ИДЕНТИФИКАЦИЯ ВЕЩЕСТВ ПО СИЛЬНО ИСКАЖЁННЫМ ОШИБКАМИ ИЗМЕРЕНИЯ СПЕКТРАМ

Васильев Н.С., Морозов А.Н.
Московский государственный технический университет им. Н. Э. Баумана

Аннотация

В работе обсуждаются вопросы обнаружения веществ по их спектрам люминесценции. Рассмотрены случайные ошибки, возникающие в процессе работы измерительной фоточувствительной аппаратуры. Анализируется влияние этих ошибок на корректность работы устройств-анализаторов. Получены соотношения, позволяющие рассчитывать параметры распределения меры схожести SAM (Spectral Angle Mapper) как случайной величины. Проведено сопоставление полученных аналитических зависимостей на примере спектров, зарегистрированных при малых отношениях сигнал/шум. Получено удовлетворительное согласие теории и эксперимента. На основе этого предлагается новый метод идентификации веществ по спектрам, в котором для анализа используется проверка гипотез. Численным критерием в этом методе выступает условная вероятность измерить величину SAM с расхождением большим или равным реализованному в эксперименте. На основе этого метода строится алгоритм идентификации, который применялся для анализа спектров с малым отношением сигнал/шум. Определено, что предложенный способ распознавания спектров позволяет получить ряд преимуществ по сравнению с методом прямого расчёта меры SAM.

Ключевые слова: люминесценция, идентификация, мера схожести, хемометрика, корреляция, распознавание, SAM.

Введение

На сегодняшний день существует большое количество оптических приборов, позволяющих анализировать спектральные свойства света в широком диапазоне длин волн. Часто такие устройства входят в состав систем химического мониторинга окружающей среды. Регистрируемый сигнал содержит полезную информацию, анализ которой позволяет отличать одни вещества от других, тем самым выявлять их присутствие. К примерам можно отнести прибор «FirstDefender RM» фирмы «Ahura Scientific», основанный на Раман-эффекте и позволяющий осуществлять оперативный контроль присутствия опасных для человека веществ на химически вредных предприятиях.

Метод интерпретации зарегистрированных спектров, основанный на сопоставлении их с эталонными, является одним из наиболее распространённых в прикладных задачах спектроскопии. Для этого применяется функция меры схожести, которая характеризует степень совпадения форм экспериментальной и эталонной кривой. Существует большое число выражений, с помощью которых можно задать требуемую меру [1]. Одной из распространённых формул является SAM [1–7] (от англ. Spectral Angle Mapper), которая используется в задачах физики [1–3], машинного зрения [4, 6], фармацевтики [7] и аналитической химии [5].

Анализ влияния шума в зарегистрированном спектре на величину меры схожести SAM проводился ранее в работах [8, 9]. В [8] для определения параметров распределения величины SAM предлагается использовать метод Fisher z-transformation, см. [10, 11]. А в работе [9] для этих же целей предлагался t-критерий Стьюдента. Оба этих способа являются статистическими, что делает невозможным применение их для проведения экспресс-анализа по одному текущему измеренному спектру.

Цель работы заключается в создании эффективного метода идентификации веществ по спектрам в условиях низкого отношения сигнал/шум в них. Для этого решается задача вывода приближённых аналитических выражений для расчёта параметров распределения используемой меры схожести.

1. Теория

1.1. Статистические характеристики нормированного скалярного произведения

Величина нормированного скалярного произведения определяется следующим соотношением:

$$\rho = \rho(x, y) = \frac{(x, y)}{|x| \cdot |y|}. \qquad (1)$$

Если x и y – две непрерывные функции, определённые на интервале частот $[\nu_{нач}, \nu_{кон}]$, то для расчёта скалярного произведения используется выражение: $(x, y) = \int_{\nu_{нач}}^{\nu_{кон}} xy \, d\nu$. Если задано разбиение интервала $[\nu_{нач}, \nu_{кон}]$ точками:

$P = \{\nu_i |\ i = 1..N,\ \nu_{нач} < \nu_1 < \nu_2 < \cdots < \nu_N < \nu_{кон}\}$,

то вместо функций рассматриваются вектора с координатами $x_i = x(\nu_i)$ $y_i = y(\nu_i)$, для которых скалярное произведение рассчитывается по формуле: $(x, y) = \sum_{i=1}^{N} x_i y_i$. Для случая $N = 2$ выражение (1) совпадает с косинусом угла между векторами $\vec{x}$ и $\vec{y}$.

Из определения SAM (1) следует, что при $x(\nu) \equiv y(\nu),\ \forall \nu \in [\nu_{нач}, \nu_{кон}] \Rightarrow \rho = 1$. И для дискретного случая: $x_i \equiv y_i,\ i = 1..N \Rightarrow \rho = 1$. Если $\exists \nu \in [\nu_{нач}, \nu_{кон}],\ т.ч.\ x(\nu) \not\equiv y(\nu) \Rightarrow \rho < 1$. Для дис-

кретного случая аналогичное условие можно записать так: $\exists i (1 \le i \le N)\ m.ч.\ x_i \not\equiv y_i \Rightarrow \rho < 1$. Равенство SAM единице для идентичных спектров и неравенство для различающихся позволяет использовать эту величину для определения меры схожести или различия в задачах распознавания.

В случае, если в каждой точке разбиения P имеется малая погрешность измерения, то вместо строгого равенства для SAM имеем приближённое: $\rho \approx 1$, при этом погрешность в силу определения (1) может быть только в меньшую сторону.

Если регистрируемые спектры таковы, что с высокой долей вероятности можно считать различия между спектрами, обусловленные случайной погрешностью, много меньше, чем различия, обусловленные различной природой анализируемых веществ, то в этом случае ([6, 12]) можно использовать некотоой порог, при превышении которого спектры будут считаться идентичными. В общем случае определение этих порогов трудоёмкая и технически сложная задача.

Если искажения, вызванные погрешностью измерения спектра, велики, то значения SAM для идентичных веществ могут стать сопоставимыми с соответствующими значениями для неидентичных веществ. Повышением значения порогов можно избежать наличия ложных срабатываний так, чтобы сильно искажённый спектр не отождествлялся ни с одним из эталонных. В результате чувствительность обнаружения веществ прибором уменьшается.

Знание закона распределения SAM как случайной величины при воздействии случайных возмущений в спектре позволяет осуществлять распознавание с заданной оценочной достоверностью результатов.

Исследование измеренных спектров как векторов, координаты которых рассчитаны на некотором разбиении рабочего интервала частот P, требует наличия модели процесса формирования и преобразования погрешностей измерения спектра. Характерным для задач обнаружения малого количества вещества является использование светосильных спектральных приборов, таких как Фурье-спектрометры. В работе авторами рассматривается система, использующая статический Фурье-спектрометр видимого и ближнего ультрафиолетового диапазона [12]. Модель формирования и преобразования шума в спектре должна учитывать ошибки измерения регистрируемой интерферограммы и восстановления спектра.

Если интенсивность света измеряется в диапазоне $[z_{нач}, z_{кон}]$ оптической разности хода лучей в интерферометре, то согласно [13] в каждой точке разбиения $O = \{z_i |\ i = 1..K,\ z_{нач} < z_1 < z_2 < \cdots < z_K < z_{кон}\}$ интенсивности истинной интерферограммы (I') и зарегистрированной интерферограммы (I) связаны через погрешность δ_I выражением:

$$I(z_i) = I'(z_i) + \delta_I(z_i), \quad i = 1..K. \tag{2}$$

При этом система случайных $\{\delta_I(z_i)\}$ величин подчиняется уравнениям:

$$\mathbf{M}(\delta_I(z_i)) = 0, \quad i = 1..K, \tag{3}$$

$$\mathbf{M}(\delta_I(z_i)\cdot\delta_I(z_j)) = \begin{cases} d_I, & \text{если } i = j. \\ 0, & \text{если } i \ne j. \end{cases} \tag{4}$$

К зарегистрированному спектру применяется Фурье-преобразование, которое в общем случае задаётся соотношением:

$$x(\nu) = (1/2\pi)\int_{-\infty}^{+\infty} I(z)\exp(-i\nu z)\,\mathrm{d}z. \tag{5}$$

Фурье-образ от (2) с учётом (3) и (4) в соответствии с [13] позволяет связать на сетке разбиения P истинный спектр $x'(\nu_i)$, зарегистрированный спектр $x(\nu_i)$ и погрешность $\delta(\nu_i)$ выражением:

$$x(\nu_i) = x'(\nu_i) + \delta(\nu_i). \tag{6}$$

При этом система случайных величин $\{\delta(\nu_i)\}$ имеет следующие моменты:

$$\begin{aligned} \mathbf{M}(\delta(\nu_i)) &= 0, \quad i = 1..N, \\ \mathbf{M}(\delta^2(\nu_i)) &= \sigma^2, \quad i = 1..N. \end{aligned} \tag{7}$$

В [13] приводится для используемой модели соотношение, связывающее ошибку в измеренной интерферограмме с ошибкой восстановленного спектра:

$$\Delta B = h(2N)^{1/2}\Delta I. \tag{8}$$

Воспользуемся для описания ошибки спектра $\delta(\nu_i)$ в точках разбиения P моделью нормально распределённой случайной величины: $\delta(\nu_i) \sim N(0, \sigma^2), i = 1..N$. При таком допущении зарегистрированный спектр x может рассматриваться как случайная N-мерная величина ($x \sim N(x', \Sigma)$) с распределением, задаваемым соотношением:

$$p(x) = \frac{1}{(2\pi)^{(N2)}\,|\,\Sigma\,|^{1/2}} e^{-\frac{1}{2}(x-x')^T\Sigma^{-1}(x-x')},\ x \in R^N. \tag{9}$$

Выясним, какой вид может иметь ковариационная матрица $\sum$ в выражении (9). Как видно из определения, SAM – это величина, инвариантная относительно перехода в новый ортогональный базис. В двумерном случае это означает сохранение углов при переходе в новую систему координат при преобразовании векторов. Известно [14], что с помощью таких преобразований можно квадратичную форму, стоящую в степени экспоненты в (9), свести к диагональному виду. В такой системе координат ковариационная матрица принимает диагональный вид: $\sum = E\cdot\sigma^2$, где E – единичная диагональная матрица размера $N\times N$. Тогда всюду ниже считается, что ковариационная матрица диагональная и величина дисперсии шума постоянна и равна σ^2, если это не так, то указанная операция приводит её к требуемому виду.

Определим спектральную базу эталонных спектров следующим образом:

$$\mathcal{B} = \{b_i |\ i, j = 1..M\ b_i \in R_+^N\ (\forall i, j\ \rho(b_i, b_j) < 1)\}.$$

Пусть в (1) первая переменная пробегает значения эталонных спектров

$$(\rho_{\mathcal{B}} = \rho(\mathcal{B}, x) = (\rho(b_1, x), \cdots, \rho(b_M, x))^T),$$

тогда получим новое отображение:

$$\rho_{\mathcal{B}} : R^N \to I_{[-1.1]} = \underbrace{[-1,1]\times\cdots\times[-1,1]}_{M}, \quad (10)$$

где *i*-я координата правой части равна SAM, рассчитанному для *i*-го эталонного спектра.

Исходя из предложенной модели, зарегистрированный спектр выражается через истинный по формуле (6). Пусть искомое вещество является одним из эталонных с номером ξ, тогда выражение (6) примет вид: $x = b_\xi + \delta$. Используя отображение (10), получим образ измеренного спектра в $I_{[-1.1]}$:

$$\rho = \rho_{\mathcal{B}}(x) = (\rho_1, \cdots, \rho_M)^T.$$

Требуется определить вектор-столбец математических ожиданий и ковариационную матрицу случайной величины $\rho \in R^M$ по известной величине разброса ошибки спектра σ^2 в предположении, что искомое вещество совпадает с одним из эталонных. Пусть это вещество в эталонной базе спектров под номером ξ, тогда задача состоит в определении вектора $\mu_{\xi,M\times 1}$ и матрицы $\mathcal{K}_{\xi,M\times M}$, координаты которых определяются, исходя из следующих выражений:

$$\mu_{\xi\nu} = \mathbf{M}(\rho_\nu), \quad (11)$$

$$\mathcal{K}_{\xi\nu\theta} = \mathbf{M}((\rho_\nu - \mathbf{M}(\rho_\nu))\cdot(\rho_\theta - \mathbf{M}(\rho_\theta))). \quad (12)$$

В приложении (см. п. 5) представлен подробный вывод аналитических выражений для этих коэффициентов.

1.2. Учёт характеристик нормированного скалярного произведения в задаче идентификации

Известно, что заданная ковариационная матрица Σ и квадратичная форма $x^T\Sigma^{-1}x$ задаёт метрику $d_N(x,y) = (x-y)^T\Sigma^{-1}(x-y)$ в R^N. Если дан случайный вектор x, для которого известно распределение $x \sim N(b_\xi, \Sigma = E\cdot\sigma^2)$, то по заданной вероятности P^* можно указать расстояние *d** такое, что выполняется условие: $P(d_N(x,b_\xi) > d^*) = P^*$. В связи с этим обстоятельством удобно ввести систему классов $\{W_i\}$, соответствующих каждому эталонному веществу из базы спектров:

$$W_\xi = \{a \mid a \in R^N, d_N(a,b_\xi) < d^*\}. \quad (13)$$

Заметим, что чем выше вероятность *P**, тем меньше расстояние *d**.

По введённой выше системе $\{W_i\}$ задачи идентификации веществ может быть сведена к задаче определения принадлежности измеренного спектра одному или нескольким её элементам. Селективность методики идентификации веществ можно определить как минимальную величину *P**, при которой выполнено условие: $\forall i,j = 1..M, W_i \cap W_j = \varnothing$.

Вычислительная сложность задачи определения вероятностей по заданному распределению случайной величины в многомерном пространстве быстро увеличивается с ростом размерности. Предлагается использовать отображение (10) для перехода в пространство меньшей размерности, равной числу эталонных спектров.

С учётом полученных выражений для μ и $\mathcal{K}$ случайный вектор $\rho(x) \in R^M$ можно аппроксимировать нормально распределённой случайной величиной:

$$p(\rho) = \frac{1}{(2\pi)^{(N2)}\,|\mathcal{K}|^{1/2}} e^{-\frac{1}{2}(\rho-\mu)^T\mathcal{K}^{-1}(\rho-\mu)}, \rho \in R^M. \quad (14)$$

На практике такое приближение является удовлетворительным для широкого интервала значений величины погрешностей в спектре.

Ковариационная матрица так же, как и в исходном пространстве спектров, задаёт в пространстве SAM метрику:

$$d_M(\rho(x),\rho(y)) = (\rho(x)-\rho(y))^T\mathcal{K}^{-1}(\rho(x)-\rho(y)).$$

По заданному распределению ρ и данной вероятности *P** аналогичным образом можно определить расстояние *d***, удовлетворяющее условию:

$$P(d_M(\rho(x),\mu_\xi) > d^{**}) = P^*,$$

с помощью которого определяется система классов $\{W_{i'}\}$:

$$W_{\xi'} = \{a \mid d_M(\rho(a),\mu_\xi) < d^{**}\}. \quad (15)$$

Определение границ классов в этом случае требует интегрирования в *M*-мерном пространстве. Величина *M* равна количеству веществ в эталонной базе спектров, и их количество может быть велико. Удачным выбором системы координат в пространстве R^M сложность задачи интегрирования может быть уменьшена с M^L до $L\cdot M$, где *L* – количество точек разбиения для численного интегрирования вдоль одной оси. Такой системой координат будет ортогональная система собственных векторов матрицы $\mathcal{K}$. Растяжением осей можно добиться приведения квадратичной формы $(\rho-\mu_\rho)^T\mathcal{K}^{-1}(\rho-\mu_\rho)$ к каноническому виду. Преобразование $\Phi\Lambda^{-1/2}$ приводит ковариационную матрицу к единичному виду. В новых осях функция распределения случайной величины $\rho(x)$ примет вид:

$$P(\rho) = \left(\operatorname{erf}\left(\frac{\left(\Phi\Lambda^{-\frac{1}{2}}\right)^T(\rho-\mu_\xi)}{\sqrt{2}}\right)\right)^M. \quad (16)$$

Основываясь на проделанных рассуждениях и полученных выражениях для μ_ξ и $\mathcal{K}_\xi$, предлагается алгоритм идентификации веществ по спектрам, схема которого показана на рис. 1.

Если условие в цикле выполнено более одного раза, то выбирается вещество, для которого величина условной вероятности $P(a|a=b_i+\delta)$ максимальна. В роли порога срабатывания в данном случае выступает величина *P**, по которой в предложенном алгоритме рассчитываются классы $\{W_{i'}\}$.

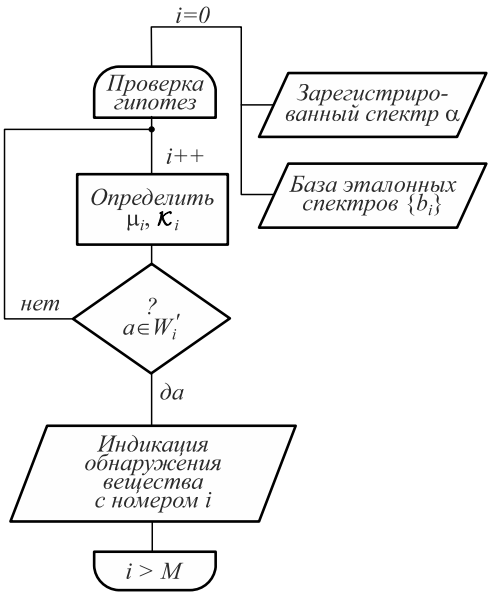

Рис. 1. Схема алгоритма идентификации веществ по спектрам люминесценции

В представленном методе расчёта меры схожести используется величина, которая имеет смысл вероятности. Это качественным образом отличает этот способ от прочих методов, в которых используется мера схожести. Для их применения требуется эмпирически определять пороги срабатывания для каждого эталонного вещества [15]. При этом добавление новых веществ в базу данных может приводить к ухудшению работы всей системы. Предложенный авторами новый способ идентификации веществ, основанный на проверке гипотез, потенциально лишён указанного недостатка. Это позволяет один раз задать порог срабатывания (вероятность ошибки) для всех эталонных элементов спектральной базы данных. При этом введение нового вещества не повлияет на надёжность работы системы в целом.

2. Эксперимент

Использовались спектры, зарегистрированные статическим Фурье-спектрометром, который входил в состав макетного образца прибора, осуществляющего экспресс-анализ присутствия жидких и твёрдых веществ на различных подстилающих поверхностях [16]. Устройство позволяет обнаруживать твёрдые и жидкие вещества в форме остаточных следов на различных поверхностях. Для триптофана обнаружительная способность в зависимости от условий эксперимента может достигать значений до 1 мг на площади 1 м^2. Схема установки показана на рис. 2.

Вторичное излучение, которое падает на объектив оптоволоконного зонда 1, собирается оптической системой линз и фокусируется на передний срез оптоволокна.

Оно подключено ко входному коллиматору спектрометра. В качестве подложки 3 использовалась специальная поверхность из непрозрачного не люминесцирующего стекла. Его рабочий диапазон чувствительности равен интервалу длин волн от 320 нм до 750 нм, который содержит видимый свет и ближнюю ультрафиолетовую область.

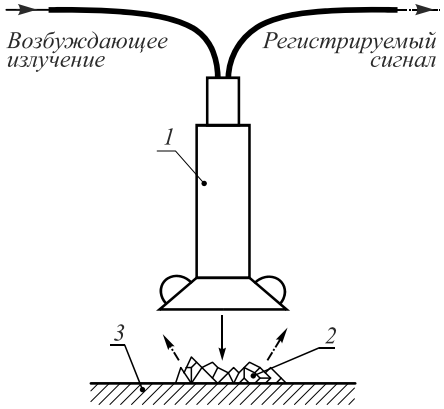

Рис. 2. Схема экспериментальной установки: 1 – оптоволоконный зонд, 2 – тестовое вещество, 3 – подложка

Прибор, помимо спектрометра, состоит из источника возбуждающего излучения и оптоволоконного световода с зондом на конце. Последнее устройство позволяет направлять и концентрировать возбуждающее излучение непосредственно на объект исследования и собирать как можно больше рассеянного излучения для анализа. В качестве источников возбуждающего излучения (см. рис. 2) использовались диоды с пиками излучения на длинах волн 280 и 310 нм, а также лазерный источник излучения с длиной волны 266 нм. В качестве тест-объектов 2 использовались вещества: антрацен, POPOP, PPO, стильбен и триптофан. Спектры этих веществ с указанием особенностей их химического строения показаны на рис. 3.

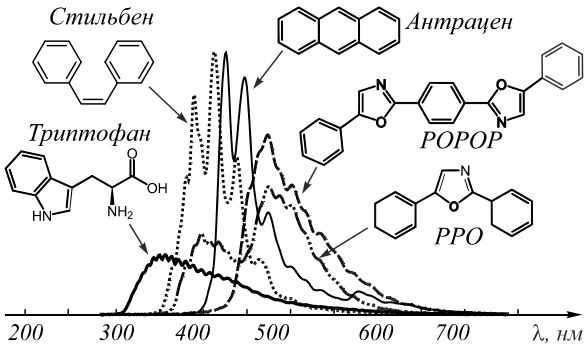

Рис. 3. Спектры люминесценции тестовых веществ, зарегистрированные с использованием источника возбуждающего излучения лазера, длина волны – 266 нм; по оси ординат откладывается интенсивность в относительных единица

В дальнейшем для краткости эти вещества будут называться «тестовыми веществами». Рассмотренные тестовые вещества являются люминофорами при использовании данных источников подсвечивающего излучения. Как видно из рис. 2, спектры этих веществ в значительной степени перекрываются.

2.1. Численный эксперимент

Для выявления возможных ошибок в расчётах, а также для проверки корректности сделанных упрощений при определении математического ожидания и

ковариационной матрицы был проведён численный эксперимент с использованием псевдослучайных величин, соответствующих рассмотренной модели формирования интерферограмм и восстановления спектра. В частности, проверялась справедливость упрощения, в результате которого учитывались 17 из 36 коэффициентов матрицы (21).

На рис. 4 показано три графика зависимости математического ожидания SAM от среднеквадратической ошибки, рассчитанные для веществ POPOP и антрацен и для веществ PPO и антрацен, а также ковариации этих величин. При этом спектр антрацена состоял из эталонного спектра антрацена и аддитивно добавленной случайной ошибки, распределённой по нормальному закону. Величина среднеквадратического отклонения этой ошибки откладывалась по оси абсцисс графиков рис. 2.

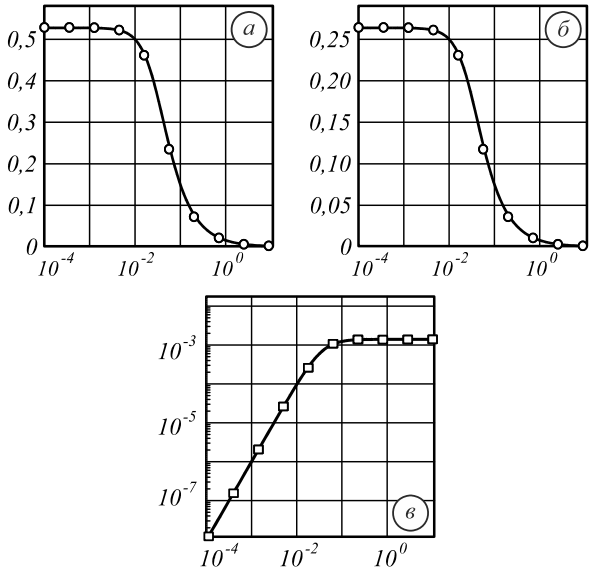

Рис. 4. Кривые зависимости математического ожидания величины нормированного скалярного произведения (ось ординат) от величины среднеквадратического отклонения шума в спектре (ось абсцисс); сплошная кривая рассчитана для веществ POPOP и антрацен (а) и для веществ PPO и антрацен (б); приведена кривая ковариации указанных величин (в); точками показаны значения, полученные при численном разыгрывании

Наблюдается совпадение теоретически рассчитанных кривых графиков и экспериментальных точек с точностью до ширины линий (см. рис. 2).

2.2. Физический эксперимент

Известно, что нормально распределённая случайная величина является моделью, которая соответствует реальному объекту с различной степенью точности. Использование алгоритмов распознавания в реальных условиях эксперимента может привести к неудовлетворительным результатам, если параметры ошибок в регистрируемом спектре будут отличаться от предусмотренных моделью.

Апробация осуществлялась на примере более 4000 спектров, зарегистрированных в ходе экспериментов, которые повторяли условия применения прибора. Для каждого из тестовых веществ был зарегистрирован набор спектров, число которых варьировалось от 50 до 300. Величина сигнал/шум, которая рассчитывалась по формуле $SNR = \overline{P}_{сигнал} \big/ \overline{P}_{шум}$, для всех спектров в наборе была постоянной. Изменение величины *SNR* осуществлялось с помощью выбора времени накопления сигнала на фоточувствительной матрице статического Фурье-спектрометра. Все прочие условия проведения эксперимента оставались постоянными для всех наборов и веществ.

По измеренным спектрам были рассчитаны статистические значения величины разброса SAM Для удовлетворительного согласия величины дисперсии с теоретически рассчитанной по формуле (23) потребовался пересчёт по формуле: $I_{\xi}^{(э)} = kI_{\xi}$, где коэффициент k – неизвестный параметр, который был определён из условия минимального расхождения теоретических кривых и экспериментальных точек. Для спектров, зарегистрированных использованным статическим Фурье-спектрометром, этот параметр равнялся 2,1. На рис. 5 показано математическое ожидание и дисперсия SAM, рассчитанная для зарегистрированного спектра вещества стильбен и эталонного спектра этого же вещества.

2.3. Апробация алгоритмов распознавания

Авторами в работе сопоставлялись два алгоритма распознавания. Первый, далее называемый «№1», предложен ранее в работах [6, 12]. В нём рассчитывалось значение меры схожести с эталонными спектрами, которое затем сравнивалось с пороговым значением. Превышение приводило к распознаванию соответствующего вещества в исследуемом образце. Второй алгоритм использует результаты, полученные в данной статье авторами, и изложен в разделе 1.2. Далее для краткости он будет называться «№2».

В табл. 1 приведены результаты для долей ложных срабатываний и верных срабатываний для спектров всех веществ, полученных от трёх источников возбуждающего излучения. Результаты представлены в виде дроби, где в числителе указана доля ложных срабатываний от общего числа возможных ошибок, а в знаменателе – соответствующая доля верных срабатываний. Верным считалось срабатывание, в результате которого обнаруживается только то вещество, которое служило источником регистрируемого прибором сигнала.

Таблица 1. Результаты распознавания в долях от максимального числа ошибок (числитель) и верных срабатываний (знаменатель)

Диапазон	Случай 1		Случай 2	
	Алгоритм №1	Алгоритм №2	Алгоритм №1	Алгоритм №2
266	0,13/0,33	0,13/0,49	0,03/0,81	0,01/0,78
280	0,11/0,48	0,11/0,56	0,02/0,78	0,01/0,62
310	0,11/0,51	0,11/0,57	0,01/0,87	0,00/0,79

Если условиям алгоритма идентификации удовлетворило хотя бы одно лишнее вещество, то такое срабатывание считалось ложным и соответствующему счётчику присваивалось число таких (лишних) веществ.

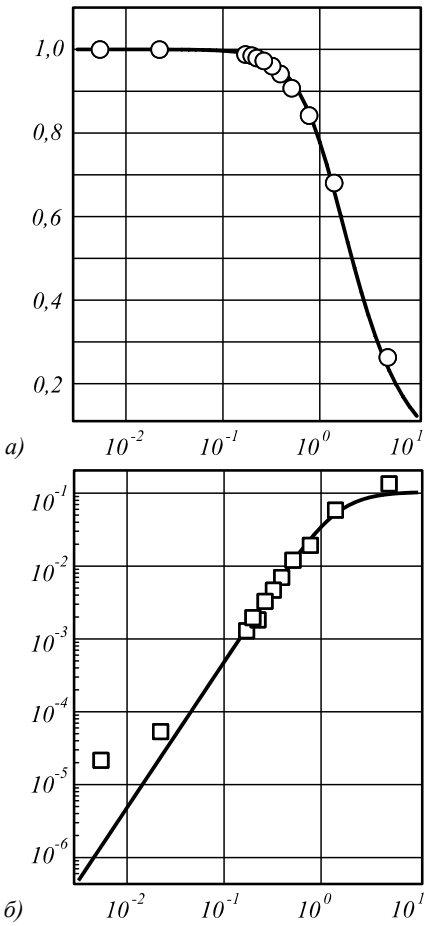

Рис. 5. Кривые зависимости математического ожидания (а) и дисперсии (б) нормированного скалярного произведения (ось ординат) от среднеквадратического отклонения ошибки экспериментального спектра (ось абсцисс); кривые рассчитаны при сопоставлении экспериментального спектра стильбена с его эталонным спектром; сплошная кривая рассчитана теоретически, маркерами отмечены результаты статистического анализа серии экспериментов

Результаты показаны для величин порогов, при которых число верных срабатываний максимально.

Из таблицы (столбец «случай 1») видно, что предложенный авторами алгоритм позволил распознавать больше веществ при одинаковом числе ошибок. Заметим, что в этих экспериментах отношение сигнал/шум варьировалось для различных веществ и находилось в интервале $[1,01-1,05]$.

Часто при работе приборов, предназначенных для автоматической идентификации веществ, в регистрируемом сигнале отсутствует полезная составляющая. Срабатывание в отсутствие полезного сигнала также считается ложным. Был определён порог срабатывания, при котором 98 % спектров, не содержащих полезного, не приводили к срабатыванию.

Результаты для такого случая показаны в таблице (столбец «случай 2»). Здесь использовались спектры с величиной отношения сигнал/шум в интервале $[1,04-1,19]$. Из данных таблицы видно, что для «алгоритма №1» доля ложных срабатываний для большей части экспериментов оказывалась больше, чем для «алгоритма №2». В то же время доля верных срабатываний для «алгоритма №1» оказалась больше. Т.е. «алгоритм №1» позволил распознать вещества в большей части случаев, в то время как «алгоритм №2» давал более надёжные результаты.

Выводы

Авторами рассмотрена одна из часто применяемых мер схожести, которая используется в задачах автоматического распознавания веществ по спектрам. Полученные выражения (22) и (23) позволяют определять параметры распределения меры схожести SAM в зависимости от величины среднеквадратического отклонения ошибок измерения в точках зарегистрированного спектра. Проведённый численный и физический эксперименты показали, что полученные соотношения можно использовать как применительно к модельному, нормально распределённому, так и применительно к реализуемому в эксперименте шуму. Для применения полученных формул к спектрам, регистрируемым в условиях эксперимента, потребовалось введение коэффициента $k=2{,}1$, учитывающего отличия шума, реализуемого в спектрах от модельного.

Предложенный алгоритм идентификации веществ по их зарегистрированным спектрам люминесценции использует для анализа условную вероятность для проверки гипотез. Для сравнения рассмотрен существующий метод, в котором рассчитывается мера схожести по формуле SAM. Оказалось: если заведомо известно, что в анализируемом сигнале есть полезная составляющая, то предложенный алгоритм даёт выигрыш в доли верных срабатываний. Если в анализируемом спектре полезный сигнал может отсутствовать, то предложенный алгоритм позволяет добиться меньшего числа ошибок распознавания. Полученные в работе результаты могут быть использованы в методиках автоматического обнаружения веществ по их спектрам в таких задачах, как мониторинг окружающей среды и беспробоотборный химический анализ.

Приложение

Ниже приводится подробный вывод аналитических выражений для расчёта коэффициентов ковариационной матрицы и вектора математических ожиданий величин SAM, применённых для сопоставления эталонных спектров и одного экспериментального. При этом считается, что шум в зарегистрированном спектре аддитивный и стационарный.

В дальнейшем индексы, относящиеся к элементам эталонной базы спектров, обозначаются символами ξ, η и θ, а i, j, k, l, m и т.д. отвечают за координаты векторов. Подставив слагаемые x и b_η в (1) и выразив величины $(x\,|\,b_\eta)$, $|\,x\,|$ и $|\,b_\eta\,|$ через их координаты, получим:

$$\rho_{\xi,\eta}=\frac{B_{\xi\eta}^2/B_{\eta\eta}+\left(1/B_{\eta\eta}\right)\cdot\sum\limits_{i=1}^{N}\delta_i b_{\eta,i}}{\sqrt{B_{\xi\xi}^2+2\sum\limits_{i=1}^{N}b_{\xi,i}\delta_i+\sum\limits_{i=1}^{N}\delta_i^2}},\qquad(17)$$

где $B_{\eta\theta} = \sqrt{\sum_{i=1}^{N} b_{\eta,i} b_{\theta,i}}$. Заметим, что $\sum_{i=1}^{N}\left(\frac{\delta_i^2}{\sigma^2}\right)$ подчиняется распределению χ^2. Как правило [12], количество точек в спектре велико и имеет порядок $\sim 10^3$ точек. Известно [13], что при $N\to\infty$ распределение χ^2 подчиняется нормальному закону, что позволяет записать приближённые выражения для его математического ожидания и дисперсии:

$$\mathbf{M}\left(\sum_{i=1}^{N}\delta_i^2\right) = N\sigma^2, \quad \mathbf{D}\left(\sum_{i=1}^{N}\delta_i^2\right) = 2N\sigma^4. \qquad (18)$$

Это позволяет записать выражения для некоторых слагаемых в (17):

$$\begin{aligned}\sum_{i=1}^{N}\delta_i b_{\eta,i} &= \alpha_\eta \sim N(0,\sigma^2 B_{\eta\eta}^2),\\ \sum_{i=1}^{N}\delta_i^2 &= \gamma \sim N(N\sigma^2, 2N\sigma^4).\end{aligned} \qquad (19)$$

Линеаризация выражения (17) осуществляется стандартным образом. Для этого введём обозначение для подкоренного выражения знаменателя:

$$Y = B_{\xi\xi}^2 + 2\alpha_\xi + \gamma.$$

Его математическое ожидание с учётом (19):

$$\mu_Y = \mathbf{M}(Y) = B_{\xi\xi}^2 + \sigma^2 N.$$

Что позволяет линеаризовать выражение $1/\sqrt{Y}$:

$$1/\sqrt{Y} \approx \mu_Y^{-1/2} - 12\mu_Y^{-3/2}\left(Y-\mu_Y\right).$$

Формула для определения нормированного скалярного произведения примет вид:

$$\rho_\eta = \left(B_{\xi\eta}^2/B_{\eta\eta} + \alpha_\eta/B_{\eta\eta}\right)\left(\mu_Y^{-1/2} - 12\mu_Y^{-3/2}\left(Y-\mu_Y\right)\right).$$

Для упрощения введём коэффициенты:

$$A_{\eta 1} = B_{\xi\eta}^2 B_{\eta\eta},\ A_{\eta 2} = 1B_{\eta\eta},\ A_3 = \frac{3}{2}\mu_Y^{-1/2} - \frac{1}{2}\mu_Y^{-3/2}B_{\xi\xi}^2,$$

$$A_4 = -\mu_Y^{-3/2},\ A_5 = -\frac{1}{2}\mu_Y^{-3/2}.$$

После раскрытия скобок получим линеаризованное выражение для (17):

$$\rho_\eta = F_{\eta 1} + F_{\eta 2}\alpha_\eta + F_{\eta 3}\alpha_\xi + F_{\eta 4}\alpha_\eta\alpha_\xi + F_{\eta 5}\gamma + F_{\eta 6}\alpha_\eta\gamma, \qquad (20)$$

где

$$F_{\eta 1} = A_{\eta 1}A_3,\ F_{\eta 2} = A_{\eta 2}A_3,\ F_{\eta 3} = A_{\eta 1}A_4,$$
$$F_{\eta 4} = A_{\eta 2}A_4,\ F_{\eta 5} = A_{\eta 1}A_5,\ F_{\eta 6} = A_{\eta 2}A_5.$$

Линеаризованное выражение для вычисления нормированного скалярного произведения (20) позволяет записать линеаризованное выражение для произведения $\rho_\eta\rho_\theta$:

$$\rho_\eta\rho_\theta = \sum\left(F_{\eta 1}, F_{\eta 2}\alpha_\eta, F_{\eta 3}\alpha_\xi, F_{\eta 4}\alpha_\eta\alpha_\xi, F_{\eta 5}\gamma, F_{\eta 6}\alpha_\eta\gamma\right)^T \times$$
$$\times\left(F_{\theta 1}, F_{\theta 2}\alpha_\theta, F_{\theta 3}\alpha_\xi, F_{\theta 4}\alpha_\theta\alpha_\xi, F_{\theta 5}\gamma, F_{\theta 6}\alpha_\theta\gamma\right),$$

где суммирование осуществляется по всем элементам матрицы размерности 6×6. Анализ этой матрицы показал, что математическое ожидание произведения $\rho_\eta\rho_\theta$ может быть рассчитано с высокой точностью (см. рис. 4) при учёте 17 из 36 коэффициентов (см. раздел 2.1). Получены аналитические зависимости для этих коэффициентов, при этом учитывались следующие приближённые тождества:

$$\mathbf{M}(\alpha_\eta\alpha_\theta) = B_{\eta\theta}^2\sigma^2,$$
$$\mathbf{M}(\alpha_\xi^2) = B_{\xi\xi}^2\sigma^2,$$
$$\mathbf{M}(\alpha_\eta\alpha_\xi) = B_{\xi\eta}^2\sigma^2,$$
$$\mathbf{M}(\alpha_\theta\alpha_\xi) = R_\theta^2\sigma^2,$$
$$\mathbf{M}(\gamma^2) = N(N+2)\sigma^4,$$
$$\mathbf{M}(\alpha_\eta\alpha_\theta\gamma^2) = B_{\eta\theta}^2(N^2+3N+11)\sigma^6,$$
$$\mathbf{M}(\alpha_\eta\alpha_\theta\gamma) = B_{\eta\theta}^2(N+2)\sigma^4,$$
$$\mathbf{M}(\alpha_\eta\gamma\alpha_\xi) = B_{\xi\eta}^2(N+2)\sigma^4,$$
$$\mathbf{M}(\alpha_\theta\gamma\alpha_\xi) = B_{\xi\theta}^2(N+2)\sigma^4.$$

Окончательное выражение для матрицы $I_{\nu,\theta}$:

$$I_{\eta\theta} = \begin{pmatrix} F_{\eta 1}F_{\theta 1} & 0 & 0 & F_{\eta 1}F_{\theta 4}B_{\xi\theta}^2\sigma^2 & F_{\eta 1}F_{\theta 5}N\sigma^2 & 0\\ 0 & F_{\eta 2}F_{\theta 2}B_{\eta\theta}^2\sigma^2 & F_{\eta 2}F_{\theta 3}B_{\xi\eta}^2\sigma^2 & 0 & 0 & F_{\eta 2}F_{\theta 6}(N+2)B_{\eta\theta}^2\sigma^4\\ 0 & F_{\eta 3}F_{\theta 2}B_{\xi\theta}^2\sigma^2 & F_{\eta 3}F_{\theta 3}B_{\xi\xi}^2\sigma^2 & 0 & 0 & F_{\eta 3}F_{\theta 6}(N+2)B_{\xi\theta}^2\sigma^4\\ F_{\eta 4}F_{\theta 1}B_{\xi\eta}^2\sigma^2 & 0 & 0 & 0 & F_{\eta 4}F_{\theta 5}(N+2)B_{\xi\eta}^2\sigma^4 & 0\\ F_{\eta 5}F_{\theta 1}N\sigma^2 & 0 & 0 & F_{\eta 5}F_{\theta 4}(N+2)B_{\xi\theta}^2\sigma^4 & F_{\eta 5}F_{\theta 5}N(N+2)\sigma^4 & 0\\ 0 & F_{\eta 6}F_{\theta 2}(N+2)B_{\eta\theta}^2\sigma^4 & F_{\eta 6}F_{\theta 3}(N+2)B_{\xi\eta}^2\sigma^4 & 0 & 0 & F_{\eta 6}F_{\theta 6}(N^2+3N+11)B_{\eta\theta}^2\sigma^6 \end{pmatrix} - \begin{pmatrix} F_{\eta 1}\\ 0\\ 0\\ F_{\eta 4}R_\eta^2 d\\ F_{\eta 5}Nd\\ 0 \end{pmatrix}\begin{pmatrix} F_{\theta 1}\\ 0\\ 0\\ F_{\theta 4}R_\theta^2 d\\ F_{\theta 5}Nd\\ 0 \end{pmatrix}^T \qquad (21)$$

для математического ожидания:

$$\mu_{\xi\eta} = F_{\eta 1} + F_{\eta 4}B_{\xi\eta}^2\sigma^2 + F_{\eta 5}N\sigma^2 \qquad (22)$$

и для корреляции:

$$K_{\xi\eta\theta} = \sum_{i,j=1}^{6} I_{\eta\theta ij}. \qquad (23)$$

Литература

1. **Guti´errez-Rodriguez, A.E.** New dissimilarity measures for ultraviolet spectra identification / A.E. Guti´errez-Rodriguez, M.A. Medina-P´erez, J.F. Martinez-Trinidad [et al.] // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2010. – V. 6256 – P. 220-229.
2. **Stephen, S.E.** Optimization and testing of mass spectral library search algorithms for compound identification / S.E. Stein, D.R. Scott // Journal of the American Society for Mass Spectrometry. – 1994. – Vol. 5(9). – P. 859-866.
3. **Kruse, F.A.** The spectral image processing system (SIPS)—interactive visualization and analysis of imaging

spectrometer data / F.A. Kruse, A.B. Lefkoff, J.W. Boardman [et al.] // Remote Sensing of Environment. – 1993. – Vol. 44(2-3). – P. 145-163.
4. **Paclik, P.** A study on design of object sorting algorithms in the industrial application using hyperspectral imaging / P. Paclik, R. Leitne, R.P.W. Duin // Journal of Real-Time Image Processing. – 2006. – Vol. 1(2). – P. 101-108.
5. **Bodis, L.** A novel spectra similarity measure / L. Bodis, A. Ross, E. Pretsch // Chemometrics and Intelligent Laboratory Systems. – 2007. – Vol. 85(1). – P. 1-8.
6. **Paclik, P.** Dissimilarity-based classification of spectra: computational issues / P. Paclik, R.P.W. Duin // Real-Time Imaging. – 2003. – Vol. 9(4). – P. 237-244.
7. **Qun, G.** Comparison of several chemometric methods of libraries and classifiers for the analysis of expired drugs based on Raman spectra / Gao Qun, L. Yan, L. Hao [et al.] // Journal of Pharmaceutical and Biomedical Analysis. – 2014. – Vol. 94(0). – P. 58-64.
8. **Hartstra, J.** How to approach substance identification in qualitative bioanalysis / J. Hartstra, J.P. Franke, R.A. Zeeuw // Journal of Chromatography B: Biomedical Sciences and Applications. – 2000. – Vol. 739(1). – P. 125-137.
9. **Tan, N.** Application of multiple statistical tests to enhance mass spectrometrybased biomarker discovery / N. Tan, W. Fisher, K. Rosenblatt, H. Garner // BMC Bioinformatics. – 2009. – Vol. 10(1). – P. 144.
10. **Fisher, R.A.** Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population / R.A. Fisher // Biometrika. – 1915. – Vol. 10(4). – P. 507-521.
11. **Fisher, R.A.** On the “probable error” of a coefficient of correlation deduced from a small sample / R.A. Fisher // Metron. – 1921. –Vol. 1. – P. 3-32.
12. **Кочиков, И.В.** Численные процедуры иден тификации и восстановления концентраций веществ в открытой атмосфере при обработке единичного измерения фурье-спектрорадиометра / И.В. Кочиков, А.Н. Морозов, И.Л. Фуфурин // Компьютерная оптика. – 2012. – Т. 36, № 4. – С. 554-561.
13. **Глаголев, К.В.** Методика получения и обработки спектральной информации с помощью статического фурье-спектрометра / К.В. Глаголев, Иг.С. Голяк, Ил.С. Голяк [и др.] // Оптика и спектроскопия. – 2011. – Т. 110, № 3. – С. 486-492.
14. Светосильные спектральные приборы / В.А. Вагин, М.А. Гершун, Г.Н. Жижин, К.И. Тарасов. – М.: Наука, 1988. – 332 с.
15. Основы Фурье-спектрорадиометрии / А.Н. Морозов, С.И. Светличный. – М.: Наука, 2014. – 456 с.
16. **Голяк, Ил.С.** Беспроботборный анализ химических веществ с использованием статического фурье-спектрометра / Ил.С. Голяк, А.А. Есаков, Н.С. Васильев, А.Н. Морозов // Оптика и спектроскопия. – 2013. – Т. 115, № 6. – С. 990-994.

References

1. **Guti´errez-Rodriguez, A.E.** New dissimilarity measures for ultraviolet spectra identification / A.E. Guti´errez-Rodriguez, M.A. Medina-P´erez, J.F. Martinez-Trinidad [et al.] // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2010. – V. 6256 – P. 220-229.
2. **Stephen, S.E.** Optimization and testing of mass spectral library search algorithms for compound identification / S.E. Stein, D.R. Scott // Journal of the American Society for Mass Spectrometry. – 1994. – Vol. 5(9). – P. 859-866.
3. **Kruse, F.A.** The spectral image processing system (SIPS) – interactive visualization and analysis of imaging spectrometer data / F.A. Kruse, A.B. Lefkoff, J.W. Boardman [et al.] // Remote Sensing of Environment. – 1993. – Vol. 44(2-3). – P. 145-163.
4. **Paclik, P.** A study on design of object sorting algorithms in the industrial application using hyperspectral imaging / P. Paclik, R. Leitne, R.P.W. Duin // Journal of Real-Time Image Processing. – 2006. – Vol. 1(2). – P. 101-108.
5. **Bodis, L.** A novel spectra similarity measure / L. Bodis, Al. Ross, E. Pretsch // Chemometrics and Intelligent Laboratory Systems. – 2007. – Vol. 85(1). – P. 1-8.
6. **Paclik, P.** Dissimilarity-based classification of spectra: computational issues / P. Paclik, R.P.W. Duin // Real-Time Imaging. – 2003. – Vol. 9(4). – P. 237-244.
7. **Qun, G.** Comparison of several chemometric methods of libraries and classifiers for the analysis of expired drugs based on Raman spectra / G. Qun, L. Yan, L. Hao [et al.] // Journal of Pharmaceutical and Biomedical Analysis. – 2014. – Vol. 94(0). – P. 58-64.
8. **Hartstra, J.** How to approach substance identification in qualitative bioanalysis / J. Hartstra, J.P. Franke, R.A. Zeeuw // Journal of Chromatography B: Biomedical Sciences and Applications. – 2000. – V. 739(1). – P. 125-137.
9. **Tan, N.** Application of multiple statistical tests to enhance mass spectrometrybased biomarker discovery / N. Tan, W. Fisher, K. Rosenblatt, H. Garner // BMC Bioinformatics. – 2009. – Vol. 10(1). – P. 144.
10. **Fisher, R.A.** Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population / R.A. Fisher// Biometrika. – 1915. – Vol. 10(4). – P. 507-521.
11. **Fisher, R.A.** On the “probable error” of a coefficient of correlation deduced from a small sample / R.A. Fisher // Metron. – 1921. – Vol. 1. – P. 3-32.
12. **Kochikov, I.V.** Numerical procedures for substances identification and concentration calculation in the open athmosphere by processing a single ftir measurement / I.V. Kochikov, A.N. Morozov, I.L. Fufurin // Computer Optics. – 2012. – Vol. 36(4). – P. 554-561. – ISSN 0134-2452.
13. **Glagolev, K.V.** Technique for obtaining and processing spectral information with static fourier spectrometer / K.V. Glagolev, Ig.S. Golyak, Il.S. Golyak [et al.] // Optics and Spectroscopy. – 2011. – Vol. 110(3). – P. 449-455.
14. High luminosity spectral instruments / V.A. Vagin, M.A. Gershun, G.N. Zhizhin, K.I. Tarasov. – Moscow: “Nauka” Publisher, 1988. – 332 p. – (In Russian).
15. Basics of Fourier spectroradiometry / A.N. Morozov, S.I. Svetlichny. – Moscow: “Nauka” Publisher, 2014. – 456 p. – (In Russian).
16. **Golyak, Il.S.** Sampling-free analysis of chemical compounds using a static Fourier-transform spectrometer / Il.S. Golyak, A.A. Esakov, N.S. Vasilev, A.N. Morozov // Optics and Spectroscopy. – 2013. – V. 115(6). – P. 884-888.

SUBSTANCE IDENTIFICATION BY ERROR DEFORMED SPECTRA

N.S. Vasil'ev, A.N. Morozov
Bauman Moscow State Technical University

Abstract

Substance identification by their luminescence spectra is a highly sensitive and non distraction method. If a signal level is low then recognition errors may occur. The aim of this work was to define the identification algorithm with error probability control. For this purpose, the value of dissimilarity measure in the form of Spectral Angle Mapper (SAM) was analyzed. The relation between errors in measured spectra and the dissimilarity measure distribution was defined. The accuracy of the statistical hypothesis was used in spectral library search. The resulting algorithm was tested on more than 4000 sample spectra. The case when the measured spectra contained a signal of unknown source was analyzed, as well as the case when the measured spectra might contain either a signal or be equal to noise.

Key words: identification; dissimilarity measure; similarity index; match factor; database retrieval; luminescence; chemometrics; spectral library search; spectral angle mapper; SAM.

Сведения об авторах

Васильев Николай Сергеевич, 1986 года. Аспирант и ассистент кафедры физики Московского государственного технического университета им. Н.Э. Баумана. Область научных интересов: спектроскопия, распознавание образов.

E-mail*:* *nickliamg@gmail.com* .

Nikolay Sergeevich Vasil'ev, born in 1986. Post graduate student, assistant professor in Physics of Bauman Moscow State Technical University. His scientific interests include spectroscopy and pattern recognition.

Морозов Андрей Николаевич, 1959 года рождения. Доктор физико-математических наук (1994 год), профессор, работает заведующим кафедрой физики Московского государственного технического университета им. Н.Э. Баумана. Область научных интересов: прецизионные измерения, физическая кинетика и спектроскопия.

E-mail: *amor59@mail.ru* .

Andrey Nikolaevich Morozov, born in 1959, PhD (ScD) (1994), prof., a head of Physics department of Bauman Moscow State Technical University. His scientific interests include precision measurements, physical kinetics and spectroscopy.

Поступила в редакцию 22 июля 2014 г.