Classification of materials using terahertz spectroscopy with principal components analysis

Y. Zou[1], P. Sun[1*], W. Liu[2]
[1] Beijing Area Major Laboratory of Applied Optics, Department of Physics, Beijing Normal University, Beijing 100875, China; *pingsun@bnu.edu.cn
[2] Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Department of Physics, Capital Normal University, Beijing 100048, China

***Abstract*—Terahertz spectroscopy has multivariable and produces large volumes of data. Principal component analysis (PCA) is a statistics analysis method of dimensionality reduction of multivariate. We studied the feasibility of PCA method for material classification in terahertz region. The results show that PCA is able to differentiate materials obviously if initial variables are chosen properly.**

I. INTRODUCTION

Terahertz time domain spectroscopy (THz-TDS) is very sensitive to vibration and rotation of biological molecules, so it is a kind of finger print spectrum which can be used in material classification. However, terahertz spectroscopy produces large volumes of data, due to the spatial and temporal components both being recorded, so it is necessary to investigate data reduction methods prior to classification. Principal components analysis (PCA) is well suited for this purpose as it provides a theoretically optimal linear reduction. PCA is a sort of data mining technology in multivariate statistics. It can choose less new variable replacing more original variable to remove the overlapping information in large coexisting data and reduce data's dimensionality without loss of information of principal variables [1-4]. The purpose of this study was to investigate the feasibility of using terahertz measurements to classify different materials.

II. RESULTS

We used typical THz time-domain transmission spectroscopy system [5]. One group of samples was D-(+)-glucose powders (Sigma co. Ltd, purity >99%). They were ground with a mortar for 1 hour and further screened through a molecular sieve to obtain particles with diameters below 25 μm. The screened particles were then pressed into 5 tablets with 13 mm in diameter, but with thicknesses of 0.362, 0.447, 0.504, 0.522 and 0.626 mm. The tablets were made with a pressure of 10 t using a tablet machine (Specac Limited, UK). The second group of samples was five different kinds of water. And the third group of sample was five kinds of Intralipid with concentration of 12%, 16%, 18%, 19% and 20%. To reduce the THz absorption of the water molecule in the air, the parts inside the dotted line in the optical path were covered and filled with dry nitrogen to keep humidity below 2.0%. The temperature was kept at 21°C.

We measured the THz-TDS of three groups of samples and then extracted their optical parameter including extinction coefficient $k(\omega)$ and refractive index $n(\omega)$ (ω is frequency) based on the Fresnel formula and calculate dielectric coefficients $\varepsilon_r(\omega)$ and $\varepsilon_i(\omega)$ according to $k(\omega)$ and $n(\omega)$. Finally, we apply PCA to classify three different materials. Fig.1 shows the Terahertz time domain spectrum of three different kinds of materials we measured. Fig.2 shows the 3D result of PCA. Fig.3 shows two kinds of PCA results based on the six initial variables which are two wave trough values, one peak value, and half-width of them in the spectrum. These three kinds of materials are separated into three groups in the 3D result shown in Fig.2. What's more, these 16 points presenting different samples are obviously located in three areas by projecting them on the PC1 and PC2 axis, the first two principle components, as can be seen in Fig. 3(a). In addition, we can judge the similarity through the distance between two groups. For example, the distance d_1 between water and Intralipid solution is less than the distance d_2 between Intralipid and Glucose, which indicates that Intralipid is similar to water. However, these points are not classified on the projection of PC2 and PC3 shown in Fig. 3(b). So it can be suspected that the first principle component carries the largest information, and the second and the third carries very small, which is also shown in Fig.4, the bar figure of principle component contribution rate describing the weight of information. It shows that PC1's contribution rate much more than another two principle components and PC3's contribution rate is less than 0.3%. This result corresponds to the PCA result shown in Fig.3 very well. So it is enough to apply the first two principle components to PCA result to classify materials. And PC3 can be omitted because of its' light information loads. And then we add another six initial variables which are $k(0.7)$, $\varepsilon_i(0.7)$, $\varepsilon_r(0.7)$, $k(0.9)$, $\varepsilon_i(0.9)$, $\varepsilon_r(0.9)$ to those six above. The values in brackets are corresponding frequencies. So we get a new PCA result as shown in Fig.3(c). Comparing with Fig.3(a) those three kinds of materials are separated into three areas more obviously which are farther from each other in Fig.3(c) and the points representing the same material gather closer. The results correspond to Fig.4 reflecting the first three principle components' contribution rates, respectively. It can be seen that contribution rate of PC1 decreases, and PC2's and PC3's both increase by adding another six new initial variables. Therefore, we can conclude that adding initial variables can decrease the contribution rate of PC1, increase those of PC2 and PC3 and make the material classification more obvious in PCA result.

And then we replace $\varepsilon_i(0.9)$ in $k(0.7)$, $\varepsilon_i(0.7)$, $\varepsilon_r(0.7)$, $k(0.9)$, $\varepsilon_i(0.9)$, $\varepsilon_r(0.9)$ with one of three full width at half maximums (FWHM) of time domain. The calculated results show that PC1's contribution rate declines from 96.33% to 94.66%, and PC2's grows from 3.6% to 5.3%. So we can conclude that

reducing correlation index between initial variables can reduce PC1's contribution rate and increase PC2's, which makes the contribution rates of principle components more average.

Furthermore, according to results above, all the contribution rates of PC1 outnumber 93%. Therefore, refractive index, extinction coefficient and dielectric coefficient, and values of peak and wave trough and peak half-width of time domain spectrum are all related to each other. But refractive index, extinction coefficient, and dielectric coefficient have closer correlation than peak and wave trough value and peak half-width of time domain spectrum.

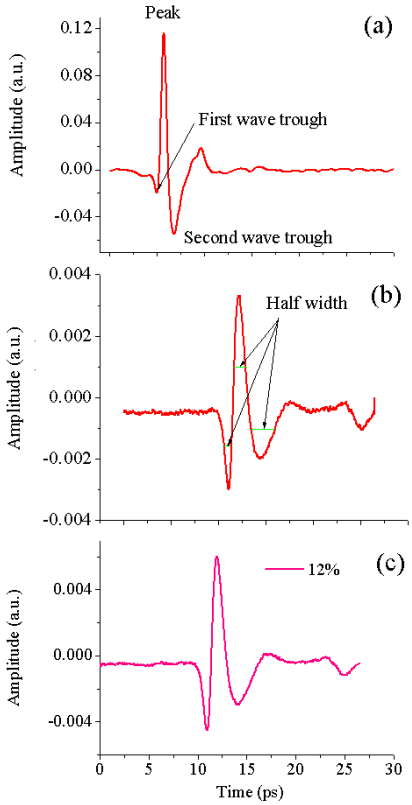

Fig. 1. Terahertz time domain spectrum: (a) Glucose tablet with thickness of 0.362mm; (b) Deion water; (c) Intralipid solution with concentration of 12%.

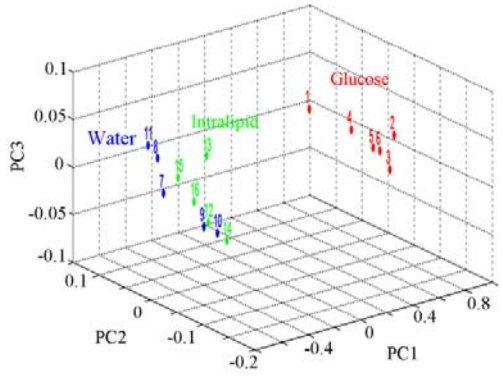

Fig. 2. The 3D result of PCA which is described by the first three principle components. Here number1-6 represent glucose tablets in six different thicknesses, 7-11 represent five kinds of water and 12-16 represent five kinds of Intralipid in different concentration

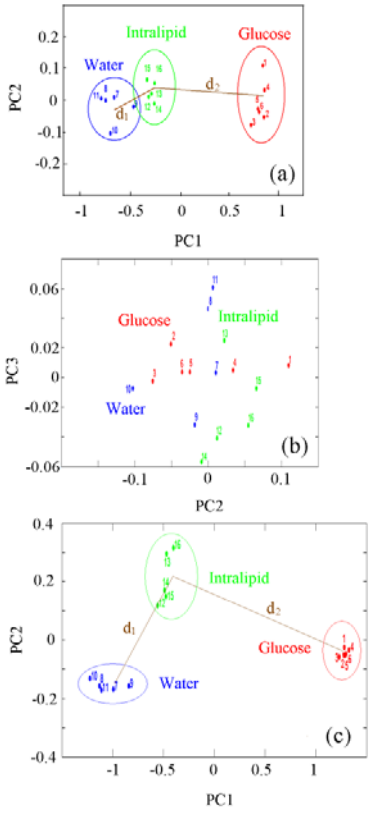

Fig. 3. The 2D result of PCA described by principle components of (a) the first two; (b) the second and the third; (c) the first two where $k(0.7)$, $\varepsilon_i(0.7)$, $\varepsilon_r(0.7)$, $k(0.9)$, $\varepsilon_i(0.9)$, $\varepsilon_r(0.9)$ are added to the six initial variables of (a).

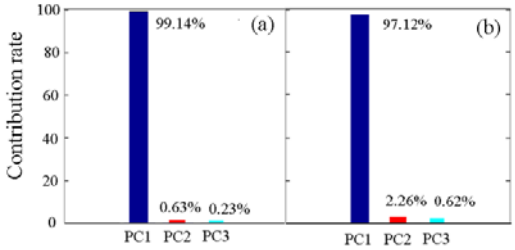

Fig. 4. Contribution rates of the first three principal components: (a) based on six initial variables chosen in Terahertz time domain spectrum; (b) adding new six initial variables to (a).

III. SUMMARY

PCA based on terahertz spectrum in time domain is able to differentiate materials obviously if initial variables are chosen properly. It also can judge the similarity of materials.

REFERENCES

[1] A. J. Fitzgerald, S. Pinder, A. D. Purushotham, P. O'Kelly, P. C. Ashworth, V. P. Wallace. Classification of terahertz-pulsed imaging data from excised breast tissue. J. Biome. Opt. 2012, 17(1), 016005-1-10.

[2] M. A. Brun, F. Formanek, A. Yasuda, M. Sekine, N. Ando, Y. Eishii, Terahertz imaging applied to cancer diagnosis. Phys. Med. Biol. 2010, 55: 4615-4623.

[3] H. Zhong, A. R. Sanchez, X.-C. Zhang. Identification and classification of chemicals using terahertz reflective spectroscopic focalplane imaging system. Opt. Express. 2006, 14(20): 9130-9141.

[4] S. Nakajima, H. Hoshina, M. Yamashita, C. Otani. Terahertz imaging diagnostics of cancer tissues with a chemometrics technique. Appl. Phys. Lett. 2007, 90: 041102-1-3.

[5] W. Zouaghi, M. D. Thomson, K. Rabia, R. Hahn, V. Blank, H. G. Roskos. Broadband terahertz spectroscopy: principles, fundamental research and potential for industrial applications. Eur. J. Phys. 2013, 34: S179–S199.