

Automatically Detecting Peaks in Terahertz Time-Domain Spectroscopy

Henrike Stephani

*Fraunhofer ITWM and Technical University, Kaiserslautern, Germany,
and Johannes Kepler University, Linz, Austria
henrike.stephani@itwm.fraunhofer.de*

Joachim Jonuscheit and Christoph Robiné

*Fraunhofer IPM, Kaiserslautern, Germany
joachim.jonuscheit, christoph.robine@ipm.fraunhofer.de*

Bettina Heise

*Johannes Kepler University, Linz, Austria
bettina.heise@jku.at*

Abstract

To classify spectroscopic measurements it is necessary to have comparable methods of evaluation. In Terahertz (THz) time-domain spectroscopy, as a new technology, neither the presentation of the data nor the peak detection is standardized yet. We propose a procedure for automatic peak extraction in THz spectra of chemical compounds. After preprocessing in the time-domain, we use a variance based algorithm for determining the valid frequency region. We furthermore propose a baseline correction using simulated THz spectra. We illustrate how this procedure works on the example of hyperspectral THz measurements of six chemical compounds. Subsequently we propose to use unsupervised classification on the thus processed data to robustly detect the characteristic peaks of a compound.

1. Introduction

In pharmaceutical quality control, technologies are necessary that identify different chemical compounds. This identification should be performed in a non invasive way. The Terahertz (THz) technology is a useful tool in that aim because in these wavelengths chemical compounds have characteristic absorption spectra while most packaging materials such as carton, plastics, and ceramics are not absorbing [3].

There are databases that contain the characteristic spectral expression of many chemical compounds in the infrared range [2]. As an emerging technique most THz

spectra are not comparably characterized yet. There are databases such as [7] but the quality and method of acquisition declaredly differ. For most specific applications this is a problem. Particularly to address the problem of comparability, we propose a procedure to detect peaks in THz measurements of solids acquired by time-domain spectroscopy. On a set of six hyperspectral imaging measurements of chemical compounds, this procedure will be presented here.

We especially propose a method to determine the Dynamic Range (DR) of spectra based on standard peak detection. Furthermore, we propose a method for baseline correction. In spite of the normalization with a reference measurement, most transmittance spectra do not have a constant baseline which makes the classification of peaks difficult. We simulate the basic shape of the spectra and propose to use this to create such a constant baseline.

After selecting peak candidates for each compounds we propose to use unsupervised learning, more particularly hierarchical clustering, to determine which of these are actual peaks. Using this method helps building comparable THz databases.

2. Data

The chemical compounds measured are Para-Aminobenzoic Acid (PABA), Acetyl Salicylic Acid, Salicylic Acid, Tartaric Acid, Glucose, and Lactose. The measurements are taken by THz time-domain spectroscopy. This technology is based on pulsed THz systems where time-domain signals (pulses) are recorded

[9]. Our samples are 36 pixel images of pressed pellets of the agents that are measured in transmission. The low resolution ($\sim 1\text{mm}$) is due to the broadness of the focused laser beam [11].

These hyperspectral images contain a whole signal in each pixel. To analyze them, the signals have to be transformed into the frequency domain and then be preprocessed such that characteristic peaks can be detected. We will visualize the proposed procedure for normalization and peak detection on the example of a PABA measurement throughout this paper. The same procedure was applied on the other measurements analogously. For these, only the results shall be presented at the end.

3. Preprocessing

All physical data can be corrupted by the environment's and the sample's conditions. These conditions have to be treated by preprocessing. Firstly, we will consider distortions within the time-domain. Secondly, we will introduce a method to automatically determine the DR of the measurement and perform a baseline correction within the frequency domain.

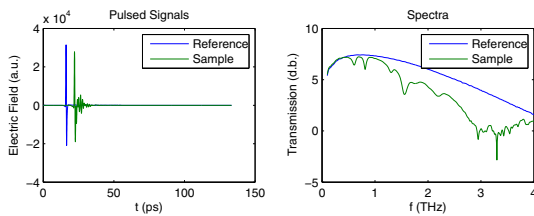


Figure 1. Sample and reference in the time- (l) and the frequency-domain (r).

3.1. Calculation of Transmittance

The acquired data consists of time-domain pulses. The left hand side of Fig. 1 shows a sample and a reference pulse of the named PABA measurement. By using windowing functions, echo pulse can be filtered away [10]. The signals are then transformed into the frequency domain as seen on the right hand side of Fig. 1. The reference pulse and the noise floor are used to normalize the spectrum and get rid of dominant environmental effects [6]. Fig 2 shows an example for the normalized transmittance spectra. They are the basis for further data analysis.

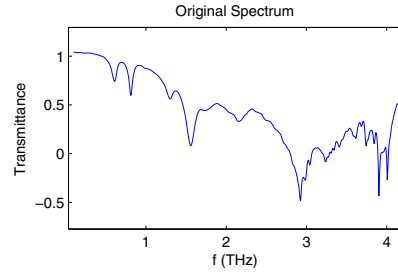


Figure 2. Normalized transmittance.

3.2. Dynamic Range Determination

The DR of a spectrum is usually determined by either taking a separate noise-floor measurement or by extracting it from the reference measurement [4]. Using the noise-floor as the indicator assumes only the system noise to be a relevant noise source. Neither the sample's thickness nor additional pulses are considered. We therefore propose a different approach.

The peaks in THz spectra of solids are rather broad (~ 100 Gigahertz). With this information we combine finding the peaks. We estimate the derivative by simply using $Der(x(n)) = x(n-1) - x(n)$. Depending on the quality of the spectrum, prior smoothing is advisable. For our DR detection we use the fact that when approaching the noise floor the spectra get noisy and the number of extrema increases. The DR ends as soon as the frequency of peaks is consistently higher than 100 GHz. Thereby we get an individually appropriate estimation of the valid frequency region of each spectrum, furthermore we get candidates for the peak positions. The DR is used in the next step for the so called baseline correction.

3.3. Baseline Correction

In Fig. 2 one can see that the baseline of the shown spectrum is not constant. Although it is possible to find the peaks, this non-constant baseline makes it difficult to determine their actual depth. For reasons of comparability of different measurements it is important, however, to know the absorption at a specific peak position. Therefore, different approaches are taken to find the basic shape of the spectrum. This can then be used for a normalization.

One way of baseline correction used in infrared spectroscopy is to apply morphological opening on the spectra [8]. A disadvantage of this method is that it is very sensitive to the choice of the structuring element. In THz spectroscopy particularly the big peaks cannot be

eliminated by this method without eliminating parts of the basic shape, too. Another popular way to find the basic shape is the application of a very coarse filtering on the spectra [12]. We use this approach with a Savitzky-Golay filter of degree 3 and half the number of samples as the window width. The result can be seen on the top level of Fig. 3. In comparison to the original transmittance the spectrum is already situated around one but still has a non constant baseline. Depending on the strength of the surrounding peaks, even deep peaks can have a transmittance around one.

We therefore propose a method that uses a basic

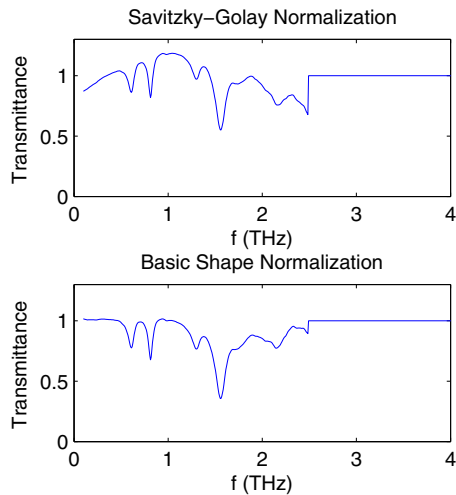


Figure 3. Savitzky-Golay baseline correction (top), simulation based baseline correction (bottom) up to DR of ~ 2.5 THz

shape which is a simulation of the basic shape of a THz spectrum. A THz pulse is generated by a polarization transient and the signal measured is then proportional to its second derivative [1]. By using the DR calculated as described in 3.2, we approximate such a polarization transient by a hyperbolic function (to be specific the hyperbolic tangent). The simulated spectrum's shape is determined by one parameter, namely the steepness of this hyperbolic function. Starting with a good estimation, we choose this parameter such that the following distance function is minimized:

$$D = \sqrt{\sum (S_i - X_i)^2} + \frac{\sqrt{\sum (-|X_i - S_i| + (X_i - S_i))^2}}{3},$$

where X is the respective simulations and S is the sample spectrum. By this distance function the values of the

simulations lying below the sample are punished harder than the ones lying above it. In this way we get an appropriate estimation of the basic shape which we now can use for the baseline correction. An illustration of the result is presented on the bottom plot of Fig. 3. It is assumed that the noise dominates the information content after the DR (at ~ 2.5 THz) and therefore the spectrum is set to one from there on. One can see that in combination with the information about the DR of the spectrum a constant baseline can be achieved.

4. Peak Detection and Classification

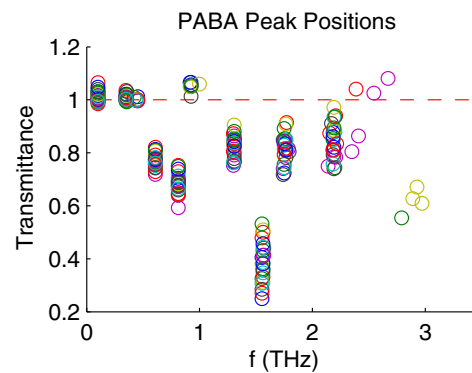


Figure 4. Possible PABA peaks.

We now have spectra that are preprocessed in such a way that we know the DR of each one and know the positions and depth of possible peaks. That means every peak candidate has a two dimensional representation by its position and its absorption. Fig. 4 shows all candidates of the 36 PABA spectra. To automatically group them and find the relevant clusters we use an unsupervised learning algorithm.

4.1. Unsupervised Classification

Out of the variety of unsupervised classification algorithms [5] we use hierarchical clustering with an average linkage function. The main advantage of hierarchical clustering is that no initial parameters are needed. As the number of peaks and their positions can vary a lot within spectra of different compounds it is essential that we do not have to determine the number of clusters or their initial positions beforehand. On the contrary: Hierarchical clustering provides us with a tree-graph that shows clusterings of different coarseness. We then use the coarseness that is most appropriate according to the number of input parameters. In our case each of the

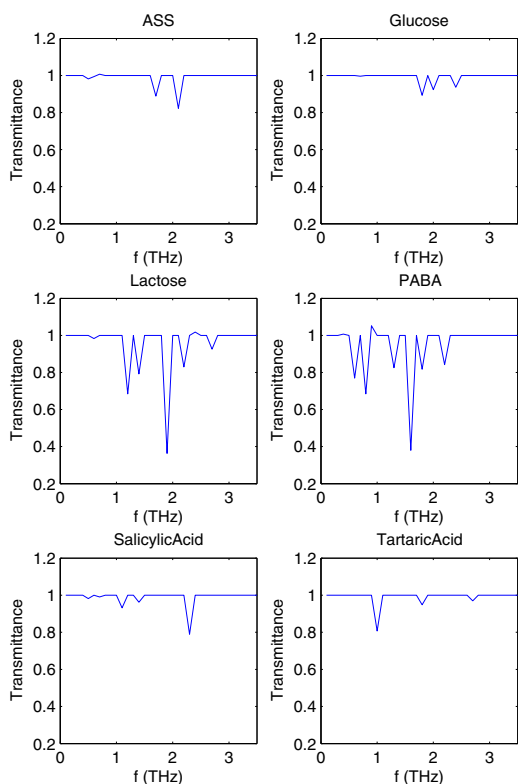


Figure 5. Mean peak positions.

hyperspectral imaging measurements consists of 36 different spectra (a 6x6-pixel measurement as introduced in section 2). We therefore determine the clustering to be complete as soon as every cluster has less or equal to 36 elements. We additionally cut off all clusters that have less than seven elements because these peaks are assumed to be noise. The result of this approach can be seen in Fig. 5. Every cluster is represented by its mean.

5. Conclusion and Further Research

We proposed a procedure that robustly detects characteristic peaks of chemical compounds when several measurements can be used. We used a method that combines finding possible peak positions with finding the valid frequency range of each spectrum. In addition, a baseline correction method predicated on a simulation of the basic shape of THz spectra was applied. The thus preprocessed spectra could well be used to classify their peaks and determine which ones appear in a critically big number of them. We thereby have a methodol-

ogy that can help build a database of peak positions of THz spectra in a comprehensible way. Further research should hence consist in continuing that work and measuring further compounds.

6. Acknowledgement

We thank the colleagues of the Department of Knowledge-Based Mathematical Systems at the JKU, especially Erich Peter Klement for discussion and support as well as Karin Wiesauer and Stefan Katletz from RECENTD GmbH. Part of this work was supported by the BMWi German Federal Ministry of Economics and Technology (THESEUS program, use case ORDO) and the BMBF German Federal Ministry of Education and Research (TEKZAS program).

References

- [1] A. Bonvalet and M. Joffre. Terahertz Femtosecond Pulses. *Femtosecond laser pulses: principles and experiments*, 1995.
- [2] P. Chu, F. Guenther, G. Rhoderick, and W. Lafferty. The NIST Quantitative Infrared Database. *Journal of Research - National Institute of Standards and Technology*, 104:59–82, 1999.
- [3] B. Ferguson and X. C. Zhang. Materials for Terahertz Science and Technology. *Nature materials*, 1(1):26–33, 2002.
- [4] J. Handley. *Time Frequency Analysis Techniques in Terahertz Pulsed Imaging*. PhD thesis, 2003.
- [5] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
- [6] P. U. Jepsen and B. M. Fischer. Dynamic Range in Terahertz Time-Domain Transmission and Reflection Spectroscopy. *Optics letters*, 30(1):29–31, 2005.
- [7] NICT and RIKEN. The terahertz database.
- [8] A. C. Sauve and T. P. Speed. Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data. *Proceedings Gensips*, 2004.
- [9] C. Schmittenmaer. Exploring dynamics in the far-infrared with terahertz spectroscopy. *Chemical Reviews*, 104(4):1759–1780, 2004.
- [10] H. Stephani, M. Herrmann, K. Wiesauer, S. Katletz, and B. Heise. Enhancing the interpretability of terahertz data through unsupervised classification. In *Proceedings of the 2009 IMEKO World Congress*, 2009.
- [11] M. Tonouchi. Cutting-Edge Terahertz Technology. *Nature photonics*, 1(2):97, 2007.
- [12] S. Wold, H. Antti, F. Lindgren, and J. Öhman. Orthogonal Signal Correction of Near-Infrared Spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2):175–185, 1998.