Subspace and Wavelet-Packet Algorithms for de-noising and classifying broadband THz transients

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

[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*—The results from a range of different signal processing schemes used for the further processing of THz transients are contrasted. The performance of different classifiers after adopting these schemes are also discussed.**

I. Introduction And Background

THIS work compares classification results of a range of solid samples, 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.

Background and sample interferograms were detrended, aligned with respect to each other and apodized in the time domain using a Mertz (asymmetric triangular) apodization window. The sample FFT was ratioed against the background FFT in order to obtain the complex insertion loss (CIL) function which is presented to the classifier.

In order to implement the subspace algorithm, the background and sample interferograms (after assuming the same pre-processing procedures, but without time-domain apodization) are treated as input and output signals respectively, so that a subspace model in state space (utilizing the n4sid algorithm) is adopted to reduce the dimensionality of the vectors presented to the classifier.

Finally, a third formulation based on an identification algorithm which is applied to several frequency sub-bands of the corresponding THz spectra is also adopted. A wavelet-packet decomposition tree is used to establish the frequency bands at which the sub-band models will be created. Each leaf node of the decomposition tree is associated to a frequency band, and the complete set of leaf nodes composes the whole frequency range. For each frequency band, a sub-band model is, therefore created. An optimization of the tree structure is performed by using a generalized cross-validation method in order to achieve a compromise between accuracy and parsimony of the overall model. Such a procedure automatically determines the most appropriate frequency partitioning for the sub-band models. The goal of this work is to demonstrate efficient and robust classification algorithms that could be adopted by the biomedical, pharmaceutical as well as security communities which provide the technology pull required for the further proliferation of THz-transient spectrometers.

II. Results

Results for the newly developed wavelet identification algorithm are presented in Fig. 1 below.

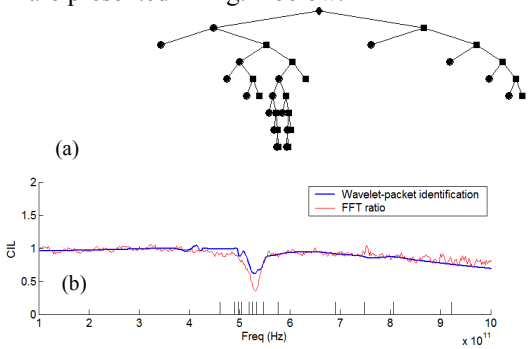

Fig. 1. a) Resulting wavelet-packet tree and b) CIL for a lactose sample obtained by wavelet-packet identification (blue line). The FFT ratio result (red line) is also presented for comparison. The frequency-domain segmentation (which is more refined in the spectral region corresponding to the absorption band) automatically defined in the identification procedure is indicated by vertical lines at the bottom of the graph.

An artificial data set of 150 sample interferograms (50 for each sample type: lactose, mandelic acid and DL mandelic acid) was generated by degrading the original interferograms with white, zero-mean Gaussian noise. No additional noise was added to the background signal. Figs. 2a and 2b present the resulting complex insertion loss function (for a particular noise realization with variance σ^2) in the range 0.1 – 1.0 THz obtained by ratioing the sample spectrum against the background.

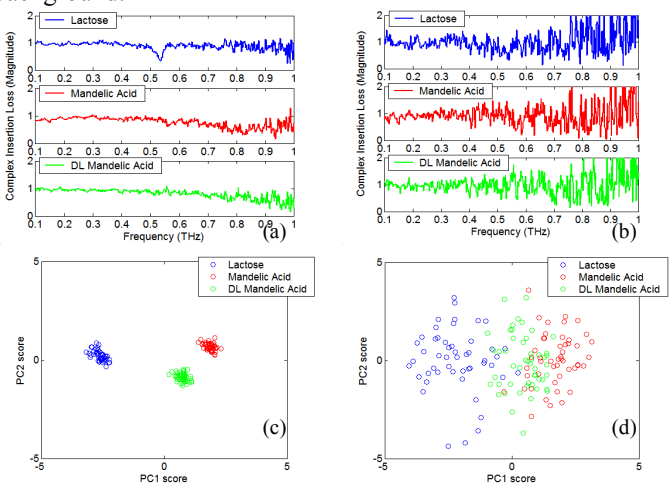

Fig. 2. Complex insertion loss functions obtained from degraded sample interferograms with (a) $\sigma = 10^{-3}$ and (b) $\sigma = 5\times10^{-3}$. The corresponding principal component score plots are presented in (c) and (d), respectively.

In this range, each CIL function comprises 500 spectral bins. The introduction of noise in the interferograms causes

978-1-4244-2120-6/08/$25.00 ©IEEE.

distortions in the spectral features of the samples (such as the absorption band around 0.52 THz for lactose) domain, which may compromise the performance of classification models. Such an effect is more apparent in the principal component score plots (for the mean-centered data set) presented in Figures 2c and 2d. As can be seen, the degree of overlapping between the three sample classes increases with the noise level. For classification purposes, let $x_{i,n}$ be the CIL magnitude of the i^{th} object (i = 1, …, 150) at the n^{th} spectral bin (n = 1, …, 500). A row vector $\mathbf{x}_i$ is defined for each object by disposing the CIL magnitude values in the form: $\mathbf{x}_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,500}]$. In what follows, classes 1, 2, and 3 will refer to lactose, mandelic acid, and DL-mandelic acid, respectively. 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:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{i\in I_1}\mathbf{x}_i \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{i\in I_2}\mathbf{x}_i \qquad \mathbf{m}_3 = \frac{1}{N_3}\sum_{i\in I_3}\mathbf{x}_i$$

where I_1, I_2, I_3 are the index sets of objects belonging to classes 1, 2, 3, respectively and $N_1 = N_2 = N_3 = 50$ are the number of objects in each class. Given a new object $\mathbf{x} = [x_1\ x_2\ \ldots\ x_{500}]$, a simple classification rule can be formulated as follows: $j^* = \arg\min_{j=1,2,3} \| \mathbf{x} - \mathbf{m}_j \|$, where $\| \cdot \|$ is the usual Euclidean norm and j^* is the class index (1, 2 or 3) assigned to $\mathbf{x}$. In order to evaluate the performance of this classification procedure, another set of 150 degraded interferograms (50 for each sample type) was generated and used as a test set. As a result, an error rate of 15% (22 errors out of 150 classified objects) was obtained. In order to improve the classification results, the $\mathbf{x}$-vectors were decomposed by using a Symlet 8 filter bank with two decomposition levels. Filtering was carried out by using circular convolution in order to have an orthogonal transform, which is isometric with respect to the Euclidean norm. The resulting wavelet coefficients were ranked according to their discriminability with respect to the three classes under consideration.

The discriminability D_k of the k^{th} wavelet coefficient c_k can be quantified as in [3]: $D_k = S_{Bk} / S_{Wk}$, where S_{Wk} and S_{Bk} are measures of the within-class and between-class dispersions for coefficient c_k, respectively. The within-class dispersion S_{Wk} is defined as: $S_{Wk} = \sum_{j=1}^{3} s_{k,j}$, where $s_{k,j}$ is the dispersion of c_k in class j, calculated as $s_{k,j} = \sum_{i\in I_j}[c_{i,k} - m_{k,j}]^2$, where $c_{i,k}$ denotes the value of c_k in the i^{th} object and $m_{k,j}$ is the mean value of c_k in class j, that is: $m_{k,j} = \frac{1}{N_j}\sum_{i\in I_j} c_{i,k}$. The between-class dispersion S_{Bk} is defined as $S_{Bk} = \sum_{j=1}^{3} N_j[m_{k,j} - m_k]^2$, where m_k is the average of c_k over all training objects.

Figure 3 presents the error rate obtained in the test by progressively adding wavelet coefficients (in decreasing order of discrimination ability) to the classification model. As can be seen, this procedure may lead to a reduction in the number of errors, as compared to the result obtained in the original spectral domain (complex insertion loss prior to the wavelet decomposition procedure).

For comparison, Figure 4 presents the principal component score plot for the complex insertion loss estimated from 75 degraded sample interferograms (25 for each sample type) by using the subspace and wavelet packet algorithms. As can be seen, the use of input-output identification algorithms does not improve the discrimination of the classes, as compared to the classic ratioing procedure (Figures 2c and 2d).

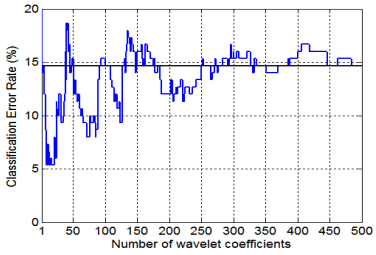

Fig. 3. Error rate in the test set as a function of the number of wavelet coefficients included in the classification model. The solid horizontal line indicates the error rate obtained in the original spectral domain.

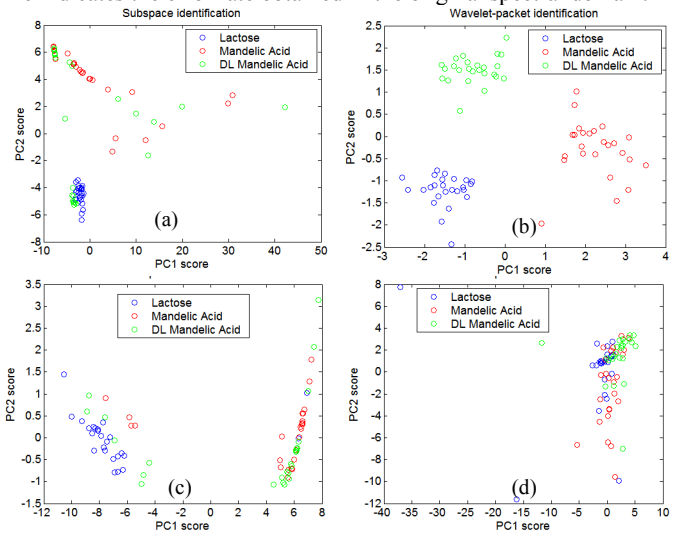

Fig. 4. Principal component score plot for the complex insertion loss estimated from degraded sample interferograms using (a, c) subspace and (b, d) wavelet packet identification. The noise level was set to $\sigma = 10^{-3}$ in (a, b) and $\sigma = 5\times10^{-3}$ in (c, d).

A leave-one-out cross-validation procedure was employed to evaluate the Euclidean-distance classifier for this reduced set of 75 degraded interferograms. Table 1 presents the resulting error rates. As can be seen, the best results were achieved by using the standard FFT-based ratioing procedure. Concerning the use of identification methods, the wavelet packet algorithm is seen to provide better results as compared to the subspace approach. This finding is in agreement with the better class discrimination achieved by using wavelet packet identification, as shown in Figure 4.

TABLE 1. Cross-validation error rates (%)

Noise level σ	FFT ratioing	Subspace identification	Wavelet Packet identification
10^{-3}	0	31	0
5×10^{-3}	24	41	37

REFERENCES

[1] S. Hadjiloucas, *et al.,* "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).

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

[3] R.O. Duda, P.E. Hart, & D.G. Stork, *Pattern Classification*, 2nd Ed. New York: John Wiley & Sons, (2001).

978-1-4244-2120-6/08/$25.00 ©IEEE.