Classification of Lactose and Mandelic Acid THz Spectra using Subspace and Wavelet-Packet Algorithms.

Xiaoxia Yin[1]*[a], Sillas Hadjiloucas[c], Bernd M. Fischer[a], Brian W.-H Ng[a],
Henrique M. Paiva[b], Roberto K.H. Galvão[b],
Gillian C. Walker[c], and John W. Bowen[c], Derek Abbott[a]

[a]Center for Biomedical Engineering and School of Electrical & Electronic Engineering The University of Adelaide, SA 5005, Australia;
[b]Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, 12228-900, Brazil;
[c]Cybernetics, School of Systems Engineering, The University of Reading, Whiteknights, Reading, RG6 6AY, UK.

ABSTRACT

This work compares classification results of lactose, mandelic acid and dl-mandelic acid, obtained on the basis of their respective THz transients. The performance of three different pre-processing algorithms applied to the time-domain signatures obtained using a THz-transient spectrometer are contrasted by evaluating the classifier performance. A range of amplitudes of zero-mean white Gaussian noise are used to artificially degrade the signal-to-noise ratio of the time-domain signatures to generate the data sets that are presented to the classifier for both learning and validation purposes. This gradual degradation of interferograms by increasing the noise level is equivalent to performing measurements assuming a reduced integration time. Three signal processing algorithms were adopted for the evaluation of the complex insertion loss function of the samples under study; a) standard evaluation by ratioing the sample with the background spectra, b) a subspace identification algorithm and c) a novel wavelet-packet identification procedure. Within class and between class dispersion metrics are adopted for the three data sets. A discrimination metric evaluates how well the three classes can be distinguished within the frequency range 0.1 – 1.0 THz using the above algorithms.

Keywords: lactose, mandelic acid, THz-transient spectroscopy, far-infrared, wavelet sub-bands, system identification, signal classification

1. INTRODUCTION

The terahertz (THz) part of the electromagnetic spectrum lying between the microwave and infrared frequencies (100 GHz -30 THz) is of significant importance to the biological sciences because complementary information to traditional spectroscopic measurements on low-frequency bond vibrations, hydrogen bond stretches and torsions may be obtained. The vibrational spectral characteristics of bio-molecules which lie in this range (wavenumbers between 33-1,000 cm^{-1}) make T-rays a promising sensing modality in future clinical diagnosis. Recent advances in T-ray sources and detectors have made it possible to image and discriminate opaque objects such as tumors[1] from normal tissue on the basis of refractive index variation. While much effort has been devoted to improving the signal to noise ratio and repeatability of measurements as well as reliability in the function of the spectrometers, the further processing of THz transients has received less attention in the literature. T-ray classification relies in observing changes in pulse amplitude, phase as well as dispersion characteristics of the tissue under study. System identification techniques[2-4] have shown that more compact parametrisation of the time domain signals can improve on the signal to noise ratio of the calculated spectra and are useful for classification purposes.

*xxyin@eleceng.adelaide.edu.au; phone +61-8-8303-5748; fax +61-8-8303-4360;

Microelectronics: Design, Technology, and Packaging III, edited by Alex J. Hariz, Vijay K. Varadan
Proc. of SPIE Vol. 6798, 679814, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.755154

The non-stationary nature of time-domain pulses obtained in T-ray spectrometry justifies their decomposition in the wavelet domain. Furthermore, compared to Fourier-based techniques, a wavelet decomposition of the experimental signal can provide better time-frequency localization characteristics facilitating subsequent classification tasks[5,6]. More recently[7], a novel technique involving the use of Auto Regressive (AR) and Auto Regressive Moving Average (ARMA) models on the wavelet transforms of measured T-ray pulse data was also presented. Wavelet-based de-noising with soft threshold shrinkage was employed to the measured T-ray signals prior to modelling.

The current work proposes the use of a novel system identification scheme implemented in the wavelet domain and contrasts its ability to extract the important features in the signal with that of the N4SID subspace identification method. The goal of this work is to demonstrate efficient and robust classification algorithms that could be adopted by the biomedical and pharmaceutical communities[9] which are envisaged to provide the technology pull required for the further proliferation of THz-transient spectrometers.

2. SYSTEM IDENTIFICATION IN THE WAVELET DOMAIN

Defining the background and sample interferograms as the input and output signals, the frequency response of an identified model would be an estimate of the complex insertion loss (CIL). A wavelet-packet formulation illustrated in Fig. 1 is adopted and sub-band models $M_{i,j}(z)$ are identified from the sample and background interferograms by following a least-squares procedure as indicated in Fig. 2.

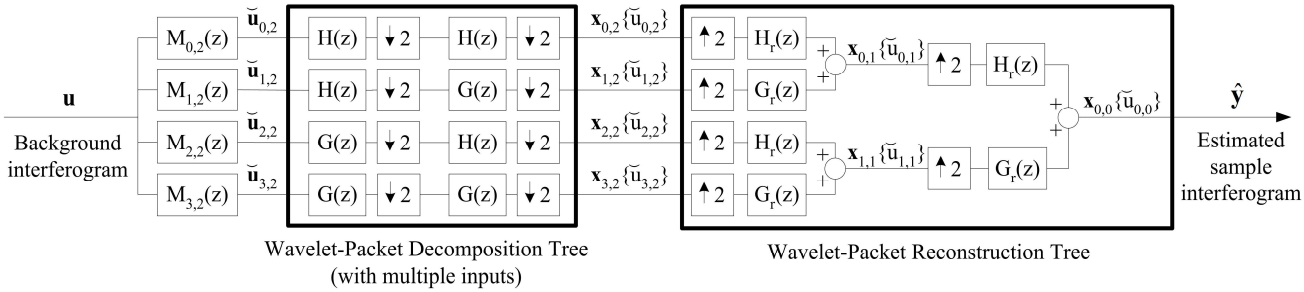

Fig. 1. Wavelet-packet model structure. In this example, a complete two-level decomposition tree, which defines four frequency sub-bands, is employed. H(z), G(z) denote low-pass and high-pass decomposition filters, respectively, with reconstruction counterparts represented by $H_r(z)$, $G_r(z)$. The four sub-band models are represented by the transfer functions $M_{0,2}(z)$, $M_{1,2}(z)$,$M_{2,2}(z)$, $M_{3,2}(z)$.

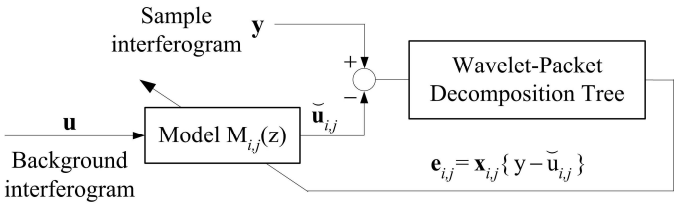

Fig. 2. Model identification of a sample interferogram for a given frequency sub-band.

Fig. 2 illustrates the procedure adopted to identify each sub-band model $M_{i,j}$. **u** is the input signal used for identification. **y** and $\breve{\mathbf{u}}_{i,j}$ are the plant and sub-band model outputs, respectively. Residue $e_{i,j}$ denotes the wavelet-packet coefficients of the difference between **y** and $\breve{\mathbf{u}}_{i,j}$, in the frequency band under consideration. The structure adopted for the subband model is a transfer function of the form:

$$M_{i,j} = P_{i,j}(z)Q_{i,j}(z) \tag{1}$$

where:

$$P_{i,j}(z) = \left(\frac{1}{1-z^{-1}}\right)^{s_{i,j}},\ s_{i,j} \in \mathrm{Z} \tag{2a}$$

$$Q_{i,j}(z) = \alpha_{i,j} + \beta_{i,j} z^{-1},\ \alpha_{i,j}, \beta_{i,j} \in \Re \tag{2b}$$

$P_{i,j}(z)$ is aimed at roughly approximating the band-limited frequency response of the plant, whereas the Finite Impulse Response (FIR) term $Q_{i,j}(z)$ provides a fine-tuning for the approximation. A least-squares adjustment for the parameters of $M_{i,j}$ can be carried out by minimizing the following cost function $J_{i,j} : \mathrm{Z} \times \Re^2 \to \Re$:

$$J_{i,j}(s_{i,j}, \alpha_{i,j}, \beta_{i,j}) = \mathrm{e}_{i,j}(\mathrm{e}_{i,j})^{\mathrm{T}} \tag{3}$$

where $\mathrm{e}_{i,j}$ denotes the row vector of residues for the identification data, as shown in Fig. 2. As discussed elsewhere[8], if $s_{i,j}$ is fixed, the optimal real-valued parameters $\alpha^*_{i,j}$ and $\beta^*_{i,j}$ are obtained by imposing:

$$\frac{\partial J_{i,j}}{\partial \alpha_{i,j}} = \frac{\partial J_{i,j}}{\partial \beta_{i,j}} = 0 \tag{4}$$

To find the optimal value of $s_{i,j}$ the following search algorithm in Z is used: the value of $s_{i,j}$ is varied in a specified range. For each value of $s_{i,j}$, the optimal values $\alpha^*_{i,j}$ and $\beta^*_{i,j}$ are calculated using

$$\begin{bmatrix} \alpha^*_{i,j} \\ \beta^*_{i,j} \end{bmatrix} = M^{-1} \begin{bmatrix} \mathrm{x}_{i,j}\{y\}\left(\mathrm{x}_{i,j}\{u^{p}_{i,j}\}\right)^{\mathrm{T}} \\ \mathrm{x}_{i,j}\{y\}\left(\mathrm{x}_{i,j}\{u^{pd}_{i,j}\}\right)^{\mathrm{T}} \end{bmatrix} \tag{5}$$

with

$$M = \begin{bmatrix} \mathrm{x}_{i,j}\{u^{p}_{i,j}\}\left(\mathrm{x}_{i,j}\{u^{p}_{i,j}\}\right)^{\mathrm{T}} & \mathrm{x}_{i,j}\{u^{pd}_{i,j}\}\left(\mathrm{x}_{i,j}\{u^{p}_{i,j}\}\right)^{\mathrm{T}} \\ \mathrm{x}_{i,j}\{u^{p}_{i,j}\}\left(\mathrm{x}_{i,j}\{u^{pd}_{i,j}\}\right)^{\mathrm{T}} & \mathrm{x}_{i,j}\{u^{pd}_{i,j}\}\left(\mathrm{x}_{i,j}\{u^{pd}_{i,j}\}\right)^{\mathrm{T}} \end{bmatrix} \tag{6}$$

provided M^{-1} exists. In the above equations, $\mathrm{u}^{pd}_{i,j}$ denotes the value of $\mathrm{u}^{p}_{i,j}$ which is the output term of $P_{i,j}(z)$ delayed by one sample. The value $s_{i,j}$ for which $J_{i,j}$ is minimum is then adopted, as well as the corresponding values of $\alpha^*_{i,j}$ and $\beta^*_{i,j}$.

The structure of the wavelet decomposition tree can be optimized in order to achieve a compromise between the parsimony and accuracy of the overall model. For this purpose, a generalized cross-validation procedure can be employed to determine whether the reduction in identification error is large enough to justify the further decomposition of any given tree node[8]. This approach is also adopted in the present paper. In what follows we present results of the calculated complex insertion loss function using a) standard FFT-based procedures, b) subspace identification and c) the wavelet-packet identification procedure.

3. EVALUATION OF COMLEX INSERTION LOSS, SUBSPACE AND WAVELET PACKET IDENTIFICATION

Time domain interferograms of Lactose, Mandelic Acid and DL Mandelic acid were recorded using a THz-transient spectrometer. Typical background and sample signatures are shown in Fig. 3. Linear detrending of the co-averaged experimental data sets using the detrend.m routine in MATLAB is also shown.

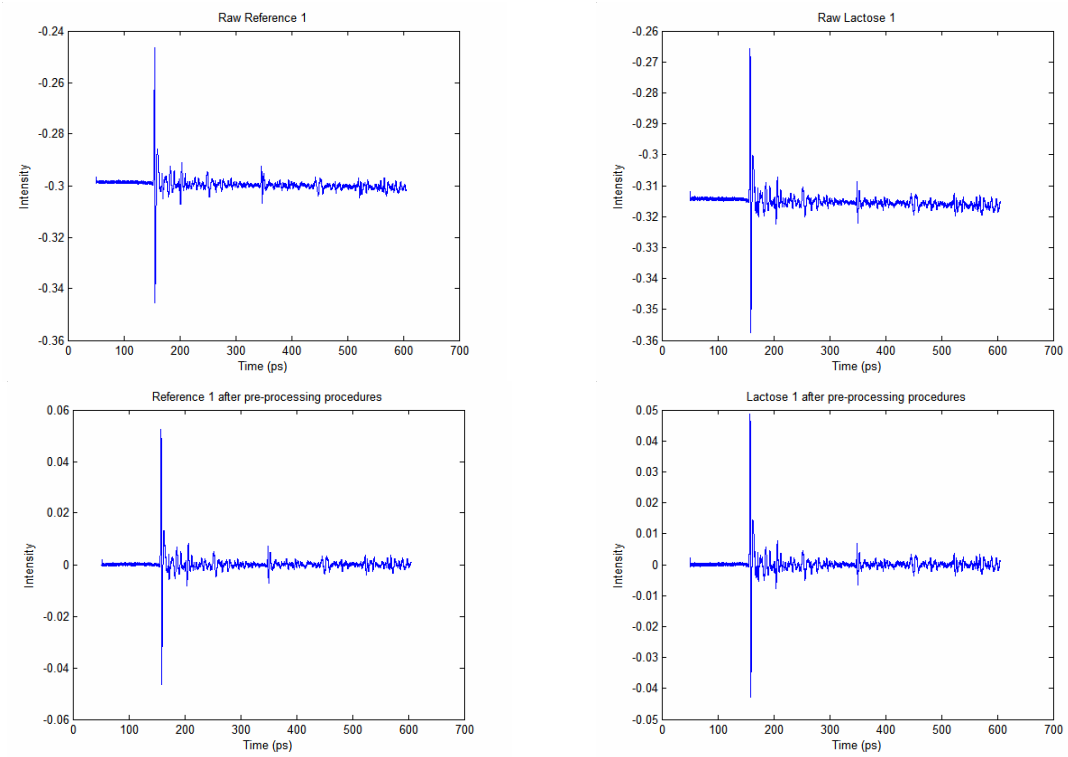

Fig. 3. Background (reference) and sample (lactose) interferograms before and after the pre-processing procedures (detrending and alignment).

After the pre-processing procedures, the background and sample interferograms were employed as input **u** and output **y** signals, respectively. The singular value plot generated in the subspace identification procedure is presented in Fig. 4. Following the default recommendation of the n4sid function, a 4th order model was adopted. It is worth noting that when using MATLAB's function n4sid.m, different results are obtained by pre-establishing an order of 4 or by making such a choice after testing orders 1 to 20. The results presented in this work were obtained by testing orders 1 to 20. The resulting CIL, which corresponds to the frequency response of the 4th order model, is presented in Fig. 4b in the spectral range 0.1 – 1.0 THz. The result obtained by ratioing the sample spectrum against the background spectrum is also shown. As can be seen, by using the subspace algorithm, the estimated position of the absorption band is slightly biased towards higher frequencies, and its magnitude seems to be over-estimated.

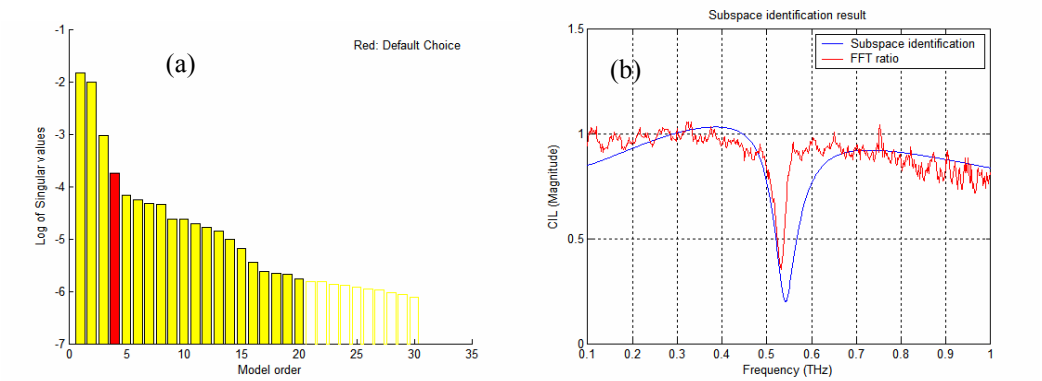

Fig. 4. a) Singular value plot for the subspace identification and b) calculated magnitude of the complex insertion loss for the lactose sample as a function of frequency obtained by ratioing the sample spectrum against the background spectrum (red line) and by subspace identification (blue line).

Prior to the wavelet-packet identification procedure, a 6th order Butterworth band-pass filter was employed to band-limit the interferograms to the 100GHz – 1.0 THz range. The interferograms were then re-sampled in order to reduce the

sampling frequency by a factor of 8. This procedure was employed to reduce the number of wavelet decomposition levels required to attain an appropriate frequency resolution. The settings for the wavelet packet identification procedure used a db12 wavelet with a maximum tree depth of 9 decomposition levels (including the root node); values tested for the *s* parameter (exponent of the integrator term in the sub-band models) were –1, 0, +1.

Fig. 5a presents the resulting wavelet-packet tree obtained by the generalized cross-validation procedure. The tree is deeper in a particular frequency range, which actually corresponds to the absorption valley, as shown in Fig. 5b (the segmentation is more refined in the frequency region corresponding to deeper levels of the tree). It is worth noting that the tree structure was automatically defined by the identification algorithm, with no prior knowledge of the spectral features of the sample under consideration.

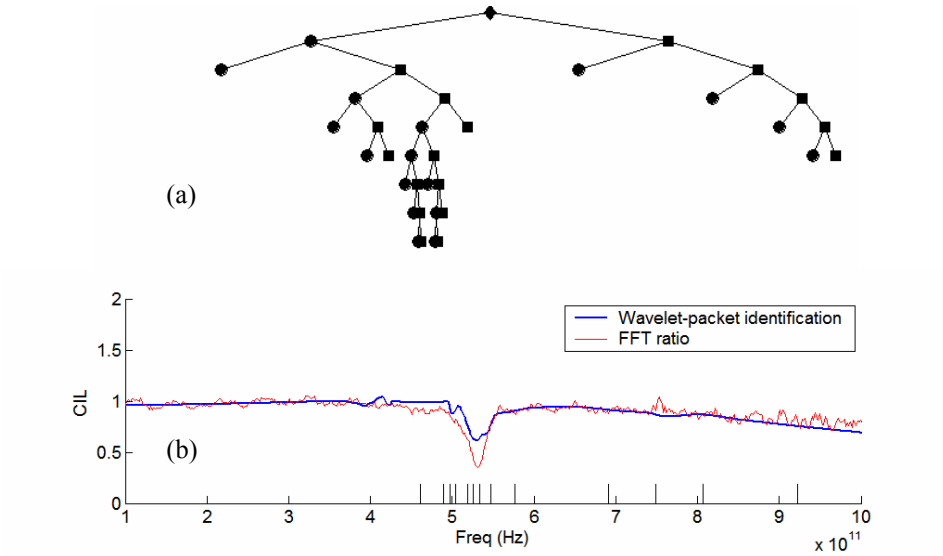

Fig. 5. a) Resulting wavelet-packet tree and b) CIL for a lactose sample obtained by wavelet-packet identification (blue line). The FFT ratio result (red line) is also presented for comparison. The frequency-domain segmentation automatically defined in the identification procedure is indicated by vertical lines at the bottom of the graph. As can be seen, the segmentation is more refined in the spectral region corresponding to the absorption band.

4. SIGNAL PROCESSING ASSUMING NOISY BACKGROUND AND SAMPLE INTERFEROGRAMS

The standard deviation of the noise (white, zero-mean Gaussian) was varied from 10^{-4} to 10^{-3} to evaluate the discrimination metric (described in section 5 below) for different signal-to-noise ratios. Ten noisy sample/background interferogram pairs were generated for each species (Lactose, Mandelic acid, DL Mandelic acid). Therefore, an overall set of 30 complex insertion loss (CIL) functions were calculated for each noise level and for each processing technique (FFT, subspace, wavelet-packet). Each of these calculated CIL functions will be termed an "object" in this study. As part of the pre-processing procedure, the time-domain interferograms were aligned with respect to each other. Figure 6 compares the noisy interferograms (noise standard deviation of 10^{-3}) before and after the pre-processing procedures. Furthermore, an asymmetric (Mertz) triangular apodization window was used for the FFT calculations.

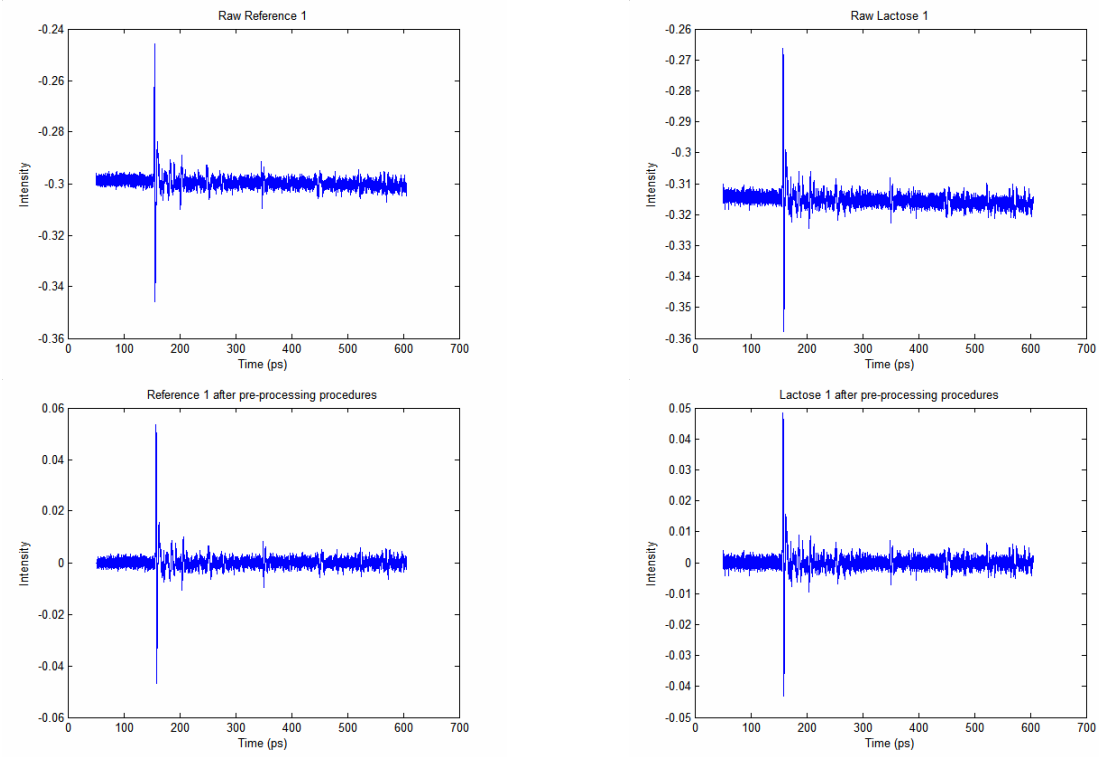

Fig. 6. Noisy background and sample interferograms before and after the pre-processing procedures.

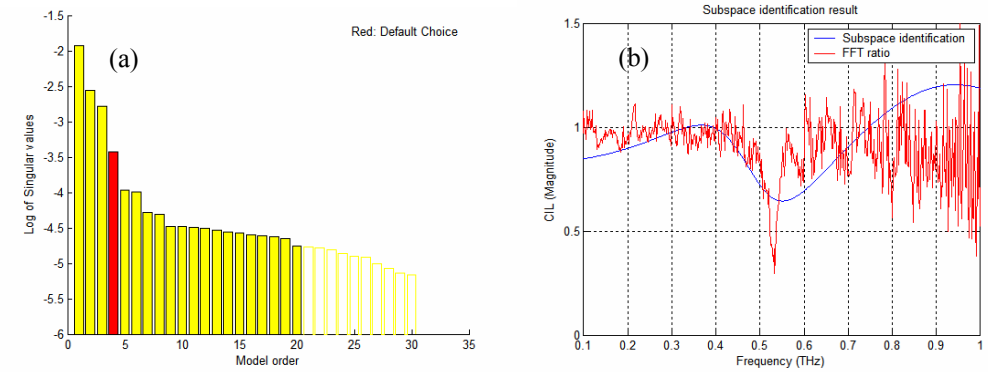

Fig. 7. a) Singular value plot for the subspace identification after the inclusion of artificial noise and b) the corresponding calculated magnitude of the complex insertion loss for the lactose sample as a function of frequency obtained by ratioing the sample spectrum against the background spectrum and by subspace identification (4^{th} order model).

The singular value plot generated in the subspace identification procedure is presented in Fig. 7a. A 4^{th} order model was adopted, as recommended by the n4sid function. The resulting CIL, which corresponds to the frequency response of the 4^{th} order model is presented in Fig. 7b. As can be seen, the identification result is very sensitive to the additional noise present in the time domain signatures.

Fig. 8 presents the resulting wavelet-packet tree obtained after the inclusion of artificial noise. As can be seen, the tree has much fewer nodes as compared to the tree obtained in the previous case (Fig. 5). Such a result was obtained because the generalized cross-validation procedure tends to generate more parsimonious models (i.e. tends to group frequency segments together in the identification procedure) when the signal-to-noise ratio is worse. Again, it is worth noting that the segmentation in the frequency domain is established in an automatic manner, and no prior knowledge of

the signal-to-noise ratio is required. This result is more clearly demonstrated in Fig. 9, which presents trees obtained for different realizations of noise with standard deviations of 10^{-3} and 10^{-4}. The structure of nodes corresponding to the absorption feature is always present, but the increase in the noise level leads to the pruning of other parts of the tree.

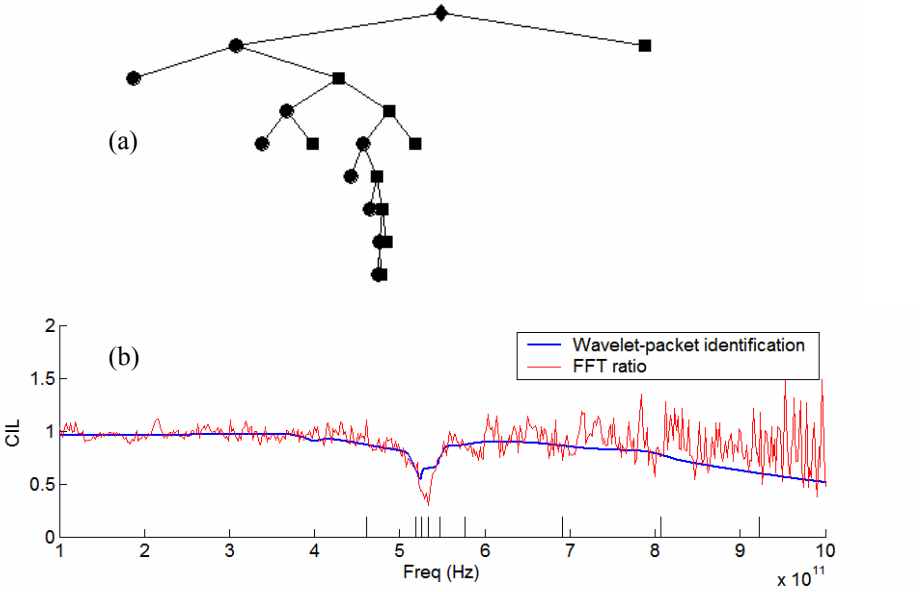

Fig. 8. a) Resulting wavelet-packet tree after the inclusion of artificial noise and b) CIL for the lactose sample obtained by wavelet-packet identification after the inclusion of artificial noise (blue line). The CIL result calculated using the ratio of sample and background FFTs is also presented for comparison (red line). The frequency-domain segmentation automatically defined in the identification procedure is indicated by vertical lines at the bottom of the graph. As can be seen, fewer frequency segments were employed, compared to the results in Fig 5.

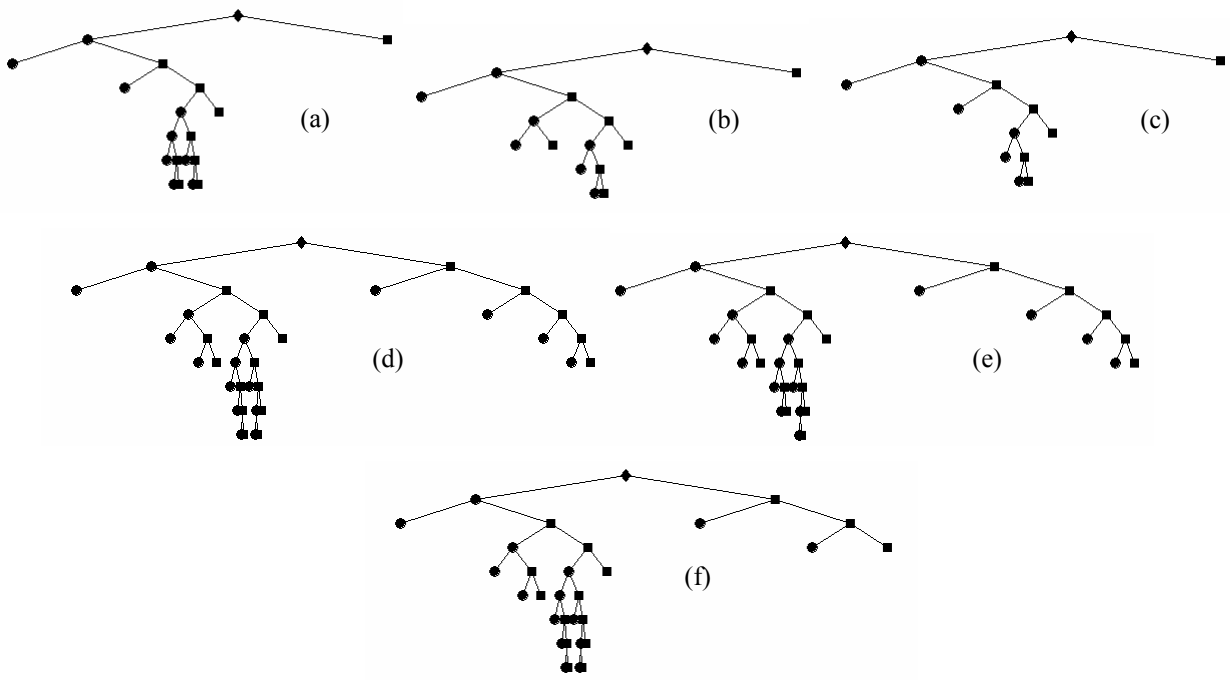

Fig. 9. Trees obtained with different realizations of noise with standard deviation of 10^{-3} (a, b, c) and 10^{-4} (d, e, f).

5. DESCRIPTION OF THE DISCRIMINATION METRIC

In what follows, classes 1, 2, and 3 will refer to the objects corresponding to lactose, mandelic acid, and dl-mandelic acid, respectively. For each noise level and for each processing technique (FFT, subspace, wavelet-packet), a discrimination metric was calculated on the basis of the estimated CIL magnitude in the range 0.1 – 1.0 THz. To do so, we let $x_{i,n}$ be the CIL magnitude of the i^{th} object (i = 1, …, 30) at the n^{th} spectral bin, and assume 500 spectral bins uniformly distributed in the range 0.1 – 1.0 THz (that is, n = 1, …, 500). A row vector $\mathbf{x}_i$ is defined for each object by disposing the CIL magnitude values in the form:

$$\mathbf{x}_i = [x_{i,1}\, x_{i,2} \ldots x_{i,500}] \tag{7}$$

Let $\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$ denote the mean value of the objects belonging to classes 1, 2, and 3, respectively, that is:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{i\in I_1}\mathbf{x}_i \tag{8a}$$

$$\mathbf{m}_2 = \frac{1}{N_2}\sum_{i\in I_2}\mathbf{x}_i \tag{8b}$$

$$\mathbf{m}_3 = \frac{1}{N_3}\sum_{i\in I_3}\mathbf{x}_i \tag{8c}$$

where I_1, I_2, I_3 are the index sets of objects belonging to classes 1, 2, 3, respectively and $N_1 = N_2 = N_3 = 10$ are the number of objects in each class. A between-class dispersion metric D_B is calculated as:

$$D_B = \frac{1}{3}\sum_{j=1}^{3}\|\mathbf{m}_j - \mathbf{m}\|^2 \tag{9}$$

where

$$\mathbf{m} = \frac{1}{3}(\mathbf{m}_1 + \mathbf{m}_2 + \mathbf{m}_3) \tag{10}$$

A within-class dispersion metric is calculated for each class j as:

$$D_{W,j} = \frac{1}{N_j}\sum_{i\in I_j}\|\mathbf{x}_i - \mathbf{m}_j\|^2 \tag{11}$$

where j ranges from one to three. An overall within-class dispersion metric D_W is calculated as:

$$D_W = \frac{1}{3}(D_{W,1} + D_{W,2} + D_{W,3}) \tag{12}$$

Finally, the discrimination metric F, which evaluates how well the three classes can be distinguished within the frequency range under consideration, is defined as:

$$F = \frac{D_B}{D_W} \tag{13}$$

Fig. 10 presents a plot of the adopted discrimination metric F for the three techniques under consideration as a function of the level of noise added to the interferograms. According to this metric, the identification methods are seen to be more robust with respect to noise than the standard ratioing procedure. In particular, the proposed wavelet-packet identification technique becomes slightly superior to the subspace method at larger noise levels.

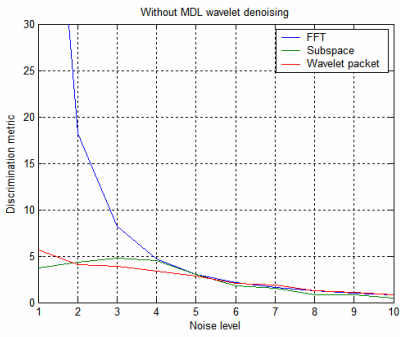

Fig. 10. Plots of the adopted discrimination metric F for different noise levels ($\times 10^{-4}$) superimposed in the THz-transient datasets.

6. CONCLUSION

A wavelet-packet identification scheme for discriminating between lactose, mandelic acid and dl-mandelic acid THz transients was proposed. After the subsequent evaluation of the complex insertion loss using the output of the model as opposed to the direct ratioing of the spectra, we observed that a ratio composed of the model output has smoothed out the calculated value of the complex insertion loss function across the frequencies range of interest and identified more correctly the absorption bands of the samples than the subspace algorithm. This was the case even when the time-domain signatures were corrupted by additional noise. Within class and between class dispersion discrimination metrics were adopted to evaluate the benefits of the proposed algorithm in classification tasks. The results were more robust with respect to noise than those obtained by the standard ratioing procedure, but the advantages of using the identification algorithm for small noise levels were unclear. It is possible that greater benefits would be obtained for more spectrally rich samples, as the wavelet-packet technique has been shown to be particularly suited to the identification of systems with several spectral resonance features[8].

REFERENCES

[1] T. Löffler, K. Siebert, S. Czasch, and H.G. Bauer, T. and Roskos, "Visualization and classification in biomedical terahertz pulsed imaging," Physics in Medicine and Biology, **47,** 3847-3852, (2002).

[2] S. Hadjiloucas, R. K. H. Galvão, V. M. Becerra, J. W. Bowen, R. Martini, M. Brucherseifer, H. P. M. Pellemans, P. Haring Bolívar, H. Kurz, J. M. Chamberlain, "Comparison of state space and ARX models of a waveguide's THz transient response after optimal wavelet filtering," IEEE Transactions on Microwave Theory and Techniques MTT, **52**(10), 2409-2419 (2004).

[3] D. M. Mittleman, R. H. Jacobsen, R. Neelamani, R. G. Baraniuk, and M. C. Nuss, "Gas sensing using terahertz time-domain spectroscopy," Appl. Phys. B, Photophys. Laser Chem., **67**, 379–390, (1998).

[4] R.K.H. Galvão S. Hadjiloucas, V.M. Becerra and J.W. Bowen, "Subspace system identification framework for the analysis of multimoded propagation of THz-transient signals, " Measurement Science and Technolgy, **16**(3), 1037-1053, (2005).

[5] X. X. Yin, B. W.-H. Ng, B. Ferguson, S. P. Mickan, and D. Abbott, "One dimensional wavelet transforms and their application to T-ray pulsed signal identification," In Proc. SPIE Photonics: Design, Technology and Packaging, volume 6038, pages 499-509,Brisbane, Australia, 2006.

[6] R.K.H. Galvão S. Hadjiloucas, J.W. Bowen and C.J. Coelho, "Optimal discrimination and classification of THz spectra in the wavelet domain," Optics Express, **11**, 1462-1473 (2003).

[7] X. Yin, B. W.-H. Ng, B. Ferguson, D. Abbott and S. Hadjiloucas, "Auto-regressive Models of Wavelet Sub-bands for Classifying Terahertz Pulse Measurements, " Journal of Biological Systems **15**(4), (2007) *Scheduled for publication December 2007*

[8] H. M. Paiva and R. K. H. Galvão, "Wavelet-packet identification of dynamic systems in frequency subbands," Signal Processing, **86**, 2001-2008 (2006).

[9] J. A. Zeitler, P. F. Taday, D. A. Newnham, M. Pepper, K. C. Gordon, and T. Rades, "Terahertz pulsed spectroscopy and imaging in the pharmaceutical setting, " Journal of Pharmacy and Pharmacology, **59**(2), 209-223, (2007).