Continuous Wavelet Transform Applied to Removing the Fluctuating Background in Near-Infrared Spectra

Chaoxiong Ma and Xueguang Shao*

Department of Chemistry, University of Science and Technology of China, Hefei, Anhui, 230026, P. R. China

Received September 19, 2003

A novel method based on continuous wavelet transform (CWT) was proposed as a preprocessing tool for the near-infrared (NIR) spectra. Due to the property of the vanishing moments of the wavelet, the fluctuating background of the NIR spectra can be successfully removed through convolution of the spectra with an appropriate wavelet function. The vanishing moments of a wavelet and the scale parameter are two key factors that govern the result of the background elimination. The result of its application to both the simulated spectra and the NIR spectra of tobacco samples demonstrates that CWT is a competitive tool for removing fluctuating background in spectra.

1. INTRODUCTION

As a powerful analytical tool, near-infrared (NIR) spectroscopy has been gaining popularity during the past decade. Its applications can be found in many fields, including the food science,[1,2] the pharmaceutical industry,[3−5] and the analyses of biological and biomedical materials,[6,7] etc. In most cases, background of the NIR spectra is highly fluctuating, thus pretreatment is generally required. Many methods such as the derivative method,[8,9] the genetic regression,[10,11] and the orthogonal signal correction (OSC)[12,13] have been developed to remove the background in the spectra. In addition, wavelet transform (WT), a technique that is popularly used in the processing of analytical signals, recently[14−19] was also used as an alternative preprocessing tool.[20,21] In these studies, one dimension discrete wavelet transform (DWT) was generally adopted to split the spectra into different frequency bands. Combined with other algorithms, the selected bands that represent the signal information are separated and used for reconstruction or directly used in regression.

In this study, one dimension continuous wavelet transform (CWT) was adopted instead of DWT to eliminate a varying background because it has better space-time resolution, which makes it a more precise method, compared with DWT. In addition, CWT is relatively simpler to perform since no reconstruction is required as in DWT. Although the property of the vanishing moments has been used previously,[21] only DWT with a specific wavelet (Symmlets No. 8) with certain vanishing moments was discussed. In our method, the result was improved by adjusting the parameters within the wavelet transform, instead of using a specific wavelet function and combining it with other techniques. It was found that the vanishing moments and the scale parameter are two key factors that affect the performance of the background elimination. Accordingly, why the property of the vanishing moments can be used for removing background was explored. In addition, how these two parameters influence the result was fully investigated, and a method to obtain the optimal parameters was proposed. To demonstrate the applicability of the proposed method, background removal from the NIR spectra of tobacco samples was investigated.

2. THEORY AND ALGORITHM

Wavelet is defined as a series of functions $\psi_{a,b}(t)$ derived from a function $\psi(t)$ by dilation and translation[22,23]

$$\psi_{a,b}(t) = |a|^{-1/2}\psi\left(\frac{t-b}{a}\right),\ a \neq 0,\ a,b \in R \qquad (1)$$

where a is the scale parameter that controls the dilation, b is the shift parameter that controls the translation, and $\psi(t)$ is the basis function of a wavelet. If $f(t)$ is a signal, then the continuous wavelet transform is defined as

$$Wf(a,b) = |a|^{-1/2}\int_{-\infty}^{+\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt \qquad (2)$$

One of the important properties of the wavelet is its vanishing moments. A wavelet with n vanishing moments is orthogonal to a polynomial of degree n-1, i.e., it can be used to suppress a polynomial of degree n-1 through convolution[24]

$$\int_{-\infty}^{+\infty} t^k\psi(t)dt = 0 \ \text{ for } \ 0 \le k < n \qquad (3)$$

The background of spectra is generally assumed to be broader, change slowly, and be represented as a polynomial with a lower degree compared with the signal.[21,25] Therefore, when convoluted with a wavelet function that has appropriate vanishing moments, the background can be well eliminated without significant loss of the signal information. For example, it is easy to deduce from eqs 2 and 3 that when the polynomial $f(t) = k_1t^2 + k_2t^6$ is convoluted with a wavelet function Daubechies No. 4 (db4) that has 4 vanishing moments, the background part (k_1t^2) will be suppressed. The convolution can be illustrated very well in Figure 1 (a) and (b). The intensity of background is the same as the signal

* Corresponding author phone: +86-551-3606160; fax: +86-551-3601592; e-mail: xshao@ustc.edu.cn.

10.1021/ci034211+ CCC: $27.50
Published on Web 03/11/2004

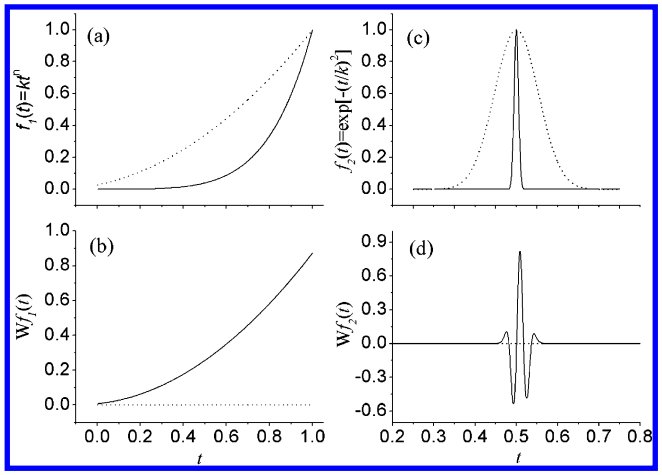

Figure 1. Illustration of the background eliminated by CWT. Parts (a) and (c) are the signal (solid line) and background (dotted line) simulated with polynomial and Gaussian, respectively. Parts (b) and (d) are respectively corresponding to the transform of parts (a) and (c) with a wavelet function db4.

before transform, but it is negligible after transform. This is also true for the Gaussian signal, as illustrated in Figure 1(c), and (d). Furthermore, wavelet transform is a linear operation, and the information of the signal will remain in the transformed coefficients for further regression analysis.

Therefore, after preprocessing with CWT, the obtained coefficients of the spectra can be used for calibration or regression with PLS. The theory and calculation of PLS can be found in the literature.[26] The PLS2 algorithm was adopted for calibration in both simulated and experimental spectra in this study. The root-mean-square error of prediction (RMSEP) is used to evaluate the performance of the proposed method, which is defined by

$$\mathrm{RMSEP} = (\sum_i (y^i_{pred} - y^i_{meas})^2/m)^{1/2} \text{ for } i = 1, \cdots, m \quad (4)$$

where m denotes the number of the spectrum in the prediction set, and y^i_{pred} and y^i_{meas} are the predicted and measured concentrations of the component in the ith sample in the prediction set, respectively.

The detail procedures that are used for background removal from the NIR spectra and the concentration prediction can be summarized as the following steps:

(1) Select a wavelet function with n vanishing moments and set an initial value for the scale parameter.

(2) Perform convolution with the wavelet function and the scale parameter on all the spectra. The boundary effect that encountered in the convolution was coped with the method of symmetric extension.[27]

(3) Divide the transformed spectra randomly into two sets of the same size, the calibration set and the prediction set. By using the PLS2 algorithm, build the calibration model with the calibration set and predict the concentrations of the samples corresponding to the prediction set. Then, based on eq 4, calculate the RMSEP using the predicted and the measured concentrations.

(4) Repeat steps (1)−(3) to optimize the parameters with the steepest descents method by using RMSEP as the criterion.

3. EXPERIMENTAL SECTION

3.1. Simulated Data. Because the background is generally considered as a broader and slowly changing curve compared with the signal, it can be simulated either with the polynomial of low degree or broad Gaussian peak.[21,25] To test the applicability of the proposed method, the summation of a broad Gaussian peak and a polynomial was used in this study. On the other hand, the signal of interest was simulated with a narrow Gaussian peak. The spectrum was thus obtained through the summation of the three sections, narrow signal, broad background, and noise with normal distribution.

The Gaussian peak of both the background and the signal was obtained by

$$f(t) = h \exp\left[-4\ln(2)\left(\frac{t - t_0}{W_{1/2}}\right)^2\right] \quad (5)$$

where h, t_0, and $W_{1/2}$ are the peak high, the center position, and the width at half-height of the simulated peak, respectively. $W_{1/2}$ of the background is 10 times that of the signal, while its h ranges from 0.1 to 10 times that of the signal.

The polynomial section of the background was obtained by

$$b(t) = \sum_{i=0}^{n} k_i t^i \quad (6)$$

The highest degree of the polynomial, n, is set to 5, and the coefficient k_i is set randomly within a certain scope to satisfy that $b(t)$ ranges from 0.1 to 10 times the peak height of the signal.

3.2. Experimental NIR Data. Fifty tobacco lamina samples were measured on a Bruker Verctor 22/N FT-NIR System. The spectra were recorded in the wavenumber 4000−9200 cm^{-1} with the digitization interval 4 cm^{-1}. Four components, alkaloids (AS), total sugars (TS), total nitrogen compounds (TN), and total volatile alkali (TVA), of the samples were measured with an AutoAnalyzer III instrument (Pulse Instrumentation Ltd.) following the standard procedures.

4. RESULTS AND DISCUSSION

To illustrate how the CWT can be used to remove the background, two groups, both containing 20 spectra, were simulated. As shown in Figure 2, spectra in both groups consist of a narrow signal peak with $W_{1/2} = 30$ (region 800−1000) and a broad and highly varying background. In group (a), all component concentrations used for the simulated spectra are 1.0, though the peak height of the narrow peaks looks a little different because of the variation of the background. In group (b), concentrations are in the range randomly from 0.1 to 2.0.

Figure 3 shows the results of both groups transformed at scale parameter 20, with a wavelet function Symmlets No. 6, which has 6 vanishing moments. It can be seen from Figure 3 that the random background in the spectra has been completely eliminated, and only the information corresponding to the narrow signal peak remained. For group (a), all spectra turn into a single line regardless of the different background in the original spectra. The reason is that the background is well eliminated when convoluted with the

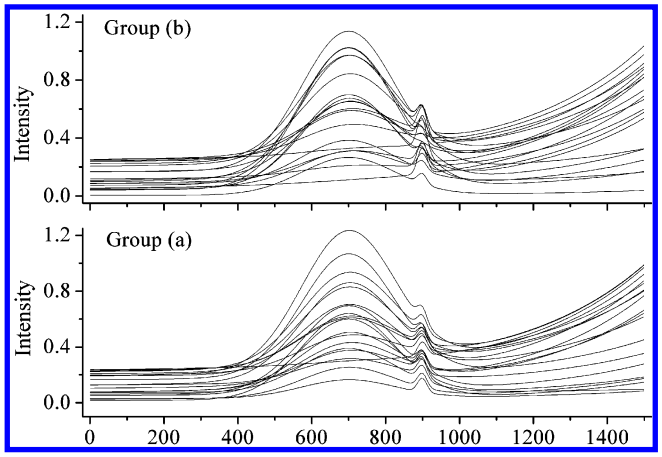

Figure 2. Simulated NIR spectra with fluctuating background. (a) Component concentrations are the same. (b) Component concentrations are different.

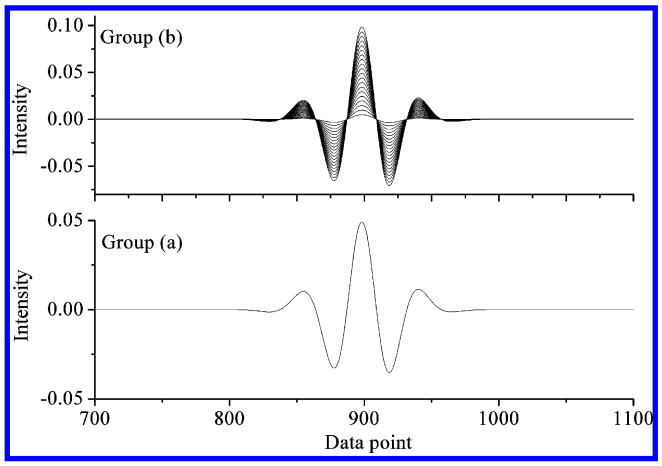

Figure 3. The transformed coefficients of the spectra with Symmlets No. 6. Group (a) and (b) are the results of (a) and (b) in Figure 2, respectively.

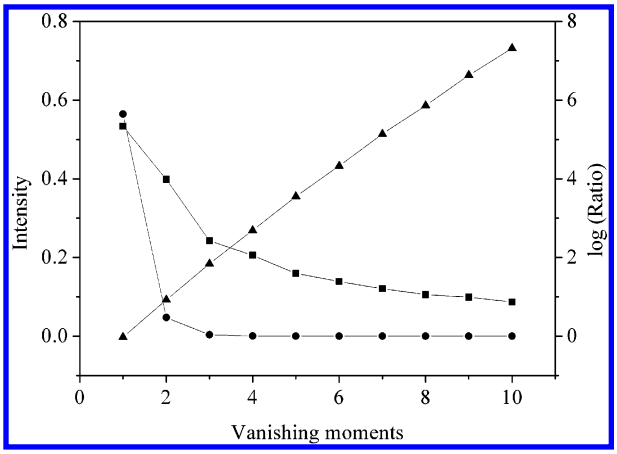

Figure 4. The intensity of the signal (■), background (●), and their ratio (▲) changing with the vanishing moments during the transform.

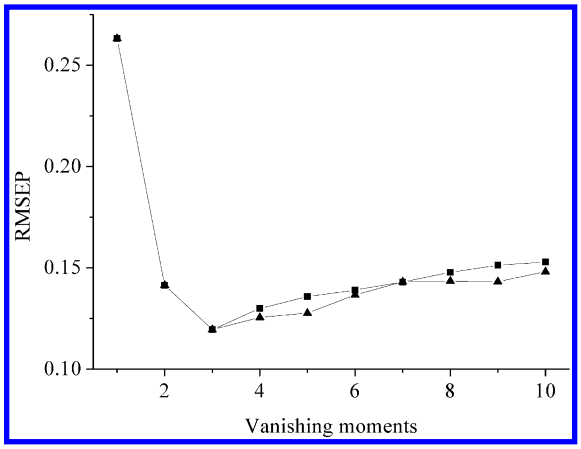

Figure 5. The trend of RMSEP of the prediction set changing with the vanishing moments. The line with symbol of ■ and ▲ are corresponding to the results obtained by Daubechies and Symmlets wavelet, respectively.

wavelet function, while the signal part still remains without much loss of information. For the same reason, all obtained spectra in group (b) have the same shape but different intensity. The correlation coefficient (R) between the intensities of the obtained coefficients and those of the original signals in the simulation was calculated. The result is $R = 1.0$, which demonstrates that the transform is a linear operation. Thus, the obtained coefficients should be eligible for calibration and regression.

4.1. Effect of the Vanishing Moments on the Result. To understand why the background can be eliminated through CWT and the important role of the vanishing moments, the transforms of the background and the signal were done separately. A broad and a narrow Gaussian peak are simulated, respectively, to represent the background and the signal part in a spectrum. Both the peak width and the peak height of the broad peak are 10 times the narrow one. The spectrum was transformed with wavelet Symmlets No. 1−10, which has 1 to 10 vanishing moments, respectively. The intensities of both peaks in the spectrum and the ratio of the narrow peak to the broad one were calculated.

Figure 4 shows the changing trend of the intensities of the two parts and their ratio. It can be seen that the intensities of both peaks decrease as the vanishing moments increase. It is also obvious that the intensity of the broad one drops more rapidly. As a result, the ratio of the narrow peak to the broad one increases dramatically with the vanishing moments. With certain vanishing moments when the ratio is large enough, the intensity of the background, compared to that of the signal, is so small that its interference is negligible.

One might expect that the higher the vanishing moments, the lower the interference from the background, and thus the better the result of background elimination would be. This is true only for the ideal spectra without noise. For the real spectra, however, the interference from the noise must be taken into account. When the vanishing moments increase to a certain value, the interference of the noise may become a significant factor. To give more of an explanation on this point, a data set consisting of 50 spectra as the group (b) in the above section was simulated, and 5.0% noise was added. Twenty-five spectra were randomly selected for the calibration set, while the other 25 spectra were selected for the prediction set. The spectra were transformed with the wavelet function of Symmlets and Daubechies that has 1−10 vanishing moments, respectively. The scale parameter is set as 10 in both transforms. The RMSEP was calculated and shown in Figure 5.

It can be seen that the trend and values of RMSEP obtained with Symmlets and Daubechies wavelet are very similar, which means that the vanishing moments but not the wavelet family is the key factor that determines the effects of the

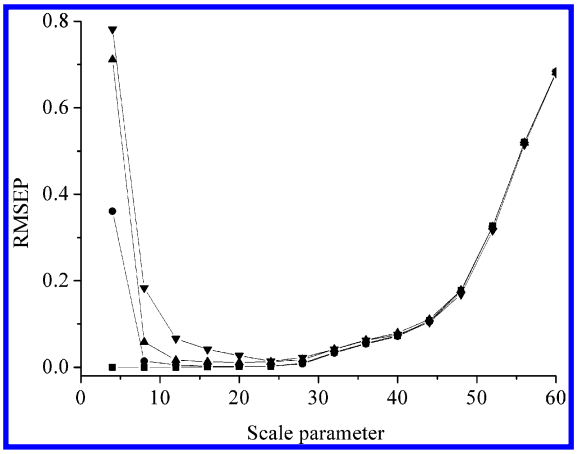

Figure 6. The trend of RMSEP of the prediction set changing with scale parameter. The lines with the symbols ■, ●, ▲, and ▼ correspond to the results of the spectra with a noise level of 0.0%, 1.0%, 3.0%, and 5.0%, respectively.

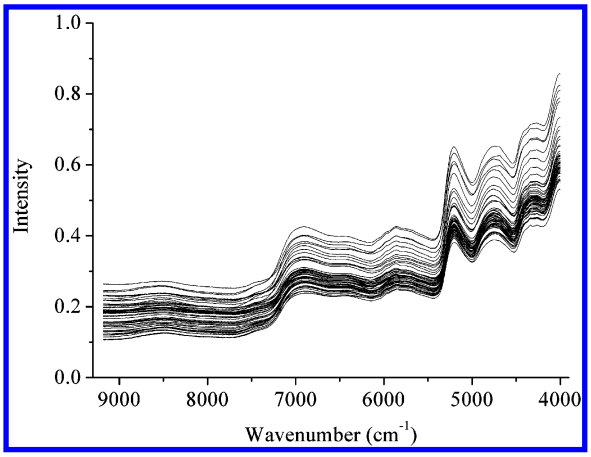

Figure 7. Near-infrared spectra of 50 tobacco samples.

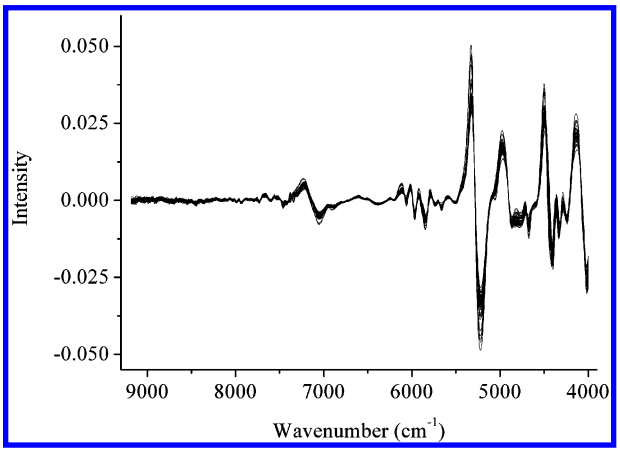

Figure 8. The coefficients of the spectra of the tobacco samples transformed with Symmlets No. 2 at scale parameter 20.

background elimination. Therefore, in the following sections, only the Symmlets wavelet functions with different vanishing moments were used. In addition, instead of the higher the vanishing moments, the better the result will be, the best result was obtained with the wavelet that has 3 vanishing moments. The recoveries of the prediction set at the optimal vanishing moments were also calculated, which range from 96.1 to 103.5%.

4.2. Effect of the Scale Parameter on the Result. As mentioned above, the effect of noise on the result cannot be neglected, especially when a wavelet with large vanishing moments is used for the transform. One way to overcome the problem is to introduce a smoothing process. From the theory of continuous wavelet transform, the effect of smoothing can be controlled by the scale parameter, which acts as the size of the smoothing window in the transform. Although the smoothing can suppress the noise, it might also cause the loss of the signal information if the smoothing window is too large. To show the influence of a scale parameter on the result, the same data set as in the above section was used and transformed with Symmlets No. 4 and a series of scale parameters. The RMSEP was calculated and shown in Figure 6.

It can be seen that, when the noise level is 0.0%, the RMSEP increases with the scale parameter due to the loss of the signal information. Different trends are obtained when the noise level is not zero. Each line has a different lowest RMSEP, which represents the optimal scale parameter. For the noise level at 1.0%, 3.0%, and 5.0%, the optimal scale parameters are 16, 20, and 24, respectively, and the corresponding RMSEP are 0.0024, 0.014, and 0.030, respectively. When the scale parameter is larger than 40, the RMSEP does not relate to the noise level any more. The possible reason is that the noise has been completely suppressed at that point and afterward.

4.3. Selection of the Optimal Parameters. It can be concluded from the above discussions that the broad background can be eliminated through CWT. The higher the vanishing moments of the wavelet function used, the more cleanly the background will be eliminated. However, the best vanishing moments depend on the noise level of the spectra. For the experimental signal, the smoothing process is required to suppress the noise. The larger the scale parameter, the better the smoothing will be. On the other hand, the scale parameter has its limitations, because smoothing will result in the loss of signal information. Therefore, the optimal vanishing moments and the scale parameter closely depend on the properties of the signal itself. In practice, the selection of the best parameters can be based on a certain criterion, e.g., RMSEP, recoveries, or the correlation coefficient of the calibration model, etc. For the purpose of a different task, the criterion might also be different. In this paper, RMSEP is used as the criterion to search for the optimal parameters. Based on the criterion, the data set in the above section with 5.0% noise was analyzed, and the optimal vanishing moments and the scale parameter are 4 and 24, respectively. The RMSEP is 0.019, and the recovery range is from 98.3 to 101.8%.

4.4. An Application to Experimental NIR Spectra. To further test the performance of the proposed method, the NIR spectra of 50 tobacco samples as shown in Figure 7 were investigated. The PLS2 algorithm was adopted to simultaneously determine four ingredients, AS, TS, TN, and TVA, in the samples. Twenty-five spectra were randomly selected for the calibration set and the other 25 for prediction. The latent variables were empirically set at 5. By using the RMSEP as the criterion, the best result was obtained with a wavelet function Symmlets No. 2 and a scale parameter of 20.

Figure 8 shows the coefficients obtained through the CWT with the optimal parameters. It can be seen that the fluctuating background of the spectra was successfully

Table 1. Results of the Calibration with Different Pretreatment Methods

	AS		TS		TN		TVA	
method	RMSEP	R	RMSEP	R	RMSEP	R	RMSEP	R
CWT	0.194	0.985	1.321	0.950	0.186	0.956	0.038	0.986
DWT	0.276	0.967	1.273	0.950	0.201	0.951	0.050	0.975
derivative	0.219	0.982	1.400	0.945	0.198	0.949	0.043	0.982
raw data	0.577	0.798	1.943	0.910	0.290	0.818	0.120	0.851

removed. The correlation coefficient (R) between the measured value obtained by the chemical method and the predicted value by PLS2 was calculated and shown in Table 1 (line 1).

For comparison, DWT and the derivative method of Savitzky-Golay (DSG) were also used for preprocessing the spectra. Parameters in both methods were optimized with the same procedure and criterion as in CWT. In the DWT method, the technique of symmetric extension as in CWT was used to treat the boundary effect, and the optimal parameter, wavelet Symmlets No. 2, and the coefficients at level 4 were used. In the derivative method of Savitzky-Golay, the window length and the polynomial order are 21 and 3, respectively. The results obtained by DWT (line 2), DSG (line 3), and the calibration directly on the raw data (line 4) were listed in the table. It can be seen that the results of the calibration were significantly improved with all three pretreatment methods. With lower RMSEP and higher correlation coefficients, CWT produced a slightly better result than DWT and DSG, since it is a more precise method. In addition, to obtaining the best result, only two parameters need to be optimized; CWT is simpler to perform than the other two methods.

5. CONCLUSION

Continuous wavelet transform was adopted as a pretreatment tool for the NIR spectra. A fluctuating background can be eliminated due to the property of the vanishing moments of the wavelet. On the other hand, the noise can also be removed in the transform by choosing an appropriate scale parameter. Results on both the simulated and real spectra proved that CWT can be used as a convenient tool for background elimination. Compared with other methods, it shows several advantages. First, wavelet transform, including DWT and CWT, is an operation that removes the background and noise simultaneously. In addition, it is comparatively simple to perform, because only the selections of the vanishing moments of the wavelet and the scale parameter are involved. Since the scale parameter in the transform can be adjusted continuously, it is a finer method than DWT. Furthermore, the proposed method can also be applied to other analytical signals, such as UV, FTIR, and NMR, etc. It may become a competitive tool for spectral analysis.

ACKNOWLEDGMENT

This study is supported by the Teaching and Research Award Program for Outstanding Young Teachers (TRAPOYT) in higher education institutions of Ministry of Education (MOE), P. R. China.

REFERENCES AND NOTES

(1) Garcia-Alvarez, M.; Huidobro, J. F.; Hermida, M.; Rodriguez-Otero, J. L. Major Components of Honey Analysis by Near-infrared Transflectance Spectroscopy. *J. Agric. Food Chem.* **2000**, *48*, 5154−5158.
(2) Fontaine, J.; Schirmer, B.; Horr, J. Near-infrared Reflectance Spectroscopy (NIRS) Enables the Fast and Accurate Prediction of Essential Amino Acid Contents. 2. Results for Wheat, Barley, Corn, Triticale, Wheat Bran/Middlings, Rice Bran, and Sorghum. *J. Agric. Food Chem.* **2002**, *50*, 3902−3911.
(3) Radtke, G.; Knop, K.; Lippold, B. C. Near Infrared (NIR) Spectroscopy: Fundamentals and Application from a Pharmaceutical Point. *Pharm. Ind.* **1999**, *61*, 848−857.
(4) Laasonen, M.; Harmia-Pulkkinen, T.; Simard, C.; Rasanen, M.; Vuorela, H. Development and Validation of a Near-Infrared Method for the Quantitation of Caffeine in Intact Single Tablets. *Anal. Chem.* **2003**, *75*, 754−760.
(5) Laasonen, M.; Harmia-Pulkkinen, T.; Simard, C. L.; Michiels, E.; Rasanen, M.; Vuorela, H. Fast Identification of Echinacea Purpurea Dried Roots Using Near-Infrared Spectroscopy. *Anal. Chem.* **2002**, *74*, 2493−2499.
(6) Riley, M. R.; Crider, H. M.; Nite, M. E.; Garcia, R. A.; Woo, J.; Wegge, R. M. Simultaneous Measurement of 19 Components in Serum-Containing Animal Cell Culture Media by Fourier Transform Near-Infrared Spectroscopy. *Biotechnol. Prog.* **2001**, *17*, 376−378.
(7) Wu, Y. Q.; Czarnik-Matusewicz, B.; Murayama, K.; Ozaki, Y. Two-Dimensional Near-Infrared Spectroscopy Study of Human Serum Albumin in Aqueous Solutions: Using Overtones and Combination Modes to Monitor Temperature-Dependent Changes in The Secondary Structure. *J. Phys. Chem. B* **2000**, *104*, 5840−5847.
(8) Karstang, T. V.; Kvalheim, O. M. Multivariate Prediction and Background Correction Using Local Modeling and Derivative Spectroscopy. *Anal. Chem.* **1991**, *63*, 767−772.
(9) Lau, O. W.; Luk, S. F.; Cheng, O. M.; Chiu, T. P. Y. Background-Correction Methods for the Determination of Caffeine in Beverages, Coffee and Tea by Using 2nd-Derivative Ultraviolet Spectrophotometry. *Analyst* **1992**, *117*, 777−783.
(10) Paradkar, R. P.; Williams, R. R. Genetic Regression as a Calibration Technique for Solid-Phase Extraction of Dithizone-Metal Chelates. *Appl. Spectrosc.* **1996**, *50*, 753−758.
(11) Paradkar, R. P.; Williams, R. R. Correcting Fluctuating Baselines and Spectral Overlap with Genetic Regression. *Appl. Spectrosc.* **1997**, *51*, 92−100.
(12) Sjoblom, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S. An Evaluation of Orthogonal Signal Correction Applied to Calibration Transfer of Near Infrared Spectra. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 229−244.
(13) Svensson, O.; Kourti, T.; MacGregor, J. F. An Investigation of Orthogonal Signal Correction Algorithms and Their Characteristics. *J. Chemom.* **2002**, *16*, 176−188.
(14) Shao, X. G.; Leung, A. K. M.; Cau, F. T. Wavelet: A New Trend in Chemistry. *Acc. Chem. Res.* **2003**, *36*, 276−283.
(15) Jetter, K.; Depczynski, U.; Molt, K.; Niemöller, A. Principles and Applications of Wavelet Transformation of Chemometrics. *Anal. Chim. Acta* **2000**, *420*, 169−180.
(16) Walczak, B.; Massart, D. L. Wavelets − Something for Analytical Chemistry? *TrAC-Trends Anal. Chem.* **1997**, *16*, 451−463.
(17) Shao, X. G.; Pang, C. Y.; Su, Q. D. A Novel Method to Calculate the Approximate Derivative Photoacoustic Spectrum Using Continuous Wavelet Transform. *Fresen. J. Anal. Chem.* **2000**, *367*, 525−529.
(18) Dinc, E.; Baleanu, D.; Ustundag, O. An Approach to Quantitative Two-Component Analysis of a Mixture Containing Hydrochlorothiazide and Spironolactone in Tablets by One-Dimensional Continuous Daubechies and Biorthogonal Wavelet Analysis of UV−Spectra. *Spectrosc. Lett.* **2003**, *36*, 341−355.
(19) Dinc, E.; Baleanu, D. Multideterminational of Thiamine HCl and Pyridoxine Hcl In Their Mixture Using Continuous Daubechies And Biorthogonal Wavelet Analysis. *Talanta* **2003**, *59*, 707−717.
(20) Tan, H. W.; Brown, S. D. Wavelet Analysis Applied to Removing Nonconstant, Varying Spectroscopic Background in Multivariate Calibration. *J. Chemometr.* **2002**, *16*, 228−240.
(21) Mittermayr, C. R.; Tan, H. W.; Brown, S. D. Robust Calibration with Respect to Background Variation. *Appl. Spectrosc.* **2001**, *55*, 827−833.
(22) Grossmann, A.; Morlet, J. Decomposition of Hardy Function into Square Integrable Wavelets of Constant Shape. *SIAM. J. Math. Anal.* **1984**, *15*, 723−736.
(23) Daubechies, I. *Ten Lectures on Wavelets*; SIAM Press: Philadelphia, 1992.
(24) Mallat, S.; Hwang, W. L. Singularity Detection and Processing with Wavelets *IEEE. Trans. Inform. Theory* **1992**, *38*, 617−643.
(25) I. Schechter, Correction for Nonlinear Fluctuating Background in Monovariable Analytical Systems. *Anal. Chem.* **1995**, *67*, 2580−2585.
(26) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(27) Mallat, S. G. *A Wavelet Tour of Signal Processing*; Academic Press: New York, New York, 1998.

CI034211+