CLASSIFICATION METHOD OF COAL AND GANGUE USING TERAHERTZ TIME-DOMAIN SPECTROSCOPY, CLUSTER ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

D. Shao [1], Sh. Miao [1*], Q. Fan [1], X. Wang [1], Zh. Liu [2,3], E. Ding [2,3]

[1] *School of Physics and Electronic Information at Huaibei Normal University, Huaibei, China; e-mail: msgmcu@126.com*
[2] *School of Information and Control Engineering at China University of Mining and Technology, Xuzhou, China*
[3] *IOT Perception Mine Research Center, China University of Mining and Technology, Xuzhou, Jiangsu, China*

The process of coal mining generates high amounts of coal gangue. Accordingly, coal-gangue separation is a key problem limiting coal production and quality. Terahertz time-domain spectroscopy was combined with multivariate statistical analyses to identify different kinds of coal and gangue. First, the terahertz spectrum and power spectrum of the sample were measured, and the refractive index and absorption coefficient of the sample were calculated from the terahertz time-domain spectrum of the sample. Significant differences in the power spectrum, refractive index, and absorption coefficient were found between different kinds of coal and gangue. After combining multivariate statistical methods – cluster analysis (CA) and principal component analysis (PCA) – a model based on THz parameters and different types of coal and gangue was established. During cluster analysis, the Euclidean distance of two types of samples and the score of the first principal component in the principal component analysis could reflect the similarity and dissimilarity between coal and gangue samples, and consistent results were obtained for CA and PCA. The experimental results showed that the combination of terahertz technology and multivariate statistical methods yielded an accurate approach to distinguishing between coal and gangue.

Keywords: *terahertz time-domain spectroscopy, coal-gangue classification, principal component analysis, cluster analysis.*

МЕТОД КЛАССИФИКАЦИИ УГЛЯ И ПУСТОЙ ПОРОДЫ С ИСПОЛЬЗОВАНИЕМ ТЕРАГЕРЦОВОЙ СПЕКТРОСКОПИИ ВО ВРЕМЕННОЙ ОБЛАСТИ И КОМБИНАЦИИ КЛАСТЕРНОГО АНАЛИЗА И АНАЛИЗА ГЛАВНЫХ КОМПОНЕНТ

D. Shao [1], Sh. Miao [1*], Q. Fan [1], X. Wang [1], Zh. Liu [2,3], E. Ding [2,3]

УДК 543.42:622.33

[1] *Школа физики и электронной информации Хуайбэйского педагогического университета, Хуайбэй, Китай; e-mail: msgmcu@126.com*
[2] *Школа информатики и управления Китайского горно-технологического университета, Сюйчжоу, Китай*
[3] *IOT Исследовательский центр Perception Mine Китайского горно-технологического университета, Сюйчжоу, Цзянсу, Китай*

(Поступила 2 сентября 2021)

Терагерцовая спектроскопия во временной области объединена с многомерным статистическим анализом для выявления различных видов угля и пустой породы. Измерены терагерцовый спектр и спектр мощности образца, показатель преломления и коэффициент поглощения образца рассчитаны по терагерцовому спектру образца во временной области. Обнаружены значительные различия в спектре мощности, показателе преломления и коэффициенте поглощения для разных видов угля и пустой породы. После объединения многомерных статистических методов – кластерного анализа (CA) и анализа главных компонент (PCA) – создана модель, основанная на параметрах ТГц, *и исследованы различные типы угля и пустой породы. При кластерном анализе евклидово расстояние между образцами двух типов и оценка первой главной компоненты при анализе главных компонент отражают сходство и различие между образцами угля и породы. Результаты, полученные этими методами, согласуются между собой. Показано, что сочетание терагерцовой технологии и многомерных статистических методов дает точный подход к различению угля и пустой породы.*

Ключевые слова: *терагерцовая спектроскопия во временной области, классификация угольных пород, анализ главных компонент, кластерный анализ.*

Introduction. The coal industry is an important basic industry for national energy security, acting as an economic lifeline. It is widely acknowledged that coal is an important energy source, accounting for 90% of China's fossil energy resources. In recent years, the use of coal for primary energy consumption has gradually decreased; however, its status as an important source of energy has not changed. Coal gangue is a solid waste discharged during coal mining and coal washing. It is a dark-gray rock with low carbon content and is harder than coal. The coal-gangue yield is high as its emission accounts for about 15–20% of the raw coal output. During the process of coal mining, it is essential that the shearer can cut along the interface between coal and gangue for optimal use of resources and complete underground gangue separation. However, the working environment of coal mining is often complex, and it is difficult for coal miners to identify the coal–gangue interface by sight, smell, or touch. Accordingly, the study of automated coal identification technology has significant value and has attracted much interest domestically and abroad.

In recent years, terahertz (THz) spectroscopy has been applied in agriculture [1, 2], medicine [3, 4], food safety [5, 6], and materials sciences [7]. Terahertz wave usually refers to the electromagnetic wave with frequencies between 0.1 and 10 THz (wavelength between 0.03 and 3 mm), spectrally located between microwaves and infrared light, belonging to the far-infrared spectrum [8]. The THz spectrum contains abundant amounts of information on samples. Moreover, the refractive index and absorption coefficient of samples can be obtained by studying the spectrum. In this study, two multivariate statistical analysis methods – principal component analysis (PCA) and cluster analysis (CA) – were combined to reprocess the obtained terahertz data. The Euclidean distance of two kinds of samples and the score of the first principal component during PCA were used to demonstrate the similarities and dissimilarities between coal and gangue samples.

Experimental and methods. *Experiment instrument.* The terahertz time-domain spectroscopy (THz-TDS) system is usually divided into two types: reflection system and transmission system. The Advantest TAS7400SU transmission terahertz spectrum system was used in this experiment with a frequency spectrum range of 0.5–7.0 THz, a scanning time of 200 ms for each spectrum, and a frequency resolution of 7.6 GHz. The structure diagram is shown in Fig. 1. The system consisted of a femtosecond laser, delay line, THz transmitter, sample test area, and THz detector. The beam of light emitted from the femtosecond laser was divided into two beams by a splitter. The beam of light acted on the terahertz emitter through multiple reflections to generate terahertz pulses focused on a sample after passing through the lens and M4. After passing through the delay region, another beam of light met with the terahertz pulse carrying the sample information on the THz detector, producing a photocurrent. Importantly, the delay line can eliminate the time delay between the femtosecond laser and THz pulse to ensure that the system can get a complete time-domain waveform [9, 10].

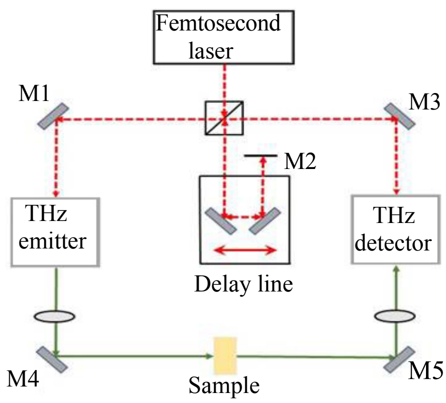

Fig. 1. The basic structure of the terahertz time-domain spectroscopy system, M1–M5 are mirrors.

Experimental. Five kinds of coal and gangue powder ZBM111C (bituminous coal), ZBM111A (bituminous coal), ZBM102 (bituminous coal), ZBM097A (anthracite), ZBM110B (gangue), used as samples in the experiment, were obtained from the National standard material center network. The codes of sample standard materials and corresponding property parameters are shown in Table 1. To prevent the sample from being broken during the measurement process because of its brittleness, polyethylene powder with low absorption in terahertz was added before the tablet press pressed the coal and gangue powder. The coal and gangue powder and polyethylene were mixed at a 1:2 ratio and then poured into the grinding tool of the tablet press. A sample with a thickness of 1.1 mm was obtained by pressing it for 5 min at a pressure of 10 MPa. To eliminate the influence of ambient temperature and humidity, measurements were taken at room temperature (24°C), keeping the relative humidity below 1%, and the average of three measurements was taken for each sample as the sample signal.

TABLE 1. The Property Parameters of Coal/Gangue Samples

Sample	Category	Carbon, %	Hydrogen, %	Nitrogen, %	Volatile, %	Ash, %
ZBM102	Bituminous coal	57.82	3.40	1.02	30.43	25.88
ZBM111A	Bituminous coal	74.16	4.44	1.38	33.64	9.62
ZBM111C	Bituminous coal	77.14	4.59	1.23	31.29	8.00
ZBM097A	Anthracite	79.96	3.31	1.12	9.10	11.75
ZBM110B	Gangue	5.80	1.03	0.27	9.37	85.58

Mathematical methods. In our everyday life, the vibration energy levels of many materials and dielectrics fall within the terahertz spectrum. Interestingly, when an electromagnetic wave within the terahertz frequency range is irradiated on a sample, the measured sample will absorb the photon energy, and thus generate terahertz spectra containing the physical information of the sample. In this study, the terahertz spectrum of the terahertz spectrometer was transformed into fast Fourier transform (FFT), and the refractive index, extinction coefficient and absorption coefficient of the sample were calculated using the formula below [11, 12]:

$$n(\omega) = \frac{\varphi(\omega)c}{\omega d} + 1, \tag{1}$$

$$k(\omega) = \frac{c}{\omega d} \ln\left\{ \frac{4n(\omega)}{\rho(\omega)\left[n^2(\omega)+1\right]^2} \right\}, \tag{2}$$

$$\alpha(\omega) = \frac{2\omega k(\omega)}{c}, \tag{3}$$

where $\varphi(\omega)$ is the phase difference between the test sample signal and the reference signal; $\rho(\omega)$ is the ratio of the amplitude of the test sample signal to the reference signal; d is the thickness of the sample in M; c is the speed of light in M/s; ω is the angular frequency in rad/s.

Cluster analysis and PCA were conducted to process terahertz data such as absorption coefficient. The results could reflect the similarity between various coal and gangue samples. CA is a well-recognized mathematical analysis method that divides samples into several categories according to the internal similarity of different kinds of samples based on "birds of a feather flock together" [13]. During the analytical process, the Euclidean distance between various samples is calculated according to the characteristics of the sample data, and the two samples with the smallest Euclidean distance are formed into a new class. Finally, all samples are combined into one class. Indeed, samples belonging to the same type exhibit similarities. PCA is a multivariate statistical method that transforms multiple variables into core variables by linear transformation [14, 15]. These core variables are called principal components and are not related to each other. The information in the principal component is not repeated, and the first principal component contains the largest amount of information. By analyzing the histogram for the first principal component, the relationship between the samples can be assessed and further used for classification. Overall, CA and PCA, based on logical operations, can be used to establish simple models to extract important information that cannot be directly expressed in the original data. Importantly, these two methods can simplify complex sample data.

Results and discussion. *Terahertz time-domain spectroscopy of coal and gangue.* The experimental samples were measured according to the predefined experimental methods, and the terahertz time-domain spectral data of several coal and gangue samples were obtained. As seen in Fig. 2a, the terahertz spectra of different samples exhibited different delays and amplitudes. Among these, the main peak of ZBM111A appeared the earliest at 17.69 ps, with an amplitude of 0.089 V. Compared with ZBM111A, ZBM097A exhibited a longer time delay, with a main peak time of 18.09 ps and an amplitude of 0.082 V. The wave velocity and optical path of terahertz pulses passing through the measured object were different owing to the different refractive index of each sample, resulting in different times for the main peak of each sample to appear. Figure 2b shows the variation in the power of several samples over different frequencies. The peak energy consumption was observed when the pulse passed through ZBM111C within the frequency range 0.5–3.5 THz, whereas the energy consumption of ZBM111A, ZBM102, ZBM097A, and ZBM110B decreased gradually. According to Fig. 2, the time-domain spectral waveforms of five kinds of samples exhibited different delays. Given that the energy consumption of the THz pulse was different in all types of samples, their amplitudes exhibited different degrees of attenuation. Most importantly, our findings demonstrate that it is feasible to identify coal and gangue by terahertz time-domain spectroscopy.

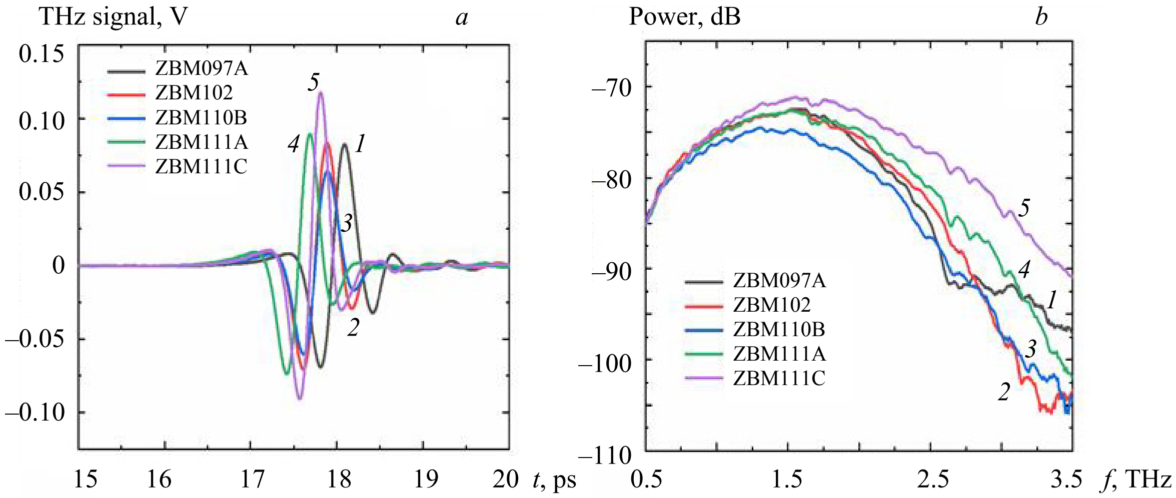

Fig. 2. Time-domain spectra (a) and power spectra (b) of the five samples:
ZBM097A (1), ZBM102 (2), ZBM110B (3), ZBM111A (4), and ZBM111C (5)

The absorption coefficient of the sample was obtained by substituting the time domain spectral data of the sample into the calculation Eq. (3). The waveforms of absorption coefficients of five samples within the frequency range 0.5–3.5 THz are shown in Fig. 3 in descending order. The absorption coefficient of sample ZBM110B was the largest, followed by that of ZBM102. The absorption coefficients for ZBM097A, ZBM111A, and ZBM111C were the lowest, with no significant differences among them.

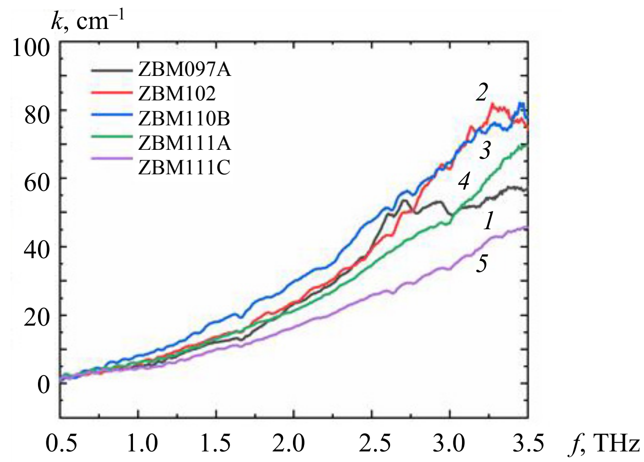

Fig. 3. Absorption coefficient (k) of the five samples: ZBM097A (1), ZBM102 (2), ZBM110B (3), ZBM111A (4), and ZBM111C (5).

Classification of coal and gangue. CA and PCA were used to classify different types of coal and gangue. During CA, each sample was divided into groups, and the distance between each group of samples was calculated. Then, the two samples with the smallest distance were merged into a new class according to the predefined experimental protocol, and the distance between the new class and other samples was calculated, and the two groups with the smallest distance were merged. This process was repeated until all samples were merged into one group. In this study, the five samples were clustered in four steps, and a dendrogram reflecting the similarity and dissimilarity of the five samples was obtained.

The power spectra of various samples within the frequency range 0.5–2.5 THz were used as the input set, and a tree diagram was obtained based on the power spectrum by CA (Fig. 4a). The Euclidean distance between ZBM097A and ZBM102 was the smallest among the five samples (10.82). They exhibited the highest similarity and were divided into one category. The Euclidean distance between ZBM111A and the new class was slightly increased. Moreover, as seen in Fig. 2b, the power spectrum curves corresponding to ZBM097A, ZBM102, and ZBM111A almost overlapped. ZBM110B and ZBM111C were also clustered with the new class obtained from the previous step.

The power spectrum of the sample was analyzed by PCA, and the first three principal components extracted from the power spectrum were retained. The contribution rate of the first three principal components to the original data was 98.64% and the first principal component accounted for 81.18%. Figure 4b shows the PC1 scores histogram of power spectra of various samples. During PCA, the importance of each principal component was characterized by the corresponding contribution rate. The greater the contribution rate, the more information the principal component reflects. The similarity between the two samples was related to the PC1 score of the samples. Two samples with small differences in PC1 score were similar. Conversely, samples with large differences in PC1 score tended to be different. According to Fig. 4b, the PC1 deviation between ZBM097A and ZBM102 was the smallest, and the PC1 score of ZBM111A was slightly higher than for the former two. ZBM111C exhibited the largest positive PC1 score, and ZBM110B the smallest negative PC1 score. Combining CA and PCA results, we found that the Euclidean distance and PC1 deviation between ZBM097A and ZBM102 was the smallest, and they were merged to form a new class. The Euclidean distance and PC1 score deviation between ZBM111A and the new class increased slightly. The Euclidean distance and PC1 score deviation between ZBM111C and the new class was the largest, followed by ZBM110B. After CA and PCA were applied, the five samples could be divided into three categories; ZBM097A, ZBM102, and ZBM111A belonged to the same class, whereas ZBM110B and ZBM111C could be classified individually. At the same time, we found that the similarity of the samples shown by the dendrogram was consistent with the PC1 scores.

Similarly, the absorption coefficients of various samples within the frequency range 0.5–2.5 THz were taken as the input data, and the CA and PCA methods were used to analyze them. The tree diagram and its PC1 score histogram based on the absorption coefficient spectrum were obtained (Fig. 5). As seen in Fig. 5a, the minimum distance between ZBM097A and ZBM102 was 22.4, and the Euclidean distance between the

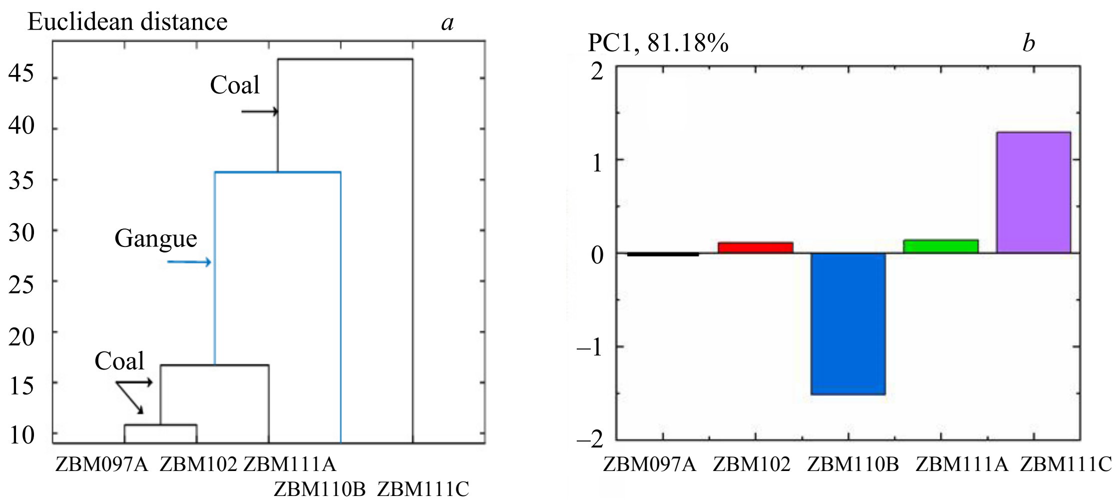

Fig. 4. The Euclidean distance dendrogram (a) and PC1 score histogram (b) with power spectra as the input.

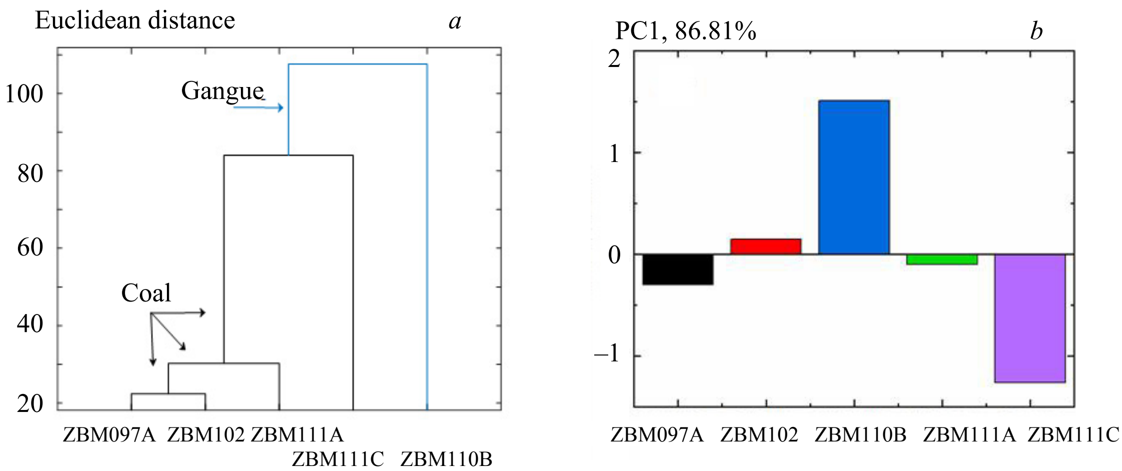

Fig. 5. The Euclidean distance dendrogram (a) and PC1 score histogram (b) with the input of absorption coefficient spectra.

new class and ZBM111A slightly increased. ZBM111C and ZBM110B were merged with the new class and the Euclidean distance between the new class and ZBM110B was the largest. Our results suggest that CA analysis based on absorption coefficient can clearly aggregate ZBM097A, ZBM102, and ZBM111A and distinguish between ZBM111C and ZBM110B.

For the PCA of the absorption coefficient of samples, the contribution rate of the first three principal components to the original data was 98.06%, and the contribution rate of the first principal component was 86.81%. The PC1 score for ZBM110B (Fig. 5b) differed from the other samples. By comparing Fig. 5, we found that ZBM110B was an individual class during CA and PCA analyses. Given that the PC1 scores for ZBM097A and ZBM102 were very close, they had some similarities, consistent with the results of CA analysis. Accordingly, the five samples could be divided into three categories using the absorption coefficient as the input data, combined with the CA and PCA analysis model. ZBM110B was the first group, ZBM111C was the second group, and the other three samples were classified into the third group. These results provide compelling evidence for the differences between ZBM110B and coal, and they could be separated to enable coal and gangue classification. The classification results based on the power spectrum and the absorption coefficient are shown in Table 2.

TABLE 2. Classification Results Based on Power Spectrum and Absorption Coefficient

The input data of the CA-PCA model	Classification results		
	Category 1	Category 2	Category 3
Power spectra	ZBM111C (Coal)	ZBM110B (Gangue)	ZBM097A (Coal), ZBM102 (Coal), ZBM111A (Coal)
Absorption coefficient spectra	ZBM110B (Gangue)	ZBM111C (Coal)	ZBM097A (Coal), ZBM102 (Coal), ZBM111A (Coal)

For different kinds of coal and gangue samples, the response in the terahertz spectrum was different. The similarity of coal and gangue was classified using CA and PCA methods. The results showed that the Euclidean distance and PC1 score corresponded with the properties of different samples. Although the power spectrum and absorption coefficient spectra of coal and gangue were different, the statistical method used improved our recognition ability and provided a more direct and effective method of identifying different types of coal and gangue. Based on the above results, a large amount of coal and gangue can be measured by terahertz time-domain spectroscopy technology; then, CA-PCA clustering can be carried out to build a rich database to provide a standard for coal and gangue identification in different areas. In a nutshell, we provided compelling evidence that terahertz time-domain spectroscopy is a fast and effective method of identifying coal and gangue and of classifying coal types.

Conclusions. It is feasible to use terahertz technology and multivariate statistical methods to identify coal and gangue. Within the range 0.5–2.5 THz, differences in the power spectrum and absorption coefficient exist in all kinds of samples, and these differences can be amplified by principal component analysis and cluster analysis methods to enable the recognition of different kinds of coal and gangue samples. Importantly, the Euclidean distance in the cluster analysis dendrogram corresponds to the PC1 score in the principal component analysis histogram, which can be harnessed to reflect the similarity and difference between various samples. However, this method exhibits limited ability to distinguish between different kinds of coal and gangue, emphasizing the need to develop and study more kinds of coal and gangue and to establish a huge database of coal and gangue based on the terahertz spectrum for future studies.

Acknowledgments. This work was supported by financially by The National Natural Science Foundation of China (Grant No. 52074273), the Key Natural Science Research Project for Colleges and Universities of Anhui Province (Grant No. KJ2020A0027), and the Natural Science Foundation of Anhui Province (Grant No. 2008085MF211).

REFERENCES

1. H. Y. Ge, Y. Y. Jiang, Z.H. Xu, F. Y. Lian, Y. Zhang, S. H. Xia, *Opt. Express*, **22**, 12533–12544 (2014).
2. Y. F. Hua, H. J. Zhang, *IEEE T. Microw. Theory*, **58**, 2064–2070 (2010).
3. C. B. Reid, G. Reese, A. P. Gibson, V. P. Wallace, *IEEE J. Biomed. Health*, **17**, 774–778 (2013).
4. Y. B. Ji, C. H. Park, H. Kim, S. H. Kim, G. M. Lee, S. K. Noh, T. I. Jeon, J. H. Son, Y. M. Huh, S. Haam, S. J. Oh, S. K. Lee, J. S. Suh, *Biomed. Opt. Express*, **6**, 1398–1406 (2015).
5. A. A. Gowen, C. O'Sullivan, C. P. O'Donnell, *Trends Food Sci. Tech*., **25**, 40–46 (2012).
6. H. L. Zhan, J. F. Xi, K. Zhao, R. M. Bao, L. Z. Xiao, *Food Control*, **67**, 114–118 (2016).
7. S. Wietzke, C. Jansen, T. Jung, M. Reuter, B. Baudrit, M. Bastian, S. Chatterjee, M. Koch, *Opt. Express*, **17**, 19006–19014 (2009).
8. J. Neu, C. A. Schmuttenmaer, *J. Appl. Phys*., **124**, 231101 (2018).
9. H. L. Zhan, K. Zhao, L. Z. Xiao, *Energy*, **93**, 1140–1145 (2015).
10. M. Mumtaz, M. A. Mahmood, A. Shahzad, S. D. Khan, M. A. Zia, M. Ahmed, I. Ahmad, *J. Infrared, Millim. Te*., **41**, 1181–1188 (2020).
11. D. Wang, Y. Zhang, J. Han, X. Li, X. F. Chen, T. Z. Qiu, H. Chen, *Sci. Rep.*, **11**, 13209 (2021).
12. T. D. Dorney, R. G. Baraniuk, D. M. Mittleman, *J. Opt. Soc. Am. A*, **18**, 1562–1571 (2001).
13. F. Wang, S. Guo, J. F. Zhao, H. Y. Xia, R. M. Bao, H. L. Zhan, J. N. Wang, *Spectrosc. Spectr. Anal*., **38**, 3638–3644 (2018).
14. H. L. Zhan, S. X. Wu, R. M. Bao, L. N. Ge, K. Zhao, *Fuel*, **143**, 189–193 (2015).
15. H. Park, J. H. Son, *Sensors*, **21**, 1186 (2021).