IOSR Journal Of VLSI And Signal Processing (IOSR-JVSP)
Volume 15, Issue 2 (March. - April. 2024), Pp.01-09
E-ISSN: 2319-4200, P-ISSN: 2319-4197
www.Iosrjournals.Org

Qualitative Analysis Of Fourier Infrared Spectroscopy Based On The EMD Algorithm

HU Yi-Ge, YU Bo, Sui Feng, YAO Zhi-Wen, Li Wen-Bo
(College Of Science, China Three Gorges University, China)
(China Shipbuilding Industry Group Co. Ltd Alphapec Instrument (Hubei), China)

Abstract:
In recent years, the application of Fourier infrared spectroscopy has become increasingly widespread. However, there is still room for improvement in its qualitative analysis for substance identification. Therefore, this paper proposes an infrared spectroscopy qualitative analysis method based on the Empirical Mode Decomposition (EMD) algorithm. The EMD algorithm is used to denoise the data, cubic spline interpolation is applied for baseline correction, and the data is normalized using the 2-norm. The error is then defined using the 1-norm, and accurate substances are identified. Experimental results show that this method achieves a high accuracy in substance identification of the detected data, reaching nearly 100%.

Key Word: *Fourier Infrared Spectroscopy; Substance Identification; EMD Algorithm; Baseline Correction; Cubic Spline Interpolation*

Date of Submission: 24-03-2025 Date of Acceptance: 04-04-2025

I. Introduction

In recent years, the application of Fourier Transform Infrared (FTIR) spectrometers has become increasingly widespread. FTIR spectroscopy is not only extensively used in the field of chemistry but also plays a significant role in related disciplines such as medicine[1]-[3], food science[4], and environmental science[5]. FTIR technology can be used for qualitative or quantitative analysis of samples. Its primary principle[6] is based on the absorption characteristics of substances to infrared radiation at different wavelengths, enabling the analysis of molecular structures and chemical compositions. As a modern analytical instrument, the FTIR spectrometer is highly characteristic for most substances, capable of analyzing changes in the chemical structure of materials and the presence of characteristic functional groups. Moreover, it is not limited by the state of the sample, meaning that the sample can be solid, liquid, or gas, and the compounds can be organic, inorganic, or polymers. With advantages such as ease of operation and rapid analysis speed, FTIR spectroscopy holds significant advantages in the identification of material components and quantitative analysis.

Recent advancements in infrared spectroscopy have improved qualitative and quantitative substance analysis. For instance, Ye Shubin et al. [7] developed a method combining the Least Absolute Shrinkage and Selection Operator (LASSO) with Linear Cyclic Least Squares (LCLS) to identify unknown gas components. Their experiments demonstrated the method's effectiveness even in spectral bands with significant interference. Similarly, Ye Shubin et al.[8] applied Principal Component Analysis (PCA) combined with Probabilistic Neural Network (PNN) and Back Propagation Artificial Neural Network (BPANN) to classify fumes from different edible oils. Zha Lixia et al.[9] addressed wavenumber shifts using water peak data and studied model transfer effects, emphasizing the importance of wavenumber accuracy.

Despite these advancements, challenges remain in distinguishing subtle signal differences between substances, particularly in large databases. Moreover, proprietary data processing methods used by foreign instrument companies necessitate independent research into FTIR data analysis. This paper explores alternative spectral methods, such as Raman spectroscopy[10] and terahertz time-domain spectroscopy [11]-[13] and proposes a novel recognition method based on the EMD algorithm for identifying chemical components in various environments.

During the data acquisition process, interference from factors such as the instrument's preheating state, ambient temperature, and humidity is inevitable, leading to the inclusion of noise in the test data. To mitigate the contamination of the true information by noise, we first preprocess the data [14]. Traditional noise reduction methods, such as Fourier transform and wavelet transform, are not suitable for the data addressed in this paper. Therefore, this paper adopts the EMD method proposed by Huang et al.[15] for analyzing nonlinear and non-stationary data. This method decomposes complex datasets into a finite and often small number of Intrinsic Mode

Functions (IMFs), which possess good Hilbert transform properties. The decomposition is adaptive, making it highly efficient for signal processing. Since it is based on the local characteristic time scale of the data, it is applicable to nonlinear and non-stationary processes. Through the Hilbert transform, the IMFs produce instantaneous frequencies that vary over time, clearly identifying the embedded structures. The final result is presented as an energy-frequency-time distribution, known as the Hilbert spectrum.

The main conceptual innovation of this method is the introduction of IMFs based on the local characteristics of the signal, giving practical significance to the instantaneous frequency. Additionally, the use of instantaneous frequency for complex datasets eliminates the need for spurious harmonics to represent nonlinear and non-stationary signals. This method significantly aids in denoising the data. After denoising, the data is normalized using the L2 norm to ensure consistent energy levels across all analyzed data and the database, facilitating easier comparison. Baseline correction is then performed using cubic spline interpolation to determine the baseline and address baseline drift in the collected data. Finally, the L1 norm is used to define the error, and based on the magnitude of the error, the most similar substance is identified from the database.

In the second section of this paper, the methods applied in data preprocessing are introduced. First, the principle of EMD [16] and its application steps are presented. This is followed by an explanation of the basic principles of normalization. Next, the concepts of cubic spline interpolation [17] and baseline correction are discussed. Finally, the error is defined using the L1 norm, and the algorithmic principles and steps are outlined. The third section provides numerical experiments to demonstrate the high accuracy of the substance identification method proposed in this paper. The fourth section offers a brief summary of the paper.

II. Problem Description

Spectral data acquisition is often affected by factors such as instrument preheating, ambient temperature, and humidity, leading to baseline drift, random noise, and light scattering. These issues significantly impact the accuracy of qualitative analysis. In this study, the database contains 475 substances, some of which exhibit minimal differences in their spectral characteristics. For example, Figure 1 compares the spectral characteristics of "Methyl Propionate" and "Ethyl Propionate," while Figure 2 compares "Ethyl Acetate" and "Propyl Acetate." These figures reveal nearly identical peak positions and side lobe sizes, making substance identification highly challenging.

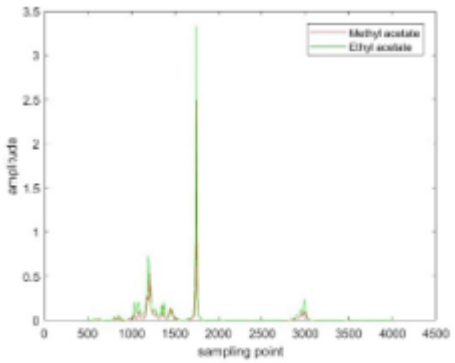

Figure.1 A comparison diagram of methyl propionate and ethyl propionate in the database

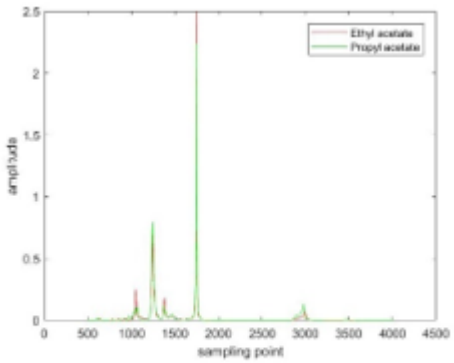

Figure.2 A comparison diagram of ethyl acetate and ethyl propionate in the database

To address these challenges, this paper proposes a qualitative analysis method based on the EMD algorithm. The first step is to standardize the data lengths of all target substances in the database and the test data to ensure consistency. Next, the EMD algorithm is used for noise reduction. EMD decomposes the signal into multiple IMFs, with noise primarily concentrated in the high-frequency IMFs and the effective signal in the low-frequency IMFs. By removing the high-frequency IMFs, the signal is effectively denoised.

EMD Algorithm and Its Noise Reduction

The EMD algorithm decomposes complex signals into multiple IMFs, enabling detailed analysis of signal characteristics. It is an adaptive signal decomposition method that does not require predefined basis functions.

The decomposition steps are as follows:

(1) Identify all the extreme points (maxima and minima) of the original signal.
(2) Use cubic spline interpolation to fit the upper and lower envelopes, denoted as e_{max} and e_{min}.
(3) Calculate the mean of the upper and lower envelopes to obtain the mean envelope m.
(4) Subtract the mean envelope from the original signal to obtain the intermediate signal: $h = x - m$.

(5) Determine whether the intermediate signal satisfies the two conditions of an IMF:

a) The number of extreme points and the number of zero crossings must be equal or differ by at most one over the entire data segment.

b) At any point in time, the mean value of the upper envelope formed by local maxima and the lower envelope formed by local minima must be zero, meaning the upper and lower envelopes are locally symmetric with respect to the time axis.

If the conditions are satisfied, the signal is considered an IMF; if not, repeat steps (1) to (4) until the two IMF conditions are met, and denote this IMF as h_1. Subtract this IMF from the original signal to obtain a new original signal, and return to step (1). Repeat steps (1) to (5) k times until the IMF screening stopping criterion is satisfied. The stopping criterion for IMF screening is as follows:

$$\frac{\sum_{n=1}^{N} | h_k(n) - h_{k-1}(n) |^2}{\sum_{n=1}^{N} | h_k(n) |^2} \le C_r \qquad (1)$$

Where N is the length of the data, and C_r is a constant set as the stopping condition, typically $C_r \in (0.2,0.3)$.

After decomposing the signal using EMD, we obtain several IMFs. By removing the first two IMFs, which are the high-frequency IMFs, and then summing the remaining IMFs and the residual signal, we obtain the denoised signal.

Baseline Correction and Cubic Spline Interpolation

Baseline drift in spectral data, caused by environmental factors, distorts spectral peaks and reduces analysis accuracy. This paper uses cubic spline interpolation for baseline correction. After denoising, valid local minima points are identified, and a smooth baseline curve is generated using cubic spline interpolation. Subtracting this baseline from the original data corrects the baseline drift.[18]

Given the significant role of cubic spline interpolation in both EMD denoising and baseline correction, we provide a brief overview of the basic principles of cubic spline interpolation for the convenience of readers. Specifically, cubic spline interpolation uses low-degree polynomials to interpolate data within subintervals and ensures the smoothness of the entire function by maintaining continuity and differentiability at the endpoints of the subintervals. Its definition is as follows: For a set of known data points (x_i, y_i),where $i = 0,1, \dots, n$,cubic spline interpolation finds a function $S(x)$ that satisfies the following conditions:

(1) $S(x)$ is a cubic polynomial on each subinterval;

(2) At the endpoints of two adjacent subintervals, the first and second derivatives are equal, i.e.,$S'_+(x_i) = S'_-(x_i)$,$S''_+(x_i) = S''_-(x_i), i = 2, \dots, n-1$. Here, we have two degrees of freedom, and different uses of these degrees of freedom yield different types of cubic splines, such as perfect splines;

(3) $S(x)$ passes through all the given data points, i.e., $S(x_i) = y_i, i = 1, \dots, n$.

The main steps of the cubic spline interpolation method are as follows:

(1) Fit a cubic polynomial within each small interval;

(2) Set boundary conditions to obtain a system of equations;

(3) Solve the system of equations to determine the coefficients for each small interval;

(4) Given an interpolation point, use the coefficients to calculate the corresponding function value.

Secondary Denoising and Normalization

After baseline correction, the test data undergoes a second round of EMD denoising, where the first IMF (high-frequency noise) is removed. The data is then normalized using the L2-norm to eliminate amplitude variations, ensuring consistent energy levels for comparison with the database.

A common approach to normalization is to use the L2 norm of a vector. The idea is to divide each element of the original vector by the L2 norm of that vector, resulting in a new vector. Each element of this new vector is a normalized value of the corresponding element in the original vector, and the L2 norm of this new vector is 1. Specifically, if we have a vector x with elements $x_i (i = 1,2,..,n)$ the L2 norm of vector x can be calculated using the following formula: $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.The elements of the normalized vector $x'$are obtained by dividing each element of the original vector by its L2 norm: $x'_i = \frac{x_i}{\|x\|_2}$, Thus, each element of the new vector x' is a normalized value of the corresponding element in the original vector, and the L2 norm of the new vector is 1. By normalizing the data using the L2 norm, different features or samples become comparable in magnitude, and the impact of significant weight differences between features on the qualitative analysis results is mitigated.

Definition of Similarity

After processing the test data through the aforementioned steps, we compare the processed test data with the data in the database to select the most similar substance and determine the target for qualitative analysis. To achieve this, we introduce two methods for defining similarity. The first method is the Pearson correlation coefficient, commonly used in engineering, with its formula given as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i y_i - n\bar{x}\cdot\bar{y})}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}\sqrt{(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}. \quad (2)$$

The Pearson correlation coefficient is one of the most widely used correlation coefficients, extensively applied in natural and social sciences, particularly in regression analysis and data analysis. The value of the correlation coefficient ranges between -1 and 1. When $|r|$ is close to 1, it indicates a strong relationship between the two variables; when r is close to 0, it suggests a weak relationship. A positive correlation means that as one variable increases, the other also increases, while a negative correlation means that as one variable increases, the other decreases. When using the Pearson correlation coefficient, the higher the correlation value between two substances, the more similar they are considered to be.

The second definition of similarity is based on the L1 norm error. The L1 norm (also known as the Lasso norm) of a vector is defined as the sum of the absolute values of its elements. For a vector $x = [x_1, x_2, x_3, \dots, x_n]$, its L1 norm is defined as: $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$.The L1 norm is often used to describe the overall magnitude of the elements in a vector, without considering their signs or order. After preprocessing the data, the characteristic peaks and other feature information are well preserved. Using the L1 norm formula, we calculate the sum of the absolute differences between the corresponding vertical coordinates of the processed test data and the database data at the same horizontal coordinates. This sum represents the error between the substance data in the database and the test data under the L1 norm. The substance with the smallest error is considered the most similar, allowing us to identify the substance in the database that corresponds to the test data. Specifically, let g denote the column vector of the vertical coordinates of the processed test data, and $\tilde{g}$ denote the column vector of the vertical coordinates of the processed data corresponding to a substance in the database. The L1 norm error E_1 between the test data and the database data is calculated as:

$$E_1 = \sum_{m=1}^{n} | g(m) - \tilde{g}(m) | . \quad (3)$$

where m is the number of data points. After calculating the L1 norm error values between the test data and all the data in the database, we sort these errors in ascending order. The smaller the value of E_1, the more similar the two substances are considered to be. This allows us to identify the substance in the database that is most similar to the test data.

In the next subsection, we summarize the above process in the form of an algorithm.

Algorithm Description

After obtaining the database data and the test data, the first step is to standardize the lengths of both datasets. Due to differences in the actual collected data and variations among different substances in the database, the lengths of the data corresponding to each target substance in the database are inconsistent and do not match the length of the test data. Therefore, it is necessary to standardize their lengths to facilitate subsequent processing and analysis. Once the lengths of the test data and the database data are unified, the test data can be denoised to reduce noise interference. In this process, the EMD algorithm is used to remove noise while preserving the effective components of the original data. During the first denoising step, the EMD algorithm decomposes the signal into several IMFs, and the first two IMFs (high-frequency signals) are removed. The remaining signals are summed to form a new signal, completing the first denoising process.

After the first denoising step, baseline correction is performed on the denoised test data. Due to the changes introduced by baseline correction, the test data undergoes a second round of denoising. In this step, the EMD algorithm is applied again to decompose the signal into several IMFs, and the first IMF (high-frequency signal) is removed, completing the second denoising process. Since the database data lies above the zero axis, the processed data is adjusted to align near the zero axis for better comparison. Following these steps, the data is normalized using the L2 norm, which facilitates a more accurate comparison between the test data and the database data. Finally, similarity is defined using the L1 norm and the Pearson correlation coefficient, and the data with the highest similarity is selected from the database as the target substance. The specific algorithm is as follows:

Algorithm: Target Substance Identification Algorithm

Input: Test data x to be analyzed.

Output: The top ten substances in the database with the highest similarity to the test data after comparison.

1) Process the database data to standardize the lengths of the database entries and align their starting and ending points.
2) Process the test data by determining the starting and ending points of both the database data and the test data based on different scenarios, resulting in the processed test data x_{new1}.
3) Resample the processed database data and the test data x_{new1} to N points using cubic spline interpolation, resulting in the new test data x_{new2} (in this paper, we select $N = 8192$).
4) Perform baseline correction on the test data x_{new2} . First, decompose x_{new2} using the EMD algorithm to determine the number of IMFs. Remove the corresponding number of IMFs to complete the denoising process, resulting in a column vector g. Combine the first column of x_{new2} with g to form an $N \times 2$ matrix. Perform baseline correction using cubic spline interpolation to obtain x_{new3}.
5) Perform EMD denoising, zero-centering, and L2 normalization on the processed test data $x_{new3}$to obtain x_{new4}.
6) Compare the standardized processed test data $x_{new4}$with the database data to identify the top ten most similar substances.

III. Experimental Results And Analysis

In this paper, we used the portable Fourier transform infrared spectrometer ALPHAPEC5010 produced by China Shipbuilding Industry Group Co. Ltd Alphapec Instrument (Hubei) (spectral range: 500 cm^{-1} to 5000 cm^{-1}, maximum resolution: 1 cm^{-1}, signal-to-noise ratio: 45000:1, transmittance repeatability better than 0.5%T) to sample 30 groups of samples. These samples were used as the test data for spectral analysis.

To demonstrate the effectiveness of the algorithm proposed in this paper, we selected six representative groups of substances to visually illustrate the matching degree between the preprocessed data and the database data, as well as the matching degree of substances calculated using different methods. The test data are labeled as 1, 2, 3, 4, 5, and 6, respectively. After conducting the data experiments, we identified that test data 1 corresponds to ethyl acetate, test data 2 to salicylic acid, test data 3 to phenylbutazone, test data 4 to probenecid, test data 5 to benzene sulfonamide, and test data 6 to ibuprofen. Due to space limitations, this paper presents the final comparison results for these six test datasets but provides a detailed step-by-step comparison and analysis only for test data 1 and test data 6.

First, we standardized the lengths of the database data to ensure that the starting and ending positions of the database entries were consistent. Next, we aligned the length of the test data with that of the database data. After standardizing the lengths, we performed the first round of denoising on the test data. Using Empirical Mode Decomposition (EMD), we decomposed the test data signal into several Intrinsic Mode Functions (IMFs) and removed the first and second IMFs, which correspond to high-frequency signals 1 and 2. The comparison images of the data before and after processing are shown below.

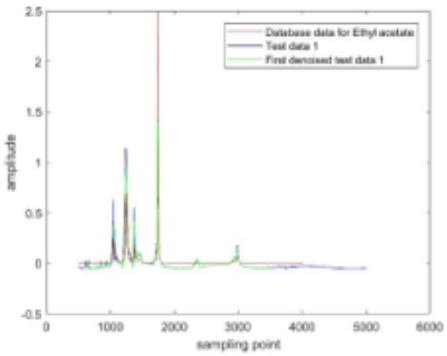

Figure 3: Comparison chart of original test data 1, first denoised test data 1, and database data of ethyl acetate

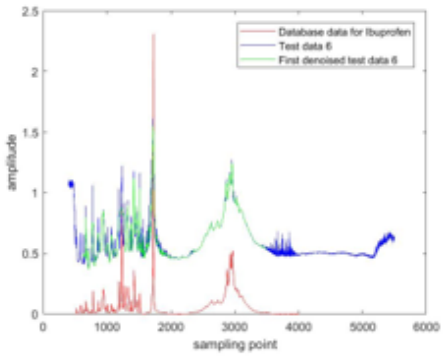

Figure 4: Comparison chart of original test data 6, first denoised test data 6, and database data of ibuprofen

Next, we performed baseline correction on the test data that had undergone the first round of denoising. During the baseline correction process, we used cubic spline interpolation along with valid maxima and minima to process the test data, resulting in baseline-corrected test data. Below is a comparison of the images before and after this processing step.

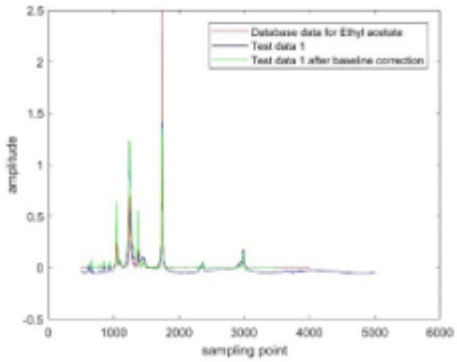

Figure5: Comparison chart of original test data 1, baseline corrected test data 1, and database data of ethyl acetate

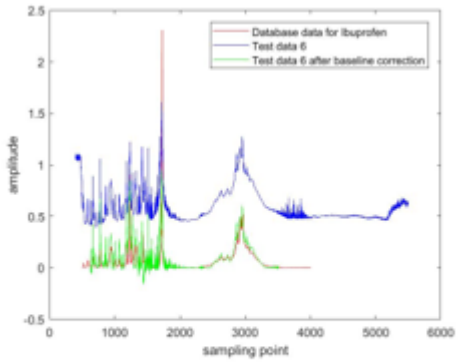

Figure 6: Comparison chart of original test data 6, baseline corrected test data 6, and database data of ibuprofen

After obtaining the baseline-corrected test data, we applied the EMD decomposition algorithm again to perform a second round of denoising on the test data. This time, we only removed the first IMF signal.

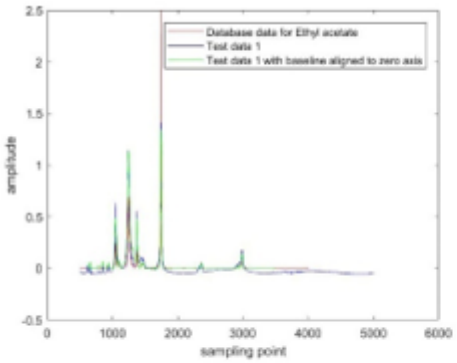

Figure 7: Comparison chart of original test data 1, second denoised test data 1, and database data of ethyl acetate

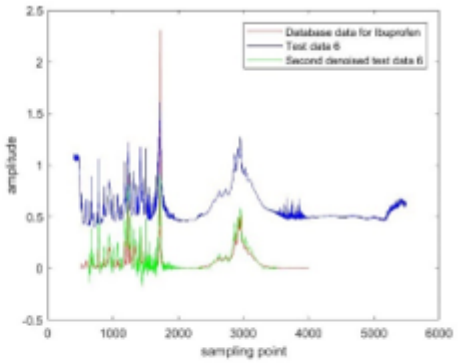

Figure 8: Comparison chart of original test data 6, second denoised test data 6, and database data of ibuprofen

After the second round of denoising, this paper performs non-negative processing on the obtained signal, which involves setting all negative values in the signal to zero. Following this processing, the most similar target substance data to the test data can be obtained using formulas (2) and (3), and the substance can be identified from the database.

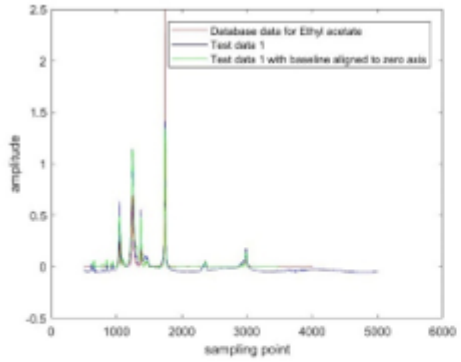

Figure 9: Comparison chart of original test data 1, zeroed test data 1, and database data of ethyl acetate

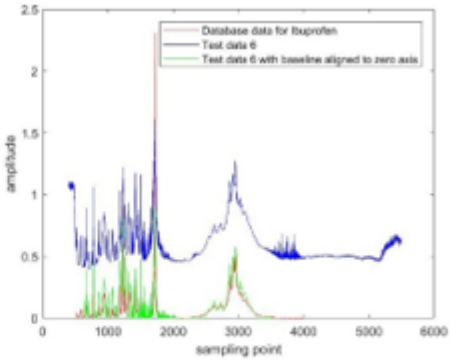

Figure 10: Comparison chart of original test data 6, zeroed test data 6, and database data of ibuprofen

Below are the comparison images of the processed test data for the six substances and the corresponding database data.

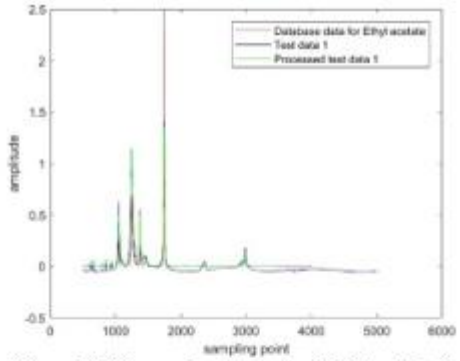

Figure 11: Comparison chart of original test data 1, processed test data 1, and database data of ethyl acetate

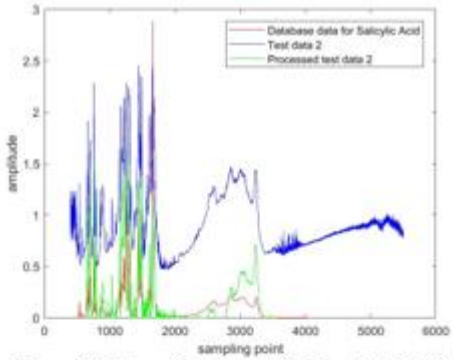

Figure 12: Comparison chart of original test data 2, processed test data 2, and database data of salicylic acid

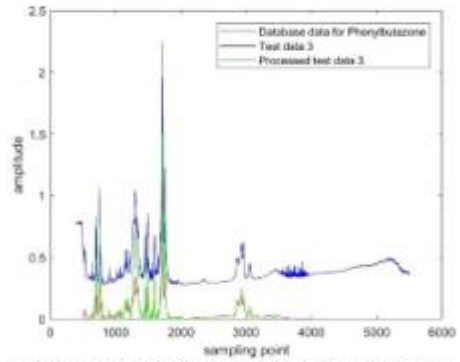

Figure 13: Comparison chart of original test data 3, processed test data 3, and database data of betamethasone

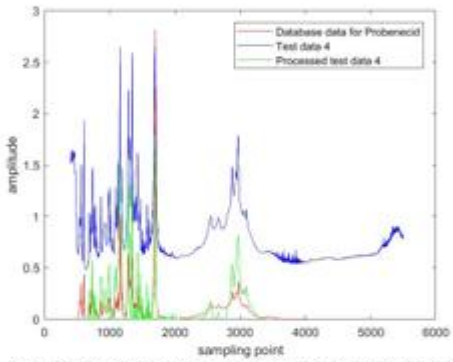

Figure 14: Comparison chart of original test data 4, processed test data 4, and database data of probenecid

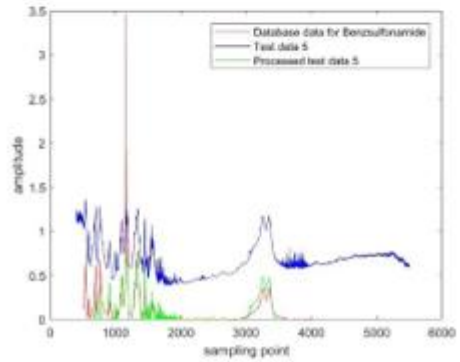

Figure 15: Comparison chart of original test data 5, processed test data 5, and database data of benzene sulfonamide

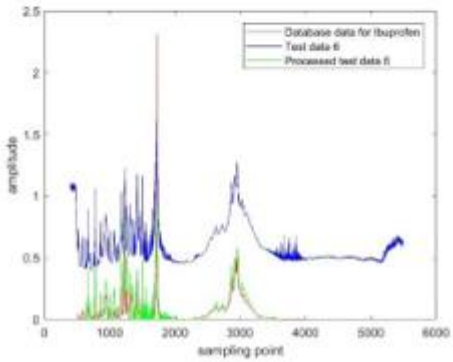

Figure 16: Comparison chart of original test data 6, processed test data 6, and database data of ibuprofen

From the data images, it can be observed that the algorithm proposed in this paper not only effectively removes noise interference but also preserves the characteristic signals of the data quite well. This lays a solid foundation for subsequent substance identification, ensuring that the follow-up work can proceed more effectively and thereby improving the accuracy of substance identification. Below, we present the results of the substance identification work in tabular form for analysis.

Tab.1 The one-norm error of test data 1 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Ethyl Acetate	14.91	0.8068
Propyl Acetate	16.17	0.8282
Benzyl Acetate	16.73	0.8424
Pentyl Acetate	16.99	0.8896
Methyl Acetate	17.26	0.7676
Diethyl Dimethylmalonate	20.51	0.7117
Diethyl Malonate	21.39	0.6528
Pentyl Acetate	23.11	0.5823

Tab.2 The one-norm error of test data 2 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Salicylic Acid	29.51	0.7655
Vanillin_121_33_5	38.36	0.7924
Probenecid	44.50	0.4324
Glutaric Acid	44.53	0.3746
Pyromellitic Acid	44.63	0.4143
Ascorbic Acid	44.70	0.5858
Atropine (free base)	45.02	0.4815
Succinic Acid	45.79	0.4141

Ethyl Butyrate	23.31	0.6593	Sebacic Acid	45.88	0.3336
Ethyl Propionate	23.81	0.5819	2-Amino-2-methyl-1-propanol	45.92	0.3714

Tab.3 The one-norm error of test data 3 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Phenylbutazone	21.28	0.8993
Di cyclohexyl Phthalate	26.71	0.7258
4-Phenylcyclohexanone	27.58	0.6977
9-Fluorenone	30.20	0.6923
Isopropyl Methyl Ketone	30.22	0.6429
Methyl propyl ketone	30.25	0.6280
1,3-Diphenylacetone	31.02	0.5880
2-Hexanone	31.77	0.5965
Diethyl Ketone	31.89	0.5611
Cocaine	32.28	0.6591

Tab.4 The one-norm error of test data 4 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Probenecid	26.83	0.8182
Sebacic Acid	36.54	0.6149
Caramel Furanone 54277 A	39.05	0.3747
Terephthalic Acid	39.18	0.7051
Phenylacetic Acid	39.63	0.5062
4-Isopropylbenzaldehyde	40.05	0.4250
Sulfolane	40.46	0.4292
2-Hexylacetophenone	41.17	0.2682
Isophorone	41.27	0.4102
Diethylene glycol dimethyl ether	41.30	0.3437

Tab.5 The one-norm error of test data 5 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Benzene sulfonamide	26.05	0.8516
Diphenyl sulfone	43.26	0.5223
Thiourea	43.27	0.2882
3-Aminopyridine	43.60	0.5887
Sulfolane	44.07	0.4795
Furfuryl alcohol	44.58	0.3865
Dimethyl sulfone	44.61	0.4759
4-Nitrophenol	44.65	0.6147
Benzhydrol	45.47	0.3593
1-Phenylethanol	45.62	0.5680

Tab.6 The one-norm error of test data 6 and the top ten substances in the database and the correlation coefficient value

Substance Name	E_1	r
Ibuprofen	22.69	0.8300
Atropine	35.17	0.6987
Glutaric acid	36.46	0.5730
Sebacic acid	36.62	0.5545
Pyromellitic acid	36.99	0.6581
Succinic acid	38.95	0.5548
Phenylacetic acid	40.03	0.4762
4-Isopropylbenzaldehyde	40.28	0.4690
Malonic acid	41.24	0.5563
Ammonium acetate	41.31	0.4742

From Tables 1 to 6, it can be seen that using the error value $E1$ defined by the L1 norm to compare magnitudes can accurately identify the substance corresponding to the test data in the database. The smaller $E1$ is, the more similar the test data is to the target substance in the database. From the data listed in the six tables, it is clear that using the defined error $E1$ to compare magnitudes achieves a 100% accuracy rate in substance identification. At the same time, when using the correlation coefficient formula to calculate the similarity between substances, the principle is that the closer the correlation coefficient value is to 1, the more similar the two substances are. However, the result analysis shows that the substances corresponding to the highest correlation coefficient values for test data 1 and test data 2 are not the actual substances matched in the database. Therefore, using the correlation coefficient to calculate substance similarity introduces a certain degree of error. This insight leads us to choose the error $E1$ defined by the L1 norm to determine the similarity between substances. From Figures 3 to 8, it can be observed that after denoising and baseline correction, the test data retains the signal characteristics well, helping us to better compare it with the database data and identify the matching target substance.

In summary, the EMD algorithm proposed in this paper demonstrates excellent noise reduction performance. It effectively removes both random and correlated noise while preserving characteristic signals. Additionally, the use of cubic spline interpolation for baseline correction not only fits the baseline but also further reduces noise, improving the signal-to-noise ratio. Therefore, this method proves to be highly effective in processing data for substance identification.

IV. Conclusion

This paper proposes a substance identification method based on the EMD algorithm for Fourier transform infrared spectroscopy. First, the database and test data are uniformly standardized. Next, the EMD method is used to remove noise interference from the test data, followed by baseline correction using cubic spline interpolation. Finally, an error metric is defined to identify the corresponding substance in the database. The data experiments conducted so far indicate that the accuracy of substance identification using this method can reach up to 100%.

References

[1]. CEHN Ya, JIANG Bin, ZENG Yuan-Er. Application Of Infrared Spectroscopy In The Identification Of Traditional Chinese Medicine[J]. Journal Of Guangzhou University Of Traditional Chinese Medicine (In Chinese), 2004, 21(3): 237-240.

[2]. LI Yan, WU Ran-Ran, YU Bai-Hua, Et Al. A Review On Applications Of Infrared Spectroscopy To The Study Of Traditional Chinese Medicine [J]. Spectroscopy And Spectral Analysis (In Chinese), 2006, (10):1846-1849.

[3]. DU Xiao-Wei, SONG Ping-Shun, NI Lin. Based On FT-IR Research On The Rapid Identification Method Of The Origin Of Traditional Chinese Medicine Codonopsis [J/OL]. Chinese Archives Of Traditional Chinese Medicine (In Chinese), 2024.

[4]. NIU Zhi-You, LIN Xin. Qualitative And Quantitative Analysis Method Of Tea By Near Infrared Spectroscopy [J]. Spectroscopy And Spectral Analysis (In Chinese) ,2009,29(09):2417-2420.
[5]. MENG Chao, MENG Guang-Zhen. Application Of FTIR Spectroscopy For The Detection Of Occupation Hygiene[J]. Spectroscopy And Spectral Analysis (In Chinese) ,1996,(02):123-127.
[6]. YANG Kun. Research And Application On Certain Number Of Core Technologies Of Fourier Transform Infrared Spectrometer[D]. Wuhan University (In Chinese), 2010.
[7]. YE Shu-Bin, SHEN Xian-Chun, XU Liang, Et Al. A Fast Qualitative Analysis Method Of Fourier Transform Infrared Spectra Base On LASSO Method[J]. Spectroscopy And Spectral Analysis (In Chinese) ,2017,37(10):3037-3041.
[8]. YE Shu-Bin, XU Liang, LI Ya-Kai, Et Al. Study On Recognition Of Cooking Oil Fume By Fourier Transform Infrared Spectroscopy Based On Artificial Neural Network[J]. Spectroscopy And Spectral Analysis (In Chinese) ,2017,37(03):749-754.
[9]. ZHA Li-Xia, ZHOU Xin-Qi, CHEN Lei, Et Al. Study On Model Transfer Of Fourier Transform Infrared Spectrometer Based On Wavenumber Calibration[J]. Analysis And Technology And Instruments (In Chinese) ,2022,28(01):37-44.
[10]. BAI He-Xuan, YANG Feng, LI Dan-Yang, Et Al. Multi-Component Substance Classification And Recognition Base On Surface-Enhanced Raman Spectroscopy[J]. Acta Optica Sinica (In Chinese), 2021, 41(20): 2024001.
[11]. CHEN Yan-Ling, CHENG Liang-Lun, WU Heng, Et Al. A Method Of Terahertz Spectrum Material Identification Based On Wavelet Coefficient Graph[J]. Spectroscopy And Spectral Analysis (In Chinese), 2021, 41(12): 3665-3670.
[12]. Platte F, Heise H M. Substance Identification Based On Transmission Thz Spectra Using Library Search[J]. Journal Of Molecular Structure (In Chinese), 2014, 1073: 3-9.
[13]. ZHANG Wen-Tao, LI Yue-Wen, ZHAN Pingping, Et Al. Recognition Of Transgenic Soybean Oil Based On Terahertz Time-Domain Spectroscopy And PCA-SVM, Infrared And Laser Engineering (In Chinese), 2017, 46(11): 1125004.
[14]. YE Wan-Ling, RAO Rui. Preprocessing Method For Vibration Signal Data [J]. Guangzhou Architecture (In Chinese),2024,52(05):11-15.
[15]. Huang N E, Shen Z, Long S R, Et Al. The Empirical Mode Decomposition And The Hilbert Spectrum For Nonlinear And Non-Stationary Time Series Analysis[J]. Proceedings Of The Royal Society Of London. Series A: Mathematical, Physical And Engineering Sciences, 1998, 454(1971): 903-995.
[16]. Boudraa, Abdel-Ouahab, And Jean-Christophe Cexus. EMD-Based Signal Filtering. IEEE Transactions On Instrumentation And Measurement 56.6 (2007): 2196-2202.
[17]. XU Xiao-Yong, ZHONG Tai-Yong. Construction And Realization Of Cubic Spline Interpolation Function [J]. Automatic Measurement And Control (In Chinese),2006, (11):76-78.
[18]. Zhu Miao-Miao, Yu Bo, Yao Zhiwen. Baseline Correction Of Fourier Transform Infrared Spectroscopy Signals Based On Cubic Splines, Journal Of Applied Mathematics And Computation，Http://Dx.Doi.Org/10.26855/Jamc.2024.12.010