Terahertz Spectroscopic Material Identification Using Approximate Entropy and Deep Neural Network

Yichao Li[1], Xiaoping A. Shen[2], Robert L. Ewing[3], and Jia Li[4]
[1]School of Electrical Engineering and Computer Science, Ohio University Athens, OH 45701 USA
[2]Department of Mathematics, Ohio University, Athens, OH 45701, USA
[3]Sensors Directorate, Wright-Patterson Air Force Base, Dayton, Ohio 45433-7022, USA
[4]School of Engineering and Computer Science, Oakland University, Rochester, MI 48309
Emails: Xiaoping Shen<shenx@ohio.edu>, Jia Li< li4@oakland.edu>

***Abstract*— Terahertz spectroscopy and imaging are a rapidly developed technique with important applications in many areas, such as medical imaging, security, chemistry, biochemistry, astronomy, communications, and manufacturing, to name a few. However, terahertz spectroscopy and imaging produce excessively high dimensional data which is prohibitive for common methods developed in the area of image processing. In this paper, we report our recent study on a novel classifier based on feature extraction using approximate entropy (ApEn). The classifier is initiated by analyzing the complexity of the terahertz spectrum, which is then combined with a deep neural network for material classification. Experimental results show that approximate entropy based features have high sensitive for detecting metal matrix composites, the accuracy of identification is up to 96.3%. Related algorithms for ApEn feature extraction and material classification are illustrated. An optimal parameter-embedding dimension, subject to classification accuracy for ApEn is studied.**

Keywords—Terahertz Spectroscopy; Approximate Entropy; Deep Neural Network; Material Classification; Metal Identification.

I. INTRODUCTION

Traditionally the three major classes of materials are metals, polymers, and ceramics. Examples of these are steel, cloth, and pottery. These classes usually have quite different sources, characteristics, and applications. Materials are usually classified by their performance, physical/chemical properties, composition/structure and processing/synthesis. Nowadays, terahertz (Thz) spectroscopy and imaging have emerged as an innovative technology for material classification with unprecedented sensitivity [1]. Thz (1 Thz = 1012 Hz) radiation covers a part of electromagnetic (EM) spectrum where wavelengths range from 3 mm. to 30 μm. or frequency bands range from 0.1 to 10 Thz [2]. Thz waves can penetrate most of dielectric materials, e.g. polymers, paper, wood, and organic films [3], [4]. In the contrast, Thz radiation shows strong absorption by water, which makes it ideal for classifying normal tissues and cancer tissues since their water content are different [5].

Previous material classification researches based on Thz spectroscopy were studied in a small scale in terms of number of kinds of materials. In [4], Zhong et al. used minimum distance classifier and neural network methods to classify 4 explosive and biochemical materials with an average accuracy above 80%. In [6], Fitzgerald et al. used principal components analysis (PCA) and support vector machine (SVM) to classify normal breast tissue and breast cancer tissue with an accuracy of 92%. In [7], Zou et al. showed that PCA is a feasible method to classify materials (water, intralipid, and glucose were used; however, classification accuracy was not reported).

This paper presents a new methodology for material classification based on Thz spectroscopy, we employ Approximate Entropy (ApEn; it will be introduced in next section) as a dimensionality reduction and non-linear feature extraction method for Thz spectrum data. We then use the learnt features as inputs to a non-linear learning machine – deep neural network – in the hope of achieving state of the art accuracy for material classification problems.

In this preliminary study, we have analyzed 107 samples from 14 kinds of materials including 6 metal-containing materials (e.g. oxides) and 8 non-metal materials. We report that ApEn provides powerful features for detecting metal-containing materials with an accuracy of 96.3%. We also show, empirically, that an optimal window length (embedding dimension) for ApEn is 4, which can generally capture key properties of Thz spectrum given an overall 14 materials classification accuracy of 80.4%.

The paper is organized as the follows, Section II provides simple mathematical background of ApEn and deep neural network. Section III describes the experiment design and a brief introduction of the data sets used in these experiments. After that, the results of material classification are presented in Section IV. We conclude the paper and briefly discuss future works in Section V.

II. BACKGROUND

A. Approximate Entropy (ApEn)

The approximate entropy ApEn(l, r) was introduced by Pincus [8], which is a statistical measure quantifying the change of complexity and regularity for time-series data. For an N-dimensional time series, its ApEn depends on two parameters: the windows length (embedding dimension l), and the filtering

level threshold (r). A time series containing many repetitive patterns has a relatively smaller ApEn; a less predictable (i.e., more complex) process has a higher ApEn.

ApEn can be summarized as follows (for detailed discussion, we refer readers to [8]–[11]).

Step 1: Form a time series of data $\{u_{1j}, u_{2j}, \cdots u_{kj}\}$, where k is the length of time series.

Step 2: Assume integer l represent the length of compared run of data, and positive real number r repesent a filtering level.

Step 3: Form a sequence of vectors $v_{1j}, v_{2j}, \cdots v_{(k-l+1)j}$ in R^l, which is real l dimensional sapce defined by

$$v_{ij} = \{u_{ij}, u_{(i+1)j}, \cdots, u_{(i+l-1)j}\}$$

Step 4: Use the sequence $v_{1j}, v_{2j}, \cdots v_{(k-l+1)j}$ to construct, for each i, $1 \le i \le k-l+1$

$$C_i^l(r) = \frac{number\ of\ v_{hj}\ such\ that\ d[v_{ij} - v_{hj}] < r}{k-l+1},$$

where $d[v_{ij} - v_{hj}]$ is defined as

$$d[v_{ij} - v_{hj}] = \max_{q=0,\cdots,l-1} \left| u_{(i+q)j} - u_{(h+q)j} \right|.$$

The signal d represents the distance between the vectors v_{ij}, v_{hj}, given by the maximum difference in their the maximum difference in their respective scalar components.

Step 5: Define

$$\Phi^l(r) = (k-l+1)^{-1} \sum_{i=1}^{k-l+1} \log(C_i^l(r)).$$

Step 6: Define approximate entropy as

$$ApEn = \Phi^l(r) - \Phi^{l+1}(r),$$

Algorithms used to calculate the approximate entropy are developed in [8]–[11]. It can be formatted as Algorithm 1 shown below. The algorithm takes 3 inputs: a finite series U, a window length l, and a filtering (thresholding) level r. The output is a real number which can be used as measurement of complexity and randomness of the input series.

Algorithm 1: Algorithm for calculating approximate entropy

Function: ApEn (U, l, r)

Input : A vector U, a positive integer l, and a positive real number r

1 Construct a set of vectors $x_1, x_2, ..., x_{N-l+1}$, where $x_i = [U_i, U_{i+1}, ..., U_{i+l-1}]$, and N is the length of U.

2 Define

$$C_i^l(r) = \frac{\text{number of } x_j \text{ such that } dist(x_i, x_j) \le r}{N-l+1}$$

, where

$$dist(x, x^*) = \arg\max_a |U(a) - U(a^*)|$$

3 Define

$$\Phi^l(r) = (N-l+1)^{-1} \sum_{i=1}^{N-l+1} log(C_i^l(r))$$

Return : $\Phi^l(r) - \Phi^{l+1}(r)$

B. Deep Neural Network (DNN)

Deep learning architectures have successful applications in both visual and speech recognitions[12]. Deep learning methods allow a machine to automatically discover intrinsic relationships needed for classification. Deep neural networks are one of the basic types of deep learning architectures, which consists of an input layer, 2 or more hidden layers, and an output layer. Each layer consists of a list of neurons. Learning process is through backpropagation iterations. In each backpropagation iteration, gradients of the objective function with respect to each neuron can be calculated using the chain rule for derivatives. Therefore stochastic gradient descent (SGD) are used by most practitioners[12]. In this preliminary study, we used Adam[13], instead of SGD, to do objective optimization. Adam[13] requires little fine tuning and it is computationally more efficient than SGD.

Many well-known deep learning frameworks are available in Python, such as Theano [14], Torch, Caffe [15], TensorFlow [16]. In this study, we used **Keras** [17], which is a high-level deep learning API (application programming interface) on top of Theano or TensorFlow. The best feature about Keras is that the learning architecture can be implemented within a minute. The algorithm used in this study is shown below as Algorithm 2. The leave-one-out cross validation is implemented using scikit-learn [18].

Algorithm 2: Algorithm for material classification

Function: DNN_classification (X, Y, n)

Input : X is a feature matrix, Y is a label vector,n is the number of classes

import : keras,numpy,sklearn

```
loo = LeaveOneOut()
ACC = [ ]
for train_index, test_index in loo.split(X) do
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = Y[train_index], Y[test_index]
    model = Sequential()
    Add layers to model
    model.compile(loss =' categorical_crossentropy', optimizer =
      Adam(), metrics = ['accuracy'])
    model.fit(X_train, y_train)
    testing_acc = model.evaluate(X_test, y_test)
    ACC.append(testing_acc)
end
return mean(ACC)
```

III. Experiment Design

A. Terahertz spectroscopy dataset

We have collected 107 Thz spectrum samples of 14 materials (6 metal matrix composites: Zinc, Iron, Cobalt, Cadmium, Mercury, and Copper; 8 non-metal materials: Tetracene, Polyfilm, Polyethylene, Terephthalic Acid, Asphalt, Trehalose, Milk powder, Coffee) from the Thz database (http://www.thzdb.org/). The dimensionality of Thz spectrums ranges from 795 to 6349. The Thz database is currently the most comprehensive Thz spectroscopic database established by the NICT and RIKEN in Japan[19].

B. *Material classification*

Most of the samples have a large range of electromagnetic spectrum, from 0 – 19 Thz (10^{12} Hz). To identify optimal window size (denoted as L_{opt}) for material classification, we study the impacts of the parameter L with 36 different choices of r. Specifically, for each of the window size, $L=2^j$, j=0, 1, 2, 3, there are 36 different filtering levels $k*10^{i-4}$, i=0, 1, 2, 3, k=1,…,9. For each of the 4 settings, we apply the same deep neural network to do classification. We use leave one out cross validation to evaluate the classification accuracy. An overall workflow is shown in Figure 1.

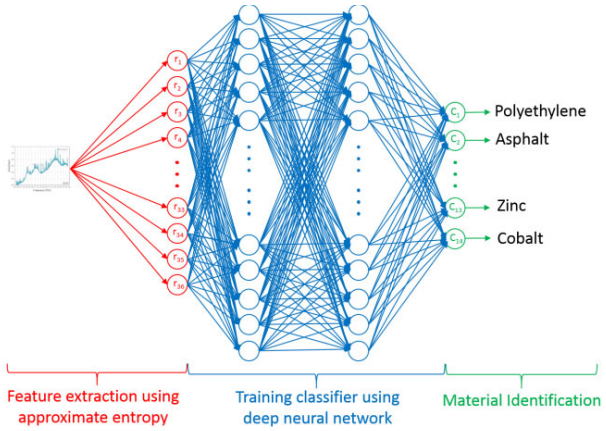

Figure 1: A deep neural network scheme is shown. The terahertz spectrum is first analyzed using approximate entropy, which is a feature extraction method to capture key properties (e.g. complexity, periodicity) of the spectrum. Totally 36 features are obtained by fixing the window length ($L=2^j$, j=0,1,2,3) and varying the filtering levels to $k*10^{i-4}$, i=0,1,2,3, k=1,…,9. Two hidden layers are used, each with 100 neurons and 0.1 dropout rate. The output layer contains 14 neurons for each class.

C. *Parameter Selection for ApEn*

There are two parameters for ApEn. One is the window length l, which is related to the embedding dimension of a time series. An overlapping partition of a time series with interval size l is determined. The complexity (use entropy as a measure) on each subinterval is calculated and compared to the consecutive subinterval to detect the change of complexity. The other one is the filtering level r, which controls the threshold of statistical bias. In this study, we focused on identifying the optimal window length, denoted as, L_{opt}. A simplified parameter selection flowchart is shown in Figure 2. Experiment results and discussion for threshold parameter will be reported in near future.

D. Equipment

A single PC was used for both approximate entropy feature extraction and deep neural network classification. The operating system is Windows 7 Service Pack 1. The CPU is Intel Core i7-4770K @ 3.50 GHz * 8. The memory is 16GB. Deep neural network classification was performed using Graphics Processing Unit (GPU) in GeForce GTX 770. The whole leave-one-out cross validation process can be done within 2 hours.

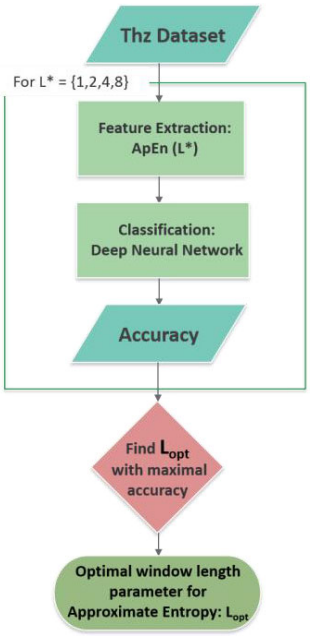

Figure 2: ApEn parameter selection scheme is shown. Four different values of window length are used. The same deep neural network is used. Accuracy is evaluated based on leave-one-out cross validation. Optimal window length is subject to maximal accuracy.

IV. RESULTS

Experimental results show that approximate entropy based features have high sensitive for detecting materials that contain metals, the accuracy of identification is up to 96.3%. Related algorithms for the classifier are illustrated above. Other experimental results related to thresholding parameter r, algorithmic development and computation complexity analysis will be reported in future.

A. *Identification of optimal window length for ApEn*

Our results show that the optimal window length is 4 for Thz spectrums (Figure 3). The overall best accuracy for 14 material classification is 80.4%.

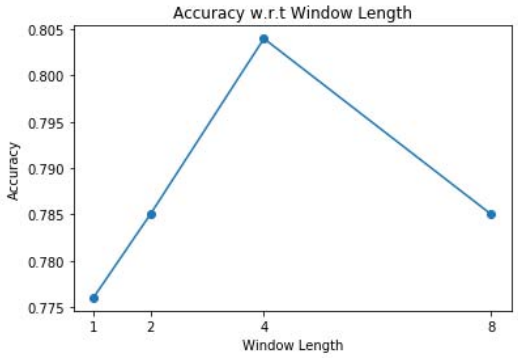

Figure 3: Classification accuracy when different ApEn window length is used. The accuracy is 77.6%, 78.5%, 80.4%, 78.5% for window length of 1, 2, 4, 8, respectively.

	Cobalt	Zinc	Cadmium	Mercury	Iron	Copper	Tetracene	Poly film	Polyethylene	Terephthalic Acid	Asphalt	Trehalose	Milk powder	Coffee
Cobalt	**24**	0	0	0	0	**1***	0	0	0	0	0	0	0	0
Zinc	0	**2**	**1***	**1***	0	0	0	0	0	0	0	0	0	0
Cadmium	0	**1***	**2**	0	0	0	0	0	0	0	0	0	0	0
Mercury	**2***	0	0	**0**	0	0	0	**2***	0	0	0	**1***	0	0
Iron	**2***	**1***	0	0	**0**	0	0	0	0	0	0	0	0	0
Copper	**1***	0	**1***	**2***	0	**0**	0	0	0	0	0	0	0	0
Tetracene	0	0	0	0	0	0	**19**	0	0	0	0	0	0	0
Poly film	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0
Polyethylene	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0
Terephthalic Acid	0	0	**1***	0	0	0	0	0	0	**9**	0	0	0	0
Asphalt	0	0	0	0	0	0	0	0	0	0	**6**	0	0	**1***
Trehalose	0	0	0	0	0	0	**1***	0	0	0	0	**5**	0	0
Milk powder	0	0	0	0	0	0	0	0	0	0	0	0	**2**	**1***
Coffee	0	0	0	0	0	0	0	0	0	0	**1***	0	0	**2**

Table 1: Confusion matrix for classification of 14 materials using approximate entropy with optimal window length of 4 and filtering level ranging from $k*10^{i-4}$, i=0,1,2,3, k=1,…,9. Rows are actual classes. Columns are predicted classes. Grey filled cells are materials containing metal. Double line cells are nonmetal materials. Diagonals are the number of correctly identified materials (bolded). Misclassified materials are marked using asterisk. The accuracy for detecting metals is 96.3%. The overall material classification accuracy is 80.4%.

B. *Approximate Entropy provides high accuracy for detecting metal-containing materials*

To illustrate the misclassified materials, a classification confusion matrix based on optimal window length is shown in Table 1. Fourteen materials are shown; the rows are actual classes and the columns are predicted classes. Diagonal cells are bolded, indicating that they are correctly identified materials. Misclassified materials are labeled using asterisk.

In terms of classification of metals vs. non-metals, our results show that ApEn features are very powerful, with an accuracy of 96.3%. Indeed, ApEn can capture the regularity and complexity of the spectrum, and metals and nonmetals do have large differences in their physical and chemical properties.

Notably, no samples from metal group (mercury, iron, or copper) are correctly separated from other members in the same group. In fact, most of them are predicted as another metal composite; while only 3 out of 5 mercury materials are classified as polyfilm and trehalose. Intriguingly, one terephthalic acid sample is classified as cadmium. These observation suggest that there is a great potential improvement for classifying among materials within the metal group.

C. *Classification accuracy varies with respect to number of classes*

Since multi-class classification problem is generally a much harder problem than binary classification problem, we ask how the accuracy will change with respect to the number of classes (by random grouping of existing classes). Figure 4 shows that, as expected, the accuracy drops when the number of classes increases. Based on this observation, we are working on developing a multiscale classifier with higher sensitivity for identify multiple classes. In order to benefit from the Thz images technology, the new multiscale identifier will be constructed to reflect information at molecular level.

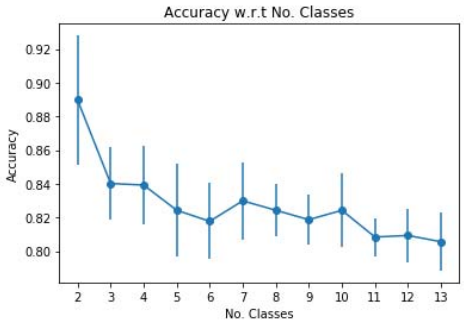

Figure 4: Classification accuracy varies with respect to the number of classes. Error bars are shown for each point. When number of classes increases, the accuracy will general go down.

V. CONCLUSION

Terahertz spectroscopy has found applications in many areas, such as material science, biology and medicine to name a few. Thz spectroscopy has advantages over other imaginary -domains by providing a measurement of intensity for the electric field produced by Thz waves, which can be used to identify sample structure [20]. This paper reports our initial study on the material classification using the Thz spectroscopy using non-linear learning algorithm –DNN. The novelty lays in developing a supervised feature selection with validation framework based on ApEn for DNN classifier. ApEn has two unknown parameters, l, window length, which is influenced by the length of the sample, and r, thresholding parameter, which depends on randomness and quality of data. In most of practical problems, the recommended values of r, in the range of 0.1-0.2 times the standard deviation of the signal, have been shown to be applicable for a wide variety of signals. However, in certain cases, r values within this prescribed range can lead to an incorrect assessment of the complexity of a given signal [21]. Our future work will focus on data preprocessing and developing a section rule for parameter r tailored for the material classification problems using Thz spectroscopy imaging to enhance the accuracy of the classifier.

ACKNOWLEDGMENT

The second author is sponsored by Air Force Office of Scientific Research (AFOSR) summer faculty fellowship program.

REFERENCES

[1] C. Balas, "Review of biomedical optical imaging—a powerful, non-invasive, non-ionizing technology for improving in vivo diagnosis," *Meas. Sci. Technol.*, vol. 20, no. 10, p. 104020, 2009.

[2] D. Abbott and X. C. Zhang, "Special Issue on T-Ray Imaging, Sensing, and Retection," *Proc. IEEE*, vol. 95, no. 8, pp. 1509–1513, Aug. 2007.

[3] N. FUSE, T. FUKUCHI, M. MIZUNO, and K. FUKUNAGA, "High-Speed Underfilm Corrosion Imaging Using a Terahertz Camera," *Electron. Commun. Japan*, vol. 99, no. 8, pp. 86–92, 2016.

[4] H. Zhong, A. Redo-Sanchez, and X.-C. Zhang, "Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system," *Opt. Express*, vol. 14, no. 20, pp. 9130–9141, Oct. 2006.

[5] X. Yang *et al.*, "Biomedical Applications of Terahertz Spectroscopy and Imaging," *Trends Biotechnol.*, vol. 34, no. 10, pp. 810–824, 2016.

[6] A. J. Fitzgerald, S. Pinder, A. D. Purushotham, P. O'Kelly, P. C. Ashworth, and V. P. Wallace, "Classification of terahertz-pulsed imaging data from excised breast tissue," *J. Biomed. Opt.*, vol. 17, no. 1, pp. 16005–16010, 2012.

[7] Y. Zou, P. Sun, and W. Liu, "Classification of materials using terahertz spectroscopy with principal components analysis," in *2015 40th International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-Thz)*, 2015, pp. 1–2.

[8] S. M. Pincus, "Approximate entropy as a measure of system complexity.," *Proc. Natl. Acad. Sci.*, vol. 88, no. 6, pp. 2297–2301, 1991.

[9] S. M. Pincus and A. L. Goldberger, "Physiological time-series analysis: what does regularity quantify?," *Am. J. Physiol.*, vol. 266, no. 4 Pt 2, pp. H1643-56, Apr. 1994.

[10] S. M. Ryan, A. L. Goldberger, S. M. Pincus, J. Mietus, and L. A. Lipsitz, "Gender- and age-related differences in heart rate dynamics: are women more complex than men?," *J. Am. Coll. Cardiol.*, vol. 24, no. 7, pp. 1700–1707, Dec. 1994.

[11] K. K. Ho *et al.*, "Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics," *Circulation*, vol. 96, no. 3, pp. 842–848, Aug. 1997.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[13] D. P. Kingma and J. Ba, "Adam: {A} Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[14] R. Al-Rfou *et al.*, "Theano: A {Python} framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[15] Y. Jia *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," *arXiv Prepr. arXiv1408.5093*, 2014.

[16] M. Abadi *et al.*, "{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems." 2015.

[17] F. Chollet and others, "Keras." GitHub, 2015.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[19] T. Notake, R. Endo, K. Fukunaga, I. Hosako, C. Otani, and H. Minamide, "State-of-the-Art Database of Terahertz Spectroscopy Based on Modern Web Technology," *IEEE Trans. Terahertz Sci. Technol.*, vol. 4, no. 1, pp. 110–115, Jan. 2014.

[20] S. Wang and X.-C. Zhang, "Pulsed terahertz tomography," *J. Phys. D. Appl. Phys.*, vol. 37, no. 4, p. R1, 2004.

[21] S. Lu, X. Chen, J. K. Kanters, I. C. Solomon, and K. H. Chon, "Automatic Selection of the Threshold Value R for Approximate Entropy," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1966–1972, Aug. 2008.