The wavelet transform for pre-processing IR spectra in the identification of mono- and di-substituted benzenes

M. Bos *, J.A.M. Vrielink

University of Twente, Department of Chemical Technology, P.O. Box 217, 7500 AE Enschede, Netherlands

Abstract

This paper describes the wavelet transformation of IR spectra with the Daubechies analyzing wavelet functions as a feature extracting method that successfully reduces the spectral data more than 20-fold with a significant improvement in the classification process.

1. Introduction

Wavelet theory involves representing general functions or signals in terms of simpler, fixed building blocks of constant shape but different size and at different positions. This has been found a useful approach in several areas as diverse as acoustics [1], fluid mechanics [2] and chemical analysis [3] and even in the field of mathematics where it is used as a tool for fast matrix computations [4]. An introduction to the technique and some applications of this relatively new technique can be found in the proceedings of an early conference on this subject [5].

The developments in this field have taken place along different lines. In mathematics Coifman [6] and Coifman and Weiss [7] developed a method to use so-called 'atomic decompositions' in harmonic analysis. For the interpretation of seismic signals Grossmann and co-workers [8,9] devel-

oped the continuous wavelet transform. Mallat [10] developed a theory for multi-resolution signal decomposition with wavelets and Daubechies [11] connected it to quadrature mirror filtering. Recently Meyer [12] brought these developments together in a unified mathematical framework. Meanwhile public domain software has become available to carry out the calculations in an efficient manner [13,14].

Basically a wavelet decomposition is an expansion of an arbitrary function into smooth localized contributions labeled by a scale and a position parameter. This position parameter corresponds with the independent variable of the measuring technique involved. Generally this will be a time-axis, but for spectroscopic techniques it can be the wavelength axis.

In some respects the wavelet transform resembles the well-known Fourier transform (FT) in which the sine and cosine are the basic analyzing functions. Like the Fourier transform, the wavelet transform is a linear operation that operates on a data vector and transforms it from the input

---

* Corresponding author.

space, which is the time domain, to a different domain. In the Fourier transform this new domain is the frequency domain. The wavelet domain, however, is characterized by *two* axes, forming an open 'time and scale' half plane in which frequency and temporal information about the signal are retained. Just like other types of transforms the wavelet transform realizes a correspondence between the signal in its original domain to the new domain by convoluting the signal $s(t)$ to be analyzed with the analyzing function. The mathematics needed for this process in case of the wavelet transform are given below and were taken from the paper by Kronland-Martinet, Morlet and Grossman [1].

The analyzing function is defined in this case in the two-dimensional space of scale and time and is derived from a mother wavelet denoted by $g_{basic}(t)$. Two processes are used to obtain a number of 'self-similar' wavelets from this mother wavelet $g_{basic}(t)$: (1) shifts in the time variable and (2) dilations which act both on the time and the scale variable. If the variable b is used to control the shift in time and the variable a is used to control the dilation then the series of wavelets that are obtained by these two processes can be described by:

$$g_{a,b}(t) = \frac{1}{\sqrt{a}} g_{basic}\left(\frac{t-b}{a}\right) \qquad (1)$$

In this equation $1/\sqrt{a}$ is a normalizing constant that ensures that all analyzing functions have unit energy.

To ensure the wanted properties of localization in time and space, the analyzing wavelet function has to fulfil a number of conditions:

$$\int |g(t)|^2 \, dt < \infty \qquad (2)$$

for finite energy and

$$\int |\hat{g}(\omega)|^2 \frac{d\omega}{|\omega|} < \infty \qquad (3)$$

in which $\hat{g}(\omega)$ denotes the Fourier transform of $g(t)$. This condition ensures a short-wave like behaviour and generally implies that $g_{basic}(t)$ does not have a direct current component.

A third restriction imposed on the function representing the analyzing wavelet is

$$\hat{g}_{basic}(\omega) = 0 \quad \text{for} \quad \omega < 0 \qquad (4)$$

The wavelet transform is now defined as the function $S(b, a)$ on the open (b, a) half-plane (b arbitrary, $a > 0$):

$$S(b, a) = \int_{-\infty}^{\infty} [g_{b,a}(t)]^* s(t) \, dt \qquad (5)$$

where the asterisk denotes that the function g is complex. This equation can be rewritten as:

$$S(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g_{basic}^*\left(\frac{t-b}{a}\right) s(t) \, dt \qquad (6)$$

The conditions imposed on the wavelet function imply that $\hat{g}(\omega)$ is negligible above a certain frequency ω_{max}, making $S(b, a)$ insensitive to the higher Fourier components of the signal $\hat{s}(\omega)$. This eliminates the influence of small-scale features. Furthermore these conditions make g negligible outside an interval $[t_{min}, t_{max}]$ of the t axis causing $S(b, a)$ to be insensitive to the values of $s(t)$ such that $t - b$ lies outside of $[t_{min}, t_{max}]$. The latter is the origin of the desired localization in time.

A commonly used analyzing wavelet that fulfils these conditions is the Morlet wavelet:

$$g_{basic,Morlet}^*(t) = e^{i\omega_0 t} e^{-t^2/2} + \text{small corrections} \qquad (7)$$

These small corrections are numerically negligible when $\omega_0 > 5$ but have to be added since $\exp[-(\omega - \omega_0)^2]$ does not vanish for $\omega < 0$.

In practice the signals to be analyzed will be acquired by computer at discrete time intervals. This necessitates the use of a discrete form of Eq. (6):

$$S(iT_s, a) = T_s \frac{1}{\sqrt{a}} \sum_n g^*\left(\frac{(n-i)T_s}{a}\right) s(nT_s) \qquad (8)$$

where $1/T_s$ is the sampling frequency.

Let $g_a^*(iT_s) = g^*(iT_s/a)$. So, for each value of a, the analyzing wavelet is sampled, yielding the sequence $g^*(iT_s)$. Then the convolution product between $s(nT_s)$ and $g_a^*(iT_s)$ is computed. As $g(t)$ has finite support the number of sampling points of $g_a^*(t)$ is finite and grows linearly with a.
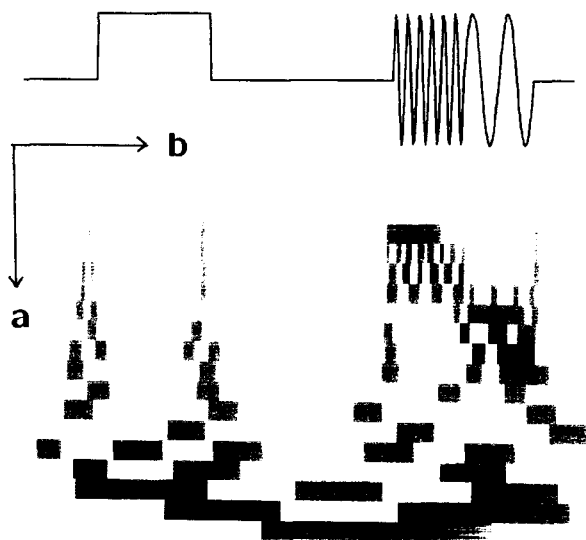
Fig. 1. Signal showing transients and local periodicity and its wavelet transform in greytones.

Fig. 1 shows the visualization of the results of the wavelet transformation with a program we developed earlier [3] of a signal with a block pulse and two connected pieces of a pure sine wave signal of which the first piece has a frequency that is three times the frequency of the second one.

The wavelet transform of this signal obtained with the use of Eq. (8) and the Morlet wavelet from Eq. (7) as the analyzing function produces a matrix of $S(iT_s, a)$ values as the result. To visualize this result the elements of this matrix are plotted in a two-dimensional picture in which the position of the elements in the matrix is retained. Generally the convention is followed to plot the time (iT_s) facing to the right and calling it the b axis, whereas the scaling factor, a, is plotted downwards. Small scales, roughly speaking, correspond to higher frequencies, so the higher frequencies are on top of the picture. The magnitudes of the $S(iT_s, a)$ elements are represented in this picture as greytones. Darker spots correspond to higher absolute values of S.

Looking at Fig. 1, one can see that at the time of the flanks of the block pulse the signal contains a large number of different frequencies, indicated by the presence of dark spots at various

scales vertically under the transitions of the signal. The same goes for the moments on which the sine wave starts, changes frequency or stops. Local periodicities can also be spotted in this picture. They show up as blocked bands for their duration at scale values corresponding with their periodicity.

If the results of a wavelet transform are inspected numerically, the first thing to note is that the majority of the coefficients have a very low absolute value. The reason for this is the dual localization, both in scale and position, that compresses the energy of the signal into a small number of consequently large coefficients. Putting the other relatively small coefficient values to zero and reconstructing the signal by inverse transformation hardly deteriorates the signal. This property can be used to compress signals to a fairly high degree without much loss of information.

These properties open up a number of possible applications in chemical analysis. The time–frequency representation of the wavelet transform can be used to extract particular features from a signal. In this respect the wavelet transform can be seen as a mathematical microscope to look at the signal. It is also possible to determine time delays from this type of representation.

The sparsity of the wavelet transform can be used to advantage in compressing spectra to be used in library search methods. Another promising application of the sparsity property is in the linear algebra needed to solve linear multivariate calibration problems.

The work described in this paper was performed to investigate whether multi-resolution analysis by wavelet transformation can be used successfully to pre-process infrared spectra for identification purposes. The relevant information of an IR spectrum is contained in the position and the shape of the absorption peaks. The purpose of this paper is to show that the wavelet transformation, due to its use of localization, both in position and scale, can extract this information in a concentrated form and thus can be used to extract the salient features of an IR spectrum effectively. Using this concentrated information instead of the full spectra for the clas-

sification of compounds greatly improves the speed with which the classifiers can be derived. Moreover it can be expected that the quality of the classifiers will improve if they are derived from smaller data sets that still contain all the relevant information.

To test this hypothesis two types of classifiers, a linear and a non-linear, were derived from the full spectra as well as from spectra processed with different wavelet analyzing functions. A comparison of the performance of these classifiers then will show whether the hypothesis is valid.

From the many wavelet analyzing functions that are available, the members of the Daubechies series were chosen for this investigation as they offer varying degrees of compactness of support and regularity [11], see Fig. 2.

The main effort of this work was spent in establishing which scale(s) of which Daubechies analyzing wavelet function are the best in preprocessing the IR spectra to be used in identifica-
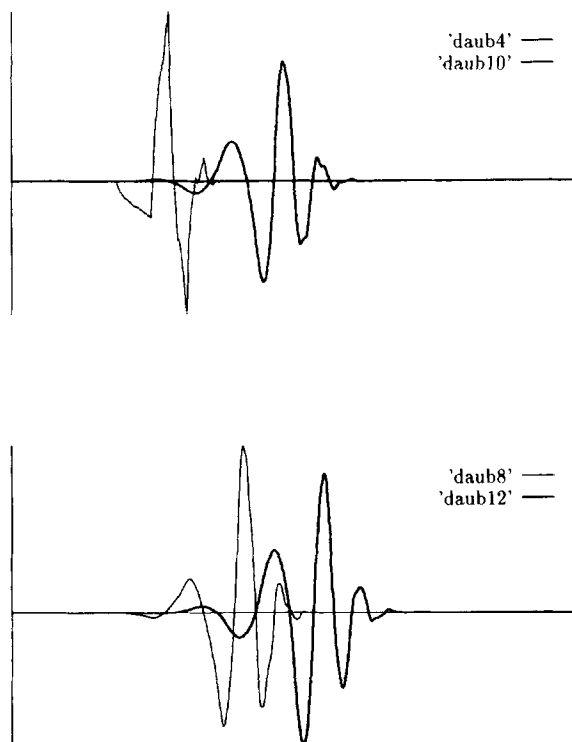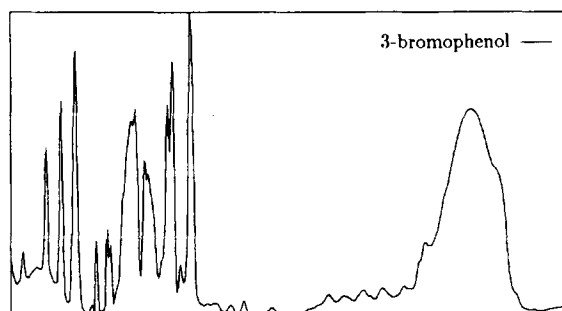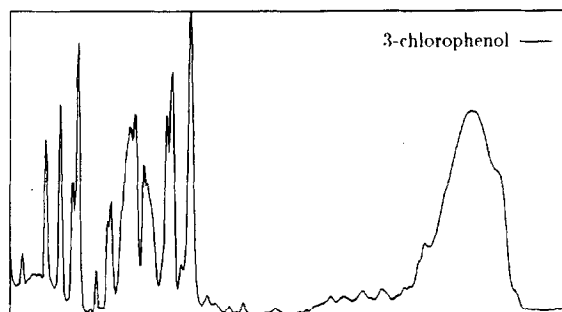

Fig. 3. 'Hard' classification problem.

tion systems based on the linear as well as the non-linear (neural network) classifiers. For this purpose a data set was chosen that comprises the IR spectra of most of the mono- and di-substituted benzenes with the substituents Cl, Br, F, NO_2, COOH, OH, CH_3 and C_2H_5. Within this data set there are a number of 'hard' classification problems as judged by experts in the field, yet the set has a size that is not too difficult to handle. Furthermore the spectra can be considered to contain quite a number of the common IR absorption bands. An example of such a 'hard' classification problem is given in Fig. 3.

The test set was chosen to contain difficult classification problems. The spectra of corresponding compounds substituted with different halogens look very similar. Therefore the test set contains a relatively large number of halogen substituted compounds. The same difficulties occur with ethyl versus methyl substitution. The ethyl substituted compounds are rather underrep-

Fig. 2. Analyzing wavelet functions from Daubechies series.

resented in the data set, but serve the purpose to harden the classification of the cresols in the test set.

2. Experimental

Tables 1 and 2 show the complete primary data set used in this investigation. The spectra

Table 1
Composition of training set

Id.	Compound	Id.	Compound
A01	Benzene	A37	Nitrobenzene
A05	2-Fluorophenol	A40	2-Bromonitrobenzene
A08	3-Chlorophenol	A43	3-Bromonitrobenzene
A11	4-Bromophenol	A56	1,2-Dibromobenzene
A14	4-Ethylphenol	A61	1,4-Difluorobenzene
A17	Bromobenzene	A70	1,2-Dimethylbenzene
A20	2-Bromotoluene	A76	2-Nitrophenol
A23	3-Bromotoluene	A81	4-Nitrotoluene
A38	2-Fluoronitrobenzene	AC85	2-Hydroxybenzoic acid
A41	3-Fluoronitrobenzene	AH77	3-Nitrophenol
A44	4-Fluoronitrobenzene	AW27	Benzoic acid
A57	1,3-Dibromobenzene	AW34	4-Fluorobenzoic acid
A62	3-Hydroxytoluene	AW45	4-Chloronitrobenzene
A71	1,3-Dimethylbenzene	AW66	1,4-Dihydroxybenzene
A79	2-Nitrotoluene	AW75	1,4-Benzenedicarboxylic acid
A02	Toluene		
A06	2-Chlorophenol	AW69	1,4-Dinitrobenzene
A09	4-Chlorophenol	AW86	3-Hydroxybenzoic acid
A12	2-Ethylphenol	AW87	4-Hydroxybenzoic acid
A15	Chlorobenzene	AH32	3-Chlorobenzoic acid
A18	2-Chlorotoluene	AH78	4-Nitrophenol
A21	3-Fluorotoluene	AW30	2-Bromobenzoic acid
A24	4-Fluorotoluene	AW35	4-Chlorobenzoic acid
A39	2-Chloronitrobenzene	AW46	4-Bromonitrobenzene
A42	3-Chloronitrobenzene	AW67	1,2-Dinitrobenzene
A54	1,3-Dichlorobenzene	AW73	1,2-Benzenedicarboxylic acid
A59	1,2-Difluorobenzene		
A63	4-Hydroxytoluene	AW83	3-Nitrobenzoic acid
A72	1,4-Dimethylbenzene	AW90	4-Methylbenzoic acid
A80	3-Nitrotoluene	AH34	4-Fluorobenzoic acid
A03	Phenol	AW25	4-Chlorotoluene
A07	2-Bromophenol	AW33	3-Bromobenzoic acid
A10	4-Fluorophenol	AW36	4-Bromobenzoic acid
A13	3-Ethylphenol	AW64	1,2-Dihydroxybenzene
A16	Fluorobenzene	AW68	1,3-Dinitrobenzene
A19	2-Fluorotoluene	AW74	1,3-Benzenedicarboxylic acid
A22	3-Chlorotoluene		
		AW84	4-Nitrobenzoic acid
		AW88	2-Methylbenzoic acid

Table 2
Composition of test set

Id.	Compound	Id.	Compound
A04	2-Hydroxytoluene	A60	1,3-Difluorobenzene
A48	3-Bromophenol	A47	3-Fluorophenol
A53	1,2-Dichlorobenzene	A50	4-Hydroxytoluene
A26	4-Bromotoluene	AW29	2-Chlorobenzoic acid
A49	3-Hydroxytoluene		

with an identification number prefix 'A' were taken from the Nicolet / Aldrich IR Spectral Library, those with prefix 'AH' from the Nicolet / Hummel Polymer IR Spectral Library, those with prefix 'AC' from the Sample Georgia State Crime Laboratory Drugs IR Spectral Library, whereas the spectra with the prefix 'AW' were recorded in house as a KBr pellet on a Nicolet 5SXC FT-IR spectrometer at a resolution of 4 cm^{-1}. After baseline correction to 100% transmission and normalizing the spectra between 100% and 10% T, the latter spectra were converted to the same library format as the other spectra by the Nicolet software on the computer system of the spectrometer. This format represents the spectra as 460 data points equidistantly divided over the range 3996.4–455.5 cm^{-1}.

Wavelet transformation of the spectra with analyzing wavelets from the Daubechies family were carried out with the fast wavelet transform algorithm described in ref. [14]. This algorithm presents the resulting wavelet coefficients not in the form of a matrix as given in Fig. 1, but as a one-dimensional array filled sequentially with the expansion coefficients of the different scales. The expansion coefficients represent the approximation of the original signal with a resolution of one point per 2^j points of the original signal in which j is the level number. Higher levels correspond to lower frequencies.

As the spectra consisted of 460 data points and the fast wavelet transform algorithm operates on a data buffer of a length that is a power of two, the spectra were padded with zeros to 512 data points.

Using either the full spectra or the wavelet transform coefficients produced by one or two chosen levels of the scale of the analyzing func-

tion as inputs, classifiers were trained on the spectra of the 73 compounds from Table 1 and their prediction ability was tested on the spectra of the 9 compounds from Table 2. The targets for the classification were the presence or absence of the 8 substituents mentioned earlier, mono- or di-substitution and the position of the substituents (*ortho-*, *meta-*, *para-*). In total 12 classifications per spectrum are produced in this way. There is some redundancy in this output target coding schema that allows some consistency checks.

Two types of classifiers were used: linear ones, trained with the Ho–Kashyap singular value decomposition (SVDHK) algorithm [15], and nonlinear ones in the form of fully feed forward connected neural network models with sigmoid neurons and one hidden layer, trained with the backpropagation algorithm [16]. The classifiers were optimized for good prediction behaviour on the test set. The neural network classifiers were trained to perfect recognition of the training set.

3. Results and discussion

Tables 3–7 show the best prediction results obtained with linear and neural network classifiers, trained with the coefficients obtained with various levels of the scale of the different analyzing wavelets from the Daubechies series, i.e., DAUB4, DAUB6, DAUB8, DAUB10 and DAUB12. For comparison purposes each of these

tables contains the results obtained with both types of classifiers derived from the full data matrix of spectra as the first entry. In the num-

Table 3
Best prediction results with classifiers trained on DAUB4 transform

Transform levels(s)	No. of coeffs. in level	Pred. errors	
		SVD	Neural network
Full spectra	(460)	10	10
9	256	16	16
8	128	14	16
7	64	16	16
6	32	11	14
5	16	13	15
6+7	96	14	12
5+6	48	9	12

Table 4
Best prediction results with classifiers trained on DAUB6 transform

Transform level(s)	No. of coeffs. in level	Pred. error	
		SVD	Neural network
Full spectra	(460)	10	10
7	64	14	13
6	32	13	12
5	16	13	10
4	8	14	12
3	4	20	
5+6	48	12	13
4+5	24	10	8
3+4	12	13	12

Table 5
Best prediction results with classifiers trained on DAUB8 transform

Transform level(s)	No. of coeffs. in level	Pred. error	
		SVD	Neural network
Full spectra	(460)	10	10
7	64	15	14
6	32	15	11
5	16	13	8
4	8	12	16
5+6	48	11	12
4+5	24	10	9

Table 6
Best prediction results with classifiers trained on DAUB10 transform

Transform level(s)	No. of coeffs. in level	Pred. error	
		SVD	Neural network
Full spectra	(460)	10	10
7	64	16	19
6	32	17	13
5	16	16	12
4	8	16	16
6+7	96		12
5+6	48	11	12
4+5	24	11	11
3+4	12	14	16

Table 7
Best prediction results with classifiers trained on DAUB 12 transform

Transform level(s)	No. of coeffs. in level	Pred. error	
		SVD	Neural network
Full spectra	(460)	10	10
7	64	16	18
6	32	14	10
5	16	15	14
4	8	13	
3	4	20	
5 + 6	48	11	13
4 + 5	24	9	12
3 + 4	12	11	11

bers given as prediction errors for a given system false positives as well as false negatives are counted. If an IR spectrum is classified as belonging to bromobenzene, while the spectrum actually belongs to chlorobenzene, this classification counts as two errors, one for the false positive of the bromo-substituent and one for the false negative classification of the chloro-substituent.

From these tables it can be observed that the linear as well as the neural network classifiers can be successfully trained on a very limited number of coefficients from the wavelet transform of their IR spectra. Compared to the use of the full spectra to derive the classifiers, there is little loss of predictive ability if all the coefficients of level four or a higher level are used. Furthermore the differences between the results of the linear classifiers and the neural network classifiers are small.

Tables 8 and 9 show a summary of the best prediction results obtained with the various analyzing wavelets. For the linear classifiers there is

Table 8
Comparison of best prediction results with linear classifiers trained on different transforms from the Daubechies series

Transform	Best level(s)	No. of coeffs.	Pred. error
DAUB4	5 + 6	48	9
DAUB6	4 + 5	24	10
DAUB8	4 + 5	24	10
DAUB10	4 + 5	24	11
DAUB12	4 + 5	24	9

Table 9
Comparison of best prediction results with neural network classifiers trained on different transforms from the Daubechies series

Transform	Best level(s)	No. of coeffs.	Pred. error
DAUB4	5 + 6	48	12
DAUB6	4 + 5	24	8
DAUB8	5	16	8
DAUB10	4 + 5	24	11
DAUB12	6	32	10

little difference. The combined wavelet transform coefficients from the levels four and five generally produced the best results and performed equal to, somewhat better or somewhat less than the full spectra. These levels together contain 24 coefficients, thus the use of the wavelet transform can reduce the data needed in the development of the classifiers almost 20-fold.

For the neural network classifiers there are somewhat more differences. Here the use of the DAUB4 wavelet transform coefficients produces results that are worse than the results obtained with the full spectra. For the DAUB8 transform, however, the results are clearly better than for the full spectra, indicating that the 16 coefficients that are obtained from the spectra with this transform have extracted the relevant features of the IR spectra in a very concentrated form.

Table 10 shows the prediction results of the neural network classifier with 16 neurons in its hidden layer that were derived with the spectra transformed with the DAUB8 analyzing wavelet

Table 10
Prediction results for test set samples with neural network classifier derived from level 5 of DAUB8 transform

Compound	Found	Error(s)
2-Hydroxytoluene	Missed CH_3	1
4-Bromotoluene	4-Chlorotoluene	2
3-Fluorophenol	Missed F	1
3-Bromophenol	3-Bromophenol	0
3-Hydroxytoluene	3-Hydroxytoluene	0
4-Hydroxytoluene	4-Hydroxytoluene	0
1,2-Dichlorobenzene	1,2-Dichlorobenzene	0
1,3-Difluorobenzene	1,3-Fluorotoluene	2
2-Chlorobenzoic acid	Benzoic acid	2

and using the 16 coefficients of the level 5 scale of this transform. Although the test set was chosen to contain 'hard' classification problems, the overall performance of the neural network classifier is more than 90% correct on this particular test set. Most errors concern halogen substitution. Nevertheless, the neural net is able to classify the spectrum of 3-bromophenol correctly, without having seen this spectrum in the training phase, whereas the closely resembling spectrum (see Fig. 3) of 3-chlorophenol was used in its training.

From the compounds from Table 10 that were not classified correctly, the redundant coding scheme only could identify the misclassification of 2-chlorobenzoic acid. In this case no *ortho, meta* or *para* substitution was found, but di-substitution was indicated which is inconsistent.

Summarizing, it can be concluded that pre-processing the IR spectra of mono- and di-substituted benzenes with the DAUB8 wavelet transform reduces the data needed to derive the classifiers for the identification of these compounds more than 20-fold. Especially for the training of the non-linear neural network classifiers, that perform better than the linear classifiers, this means a considerable saving in computer time. Moreover the performance of these classifiers derived from the pre-processed spectra is better than for the full spectra.

References

[1] R. Kronland-Martinet, J. Morlet and A. Grossmann, Analysis of sound patterns through wavelet transforms, *International Journal of Pattern Recognition and Artificial Intelligence*, 1 (1987) 273–302.

[2] J. Liandrat and F. Moret-Bailly, The wavelet transform: some applications to fluid dynamics and turbulence, *European Journal of Mechanics, B: Fluids*, 9 (1990) 1–19.

[3] M. Bos and E. Hoogendam, Wavelet transform for the evaluation of peak intensities in flow-injection analysis, *Analytica Chimica Acta*, 267 (1992) 73–80.

[4] F. Keinert, Biorthogonal wavelets for fast matrix computations, *Applied and Computational Harmonic Analysis*, submitted.

[5] J.M. Combes, A. Grossmann and Ph. Tchamitchian (Editors), *Wavelets, Time–Frequency Methods and Phase Space, Proceedings of the International Conference, Marseille – France, Dec. 14–18, 1987*, Springer, Berlin, 2nd edn., 1990.

[6] R.R. Coifman, A real variable characterization of h^p, *Studia Mathematica*, 51 (1974) 269–274.

[7] R. Coifman and G. Weiss, Extensions of hardy spaces and their use in analysis, *Bulletin of the American Mathematical Society*, 83 (1977) 569–645.

[8] A. Grossmann and J. Morlet, Decompositions of hardy functions into square integrable wavelets of constant shape, *SIAM Journal of Mathematical Analysis*, 15 (1984) 723–736.

[9] P. Goupillaud, A. Grossmann and J. Morlet, Cycle-octave and related transforms in seismic signal analysis, *Geoexploration*, 23 (1984) 85–102.

[10] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11 (1989) 674–693.

[11] I. Daubechies, Orthonormal bases of wavelets with finite support — connection with discrete filters, in J.M. Combes, A. Grossmann and Ph. Tchamitchian (Editors), *Wavelets, Time–Frequency Methods and Phase Space, Proceedings of the International Conference, Marseille — France, Dec. 14–18, 1987*, Springer, Berlin, 2nd edn., 1990, p. 38–66.

[12] Y. Meyer, *Ondelettes*, Hermann, Paris, 1990.

[13] M.A. Cody, The fast wavelet transform, *Dr. Dobbs Journal*, 17 (1992) 16–28.

[14] W.H. Press, Wavelet transforms, *Harvard-Smithsonian Center for Astrophysics Preprint, No. 3184*, 1991.

[15] M. Bos and G. Jasink, The learning machine in quantitative chemical analysis, *Analytica Chimica Acta*, 103 (1978) 151–165.

[16] M. Bos, A. Bos and W.E. van der Linden, Processing of signals from an ion-selective electrode array by a neural network, *Analytica Chimica Acta*, 233 (1990) 31–39.