СТАТИСТИЧЕСКАЯ ОЦЕНКА ВЕРОЯТНОСТИ ПРАВИЛЬНОГО ОБНАРУЖЕНИЯ ВЕЩЕСТВ В ИК ФУРЬЕ-СПЕКТРОМЕТРИИ

А.Н. Морозов[1], И.В. Кочиков[2], А.В. Новгородская[1], А.А. Сологуб[1], И.Л. Фуфурин [1]
[1] МГТУ им. Н.Э.Баумана, Москва, Россия,
[2] НИВЦ МГУ им. М.В.Ломоносова, Москва, Россия

Аннотация

В работе рассмотрена задача идентификации веществ по инфракрасным спектрам. В качестве метода идентификации используется последовательное сравнение исследуемого спектра с эталонными спектрами веществ из базы данных. В качестве меры схожести между двумя спектрами используется статистический коэффициент корреляции Пирсона. Рассмотрен случай, когда исследуемый спектр является спектром некоторого вещества в виде заданного спектра из базы данных с аддитивной добавкой белого δ-коррелированного шума с Гауссовым распределением. В этом случае найдены вероятностные характеристики статистического коэффициента корреляции. Введено понятие правильного обнаружения вещества, и найдены теоретические выражения для вероятности правильного обнаружения. Разработана методика определения пороговых значений коэффициента корреляции по заданной вероятности правильной идентификации. В численных экспериментах показана применимость описанных методик.

Ключевые слова: спектроскопия, идентификация, вероятность обнаружения, вероятностный порог обнаружения.

Цитирование: **Морозов, А.Н.** Статистическая оценка вероятности правильного обнаружения веществ в ИК Фурье-спектрометрии / А.Н. Морозов, И.В. Кочиков, А.В. Новгородская, А.А. Сологуб, И.Л. Фуфурин // Компьютерная оптика. – 2015. – Т. 39, № 4. – С. 614-621. – DOI: 10.18287/0134-2452-2015-39-4-614-621.

Введение

Проблема дистанционного контроля загрязнения атмосферы, а также химического контроля воздушной среды промышленных и других важных объектов на сегодняшний день является достаточно актуальной. Причиной этому служит всё больший рост числа загрязняющих веществ, а также производств, выбросы которых не удаётся определить в непосредственном контакте. Вследствие этого уже более двух столетий ведутся разработки по созданию технологий и методов бесконтактной идентификации. Одним из характерных свойств, являющимся сугубо индивидуальным для каждого вещества, является оптический спектр, что позволяет использовать его как некий идентификатор. Одним из распространённых методов получения спектров веществ является инфракрасная (ИК) спектрометрия. Выбор именно ИК-области обусловлен тем, что излучение этого диапазона возбуждает колебательное движение молекул или их отдельных фрагментов, вследствие чего происходит ослабление интенсивности только на частотах колебаний молекул, поэтому спектры каждого из веществ являются уникальными, а линии спектра селективными и ярко выраженными.

ИК-спектрометры разделяются на дифракционные и фурье-спектрометры на базе интерферометров. В данной работе будет рассмотрен второй класс приборов, однако описанные методы можно применять и для дифракционного метода. Особенностью фурье-спектрометров является возможность работы с более низкими интенсивностями [1] нежели в дифракционных установках, что позволяет детектировать спектры собственного излучения.

Важной задачей, возникающей при разработке ИК фурье-спектрометра, является выбор способа идентификации вещества по восстановленному спектру. Разумеется, возможен анализ химического состава по наличию некоторых характерных полос в спектре, однако более надёжным и точным методом является последовательное сравнение полученного спектра со спектрами уже известных веществ, хранящихся в базе данных эталонных спектров. Очевидно, что идентифицируемый сигнал может быть значительно зашумлён, что ещё сильнее усложняет ситуацию. В настоящее время известно множество различных методов и средств решения задач распознавания, таких как: обучаемые нейронные сети [2], использование нечёткой логики [3], методы PCA (principal component analysis) [4], а также введение различных мер схожести между данными [5 – 11]. Основные работы по поиску и сравнению спектров веществ в базе данных были сделаны Клерком [12] и Луинджи [13]. Также фундаментальная работа по расчёту мер схожести в базе данных была проделана в [14]. Одной из возможных мер схожести является статистический коэффициент корреляции Пирсона [15] (здесь и далее будем считать, что спектры представлены в виде числовых векторов конечной длины):

$$r = \frac{(\vec{x} - \overline{x})^{\mathrm{T}} (\vec{y} - \overline{y})}{\left\| \vec{x} - \overline{x} \right\| \left\| \vec{y} - \overline{y} \right\|}, \qquad (1)$$

где $\vec{x}, \vec{y}$ – векторы сравниваемых спектров,

$\overline{x}$ – среднее арифметическое компонент вектора $\vec{x}$,

$\left\| \vec{x} \right\|$ – евклидова норма вектора.

Здесь и далее выражение $\vec{x} - \overline{x}$ означает вычитание из всех компонент вектора одинаковый скаляр.

Известно, что $r \in [-1,1]$, причём равенство единице достигается только если сигналы линейно зависи-

мы, что в случае спектров даёт повод говорить об их идентичности. Чем менее «схожи» спектры, тем ближе величина (1) к нулю. Заметим, что из стремления (1) к единице следует стремление нормы разности спектров к глобальному минимуму. В работах [16–22] описан метод идентификации по коэффициенту корреляции: сначала величина (1) рассчитывается для исследуемого спектра со всеми спектрами из базы данных, далее выбирается вещество с максимальным коэффициентом корреляции, и если (1) превысили заранее заданный эмпирический порог, то вещество считается идентифицированным.

1. Вероятностные характеристики коэффициента корреляции

Фишер и Кенни в [23, 24] получили точное теоретическое выражение для функции плотности вероятности коэффициента корреляции Пирсона в случае, когда сравниваются две случайные величины с двумерным гауссовым распределением в известной корреляции. Существенным ограничением является то, что оба вектора должны быть случайными. В свою очередь, при идентификации по базе данных в ИК фурье-спектрометрии эталонные спектры могут считаться точно известными, так как были получены в лабораторных исследованиях путём множественных усреднений. Также невозможно заранее знать вероятностный коэффициент корреляции без набора множественной статистики.

Предположим, что исследуемый вектор зашумлён белым аддитивным шумом, имеющим нормальное распределение, тогда итоговый спектр будет иметь такие же характеристики шума из-за свойств преобразования Фурье. То есть исследуемый спектр может быть представлен в виде:

$$\vec{x} = \vec{\tau}^* + \vec{\xi}, \quad \vec{y} = \vec{\tau}^*, \tag{2}$$

где $\vec{\tau}^*$ – вектор эталонного спектра, $\vec{\xi}$ – шумовой вектор, каждая компонента которого имеет нормальное распределение с известными характеристиками $\mathcal{N}(0,\sigma_\xi^2)$.

Обозначим через r^j коэффициент корреляции спектра $\vec{x}$ с веществом из базы данных под номером $j = 1,...,M$, где M – число веществ в базе данных. Будем считать, что база включает в себя вещество со спектром $\vec{\tau}^*$, и пусть его номер j^*.

Получим явное выражение для коэффициента корреляции в случае, когда спектр $\vec{x}$ сравнивается со спектром $\vec{\tau}^*$. Подставив (2) в (1), получим:

$$r^{j^*} = \frac{\hat{\sigma}_{\tau^*} + \sigma_\xi r_\xi^{\,j^*}}{\sqrt{\hat{\sigma}_{\tau^*}^2 + \delta^2 + 2\hat{\sigma}_{\tau^*}\sigma_\xi r_\xi^{\,j^*}}}, \tag{3}$$

где $\hat{\sigma}_{\tau^*}^2 = \dfrac{\left\|\vec{\tau}^* - \overline{\tau}^*\right\|^2}{N}$ – среднеквадратичное отклонение спектра $\vec{\tau}^*$, характеризующее его интенсивность,

$$r_\xi^{j^*} = \frac{\vec{\xi}^T(\vec{\tau}^* - \overline{\tau}^*)}{N\sigma_\xi\hat{\sigma}_{\tau^*}} \quad - \tag{4}$$

– коэффициент корреляции чистого шума с эталонным спектром $\vec{\tau}^*$,

$$\delta^2 = \frac{\left\|\vec{\xi} - \overline{\xi}\right\|^2}{N} \quad - \tag{5}$$

среднеквадратичное отклонение шума,

N – число точек в экспериментальном спектре.

Видно, что r^{j^*} является функцией от двух случайных величин (4) и (5), имеющих нормальное и хи-квадрат распределения соответственно. Точный анализ функции (3) в этом случае представлен в [25]. Однако можно воспользоваться тем фактом, что в рассматриваемых спектрах число N достаточно велико (обычно от 200 до 800) для того, чтобы рассматривать выражение (5) как точечную оценку дисперсии σ_ξ^2. Заметим, что в знаменателе (3) складываются δ^2 и величина $2\hat{\sigma}_{\tau^*}\sigma_\xi r_\xi^{\,j^*}$. Если принять $\delta^2 = \sigma_\xi^2$, то ошибка такой оценки будет составлять $\dfrac{(2N-2)\sigma_\xi^4}{N^2}$ [26], а дисперсию величины $2\hat{\sigma}_{\tau^*}\sigma_\xi r_\xi^{\,j^*}$ можно получить по теореме о сумме случайных величин: $\dfrac{4\sigma_\xi^2\hat{\sigma}_{\tau^*}^2}{N}$.

Если выполнено условие

$$\begin{gathered}\frac{(2N-2)\sigma_\xi^4}{N^2} \ll \frac{4\sigma_\xi^2\hat{\sigma}_{\tau^*}^2}{N} \;\Leftrightarrow \\ \Leftrightarrow \frac{\hat{\sigma}_{\tau^*}^2}{\sigma_\xi^2} \gg 0{,}5 - \frac{1}{2N}\end{gathered}, \tag{6}$$

то можно принять величину δ^2 за детерминированную и равную σ_ξ^2. В левой части выражения (6) стоит отношение сигнал/шум (*SNR*). Таким образом, условие (6) можно интерпретировать как требование значительного превышения уровня сигнала над уровнем шума.

С учётом (6) выражение (3) может быть представлено в виде:

$$r^{j^*} = \frac{\sqrt{SNR} + r_\xi^{\,j^*}}{\sqrt{SNR + 1 + 2\sqrt{SNR}\,r_\xi^{\,j^*}}}. \tag{7}$$

Сходный результат был получен в [27], однако при учёте того, что $r_\xi^{\,j^*} \equiv 0$, что возможно только при $N \to \infty$.

Коэффициент корреляции (7) является функцией только одной случайной величины с известным распределением:

$$w(r_\xi^{j^*}) = \sqrt{\frac{N}{2\pi}}\exp\left(-\frac{\left(r_\xi^{j^*}\right)^2 N}{2}\right). \tag{8}$$

С использованием функции (8), могут быть получены моменты всех порядков.

Для случая $j \neq j^*$ коэффициент корреляции запишется в виде

$$r^j = \frac{r^{jj^*}\sqrt{SNR} + r_\xi^j}{\sqrt{SNR+1+2\sqrt{SNR}r_\xi^{j^*}}}, \quad (9)$$

где введено $r^{jj^*} = \frac{(\vec{\tau}^j - \overline{\tau}^j)^T(\vec{\tau}^* - \overline{\tau}^*)}{\left\|\vec{\tau}^j - \overline{\tau}^j\right\|\left\|\vec{\tau}^* - \overline{\tau}^*\right\|}$, r_ξ^j определяется аналогично (4), но для спектра $\vec{\tau}^j$, $\vec{\tau}^j$ – эталонный спектр вещества j в базе данных.

Функция (9) зависит от двух случайных величин, причём стоит учитывать, что эти величины не являются независимыми. Чтобы найти их совместное распределение, нужно знать корреляционную функцию шума. Примем, что шум является некоррелированным $\langle \xi_i \xi_j \rangle = \sigma_\xi^2 \delta_{ij}$, где δ_{ij} – символ Кронекера. Тогда с учётом (8) ковариация коэффициентов корреляции запишется

$$\begin{aligned}\operatorname{cov}(r_\xi^{j^*}, r_\xi^j) &= \langle r_\xi^{j^*}, r_\xi^j \rangle - \langle r_\xi^{j^*} \rangle \langle r_\xi^j \rangle = \\ &= \frac{(\vec{\tau}^* - \overline{\tau}^*)^T(\vec{\tau}^j - \overline{\tau}^j)}{N^2 \hat{\sigma}_{\tau^*} \hat{\sigma}_{\tau^j}} = \frac{r^{jj^*}}{N},\end{aligned} \quad (10)$$

а вероятностный коэффициент корреляции будет равен r^{jj^*}.

Таким образом, получены упрощённые формулы для коэффициентов корреляции (7) и (9), а также найдены свойства случайных величин, входящих в них.

2. Правильное обнаружение вещества

Поскольку исследуемый спектр является зашумлённым, не всегда удаётся правильно идентифицировать вещество. Возможны случаи, когда, например, будет идентифицировано другое вещество, либо когда сигнал будет расценён как чистый шум и пропущен. На практике необходимо знать, с какой вероятностью было произведено обнаружение. Конкретизируем понятие правильного обнаружения.

Пусть исследуется спектр $\vec{x}$ вещества $j*$ из базы данных. Тогда обнаружение считается правильным, если коэффициент корреляции r^{j*} превысил коэффициент корреляции со всеми другими веществами, и к тому же превысил некоторый заранее заданный порог $r_{j^*}^*$.

Математически вероятность такого события запишется следующим образом

$$P_{correct} = P\left[\left(r^{j^*} > r^j, \forall j \neq j*\right) \wedge \left(r^{j^*} > r_{j^*}^*\right)\right] \quad (11)$$

Рассмотрим отдельно каждую из скобок. Подставим в (11) полученные ранее выражения (7) и (9), тогда

$$r_\xi^{j^*} - r_\xi^j > \sqrt{SNR}\left(r^{jj^*} - 1\right). \quad (12)$$

Введем величину Φ_j:

$$\Phi_j = r_\xi^{j^*} - r_\xi^j = \frac{\vec{\xi}^T\left(\vec{\tau}^* - \overline{\tau}^*\right)}{N\sigma_\xi \hat{\sigma}_{\tau^*}} - \frac{\vec{\xi}^T\left(\vec{\tau}^j - \overline{\tau}^j\right)}{N\sigma_\xi \hat{\sigma}_{\tau^j}}.$$

Из формулы (11) видно, что вероятность правильного обнаружения можно представить как функцию только одного случайного вектора $\vec{\Omega} = \left\{\Phi_1, \Phi_2, \ldots, r_\xi^{j^*}, \ldots, \Phi_M\right\}$, имеющего длину M (число веществ в базе данных), и на позиции j^* стоит величина $r_\xi^{j^*}$. Видно, что все компоненты вектора $\vec{\Omega}$ статистически зависимы и вместе подчиняются многомерному Гауссову распределению. Чтобы найти их совместное распределение, построим ковариационную матрицу для величин Φ_i и Φ_j. Воспользуемся условиями из (10) для корреляции шума, тогда

$$\operatorname{cov}\left(\Phi_i, \Phi_j\right) = \frac{1 + r^{ij} - \left(r^{ij^*} + r^{jj^*}\right)}{N}, \quad (13)$$

где $i, j \neq j^*$.

Ковариация величин $r_\xi^{j^*}$ и Φ_j будет равна:

$$\operatorname{cov}\left(r_\xi^{j^*}, \Phi_j\right) = \frac{1 - r^{jj^*}}{N}. \quad (14)$$

Рассмотрим выражение во второй скобке в (11). Подставив (7), получим:

$$\frac{\sqrt{SNR} + r_\xi^{j^*}}{\sqrt{SNR+1+2\sqrt{SNR}\, r_\xi^{j^*}}} > r_{j^*}^*.$$

Обозначим верхнее и нижнее решение этого выражения относительно $r_\xi^{j^*}$ как Γ_{up} и Γ_{down} соответственно.

Совместив (13) и (14) и с учетом формулы (8), получим итоговую ковариационную матрицу:

$$\Sigma = \frac{1}{N}\begin{pmatrix} 2\left(1 - r^{1j^*}\right) & 1 + r^{1,2} - \left(r^{1j^*} + r^{2j^*}\right) & \cdots & 1 - r^{1j^*} & \cdots \\ 1 + r^{2,1} - \left(r^{2j^*} + r^{1j^*}\right) & 2\left(1 - r^{2j^*}\right) & \cdots & 1 - r^{2j^*} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 - r^{1j^*} & 1 - r^{2j^*} & \cdots & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Совместная плотность вероятности для вектора $\vec{\Omega}$ будет иметь вид

$$w(\Omega) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} \exp\left(-\frac{1}{2}\Omega^T \Sigma^{-1} \Omega\right). \quad (15)$$

В итоге вероятность (11) может быть найдена как интеграл от (15) по области допустимых значений:

$$P_{correct} = \int_\Gamma w(\Omega)\, d\Omega, \quad (16)$$

где Γ – область, ограниченная $\Gamma_{up}, \Gamma_{down}$ и неравенством (12).

Опишем кратко полученный алгоритм расчёта вероятности правильного обнаружения:

1. Исследуемый спектр последовательно сравнивается со всеми веществами из базы данных.
2. Выбирается вещество, коэффициент корреляции с которым оказался максимальным.
3. Если максимальный коэффициент корреляции превысил эмпирический порог, то вещество считается идентифицированным.
4. Запоминается номер идентифицированного вещества.
5. По этому номеру строится ковариационная матрица Σ и находятся границы области значений Γ .
6. По формуле (16) находится вероятность правильного обнаружения.

3. Экспериментальная проверка

Для проверки полученных результатов были проведены численные эксперименты с реальной базой данных спектральных коэффициентов пропускания, состоящей более чем из 50 спектров. Численное моделирование проводилось в MATLAB путём генерации случайных величин с последующим расчётом различных статистических характеристик. Обычно эксперимент повторялся $1000-10000$ раз с различными веществами и статистически находились несмещённые вероятностные оценки математического ожидания.

На рис.1 изображено сравнение теоретических средних значений, полученных по формулам (7) и (9), с математическими ожиданиями коэффициента корреляции (1) (на рисунке показаны точками). Видно хорошее совпадение результатов, что говорит о применимости оценки (6) для расчётов моментов первого порядка. Далее, на рис. 2, показаны эксперименты по расчёту среднеквадратичного отклонения в сравнении с дисперсиями коэффициентов корреляции для двух пар тестовых веществ. Однако, в этом случае уже наблюдается расхождение при низких значениях отношения «сигнал/шум».

Из (15) и (16) видно, что для расчёта вероятности правильного обнаружения необходимо вести интегрирование по области очень большой размерности (в данном случае $\dim(\Gamma)=58$). Этот факт существенно замедляет скорость работы реальных систем и полностью исключает возможность работы в режиме реального времени. Однако в ходе численных экспериментов было установлено, что в качестве оценки величины (16) может выступать математическое ожидание вероятностей превышения коэффициента корреляции r^{j^*} над всеми остальными r^j :

$$P_{correct}=\frac{1}{M}\left[\sum_j P\left(r^{j^*}>r^j\right)\right]P\left(r^{j^*}>r^*_{j^*}\right), \qquad (17)$$

причём каждая вероятность в сумме (17) выражается через функции ошибок:

$$P\left(r^{j^*}>r^j\right)=\frac{1}{2}\times$$
$$\times\left(erf\left(\sqrt{\frac{N}{1-r^{jj^*}}}\right)+erf\left(\sqrt{\frac{N\cdot SNR(1-r^{jj^*})}{4}}\right)\right). \qquad (18)$$

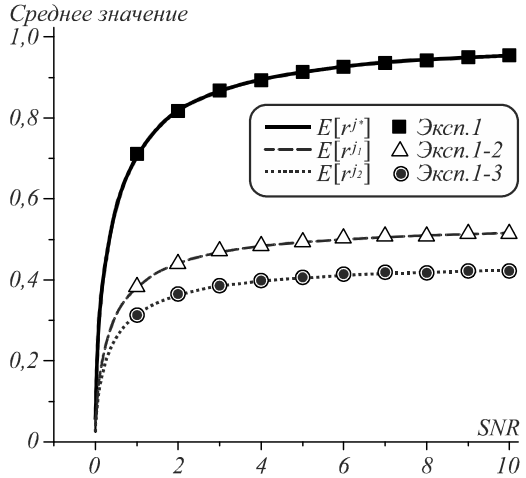

Рис.1. Зависимость среднего значения $E[r^j]$ коэффициента корреляции от отношения «сигнал/шум»

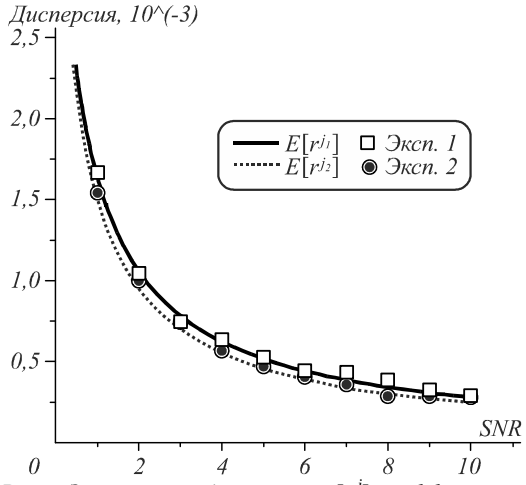

Рис.2. Зависимость дисперсии $D[r^j]$ коэффициента корреляции от отношения «сигнал/шум» SNR

Результаты сравнения (17) с экспериментом показаны на рис. 3. Каждая экспериментальная точка просчитывалась 1000 раз и затем усреднялась. Результаты совпадают даже для низких значений «сигнал/шум». Для случая $SNR > 1$ для всех веществ показано, что вероятность события $r^{j^*}<r^j$ крайне мала. Поэтому решающую роль играет правый множитель в (17):

$$P\left(r^{j^*}>r^*_{j^*}\right)=\int_{\Gamma_{down}\left(r^*_{j^*},SNR\right)}^{\Gamma_{up}\left(r^*_{j^*},SNR\right)} w\left(r_\xi^{j^*}\right)\mathrm{d}\,r_\xi^{j^*}. \qquad (19)$$

Выражение (19) удаётся представить через неэлементарные функции, однако запись является слишком громоздкой, чтобы приводить её в статье. Заметим, что с помощью (19) можно находить пороги для заданной вероятности правильного обнаружения. Хотя зависимость $r^*_{j^*}\left(P_{correct}\right)$ не выражается явно, опреде-

ление порога обнаружения при заданной вероятности может вестись с помощью алгоритма интерполяционного поиска, так как $P(r^*_{j*})$ является монотонно убывающей функцией. Трудоёмкость такого алгоритма [28] можно оценить как $O(\log_2[\log_2(1/\varepsilon)])$, где ε – допустимая погрешность. Этот результат даёт возможность находить пороги обнаружения веществ при различных отношениях «сигнал/шум», гарантируя заданную вероятность правильного обнаружения (рис. 4).

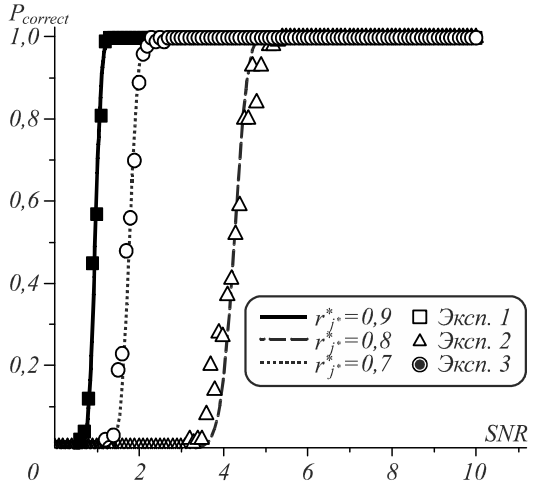

Рис. 3. Вероятность правильного обнаружения $P_{correct}$ от отношения «сигнал/шум» SNR

Заключение

В работе рассмотрен метод определения вероятностных характеристик коэффициента корреляции Пирсона в случае, когда один из сигналов имеет аддитивную добавку из гауссовского δ-коррелированного шума, а другой является незашумлённым. Данный метод предлагается применять для идентификации веществ по их спектрам в фурье-спектрометрии. Основное допущение метода основывается на замене величины с распределением хи-квадрат на её статистическую оценку. Показано, что такое приближение применимо при больших отношениях сигнал/шум.

На основе полученных вероятностных характеристик введено понятие вероятности правильного обнаружения вещества и найдено явное теоретическое выражение для этой вероятности. Также найдены упрощённые выражения, позволяющие находить вероятность гораздо быстрее при численном расчёте.

Показано, что вероятность правильного обнаружения зависит от порогового коэффициента корреляции, который ранее определялся эмпирическим путём. Предложена методика, позволяющая находить порог обнаружения по заданной вероятности правильного обнаружения и отношению сигнал/шум.

Разработанные методы проверены на реальной базе данных спектров веществ, состоящей из 58 веществ, и показана применимость предложенных методов при выполнении введённых приближений.

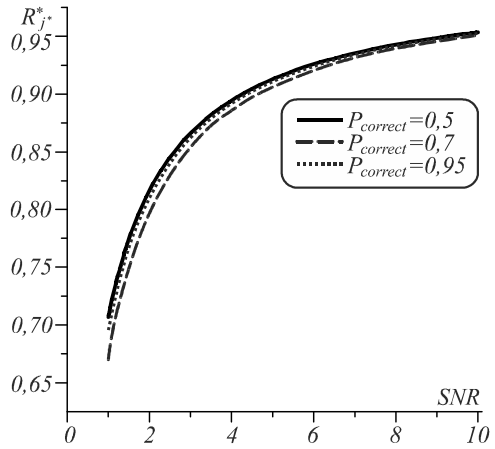

Рис. 4. Пороговый коэффициент корреляции в зависимости от отношения «сигнал/шум» SNR для заданной вероятности правильного обнаружения

Литература

1. Основы фурье-спектрорадиометрии / А.Н. Морозов, С.И. Светличный. – М.: Наука, 2006. – 275 с.
2. **Hemmer, M.C.** Prediction of Three-Dimensional Molecular Structures Using Information from Infrared Spectra / M.C. Hemmer, J. Gasteiger // Analitica Chimica Acta. – 2000. – Vol. 420(2) – P. 145-154. – ISSN 0003-2670.
3. Joined knowledge-and signal processing for infrared spectrum interpretation / F. Ehrentreich // Analitica Chimica Acta. – 1999. – Vol. 393. – P. 193-200. – ISSN 0003-2670.
4. **Schoonjans, V.** Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy / V. Schoonjjans, F. Questier // Journal of Pharmaceutical and Biomedical Analysis. – 2001. – Vol. 24. – P. 613-627.
5. **Лебедев, К.С.** Использование баз данных по ИК- и масс-спектрам для установления строения органических соединений / К.С. Лебедев // Журнал аналитической химии. – 1993. – Т. 48. – С. 851-863.
6. **Varmuza, K.** Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra / K. Varmuza, P.N. Penchev, H. Scsibrany // Journal of Chemical Information and Computer Sciences Impact. – 1998. – Vol. 38. – P. 420-427.
7. **Varmuza, K.** Large and frequently occurring substructures in organic compounds obtained by library search of infrared spectra / K. Varmuza, P.N. Penchev, H. Scsibrany // Vibrational Spectroscopy. – 1999. – Vol. 19. – P. 407-412.
8. **Penchev, P.N.** Characteristic substructures in sets of organic compounds with similar infrared spectra / P.N. Penchev, K. Varmuza // Computers&Chemistry. – 2001. – Vol. 25. – P. 231-237.
9. **Varmuza, K.** Evaluation of Hitlists from IR Library Searches by the Concept of Maximum Common Substructures / K. Varmuza, N.T. Kochev, P.N. Penchev // Analytical Sciences. – 2001. – Vol. 17. – P. 659-662.
10. **Derendyaev, B.G.** Taxonomy of structures selected from the IR spectroscopy database / B.G. Derendyaev // Journal of Structural Chemistry. – 2001. – Vol. 42(2). – P. 271-280.
11. **Ehrentreich, F.** Three-step procedure for infrared spectrum interpretation / F. Ehrentreich // Analitica Chimica Acta. – Vol. 427(2). – P. 233-244.
12. **Clerc, J.T.** Perfomance Analysis of Infrared Library Search Systems / J.T. Clerc, E. Pretsch, M. Zurcher // Microchimica Acta [Wien]. – 1986. – Vol. 2. – P. 217-242.

13. **Luinge, H.J.** Automated interpretation of vibrational spectra / H.J. Luinge // Vibrational Spectroscopy. – 1990. – Vol. 1. – P. 3-18.
14. **Zurcher, M.** General theory of similarity measures for library search systems / M. Zurcher, J.T. Clerc, M. Farkas, E. Pretsch // Analitica Chimica Acta. – Vol. 206(0). – P. 161-172.
15. **Pearson, K.** Notes on regression and inheritance in the case of two parents / K. Pearson // Proceedings of the Royal Society of London. – 1895. – Vol. 58. – P. 240-242.
16. **Кочиков, И.В.** Распознавание веществ в открытой атмосфере по единичной интерферограмме фурье-спектрорадиометра / И.В. Кочиков, А.Н. Морозов, С.И. Светличный, И.Л. Фуфурин // Оптика и спектроскопия. – 2009. – Т. 106, № 5. – С. 743-749.
17. **Harig, R.** Toxic cloud imaging by infrared spectrometry: A scanning FTIR system for identification and visualization / R. Harig, G. Matz // Field Analytical Chemistry&Technology. – 2001. – Vol. 5. – P. 75-90.
18. **Beil, A.** Remote sensing of atmospheric pollution by passive FTIR spectrometry in Spectroscopic Atmospheric Environmental Monitoring Techniques / A. Beil, R. Daum, G. Matz, R. Harig // Proceedings of SPIE. – 1998. – Vol. 3493. – P. 32-43.
19. **Clerbaux, C.** Trace gas measurements from infrared satellite for chemistry and climate applications / C. Clerbaux, J. Hadji-Lazaro, S. Turquety, G. Mégie, P.-F. Coheur // Atmospheric Chemistry and Physics. – 2003. – Vol. 3. – P. 1495-1508.
20. **Кочиков, И.** Численные процедуры идентификации и восстановления концентраций веществ в открытой атмосфере при обработке единичного измерения фурье-спектрорадиометра / И. Кочиков, А. Морозов, И. Фуфурин // Компьютерная оптика. – 2012. – Т. 36, № 4. – С. 554-561.
21. **Зайцев, К.И.** Высокоточное восстановление спектральных оптических характеристик среды с помощью импульсной терагерцовой спектроскопии / К.И. Зайцев, А.А. Гавдуш, В.Е. Карасик, С.О. Юрченко // Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки. – 2014. – № 3. – С. 69-92.
22. **Морозов, А.Н.** Физические основы расчёта интерферометра с вращающейся пластинкой / А.Н. Морозов, С.И. Светличный, С.Е. Табалин, И.Л. Фуфурин // Оптический журнал. – 2013. – Т. 80, № 8. – С. 37-41.
23. **Fisher, R.A.** On the probable error of a coefficient of correlation deduced from a small sample / R.A. Fisher // Metron. – 1921. – Vol. 1(4). – P. 3-32. – Retrieved 2009-03-25.
24. Mathematics of Statistics. Pt. 2 / J.F. Kenney, E.S. Keeping. – NY: D Van Nostrand Company, inc., 1951.
25. **Васильев, Н.С.** Идентификация веществ по сильно искажённым ошибками измерения спектрам / Н.С. Васильев, А.Н. Морозов // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 856-864.
26. Курс теории вероятностей и математической статистики для физиков / Ю.П. Пытьев, И.А. Шишмарёв. – М.: Издательство Московского университета, 1983. – 256 с.
27. **Benesty, J.** On the Importance of the Pearson Correlation Coefficient in Noise Reduction / J. Benesty, Chen Jingdong, H. Yiteng // Audio, Speech and Language Processing, IEEE Transaction on. – 2008. – Vol. 16(4). – P. 757-765.
28. **Perl, Y.** Interpolation search – a log logN search / Y. Perl, A. Itai, H. Avni // Communications of the ACM. – 1978. – Vol. 21(7). – P. 550-553.

References

[1] Morozov AN, Svetlichniy SI. Basics of Fourier Spectroradiometry. M: Science; 2006.
[2] Hemmer MC, Gasteiger J. Prediction of Three-Dimensional Molecular Structures Using Information from Infrared Spectra. Analitica Chimica Acta 2000; 420(2): 145-54.
[3] Ehrentreich F. Joined knowledge-and signal processing for infrared spectrum interpretation. Analitica Chimica Acta 1999; 393: 193-200.
[4] Schoonjans V, Questier F. Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy. Journal of Pharmaceutical and Biomedical Analysis 2001; 24: 613-27.
[5] Lebedev KS. Data base usage in IR and mass spectroscopy for the structural detection of organic substances. Journal of Analytical Chemistry 1993; 48: 851-63.
[6] Varmuza K, Penchev PN, Scsibrany H. Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra. Journal of Chemical Information and Computer Sciences Impact 1998; 38: 420-7.
[7] Varmuza K, Penchev PN, Scsibrany H. Large and frequently occurring substructures in organic compounds obtained by library search of infrared spectra. Vibrational Spectroscopy 1999; 19: 407-12.
[8] Penchev PN, Varmuza K. Characteristic substructures in sets of organic compounds with similar infrared spectra. Computers&Chemistry 2001; 25: 231-7.
[9] Varmuza K, Kochev NT, Penchev PN. Evaluation of Hitlists from IR Library Searches by the Concept of Maximum Common Substructures Analytical Sciences 2001; 17: 659-62.
[10] Derendyaev BG. Taxonomy of structures selected from the IR spectroscopy database. Journal of Structural Chemistry 2001; 42(2): 271-80.
[11] Ehrentreich F. Three-step procedure for infrared spectrum interpretation. Analitica Chimica Acta; 427(2): 233-44.
[12] Clerc JT, Pretsch E, Zurcher M. Perfomance Analysis of Infrared Library Search Systems. Microchimica Acta [Wien] 1986; 2: 217-42.
[13] Luinge HJ. Automated interpretation of vibrational spectra. Vibrational Spectroscopy 1990; 1: 3-18.
[14] Zurcher M, Clerc JT, Farkas M, Pretsch E. General theory of similarity measures for library search systems. Analitica Chimica Acta; 206(0): 161-72.
[15] Pearson K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 1895; 58: 40-242.
[16] Kochikov IV, Morozov AN, Svetlichniy SI, Fufurin IL. Substance detection in the free atmosphere by single interferogram in FTIR. Optics and Spectroscopy 2009; 106(5): 743-9.
[17] Harig R, Matz G. Toxic cloud imaging by infrared spectrometry: A scanning FTIR system for identification and visualization. Field Analytical Chemistry&Technology 2001; 5: 75-90. DOI: 10.1002/fact.1008.
[18] Beil A, Daum R, Matz G, Harig R. Remote sensing of atmospheric pollution by passive FTIR spectrometry in Spectroscopic Atmospheric Environmental Monitoring Techniques. Proceedings of SPIE 1998; 3493: 32-43.
[19] Clerbaux C, Hadji-Lazaro J, Turquety S, Mégie G, Coheur P-F. Trace gas measurements from infrared satellite for chemistry and climate applications. Atmospheric Chemistry and Physics 2003; 3: 1495-508. DOI:10.5194/acp-3-1495-2003.
[20] Kochikov I, Morozov A, Fufurin I. Numerical procedures for substances identification and concentration calculation in the open athmosphere by processing a single ftir measurement. Computer Optics 2012; 36(4): 554-61.

[21] Zaitschev KI, Gavdush AA, Karasik VE, Jurchenko SO. Highly accurate reconstruction of spectral optical characteristics of a medium using terahertz pulsed spectroscopy. Herald of the Bauman Moscow State Technical University. Ser. Natural Sciences 2014; 3: 69-92.
[22] Morozov AN, Svetlichniy SI, Tabalin SE, Fufurin IL. Physical basics of interferometer with rotating plate estimation. Journal of Optical Technology 2013; 80(8): 37-41.
[23] Fisher RA. On the probable error of a coefficient of correlation deduced from a small sample. Metron 1921; 1(4): 3-32. Retrieved 2009-03-25.
[24] Kenney JF, Keeping ES. Mathematics of Statistics. Pt. 2. NY: D Van Nostrand Company, inc. 1951.
[25] Vasiljev NS, Morozov AN. Subsnce identification by error deformed spectra. Computer Optics 2014; 38(4): 856-64.
[26] Pitjev JP, Shishmarev IA.The theory of probability for physics. Moscow State University Publishing; 1983.
[27] Benesty J, Chen Jingdong, Yiteng H. On the Importance of the Pearson Correlation Coefficient in Noise Reduction. Audio, Speech and Language Processing, IEEE Transaction on 2008; 16(4): 757-65.
[28] Perl Y, Itai A, Avni H. Interpolation search – a log logN search. Communications of the ACM 1978; 21(7): 550-3.

STATISTICAL ESTIMATION OF THE PROBABILITY OF THE CORRECT SUBSTANCE DETECTION IN FTIR SPECTROSCOPY

A.N. Morozov[1], I.V. Kochikov[2], A.V. Novgorodskaya[1], A.A. Sologub[1], I.L. Fufurin[1]
[1] Bauman Moscow State Technical University,
[2]Research Computer Center, M.V. Lomonosov Moscow State University

Abstract

In the present paper a problem of substance identification in FTIR (Fourier transform infrared) spectroscopy is considered. The spectral library hitlist search is chosen as the main tactic. In the paper the Pearson correlation coefficient as a similarity criterion between two spectra is suggested. A situation when one of the measured spectra has an additive narrowband white noise component with a Gaussian distribution is considered. In that case the probability density of the correlation coefficient is found. A concept of the probability of correct detection is proposed and a theoretical expression is found. In addition, we consider a boundary correlation coefficient search algorithm, which allows one to find a boundary value providing the required correct detection. Computational experiments have shown the applicability of the method.

Keywords: spectroscopy, identification, probability of correct detection, the boundary value of detection.

Citation: Morozov AN, Kochikov IV, Novgorodskaya AV, Sologub AA, Fufurin IL. Statistical estimation of the probability of the correct substance detection in FTIR spectroscory. Computer Optics 2015; 39(4): 614-21. DOI: 10.18287/0134-2452-2015-39-4-614-621.

Сведения об авторах

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, Ph.D. (Sc.D.) (1994), prof. is a head of Physics chair of Bauman Moscow State Technical University. His scientific interests include precision measurements, physical kinetics and spectroscopy.

Кочиков Игорь Викторович, 1959 года рождения. Доктор физико-математических наук (2003 год), работает ведущим научным сотрудником в Научно-исследовательском вычислительном центре МГУ им. М. В. Ломоносова. Область научных интересов – вычислительная электродинамика, исследование структуры молекул, распознавание образов и обработка изображений.

E-mail: *igor@kochikov.ru* .

Igor Viktorovich Kochikov, born in 1959, Ph.D. (Sc.D.) (2003) is a leading researcher in Scientific Research Computing Center of the Moscow University. His scientific interests include computational electrodynamics, molecular structure research, pattern recognition and image processing.

Новгородская Алла Викторовна, 1961 года рождения. В 1985 году окончила Московское высшее техническое училище им. Н.Э. Баумана по специальности 0609 «Гироскопические приборы и устройства», работает научным сотрудником в Центре прикладной физики МГТУ им. Н.Э.Баумана. Область научных интересов: спектроскопия, квантово-каскадные лазеры, патентоведение.

E-mail: *soullll@yandex.ru* .

Alla Viktorovna Novgorodskaya, (b.1961) graduated from Moscow Higher Technical School n.a. Bauman in 1985, majoring Giroscopic Instruments and Devices. Currently she works as researcher at the Center of Applied Physics Moscow State Technical University n.a. N.E. Bauman. Research interests are spectroscopy, QCL, patents.

Сологуб Александр Александрович, 1993 года рождения. В 2010 году поступил в Московский государственный университет им. М.В. Ломоносова на физический факультет. Работает техником в Центре прикладной физики МГТУ им. Н. Э. Баумана. Область научных интересов: спектроскопия, теория оптимизации, теория расписаний.

E-mail: *sologub10@gmail.com* .

Alexander Alexandrovich Sologub, born in 1993. In 2010 went to Lomonosov Moscow State University in Physical faculty. Works as a technician in Center of Applied Physics Moscow State Technical University n.a. N.E. Bauman. Research interests: spectroscopy, optimization theory, scheduling theory.

Фуфурин Игорь Леонидович, 1984 года рождения. Кандидат физико-математических наук (2010 год), работает доцентом кафедры физики Московского государственного технического университета им. Н.Э. Баумана. Область научных интересов – атмосферная оптика, спектроскопия и вычислительная математика.

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

Igor Leonidovich Fufurin, born in 1984, Ph.D (2010) is an associated professor of Physics chair of Bauman Moscow State Technical University. His scientific interests include atmospheric optics, spectroscopy and calculus mathematics.

Поступила в редакцию 26 июля 2015 г.
Окончательный вариант – 10 сентября 2015 г.