromatograms

B. Steffen[a,*], K.P. Müller[b], M. Komenda[b], R. Koppmann[b], A. Schaub[b]

[a] *Zentralinstitut für angewandte Mathematik, Forschungszentrum Jülich, 52425 Jülich, Germany*
[b] *Institut für Chemie und Dynamik der Geosphäre, Institut II; Troposphäre Forschungszentrum Jülich, 52425 Jülich, Germany*

Available online 15 December 2004

Abstract

Automatic peak evaluation in chromatograms and subsequent quantification of compound concentrations is still a challenge in the analysis of complex samples containing hundreds or thousands of compounds. Although a number of software packages for peak evaluation exist, baseline definition and overlapping peaks of different shapes are the main reasons which prevent reliable automatic analysis of complex chromatograms. A new mathematical procedure is presented which uses peak shapes extracted from the chromatogram itself and modified by nonlinear (in fact, hyperbolic) stretching of the peak head and tail. With this approach, the peak parameters are position, height, scale of front, scale of tail, and smoothness of transition from front to tail scaling. This approach is found to give a substantially better fit than traditional analytically defined peak shapes. Together with a good peak finding heuristic and nonlinear optimization of parameters this allows a reliable automatic analysis of chromatograms with a large number of peaks, even with large groups of overlapping peaks. The analysis matches the quality of standard interactive methods, but still permits interactive refinement. This approach has been implemented and tested on a large set of data from chromatography of hydrocarbons in ambient air samples.

Keywords: Chromatography; Automatic peak analysis; Deconvolution; Peak shape heuristics; Hyperbolic scaling; Overlapping peaks

1. Introduction

The analysis of complex mixtures of compounds both in process chemistry and environmental chemistry is routinely done by chromatographic techniques. The major goal of chromatography is to separate the compounds of a sample taking advantage of compound specific parameters such as boiling point, molecular structure, mass, charge, and diffusivity. Following the separation the next important step is an appropriate detection of the compounds applying specific detectors. Separation and subsequent detection delivers a chromatogram, which in the ideal case allows to identify individual peaks and to attribute them to individual compounds. The evaluation of peaks is the crucial point of the processing of chromatograms. Typical analytical purposes in industrial processes usually deal with a manageable amount of well known peaks of compounds at relatively high concentrations, which can easily be done automatically. A variety of software packages is offered which are suitable for this type of routine analysis of chromatograms.

However, the situation is completely different in environmental analysis. Depending on the sampling site, samples of ambient air may contain hundreds or even thousands of compounds, sometimes at moderate or low concentrations. Furthermore, the matrix contains variable amounts of major components such as carbon dioxide or water vapour, which may interfere with the compounds to be analyzed. Chromatographic separation of such complex mixtures leads to a large number of frequently overlapping peaks, which can no longer be automatically analyzed with sufficient reliability and accuracy using existing software.

Generic peak detection algorithms are sensitive to baseline variations and signal intensity. This is especially a problem in environmental analysis which has to deal with overlapping peaks of different shapes, baseline drifts due to complex temperature programs, baseline steps, or spikes induced by the switching of valves. Furthermore, the peaks of interest

* Corresponding author. Tel.: +49 2461 61 6431; fax: +49 2461 61 6656.
E-mail address: b.steffen@fz-juelich.de (B. Steffen).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2004.11.073

often have low signal to noise ratios which make an automatic analysis of complex chromatograms almost impossible. Therefore, a large number of chromatograms in environmental research have to be analysed at least partly manually, which is extremely time consumable and probably the cause of additional errors.

During the last three decades various approaches have been published towards a more or less automatic peak detection using different criteria for the identification of peaks [1–4]. These algorithms generally seek instants of rapid increase or decrease in signal intensity or intensities above a critical threshold which are used to identify peaks in a chromatogram. Other peak detection software makes use of different metrics such as nonlinear filters [5], intensity weighted variances [6], or fast Fourier transforms [7].

Today, cheap high performance desktop computers provide the advantage that elaborate mathematical procedures are within the reach of everyone. In this paper we present a new approach to resolve peak overlaps, which uses peak shapes extracted from the chromatogram to be analyzed and an adaptive method for determining the baseline in order to evaluate peak areas and thus a quantification of the individual compounds. This procedure allows automatic analysis that matches the quality of standard interactive analysis, while still permitting interactive refinement. The approach described here does not need any information except the chromatogram and possibly a calibration chromatogram for the same type of compounds analysed on the same instrument, especially no knowledge on probable peak positions and sizes.

2. Peak shape

2.1. Peak shapes from first principles

The peak shapes are determined by dynamical processes and to some extent by nonlinear distribution functions which are theoretically well known. The theories are often simplified in that they assume infinitely small phase transitions, infinitely small plate heights, and instantaneous distribution of the compounds between mobile and stationary phase. These assumptions do, of course, not match reality. More serious problems are the assumptions of constant and homogeneous conditions along the column and of non-interference of compounds. Therefore, a modelling of exact peak shapes would be demanding and laborious. It is much easier to determine the peak shape experimentally and derive all necessary parameters to describe the peak. Peak shapes result from a convolution of the characteristics of the separation process, of the detector and of the transport processes involved. Ideally, the characteristics of the separation process and the detector are known. In a well designed chromatographic system the impact of other processes (injection, transport through valves, etc.) on the peak shapes are negligible. This would allow an accurate modelling of the peaks by a low-dimensional space of analytic functions. In practice, there is a large deviation from the ideal state. Chromatographic separation is a continuous process which is theoretically divided into a multistep dynamic equilibration between stationary and mobile phase. A molecule then has a probability $\exp(-\lambda_m \Delta x)$ to travel a distance Δx in free flow before being absorbed in the stationary phase. It then stays in stationary phase for some time, the probability of staying for a time greater than Δt being $\exp(-\mu_m \Delta t)$. The average travelling distance per step, λ_m, and the average delay time μ_m depend on the compound and on the material of the stationary phase, and are difficult to measure independently. By choice of materials, most molecules of the mixture will experience many thousand steps of absorption and desorption, each a few milliseconds on average. The total delay for a molecule then has a distribution that is almost a Gaussian with an average retention time given by the average delay time μ_m multiplied by the average number of steps L/λ_m. The peak width is proportional to the square root of the number of steps multiplied by the average delay time. Felinger [8] discusses a number of theoretical approaches to modelling a column as a discrete system with a moderate number of plates which result in distributions that are almost indistinguishable from a Gaussian in practice.

Following the separation process the effluents are analyzed by an appropriate detector. If the detection process is modelled as a single step delay line with an average delay time of ν_m, this delay has a distribution more like a Poisson (single step) distribution

$$\frac{1}{\nu_m} \times \exp(-\nu_m t) \tag{1}$$

As the proper peak shape of a non-interacting sequence of devices is the convolution of the initial distribution with the pulse response of the following devices, the resulting distribution is an exponentially modified Gaussian (EMG), and this is the peak shape most available data analysis programs favour. The separation of peaks and attribution of the proper peak area then is a linear deconvolution process that would have its limits only in the noise of the signal. Unfortunately, actual peaks cannot, in general, be accurately modelled by an EMG function; obviously, the generating processes (inhomogeneous transport conditions and substance interactions) are more complicated than the (simplified) theory assumes.

2.2. Peak shape heuristics

Visual inspection of a chromatogram shows that similar (with respect to transport and detection mechanisms) substances give similar peak shapes. This is confirmed by the fact that the approximation error from fitting a standard analytical model to the peaks also gives similar shapes. There seems to be only a very small number of basic shapes—often only one—every peak being similar to one of those. Obviously shifting of peak position and scaling of height are allowed, but additionally a nonlinear distortion in width and some asymmetry is needed. This is easy to see with a change

in peak description: a peak shape may as well be given by the time Δt it takes for the signal to drop to the relative height h_{rel}, separately for front and tail. The simplest scaling of Δt(hrel) giving sufficient flexibility and smoothness is a linear scaling of the front, a different linear scaling of the tail, and a smooth transition in between. Additionally, the scaling should be invertible, if peak A can be scaled to fit peak B, than an inverse scaling should be possible to scale peak B to fit peak A. This can be achieved with the Eq. (2)

$$g(t, a) = a_1 \times g_o(\phi(t - a_2, a_3, a_4, a_5)) \quad (2)$$

where $g_o(t)$ is the peak shape typical for the column, compound type and operating conditions, and the scaling is given by Eq. (3)

$$\phi(t, a_3, a_4, a_5) = a_3 \times (t - a_4 \times \mathrm{sqrt}(t^2 + (a_5)^2) + a_4 \times a_5). \quad (3)$$

This is a hyperbola passing through the origin, where a_3, a_4 give the asymptotic slope at times far after and far before the peak maximum (the scaling of front and tail) while a_5 gives the width of the transition region. a_3, a_4, a_5 are bounded such as to give a monotonous function and to restrict tailing to about four times the amount of the standard peak tailing, fronting to somewhat less. With proper standard peak $g_o(t)$, this model results in a very good fit, better than the EMG model in most tests—in the example of Fig. 1 e.g. the best EMG fit does not show strong enough tailing—and with a fit error close to the noise level. The basic peak shape $g_o(t)$ may be obtained by averaging a number of suitably scaled peaks [9]. More descriptive parameters like peak width, skewness etc. can be easily calculated from the basic shape and the fit parameters.

What peaks this model can describe depends very much on the standard peak $g_o(t)$, it allows arbitrary changes in height and position, but changes in width, fronting, tailing (and thus excess and skewness) are restricted by heuristics.

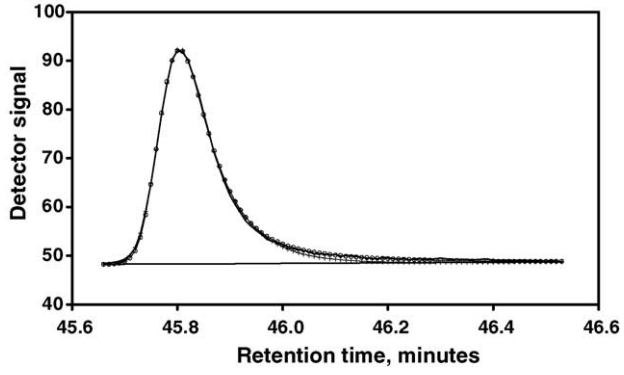

Fig. 1. Example of an individual peak (taken from the chromatogram shown in Fig. 2. The solid line is the original chromatogram, the open circles represent the fit with the new procedure the crosses represent the best EMG fit to this peak.

2.3. Automatic peak search and analysis

In interactive analysis, the peak shape is found by visual inspection. However, an automatic construction needs a sequence of iterative refinement steps to reach a result of similar quality. Each step consists of an approximate determination of a baseline, a peak shape and a fit to peaks larger than a certain threshold. The goal of the first step is to extract the typical (standard) peak shapes. As these shapes depend on equipment and operating conditions, the result may be stored and there is no need to repeat it for every single measurement. The first part is finding all large peaks. Any point where the second derivative of the signal has a local minimum that is larger than a tenth of the global minimum is a first-rate candidate for the top of a large peak. An estimate for the width of a peak is given by the distance between the adjacent strongly positive local maxima of the second derivative. The baseline is estimated by a straight line through the minima of the signal in intervals to the right and left of the suspected peak position with a length of about 10 times the peak width. This allows an estimate of peak height, every peak less then 10% of the tallest peak will be disregarded at the moment as being too small, and therefore too much influenced by noise, to enter the standard. These peaks are tested for usability. They are approximated with a standard analytical peak model for the process (EMG for chromatography). Those peaks not allowing at least a low quality approximation (least square error as indicator) are excluded—they are most likely not isolated. The peaks kept are tested for being well separated from other peaks—anything that sticks out from the neighbourhood by more than a few times noise level. What is left makes up the set from which the standard shape is defined. Another, possibly better, way to get the standard is to extract it from the calibration measurements of the instrument, if that is available. It may also be transferred from other chromatograms generated with the same instrument and operating conditions, but different load.

Fig. 2 shows an example chromatogram of volatile organic compounds in a typical air sample. The organic compounds were preconcentrated from 500 mL of air on a Silco Steel™

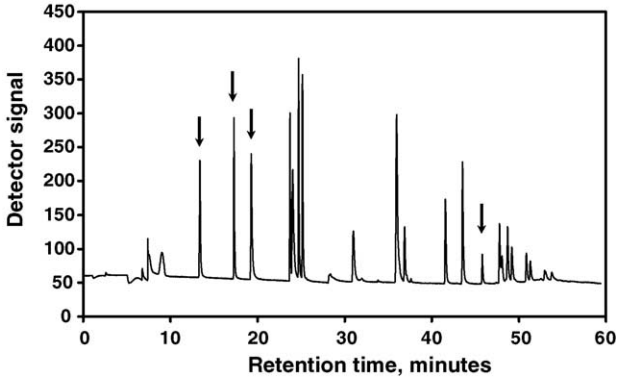

Fig. 2. Example chromatogram of an air sample (for details see text). Peaks used for construction of standard peak shape are marked by an arrow.

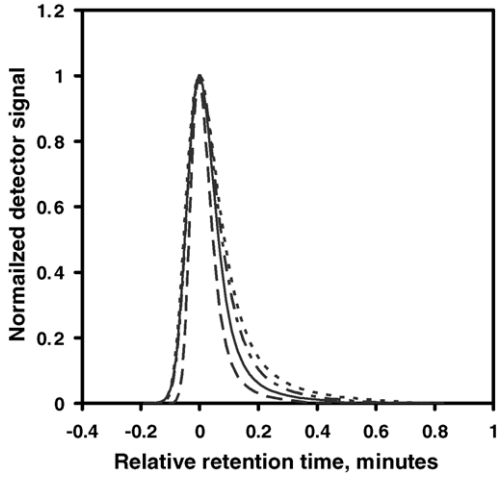

Fig. 3. Selected peaks (see Fig. 2) normalised and superimposed.

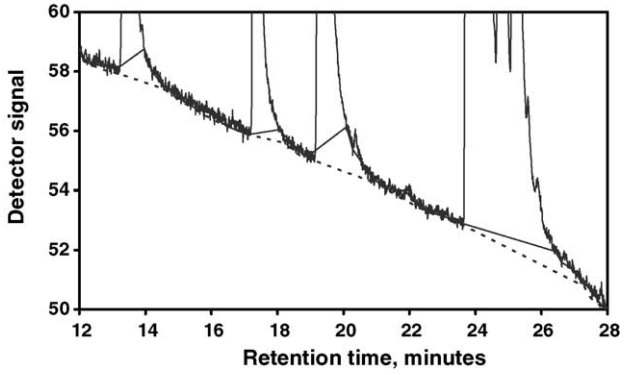

Fig. 5. Base lines created by the methods described in the text. The baseline determined by method 1 is shown as the solid line; the baseline determined by method 2 is shown as the dashed line. The section shown here is a part of the chromatogram shown in Fig. 2.

column packed with porous glass beads (length = 150 mm, i.d. = 2 mm. Volume = 0.47 mL) at −196 °C. Following the cryogenic preconcentration the sample was thermally desorbed and separated on a PLOT column (ASTEC Gas-Pro GSC, length = 60 m. i.d. = 0.32 mm). The initial temperature of the GC was held at 2 °C for 6.5 min and then ramped to 230 °C at a rate of 5 °C min^{-1}. Helium was used as a carrier gas at a flow rate of 4.2 mL min^{-1}. From this example chromatogram the peaks at retention times of 13.4, 17.3, 19.3 and 45.8 min are selected. The peak at 43.4 min is not really a single one. Fig. 3 shows these selected peaks superimposed and normalised. They look very similar, but differ in width. Averaging the width at height 0.5 to the front and back, and scaling each peak to this average, results in the peaks shown in Fig. 4, where the differences are minute except for the height of the tail. This is likely to result from errors in the basis. The standard shape is now constructed by averaging the scaled peaks.

After this procedure, it is possible to define an improved baseline for the entire chromatogram. One method is to construct the baseline from piecewise linear functions. In a sequence of overlapping intervals the highest straight line strictly below the signal is found. The baseline is then made up of the highest continuous connection of line segments, increased by one noise level. Finally, the edges situated below a peak or peak group are cut off by a straight line through the values of the baseline before and after the peak. This construction works well except for those parts of the chromatogram where operating conditions change rapidly, e.g. temperature is increased. If these points are given, they can be made an edge or even a discontinuity of the baseline, but at present the information is not included in the data. In general, this construction gives a slightly low estimate of the baseline, resulting in an overestimation of the peaks, especially the small ones (Fig. 5).

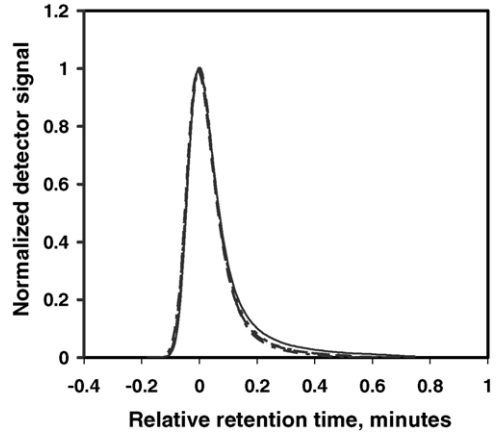

Fig. 4. Selected peaks (see Fig. 2) fully scaled for optimal similarity.

Another approach to baseline construction uses the fact that the signal shape is dominated by noise only in the small areas on top of the peaks or in the minimum between two overlapping peaks, or in possibly larger areas where no peak gives a substantial contribution. Therefore, any area that is dominated by noise for more than the average peak width is considered true baseline. These areas may be connected by straight lines or by spline functions with widely spaced knots that approximate the signal from below. While the latter has the advantage of smoothness and higher accuracy in most places, the danger of overshoot errors in difficult areas is larger, so this may not be any better than the simpler approach (Fig. 4). This baseline estimate tends to be a little too high.

After this, the final peak fitting proceeds for the reduced signal (original signal minus baseline) as usual. If there are different basic shapes, suspected peaks have to be classified. Classification may be possible using width and asymmetry, but if this is not sufficient, it may be necessary to try different shapes and choose the one giving the best fit. As the scaling function $\phi(t, a_3, a_4, a_5)$ is nonlinear and the possible values of a_3, a_4, a_5, as are bounded, a Newton method for bounded regions serves well. Peaks are searched for starting at $t = 0$ by analyzing the maxima of the reduced signal and the minima

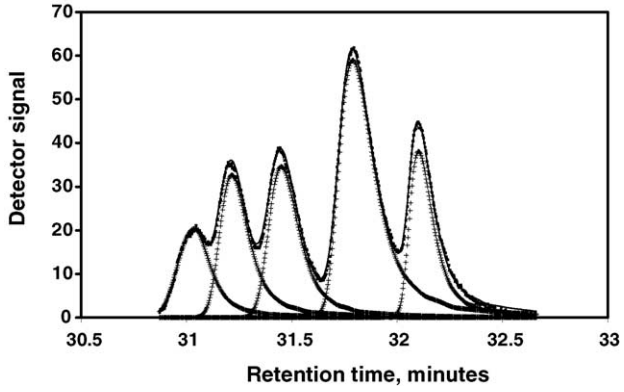

Fig. 6. Group of five peaks with measured chromatogram (black squares), the fitted individual peaks (crosses), and the sum of the fitted peaks (solid line).

of its second derivative. If a peak (or an overlapping group of peaks) is found, it is subtracted from the reduced signal and the procedure iterated. Acceptance of a peak is based on height, width and least square error of the fit. The wider it is, the higher it has to be for not to be regarded as baseline drift. Small peaks, even if sharp, that can not be fitted are considered noise. Setting the threshold is a problem, too low a threshold generates a large number of nonsense peaks, too high a threshold might miss an actual one. A list of known approximate peak positions would help, and may frequently be available. For every single peak or group of overlapping peaks found, the parameters of the peaks are optimized using an iterative nonlinear weighted least squares procedure. For examples of such a fit see Figs. 6 and 7. In all cases, the plot of the sum of the peaks differs from the smoothed measured curve by less than the thickness of the line, so it is not plotted in the two figures. This gives some confidence in the results, although it is still possible to mistake two peaks close to each other for a single one or a strongly distorted single peak for a collection of overlapping peaks. These errors, however, are just as likely with manual analysis.

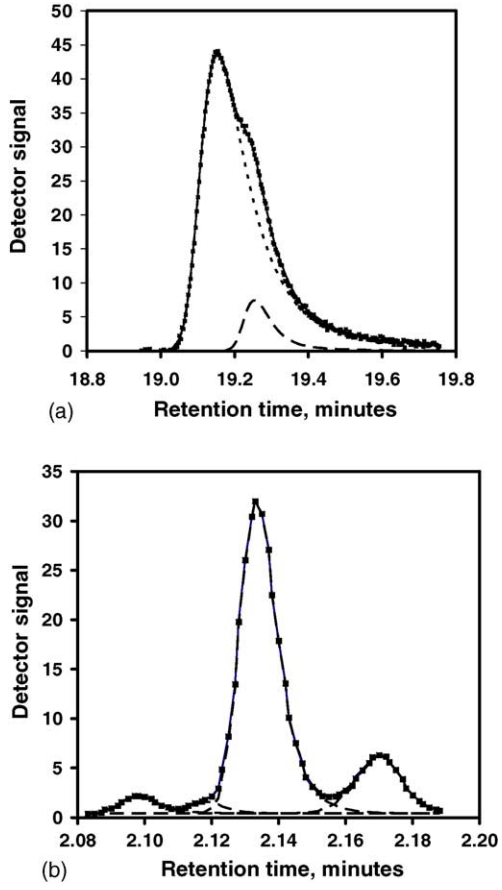

Fig. 7. Example of the fit results of a small peak on the shoulder (a) or on the front of a large peak (b). The squares and the solid line represent the measured peaks; the fitted peaks are plotted as dashed lines.

3. Results

The ideas described above have been implemented in a Scilab™ code [10]. Scilab was chosen because it is a portable (versions for all Unix derivates and Windows flavours are available), easy to install public domain system that offers all necessary mathematical functions, a simple but sufficient graphics system and a graphical user interface with dialog boxes. The code was tested heavily against the results of the traditional interactive analysis for ∼1000 chromatograms. The code gives the position and area of each peak along with an estimate of the quality of fit. Judging the quality of fit for an individual peak in a group of overlapping peaks is not always possible, so the quality indicator is defined for the group of peaks only.

The detailed example was run with an analysis for hydrocarbons in an ambient air sample. The automatic procedure found most of the peaks that the interactive analysis found, and for almost all peaks the areas agreed within the limits of accuracy of the determination of the baseline. Where there were differences, the new procedure was superior in characterizing small peaks on the sides of large ones, and inferior only for distorted peaks that happen to coincide with areas where operating conditions vary. This information is used in interactive analysis to improve the estimate, while the automatic analysis does not have the information and—at the time being—would not know how to use it. The automatic analysis also found a few very small peaks missed by the interactive analysis, but those could not be identified. It disregarded some peaks found interactively as too small. One of those, 2,3-dimethyl-butane, could be identified, but its height is only two times the noise level and it sits on the tail of 2-methyl-pentane, so the interactive detection is possible only because of a priori knowledge on the peak position. The results are summarized in Table 1. The automatic procedure for this example takes 70 s on a 1.2 GHz Athlon under Linux. The time needed depends heavily on the number and arrangement of over-

Table 1
Comparison of retention times, peak areas, and peak heights of an example chromatogram analyzed manually (with APEX software) and automatically with the method described here

Compound	Retention time, manual (min)	Retention time, automatic (min)	Area, manual (μV s)	Area, automatic (μV s)	Height, manual (μV s)	Height, automatic (μV s)	Remarks
???		6.782		204510		1792	
???		6.868		14222		2015	
Ethane	7.410	7.415	1445428	130215	33057	35177	Double
???		7.535		967943		36132	
Ethene	9.020	9.008	894799	838928	33058	33717	Disturbed[a]
???		9.269		116370		6416	
Propane	13.385	13.385	1437088	1420958	171874	173633	
Propene	17.302	17.300	1445645	1420043	237768	239522	
i-Butane	19.286	19.285	1981306	1922830	185118	185476	
Propyne	23.731	23.731	1250892	1272450	247709	247815	
1,3-Butadiene	24.037	24.037	2288166	2284181	163902	161739	
(*E*)-2-Butene	24.719	24.718	2020290	1929006	328940	326482	
2,2-Dimethyl-propane	25.155	25.154	2036822	1870640	304849	303763	
Water		28.365		490225		11424	
Cyclohexane	31.008	31.000	1112893	1133312	75078	74254	
2-Methyl-1-butene (2)	31.970	31.980	133686	70945	5119	3624	On tail
2-Methyl-pentane	33.828	33.829	19066	23725	2778	2860	Small
2,3-Dimethyl-butane	34.050		6893		609		Classified as noise
3-Methyl-hexane	35.967	35.970	3424855	3120798	247637	243432	
???		36.270		267376		12978	On tail
n-Heptane	36.884	36.885	720036	700922	82798	81314	
???	37.140	37.135	104592	38644	11481	5527	On tail
1-Hexene	37.608	37.611	53993	35316	6145	5406	On tail
Octane	41.572	41.576	1081097	1050082	125164	125375	
???	42.879		15585		1474		Classified as noise
Toluene	43.530	43.530	1696590	1652065	179516	178737	
n-Nonane	45.801	45.804	372570	359172	43650	43924	
???	47.420		16779		1528		Classified as noise
Ethylbenzene	47.790	47.774	791527	807756	85502	86334	
Styrene	48.070	48.075	480657	395524	40197	33955	On tail
???	48.490	48.540	56669	42257	4632	2778	
p,m-Xylene	48.710	48.700	812232	779928	51854	78898	
o-Xylene	49200	49.194	595389	565759	51854	50376	
???	49.650		51319		2812		Classified as noise
Cumol	50.859	50.858	431886	435799	41781	41352	
n-Propylbenzene	51.318	51.318	316576	279731	29330	28189	
???	52.520		33724		2086		Classified as noise
1,3,5-Methyl-benzene	52.980	52.974	121744	205279	14032	14036	Double
???	53.100	53.114	145643	71268	11829	4810	Double
1,2,4-Methyl-benzene	53.785	53.785	211170	190720	12010	11322	

The chromatogram is shown in Fig. 2. ??? indicates an unidentified compound.
[a] Known interference with CO_2.

lapping peaks, it is hard to predict. In the routine usage for chromatograms with 50–100 peaks, average time used was about 2 min. The program is presently not optimized for speed.

Interactive refinement is possible. For each group of peaks it is possible to interactively split the group into two groups, change the threshold for peak size, delete peaks, add peaks, and perform additional iterations of the nonlinear optimization. In practice, this has rarely been done; the verification data results from fully automatic runs with every parameter set to default, while for the Figs. 5 and 6 groups of peaks were split at a point where the overlap was too small to be of any importance—the automatic estimate of overlap is rather cautious.

4. Verification

A comprehensive verification of any method requires a trusted reference. The only way to do this for chromatography is to analyse a synthesised mixture of compounds, but this is restricted to a small number of compounds and therefore not viable for the problems addressed her. Instead we used a synthetic chromatogram to do this. Two series of 43 peaks, one of Fazer–Suzuki type, the other a bi-Gaussian one, were analysed, both having identical random positions and sizes, with similar random width and skewness, and different random noise added. Most of the peaks where found with accurate position and an error in area of less than 1%. Larger errors appeared for very small peaks and for strongly over-

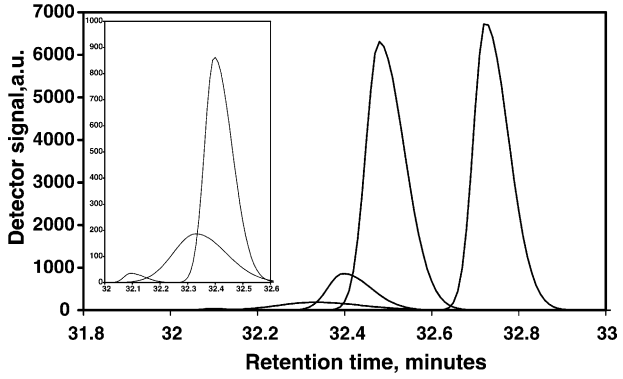

Fig. 8. Synthetically generated peaks for a test of the new procedure. The insert shows an enlarged plot of the first three peaks.

lapping multiple peaks. For the bi-Gaussian peaks, even a separation of $\sigma/2$ was enough to give fairly correct (2% error) peak areas for peaks of same size. However, small peaks within an overlapping group where not estimated reliably. For the Fazer–Suzuki type peaks the required separation is a little larger.

The result of a fit to peaks of a very difficult group of five peaks (the synthetic peaks are shown in Fig. 8) is given in Fig. 9. The fit to the first small peak and the two large peaks is almost perfect. The area estimates for the two small peaks in between have large errors, which shows the limitations of the method. The sum of these two peaks, however, fits the data well.

What is also possible is to check the new procedure against standard—though not perfect—procedures with real chromatograms and have a close look at the differences. This has been done for eight compounds from ~1000 chromatograms obtained during a field campaign in 2002. The peaks of these compounds (acetone, isoprene, methacrolein, methyl vinyl ketone (MVK), benzene, toluene, ethylbenzene, *p*,*m*-xylene, and *o*-xylene) have been integrated both interactively and with the approach described here. From this comparison, we

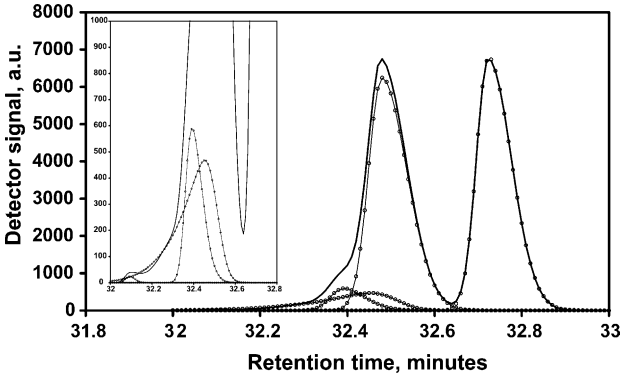

Fig. 9. Plot of the fits to the synthetic peaks shown in Fig. 8. The solid line is the sum of the five synthetic peaks. The circles represent the fitted peaks. The insert shows an enlarged plot of the first three peaks. Note the considerable different peak shape of the third peak (for details see text).

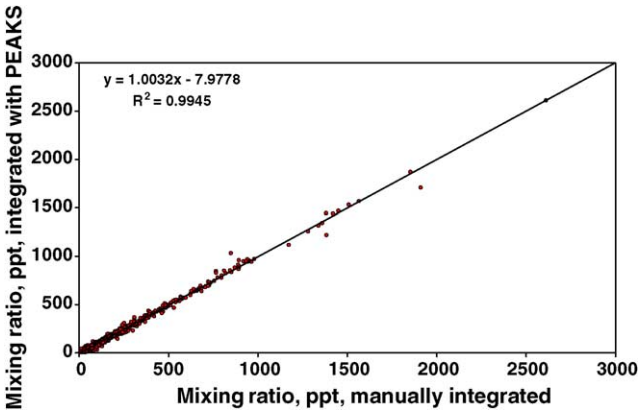

Fig. 10. Plot of the isoprene peak area obtained by automatic integration vs. the peak area obtained by manual integration.

show two examples (see Figs. 10 and 11) of a large (isoprene) and a rather small (methyl vinyl ketone) peak. Obviously, all small values have a large error margin due to the influence of noise and even more due to the uncertainty in the level of the baseline, and this error is present in the traditional method as well as in the new one. Therefore, it is no surprise that there is considerable deviation from the regression line for small concentrations. For larger values, the isoprene plot shows excellent agreement of the methods except for three points where the 'manual' values are slightly larger. The chromatographic separation of the samples has been done on a DB-5 capillary column. On this column isoprene elutes immediately after acetone in the chromatogram, and the acetone peak shows a considerable tailing on this column. The manual analysis will thus add the end of the acetone tail to the area of the isoprene peak. If the acetone peak is very large as it was in these cases, this leads to an overestimate of the isoprene peak area. The mechanism of this error may easily be seen in Fig. 6, where the manual procedure would give about 10% smaller area of the leftmost peak than the obviously correct automatic procedure, and distribute this onto the following

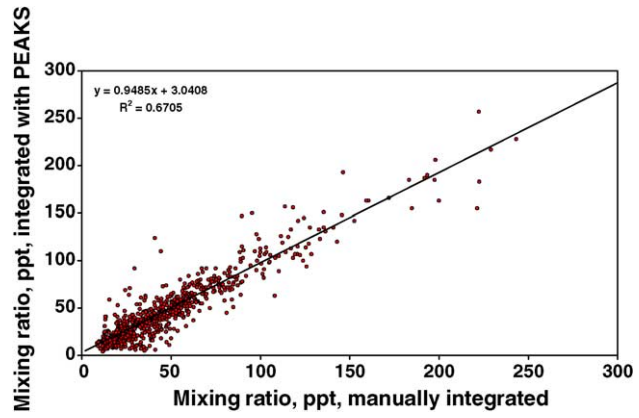

Fig. 11. Plot of the methyl vinyl ketone peak area obtained by automatic integration vs. the peak area obtained by manual integration.

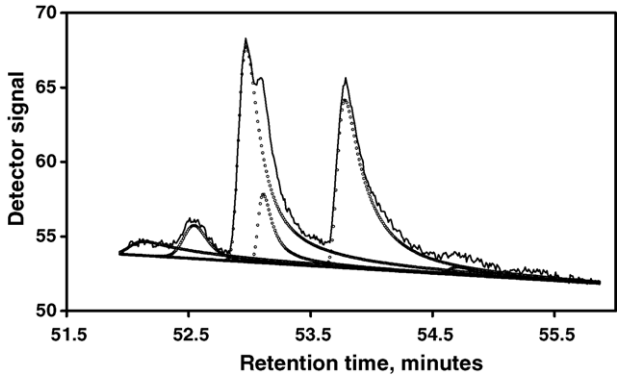

Fig. 12. Plot of the group containing 1,3,5-trimethylbenzene and 1,2,4-trimethylbenzene. The solid line is the original chromatogram, the circles represent the fit of the individual peaks (for details see text).

peaks. With MVK, the situation is more complicated. It is preceded and followed closely by peaks of about half the size of MVK, with large variations in relative size. Therefore, the manual procedure may overestimate or underestimate the concentration of MVK depending on the relative sizes of the neighbouring peaks. Of all the substances checked, MVK had the largest differences between manual and automatic analysis.

A further test was detailed comparison of the entire analysis with old and new method for a chromatogram. The result for the chromatogram shown in Fig. 2 is given in Table 1. Most peaks agree very well but there are a few deviations that require close scrutiny. The ethane peak produces the largest difference. As may be seen from Fig. 2, there is a very sharp peak at the left flank of a rather broad and strongly asymmetric one. While the APEX analysis sees only one peak, the automatic one sees two, as is obviously correct. The remaining difference may be explained by differences in the estimate for the basis, which shows strong variations in this area. The automatic analysis uses a higher—more cautious estimate. Further large deviations are for 2-methyl-1-butene, where APEX adds a part that should be attributed to cyclohexane. For 3-methyhexane again APEX combines two peaks to one area. Finally, for the three overlapping peaks of 1,3,5-methylbenzene, the sums agree, but the automatic analysis gives a different partitioning, which seems much more plausible on visual inspection (Fig. 12).

For this chromatogram a fit using EMG peak shape was done. The best EMG fit to the four selected peaks had a least square deviation between two and four times larger than the approximation used here, because at the region of highest curvature on the tail they have higher levels than EMG allows. For the overlapping peaks, the difference in fit quality varied from similar to about six times the least square deviation given by our model, always using the same base line approximation. If EMG gives a good fit to the isolated peaks, the difference between the models is very small.

5. Conclusion

We present a new mathematical approach which makes use of peak shapes extracted from the chromatogram to be analysed. This approach gives substantially better fit results than traditionally defined peak shapes. A test of this approach on a large set of chromatographic data compared with standard peak detection and integration software showed the strength of this procedure. Even for chromatograms where there is no good analytical model for the peak shape, our procedure allows a reliable automatic analysis.

The peak detection of extremely small peaks in a difficult synthetic chromatogram shows that there is much room for improvement. Especially the heuristics for determining the baseline—while obviously quite accurate in many cases—has problems treating variations in operation conditions, and has little theoretical foundation.

Further, there are some parameters (thresholds, initial guesses, etc.) of the implementation that allow tuning. The default choice tries to avoid over-fitting, even if it means missing actual peaks. The approach focuses on the problem of analysing overlapping peaks. The problem of detecting well separated tiny peaks is not addressed. This could be done by incorporating suitable methods [6] into the parameter initialisation heuristics.

Applying this procedure for the peak detection and calculation of the peak areas makes a fully automatic processing of complex chromatograms possible. Despite some remaining problems it may considerably reduce time needed for analysis and increase the throughput in environmental analysis.

Acknowledgements

We thank Dr. M. Gautrois for testing the software and providing hints to problems as well as guidance for the user interface. Part of this work was funded by the German Ministry for Education and Research within the German Atmospheric Research Program AFO 2000 under grant no. 07ATF47.

References

[1] V.G. Yakovlev, Automat. Remote Contr. 40 (3) (1979) 472.
[2] W.F. Hargrove, P. Rosenthal, Anal. Chem. 53 (3) (1981) 538.
[3] J.L. Excoffier, G. Guichon, Chromatographia 15 (9) (1982) 543.
[4] W.A. Spencer, L.B. Rogers, Anal. Chem. 52 (6) (1980) 950.
[5] C.A. Hastings, S.M. Norton, S. Roy, Rapid Comm. Mass Spectrom. 16 (5) (2002) 462.
[6] K.H. Jarman, D.S. Daly, K.K. Anderson, K.L. Wahl, Chemom. Intell. Lab. Syst. 69 (1/2) (2003) 61.
[7] P.J.P. Cardot, P. Trolliard, S. Tembely, J. Pharm. Biomed. Anal. 8 (8–12) (1990) 755.
[8] A. Felinger, Data Analysis and Signal Processing in Chromatography, Elsevier, Amsterdam, 1998, ISBN 0-444-82066-3.
[9] B. Steffen, R. Koppmann, Patent application PCT/DE 03/00560.
[10] C. Gomez (Ed.), Engineering and Scientific Computing with Scilab, Birkhauser Boston, 1999, ISBN 0-8176-4009-6.