УДК 543.42 : 004.032.26

Применение методов машинного обучения для анализа наличия вредных примесей в атмосфере по спектральным данным

Ф.А. Кожевников[1], М.Р. Конникова[1], А.С. Синько[1], А.А. Ангелуц[1,2]

[1] Физический факультет МГУ им. М.В. Ломоносова
119991, Москва, РФ, Россия, Ленинские горы, 1 стр. 2, angeluts@physics.msu.ru
[2] Институт динамики систем и теории управления
664033, Иркутск, Россия, ул. Лермонтова, 134

Работа посвящена развитию комплексного подхода к анализу наличия в атмосферном воздухе вредных примесей. Этот подход предполагает: 1) использование результатов измерений спектров поглощения воздуха, содержащего вредные примеси на исследуемой трассе, полученных с использованием методов импульсной терагерцовой спектроскопии; 2) создание и использование нейронной сети для анализа полученных данных. Для обучения нейронной сети генерируются массивы модельных спектров поглощения газовой смеси с различным качественным и количественным составом. Применение нейросети к модельным наборам спектров продемонстрировало идентификацию шести газовых компонентов с концентрациями до 0,01 ppm (1 ppm = 0,0001%). Нейронная сеть достигла 90–95% точности при распознавании газов. Проведены серии экспериментов для реальных газов, показав чувствительность метода ТГц спектроскопии к малым концентрациям газа в смеси.

Ключевые слова: терагерцовая спектроскопия, нейросети, газовый анализ; terahertz spectroscopy, neural networks, gas analysis.

Введение

Контроль состояния окружающей среды является важным компонентом системы обеспечения безопасной жизнедеятельности человека. Развитие технологий обнаружения вредных газов играет важную роль во многих областях, таких как экологический атмосферный мониторинг, медицинская диагностика, управление производственными процессами, газовый мониторинг. Качественный и количественный анализ состава атмосферного воздуха требует наличия приборов газового анализа. Существует много подходов для построения таких приборов, причем многие из них предполагают необходимость применения одного прибора для обнаружения одного-двух газов в составе атмосферы [1]. Спектроскопические методы газового анализа лишены этого недостатка, так как анализ ведется в широком диапазоне частот [2], что позволяет обнаруживать любые газы, имеющие линии поглощения в этом диапазоне. Привлекательность таких методов связана с тем, что большинство атмосферных молекул имеют спектральные «отпечатки пальцев» в ближнем, среднем и дальнем диапазонах ИК-спектра. Современное состояние развития элементов аппаратуры терагерцового (ТГц) диапазона позволяет реализовать спектроскопический подход для газового анализа состава атмосферного воздуха. Несмотря на высокое поглощение ТГц-излучения водяными парами, существуют «окна прозрачности», в которых расположены интенсивные линии поглощения вращательного движения многих молекул, что позволяет однозначно определять наличие целевых веществ [3, 4].

Низкая энергия импульсных источников ТГц-излучения ограничивает отношение сигнал/шум до трех порядков. Это отношение можно увеличить за счет накопления и усреднения детектируемого сигнала, что требует увеличения времени измерений. В современных импульсных ТГц-спектрометрах этот недостаток устранен за счет использования малошумящих волоконных лазеров, а также сверхбыстрого усреднения несколько сотен спектров, добиваясь увеличения отношении сигнал/шум до шести порядков.

Наша работа посвящена применению импульсной ТГц-спектроскопии для анализа газовых смесей, основным компонентом которых будет атмосферный воздух. Решаемая в рамках этой работы проблема состоит в том, что в многокомпонентной газовой смеси на фоне линий поглощения паров воды нужно обнаружить присутствие линий поглощения других газов. Для решения проблемы была разработана аналитическая система на основе нейросетей, которая была обучена на большом наборе модельных спектров поглощения газовой смеси с учетом вариации концентрации примесных газов. Далее работа нейросети протестирована на выборке сгенерированных данных спектров поглощения газовой смеси.

Эксперимент

Исследование спектров поглощения газовых смесей осуществлялось методом импульсной терагерцовой спектроскопии на спектрометре Terasmart (Menlo Systems GmbH, Германия), источником и детектором ТГц-излучения, в котором служили фотопроводящие антенны. ТГц-спектрометр имеет динамический диапазон 95 дБ, который обеспечивается усреднением по 1000 измерениям, а также спектральный диапазон 0,15–

3,5 ТГц (5–117 см$^{-1}$). Время сканирования составляло 440 пс, что обеспечило частотное разрешение 2,3 ГГц (0,074 см$^{-1}$). ТГц излучение фокусировалось в газовую кювету системой из двух параболических зеркал, размер ТГц пучка на образце составляет около 500 мкм на уровне $1/e^2$.

Методика получения коэффициента поглощения из спектров пропускания подробно описана в работе [5]. Амплитудный коэффициент пропускания образца на частоте ω определяется как отношение комплексной амплитуды поля $E(\omega)$, прошедшего через образец, к амплитуде поля при отсутствии образца $E_0(\omega)$: $T(\omega) = |E(\omega)|/|E_0(\omega)|$. Отсюда коэффициент поглощения газа вычисляется по формуле:

$$\alpha(\omega) = [-2 \cdot \ln T(\omega) + \ln(1 - R)^2]/d,$$

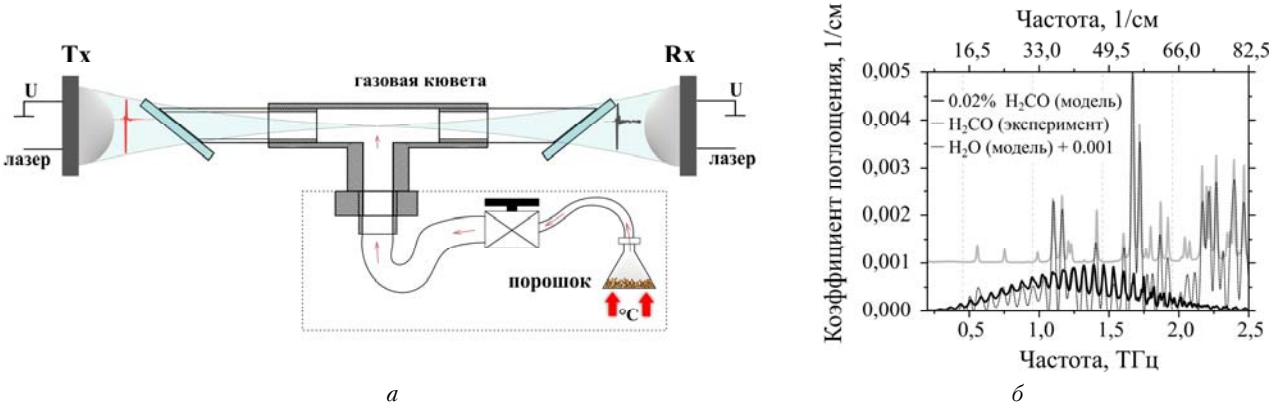
где d – длина образца (кюветы), R – коэффициент отражения поверхности образца. Вторым членом в этом выражении можно пренебречь, т.к. окна газовой кюветы установлены под углом Брюстера.

Экспериментально исследованы газовые смеси аммиака NH_3 и формальдегида H_2CO. Аммиак в виде 10% водного раствора (Genel, Россия) в объеме 150 мкл добавлялся в газовую кювету. Пары формальдегида получены методом возгонки из порошка 95–98% параформальдегида (Метафракс Кемикалс, Россия). В сосуд объемом 100 мл добавлялось 500 мг порошка параформальдегида, после чего сосуд плотно закрывался крышкой и соединялся трубкой длиной 80 см с измерительной кюветой. Сосуд с порошком устанавливался на плиту, нагретую до 150 °С, которая была изолирована от измерительной системы.

Измерительная газовая кювета, адаптированная для импульсного ТГц спектрометра, имела полипропиленовые окна, расположенные под углом Брюстера, который составляет 56° для показателя преломления ~ 1,5 [6], чтобы избежать эталонных эффектов. Также в кювете имелся разъем, позволяющий откачивать атмосферный воздух из кюветы или подключать ее к источникам газовых смесей. Объем кюветы составлял 42 мл при длине 130 мм. При измерениях газовая кювета помещалась в ТГц спектрометр между источником импульсного излучения Tx и детектором Rx (рис. 1, *а*).

Для исключения влияния поглощения ТГц излучения парами воды, присутствующими в атмосфере, спектрометр был закрыт кожухом, объем которого заполнялся сухим воздухом, что обеспечивало относительную влажность в камере до 7%.

В качестве примера на рис. 1, *б* приведены экспериментальный спектр воздушной смеси с примесями формальдегида и расчетный спектр формальдегида. Здесь же с небольшим смещением по вертикальной оси приведен модельный спектр атмосферы показывающий, что выбросы на экспериментальном спектре соответствуют парам воды. Сравнение кривых на рисунках показывает соответствие расчетов экспериментальным результатам.



а *б*

Рис. 1. Схема установки ТГц-спектрометра (*а*), ТГц-излучение генерируется в антенне источнике Tx, фокусируется в газовую кювету и регистрируется антенной детектором Rx; экспериментальный (пунктир) и модельный (линия) спектры формальдегида, для сравнения приведен спектр поглощения паров воды (серая линия) (*б*)

При нагревании порошка параформальдегида, H_2CO возгонялся по трубкам в газовую кювету, заполняя ее. В течение 2-х часов проведено 11 измерений спектров пропускания H_2CO, которые при пересчете в коэффициент поглощения $\alpha(\omega)$ показали уменьшение амплитуды $\alpha(\omega)$ со временем, что может быть частично связано с выпаданием H_2CO на стенки кюветы. На рис. 1, *б* приведены спектры коэффициента поглощения модельных спектров паров воды (80%) (серый) и формальдегида (0,02%) (черный), а также экспериментальный спектр формальдегида с наименьшей зарегистрированной концентрацией газа (пунктир).

Момент инерции или вращательная постоянная B H_2CO равна 1,29 см$^{-1}$ (38,83 ГГц) [7], что согласуется с частотой повторения эквидистантно отстоящих поглощательных линий, соответствующей 75,6 ГГц. Выражение для частот линий вращательного спектра, на которых будет происходить резонансное взаимодействие излучение с молекулой может быть записано, как

$$\nu = 2B(J + 1) - 4D(J + 1)^3,$$

где B – вращательная постоянная, D – постоянная линейного центробежного растяжения, J – вращательное квантовое число [4, 7].

Как видно, динамический диапазон спектрометра и частотное разрешение позволяет зарегистрировать газовую смесь, содержащую 0,02% формальдегида, что демонстрирует чувствительность экспериментального метода.

Нейросеть и модельные спектры

Особенностью анализируемых данных является наложение линий поглощения паров воды на спектральные особенности целевых газов, а также взаимное перекрытие линий различных примесных компонентов, что обусловило выбор методов машинного обучения для решения поставленной задачи идентификации и количественного определения.

На первом этапе анализа использовалась модель для классификации присутствующих в спектре газов, а на втором этапе – отдельная модель, которая, используя информацию о качественном составе от первой модели, определяла количественные концентрации компонентов. Такой подход направлен на повышение общей точности и надежности получаемых результатов. Для классификации газов была разработана сверточная нейронная сеть AdvancedGasCNN, построенная на основе известной архитектуры ResNet-34, адаптированной для работы с одномерными данными [8]. На вход модели подавался одномерный вектор, представляющий собой интерполированный терагерцовый спектр поглощения, состоящий из 2203 точек. Сверточная часть состояла из последовательности остаточных блоков (Residual Blocks) 1D-ResNet-34. Каждый блок включал в себя несколько одномерных сверточных слоев (Conv1d), слои пакетной нормализации (BatchNorm1d) и функцию активации ReLU. Такая архитектура позволяет эффективно извлекать сложные иерархические признаки из спектральных данных.

На выходе модели получали вектор вероятности принадлежности к каждому из предопределенных газов. Признаки, полученные от сверточной части, поступали в классификационный блок, который состоял из слоя глобального адаптивного усредняющего пулинга (AdaptiveAvgPool1d) и полносвязного линейного слоя (Linear) для финальной классификации. Линейный слой содержал 6 нейронов, по одному для каждого целевого газа (H_2S, H_2CO, NH_3, NO_2, SO_2, O_3). Финальная функция активации в модели отсутствует, так как для повышения вычислительной стабильности используется функция потерь BCEWithLogitsLoss, включающая аппарат функции активации. Также применялся механизм внимания [9], позволяющий модели взвешивать важность различных участков спектра.

Оптимизация параметров модели осуществлялась с помощью алгоритма AdamW с начальной скоростью обучения 1e-5. Для предотвращения переобучения использовался механизм ранней остановки (patience=25), а также планировщик ReduceLROnPlateau, который снижает скорость обучения при отсутствии улучшения на валидационной выборке.

Модельные спектры поглощения газов получены из базы данных HITRAN [10], а для расчетов контуров спектров поглощения атмосферы при нормальных условиях был применен пакет открытого программного обеспечения SPECTRA [11]. Спектры поглощения получались путем аддитивного смешения индивидуальных модельных спектров целевых газов (H_2CO, H_2S, NH_3, SO_2, NO_2) при различных концентрациях со спектрами паров воды, имитирующими вариации влажности атмосферы. Часть сгенерированных данных (80%) использовалось для процесса обучения, а часть (20%) для валидационной выборки, представляющей собой набор спектров поглощения для случайных концентраций примесей газов, лежащих в диапазоне от 0,5 до 10^{-9} %. Такое соотношение при разделении выборки позволяет модели эффективно обучаться на репрезентативном объеме данных, при этом обеспечивая надежную оценку ее обобщающей способности на отложенных данных.. Для валидационной выборки были выделены 4 группы по концентрациям: 0,5–10^{-3}, 10^{-3}–10^{-5}, 10^{-5}–10^{-7}, 10^{-7}–10^{-9}%, результаты точности обнаружения газов, для которых приведены на рис. 2 в виде значений F1-меры для каждого газа. F1-мера получена усреднением по 15 эпохам обучения.

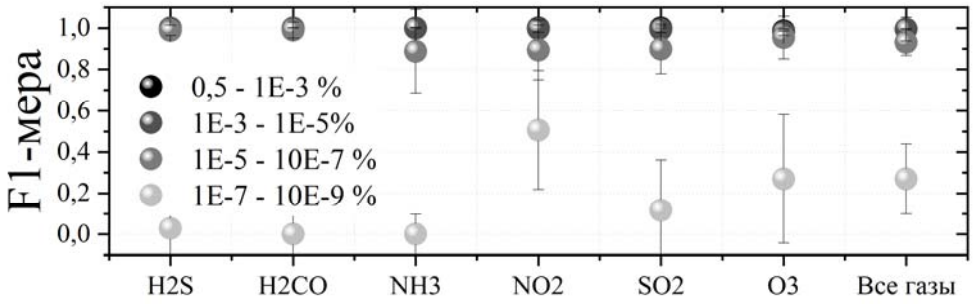


Рис. 2. Результаты идентификации газов в смеси на основе оценки по F1-мере. Рассмотрены 4 группы по концентрациям газов в смеси: 0,5–10^{-3}, 10^{-3}–10^{-5}, 10^{-5}–10^{-7}, 10^{-7}–10^{-9}%. Ошибки соответствуют усреднению F1-меры по 15 эпохам обучения

Такая метрика, как F1-мера представляет собой гармоническое среднее метрик *precision* (точность) и *recall* (полнота). При стремлении значений для каждой из этих метрик к 1, можно сделать вывод о том, что 1) при идентификации газа моделью отсутствовали ложные срабатывания, 2) модель обнаружила все присутствующие газы.

Наиболее точно модель определяет присутствие газов H_2S, H_2CO и O_3. При этом точность обнаружения для концентраций 10^{-5}–10^{-7}% составляет усредненное значение 0,93 по всем газам. Для NH_3, NO_2 и SO_2 значение F1 меньше, чем для других газов, что может быть связано с перекрытием спектральных областей с линиями поглощения паров воды. При концентрациях газов 10^{-7}–10^{-9}% точность обнаружения составляет в среднем 0,17, что меньше среднеквадратичной ошибки для этой группы. Согласно F1-мере, существует граница применимости модели, которая определена для концентраций газов от 10^{-5}–10^{-7}% и выше.

Выводы

Экспериментальные исследования на модельных наборах демонстрируют потенциал методики в идентификации шести газовых компонентов с нижней границей концентрации 0,001 ppm (10^{-7}%). Нейронная сеть достигла 89–98% точности при детектировании газов случайной концентрации в диапазоне 10^{-5}–10^{-7}% на основе сгенерированной выборки. Проведены серии экспериментов, которые демонстрируют чувствительность метода ТГц спектроскопии к идентификации 0,02% формальдегида в атмосфере. Экспериментальные данные могут использоваться для уточнения архитектуры модели, улучшения ее характеристик и генерации экспериментальной валидационной выборки для приближения метода к реальным условиям и созданию универсальной платформы газового анализа с высокой чувствительностью к микроконцентрациям атмосферных газов.

Благодарности. Работа выполнена в рамках гранта № 075-15-2024-533 Министерства науки и высшего образования РФ на выполнение крупного научного проекта по приоритетным направлениям научно-технологического развития (проект «Фундаментальные исследования Байкальской природной территории на основе системы взаимосвязанных базовых методов, моделей, нейронных сетей и цифровой платформы экологического мониторинга окружающей среды»).

Список литературы

1. *Bassous N.J., et al.* Significance of various sensing mechanisms for detecting local and atmospheric greenhouse gases: A review // Adv. Sensor Res. 2024. T. 3. P. 2300094.
2. *Dong M., et al.* Development and measurements of a mid-infrared multi-gas sensor system for CO, CO_2 and CH_4 detection// Sensors. 2017. T. 17. P. 2221.
3. *Vaks V.L. et al.* High resolution terahertz spectroscopy for analytical applications // Physics-Uspekhi. 2020. V. 63, N 7. C. 708.
4. *Высокоточная* резонаторная спектроскопия атмосферных газов в миллиметровом и субмиллиметровом диапазонах длин волн / М.Ю. Третьяков; Нижний Новгород ИПФ РАН, 2016. 320с.
5. *Nazarov M.M., et al.* // Quantum Electron. 2008. V. 38, N 7. P. 647.
6. *Wietzke S., et al.* Terahertz spectroscopy on polymers: A review of morphological studies // Journal of Molecular Structure. 2011. V. 1006. P 41–51
7. *Eliet S., Cuisset A., Guinet M., Hindle F., Mouret G., Bocquet R., & Demaison J.* Rotational spectrum of formaldehyde reinvestigated using a photomixing THz synthesizer // Journal of Molecular Spectroscopy. 2012. V. 279. P. 12–15.
8. *LeCun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. V.521(7553). P. 436-444.
9. *Woo S., Park J., Lee J.Y., Kweon I.S.* Cbam: Convolutional block attention module // Proc. of the European conference on computer vision (ECCV). 2018. P. 3–19.
10. *Gordon I.E., Rothman L.S., et al.* The HITRAN2020 molecular spectroscopic database // J. Quant. Spectrosc. Radiat. Transfer. 2022. V. 277. P. 107949.
11. *Михайленко С.Н., Бабиков Ю.Л., Головко В.Ф.* Информационно-вычислительная система "Спектроскопия атмосферных газов". Структура и основные функции // Оптика атмосферы и океана. 2005. Т. 18, № 9. С. 765–776.

Ph.A. Kozhevnikov, M.R. Konnikova, A.S. Sinko, A.A. Angeluts. **Application of machine learning methods to analyze the presence of harmful impurities in the atmosphere based on spectral data.**

The work is devoted to the development of an integrated approach to the analysis of the presence of harmful impurities in the atmospheric air. This approach involves: 1) using the measurement results of terahertz absorption spectra of air containing harmful impurities; 2) creating and using a neural network to analyze the data obtained. Sets of model absorption spectra of a gas mixture with different qualitative and quantitative compositions are generated to train the neural network. The application of a neural network to model sets of spectra demonstrated the identification of six gas components with concentrations up to 0.01 ppm. The neural network has achieved 90-95% accuracy in gas detection. A series of experiments were conducted for real gases, showing the sensitivity of the THz spectroscopy method to low concentrations of gases in atmosphere.