Identification of Coals Using Terahertz Spectroscopy Combined with Manifold Learning and SVM Discriminant Analysis

Liang LIANG¹*, Yu-xin DU¹, Zhong-wen WU², Zi-jian YANG¹ and Hong-yan ZHANG¹

¹School of Mechanical & Electrical Engineering, Xuzhou University of Technology, Xuzhou, Jiangsu 221018, China
²Foreign Language Department, Xuzhou University of Technology, Xuzhou, Jiangsu 221018, China

*Corresponding author

Keywords: THz-TDS, Pulverized coal, Manifold learning, SVM, Discriminant analysis.

Abstract. The coal quality links to the efficiency and toxic gas emission of coal combustion, which has led to the increasing demand for quick and non-destructive detection method to identify various coal. In this study, terahertz spectroscopy combined with manifold learning algorithm and support vector machine (SVM) discriminant analysis was applied to analysis six types of coal materials. To evaluate the effectiveness of the proposed method, interval PLS (iPLS) and genetic algorithm combined PLS (GA-PLS) were used for spectral variable selection, principal component analysis (PCA) and stochastic neighbor embedding (SNE) algorithm were applied for spectral dimensional reduction. The experimental result showed that the SNE algorithm combined with SVM (SNE-SVM) has a higher correlation coefficient of prediction set (0.9842), a lower root mean squared error of prediction (0.2144), and the prediction accuracy of different coal materials reached 100%. This study indicates that the terahertz spectrum analysis combined with manifold learning algorithm is a promising method for the classification of different coals.

Introduction

Given the lack of petroleum and nature gas resources, coal plays a major role of the total energy consumption in China. Coal quality has proven to be one of the most important affecting factors of burning process [1]. The identification processes of coal are time-consuming using conventional laboratory methods. Thus, for ensuring a highly efficient combustion/pyrolysis process and reducing toxic gas emissions, a reliable, quick and non-destructive detection method to identify various coals is required in coal-fired power plants and coal-gasification plants.

Owing to the rapid development of ultrafast laser technologies in recent decades, terahertz (THz) spectroscopy has made significant progress. Relying on its coherence, low-energy and fingerprints, the THz spectroscopy has been applied for mineral analysis [2]. In this study, the absorption spectra in THz band of six types of secondary reference coal material were obtained using terahertz time-domain spectroscopy (THz-TDS) system. Two types of manifold learning algorithm, including principle component analysis (PCA) and stochastic neighbor embedding (SNE), were employed for spectral data dimensional reduction. Support vector machine (SVM) algorithm was used to establish a model for coal identification. The experimental results of this study demonstrate the potential of the THz-TDS system applied in distinguishing various types of coal material.

Experiments and Methods

Experimental Setup

The THz-TDS system used in the experiment was composed by a diode-pump mode-locked Ti:sapphire femtosecond laser which provided 20 fs pulse width with a wavelength of 1560 nm and a repeating frequency of 80 MHz and a THz spectrometer (TeraKit® AIO) produced by Rainbow Photonics AG Inc., Switzerland, as shown in Figure 1. Because of highly polarized molecule
structure of water molecule, water vapor has great impact on the propagation of THz wave. Thus, dry nitrogen was used to fill the container, which is indicated using dashed lines in Figure 1(a). The humidity in the container was maintained lower than 3%, and the temperature was maintained at 23°C during the experiment.

According to the physical parameters extraction method proposed by Duvillaret and Dorney [19, 20], the refractive index \(n(\omega)\) and absorption coefficient \(\alpha(\omega)\) can be calculated as follows:

\[
n(\omega) = \phi(\omega) \frac{c}{\omega d} + 1 = 1 + \frac{c}{\omega d} \left[ \phi_{\text{sample}}(\omega) - \phi_{\text{ref}}(\omega) \right]
\]

\[
\alpha(\omega) = \frac{2k(\omega)\omega}{c} = \frac{2}{d} \ln \left[ \frac{4n(\omega)}{\rho(\omega)(n(\omega) + 1)^2} \right] = \frac{2}{d} \ln \left| \frac{E_{\text{ref}}(\omega)}{E_{\text{sample}}(\omega)} \right|
\]

where \(\omega\) denotes the angular frequency, \(n(\omega)\) is the refractive index, and \(\alpha(\omega)\) is the absorption coefficient, \(d\) represents the sample tablet thickness, \(c\) is the speed of light in vacuum.

### Sample Preparation

Six types of the secondary reference coal material were purchased from Quandong reference material research institute, Ji’nan, Shandong province, China. The proximate analysis and elemental analysis of the coal samples was carried out and the results are listed in Table 1. The raw coal was first dried to remove external moisture. The dried coal was then ground to fine powder (passing 200 mesh sieve). The final step involved sealing the manufactured coal for further processing. High-density polyethylene (HDPE) powder, which is transparent in THz band, was used as a diluter to mix with the coal powder with a coal/HDPE mass ratio of 1:4. The mixture was then pressed using a tablet machine under a pressure of 16 MPa for 3 min. After this procedure, the mixture was pressed into tablets with a diameter of 13 mm and a thickness of nearly 0.68 mm.

| No. | Samples(GBW(E)) | S(%) | A(%) | V(%) | FC(%) | H(%) | N(%) |
|-----|-----------------|------|------|------|-------|------|------|
| C1  | 110026          | 4.04 | 18.62| 31.95| 65.06 | 4.35 | 1.15 |
| C2  | 110027          | 4.03 | 16.35| 11.85| 71.72 | 3.2  | 1.02 |
| C3  | 110031          | 2.79 | 8.62 | 10.86| 81.27 | 3.55 | 1.16 |
| C4  | 110035          | 0.51 | 8.47 | 25.81| 77.01 | 4.2  | 1.17 |
| C5  | 110037          | 3.15 | 25.8 | 8.86 | 64.97 | 2.22 | 0.8  |
| C6  | 110066          | 1.49 | 41.05| 24.54| 46.06 | 3.22 | 0.88 |

### Modelling Methods

**Interval PLS.** The interval PLS (iPLS) algorithm can be used for spectral variable selection [5]. The whole spectrum is equally divided into several intervals, then PLS regression model is
established for every subinterval. The subinterval which has the lowest root mean error of cross validation (RMSECV) will be regarded as the best interval.

**PLS Combined with Genetic Algorithm (GA-PLS).** Hasegawa et al. and Leardi et al. introduced and developed GA-PLS algorithm. The variable selection process can be briefly summarized as follows [18]:

1. The spectral variables are randomly selected for the initial populations.
2. The selected variable set is used to establish PLS-R regression model. Leave-one-out cross validation is then carried out for evaluating the performance during modelling process. The variable set with higher accuracy is selected for the next generation.
3. The operation of crossover and mutation among the selected variables is then carried out to generate new spectral variable set.
4. Set the selected and newly formed variables as a new population for step 2.

Step (2)-(4) are repeated for some times, and the best variable set will be obtained.

**Principal Component Analysis (PCA).** PCA is a powerful statistical method that uses an orthogonal transformation to convert the correlated variables into linearly uncorrelated variables called PCs [6], and has been employed in dimensional reduction for THz spectroscopy analysis [7].

**Stochastic Neighbor Embedding (SNE).** The SNE algorithm is aim to find the low dimensional embedding from the raw high dimensional data. A probability distribution from pairwise distances wherein larger distance correspond to smaller probabilities and vice versa can be established, and then by minimizing the KL divergence of the two probability distributions, the low-dimensional embedding can be obtained [8].

**Support Vector Machine (SVM).** SVM is a useful tool for solving nonlinear classification and regression problems. In this study, the Lib-SVM was used to classify the investigated reference coal material. Particle swarm optimization (PSO) algorithm was applied in this study to determine the optimal parameters (c and g) for SVM modelling process.

**Results and Discussion**

**Spectra Collection**

For each coal sample, 12 measurements were taken at different position of the tablet. By using Eq. (2) and (3), the average absorption spectra, ranging from 1.8 to 4 THz, are shown in Figure 3.

Owing to same main ingredients in coal materials, such as carbon, hydrogen, nitrogen, sulphur, similar frequency dependent absorption coefficient curves of these coal sample can be found. In addition, the curves have similar absorption peaks at 2.32, 2.68, 2.88, 3.1, 3.3, 3.67 THz, indicating that the coal materials cannot be classified using absorption peak directly. What’s more, overlaps shown in Figure 3 add more fuel to difficult-to-classify. Therefore, variable selection and dimensional reduction were applied to create a subset of relevant features for use in classification model construction.

![Figure 3. Averaged frequency dependent extinction spectrum of the six coal samples in 1.8-4 THz.](image1)

![Figure 4. Results of iPLS for the THz absorption spectra.](image2)
Spectral Variable Selection

**Variable Selection Using iPLS.** Figure 4 shows the variable selection result of iPLS. The raw THz absorption spectra of coal materials were divided into 19 intervals, and the variable number of each subinterval is equal to 6. The blue line is the average spectra, and the green line is the RMSECV of the global model. The bars represent the RMSECVs of the subintervals. Interval 8 (2.66-2.76 THz) reaches the lowest RMSECV, which is 1.293, indicating that the variables in this interval have better modelling performance than others.

**Variable Selection Using GA-PLS.** The parameters of genetic algorithm were set as follows: the population was set to 25, and the maximum iterate number was 100. The crossover probability was 0.3, and the mutation probability was 0.01. Figure 5 shows the number of selected variable dependent RMSECV. The green dot represents the minimum RMSECV (1.254) when the number of selected variables equals to 35. The selected variables account for 30.7% of the whole spectral data. Figure 6 shows the frequency of selections. The histograms in dark blue denote the frequency of the variable selected in 100 iterations. The green solid line represents the selected variables when reaching the lowest RMSECV.

![Figure 5. Selected number of variables dependent RMSECV.](image1)

![Figure 6. Selection frequency of spectral variables.](image2)

**Dimensional Reduction**

By explaining the variance in the spectra data in the best way, the PCA operation can discover the internal structure of the spectral dataset. As shown in Figure 7, the first three PCs can explain 97.29% of the data variance. Therefore, in this paper, the first three PCs are adopted to take place of the raw absorption spectra. The dimension for SNE algorithm was set to 3. Figure 8 shows the data clustering visualizations of the absorption spectra using PCA and SNE, respectively.

![Figure 7. The explained variance of PCs.](image3)

![Figure 8. Visualizations of the absorption spectra using PCA and SNE.](image4)

**Performance of the Identification Model**

As mentioned above, the absorption spectra sub dataset of each coal material type was randomly selected as the calibration set (8 measurements) and the remainder as the prediction set (4 measurements). Finally, the calibration set contained 48 measurements, and the prediction set had 24 measurements. To identify the coal material, SVM regression model was established to distinguish the coal material. The pre-treatment of variable selection (iPLS and GA-PLS) and dimensional reduction (PCA and SNE) was carried out. To evaluate the performance of
pre-treatment combined with SVM regression model, correlation coefficient of calibration ($R_c$), root mean squared error of calibration set (RMSEC), correlation coefficient of prediction set ($R_p$), and root mean squared error of prediction set (RMSEP) were applied in this study. The modelling results are shown in Table 2. The SNE and PCA algorithms combined with SVM give higher $R_p$s and lower RMSEPs than others. Given an error threshold range ($\pm 0.5$), the identification results of the prediction set are shown in Table 3. It can be seen that the SNE-SVM algorithm can provide the highest total predictive accuracy (100%). Although the PCA-SVM has relatively high $R_p$ (0.9387) and low RMSEP (0.4229), it gives an unacceptable prediction accuracy (66.67%). iPLS-SVM, GA-PLS- SVM and none-pre-treatment provides poor prediction results. The total predictive accuracy (TPA) are 41.67%, 45.83%, and 50%. Figure 9 shows the prediction result of the six types of coal material using SNE-SVM model.

Table 2. Results and comparison of modelling with different pre-treatment methods.

| Pre-treatment | Variable number | $R_c$     | RMSEC  | $R_p$    | RMSEP  |
|---------------|-----------------|-----------|--------|----------|--------|
| None          | 114             | 0.9978    | 0.0798 | 0.6995   | 0.9362 |
| iPLS          | 6               | 0.8272    | 0.7099 | 0.6600   | 0.9958 |
| GA-PLS        | 35              | 0.9440    | 0.4042 | 0.7350   | 0.8792 |
| PCA           | 3               | 0.9956    | 0.1130 | 0.9387   | 0.4229 |
| SNE           | 3               | 0.9956    | 0.1132 | 0.9842   | 0.2144 |

Table 3. Prediction accuracy of six types of coal material.

| Pre-treatment | C1(%) | C2(%) | C3(%) | C4(%) | C5(%) | C6(%) | TPA(%) |
|---------------|-------|-------|-------|-------|-------|-------|--------|
| None          | 50    | 25    | 100   | 0     | 100   | 25    | 50.00  |
| iPLS          | 25    | 25    | 50    | 0     | 100   | 25    | 41.67  |
| GA-PLS        | 25    | 25    | 75    | 25    | 100   | 25    | 45.83  |
| PCA           | 100   | 100   | 75    | 50    | 100   | 75    | 66.67  |
| SNE           | 100   | 100   | 100   | 100   | 100   | 100   | 100    |

Figure 9. Plots of the predicted value as a function of actual value using the SNE-SVM model for different coals.

Conclusions

In this paper, variable selection and dimensional reduction methods combined with THz spectra were applied to distinguish six types of reference coal material. The absorption spectra of six types of the secondary reference coal material, in the frequency range of 1.8-4 THz, were obtained. There are no obvious absorption features of the coal materials in THz region. To accurate recognize the different coal samples, iPLS and GA-PLS were applied for variable selection, and PCA and SNE were applied for spectral dimensional reduction. SVM was adopted to construct a recognition model to classify different coal material. The experimental results demonstrate that the proposed SNE-SVM method was considered to have better performance than others, and the recognition rate reached 100%. In conclusion, the proposed method is fast and accurate for coal identification, indicating that the THz-TDS system could be a potential tool for the coal quality detection.
Acknowledgement
This research was financially supported by the Xuzhou Science and Technology Projects (KC16GX043) and Youth Foundation of Xuzhou University of Technology (XKY2018237).

References
[1] BP p. l. c., "Statistical Review of World Energy"-China's energy market in 2016, 2017.
[2] W. Fan, T. Chang, L. Liu, H. Cui, FUEL, 2017, 188, 246-253.
[3] L. Duvillaret, F. Garet, J. L. Coutaz, Selected Topics in Quantum Electronics, IEEE Journal of, 1996, 2, 739-746.
[4] T. D. Dorney, R. G. Baraniuk, D. M. Mittleman, J Opt Soc Am A Opt Image Sci Vis, 2001, 18, 1562-1571.
[5] A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, L. Norgaard, S. B. Engelsen, Applied Spectroscopy, 2007, 54, 413-419.
[6] H. Abdi, L. J. Williams, Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2, 433-459.
[7] Y. Xie, P. Sun, Optical & Quantum Electronics, 2018, 50, 46.