

## МЕТОДЫ И СИСТЕМЫ АВТОМАТИЗАЦИИ. ОБРАБОТКА ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ

УДК 539.534, 539.194

**А.П. Щербаков**

### **ПРИМЕНЕНИЕ МЕТОДОВ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ ДЛЯ ИДЕНТИФИКАЦИИ ЛИНИЙ В КОЛЕБАТЕЛЬНО-ВРАЩАТЕЛЬНЫХ СПЕКТРАХ**

Методы теории распознавания образов применены в задачах идентификации линий колебательно-вращательных спектров молекул. В основе предложенного способа лежит распознавание систем линий, удовлетворяющих комбинационному правилу Ридберга–Рица и наиболее «похожих» на рассчитанные с помощью некоторых оценок вращательных постоянных и моментов переходов. Создана экспертная система для анализа спектров молекул типа асимметричного волчка, способная производить поиск и отнесение линий автоматически. В системе реализованы модели эффективного вращательного гамильтониана Уотсона, аппроксимаций Падэ–Бореля, производящих функций, учтены резонансные взаимодействия. Для построения решающего правила в процессе обучения в экспертной системе реализован метод потенциальных функций, позволяющий восстанавливать решающее правило с помощью обучающей последовательности.

#### **1. Введение**

В данной статье сообщается о разработанной в лаборатории молекулярной спектроскопии ИОА СО РАН экспертной системе, способной в автоматическом режиме проводить идентификацию колебательно-вращательных линий.

Для решения задач атмосферной оптики необходимо знать спектральные характеристики линий поглощения атмосферных или загрязняющих атмосферу газов, в том числе и их квантовую идентификацию [1]. Использование спектроскопических данных, с не обоснованной или не подтвержденной точными расчетами идентификацией, может приводить к неверной интерпретации измерений, проводимых в атмосферных исследованиях. Например, в [2] был обнаружен аномально большой коэффициент сдвига давлением воздуха линии  $\text{H}_2\text{O}$   $13947,2608 \text{ см}^{-1}$ , отнесенной к переходу  $[624](301) \leftarrow [523](000)$ . Более поздний анализ показал, что в действительности линия  $13947,2608 \text{ см}^{-1}$  является дублетом, а большой по величине сдвиг обусловлен деформацией общего контура при возрастании давления [3]. Необходимо отметить, что ошибки интерпретации являются довольно частыми, особенно для слабых спектров.

В данной статье идентификация линий рассматривается как отдельная задача молекулярной спектроскопии, имеющая определенные специфические черты: «распознавание» при условии, что имеющиеся теоретические оценки положений линий, их интенсивностей недостаточно точные, а измеренные параметры линий зашумлены. В данной работе для решения задачи применяются уже хорошо разработанные методы теории распознавания образов. Необходимость в решении задачи такого рода в настоящее время вполне очевидна и обусловлена следующими обстоятельствами.

Способы идентификации линий в колебательно-вращательных (КВ) спектрах основываются на общих квантово-механических правилах, но практически определяются степенью сложности спектра. Так, в простых случаях, например в основных полосах поглощения гетероядерных двухатомных молекул ( $\text{HF}$ ,  $\text{HCl}$ ,  $\text{CO}$  и т.д.), когда имеются простые закономерности в расположении линий  $P$ - и  $R$ -ветвей, линии можно идентифицировать по их расположению относительно центра полосы, даже не прибегая к вычислению уровней энергии. В более сложных случаях, например в спектрах молекул типа асимметричного волчка, необходимы предварительный анализ взаимного расположения линий, их относительных интенсивностей и определение устойчивых (относительно малых вариаций спектроскопических постоянных) элементов спектра. Как следствие, нужны расчеты энергетических уровней и соответствующих волновых функций.

Различные способы идентификации линий в спектрах линейных молекул, молекул типа симметричного, сферического и асимметричного волчков описаны в книге Г. Герцберга [4].

При сильном колебательном возбуждении, когда энергия колебательно-вращательного взаимодействия, ангармонические поправки сравнимы с колебательной энергией, простые способы отнесения линий, следующие из модели малых колебаний, оказываются более неприменимыми. В спектрах высоких колебательных обертонов характерные особенности (положение наиболее сильных линий, серии дублетов и т.д.) могут меняться от полосы к полосе. В таких случаях приемы и способы идентификации усложняются, что, собственно, и вызывает необходимость разработки некоторого общего универсального метода идентификации.

Другая сторона рассматриваемой проблемы заключается в следующем. Многоатомные молекулы – водяной пар, метан, озон, представляющие интерес для атмосферных приложений, обладают сложными спектрами поглощения, содержащими тысячи или даже десятки тысяч линий. Анализ таких спектров, идентификация линий в них – трудоемкий процесс, и, как заметил Р. Jensen [5], «время, необходимое для регистрации спектра, является пренебрежимо малым по сравнению с временем, затрачиваемым на его теоретический анализ». В связи с этим появляется необходимость разработки соответствующих средств автоматизации, которые позволили бы быстро и эффективно работать с такими спектрами, проводить идентификацию линий, решать обратные задачи, создавать необходимые базы данных.

Представляемая экспертная система использует, в отличие от аналогичных программ [6–9], алгоритмы теории распознавания образов [13–15], «обучение» распознаванию, что позволяет достичь необходимой гибкости системы – возможности применения различных способов идентификации как с применением комбинационных правил для анализа сильных полос, так и для идентификации слабых, «одиночных» линий, соответствующих переходам на состояния с большими значениями углового момента, или линий слабых колебательных полос, для которых метод комбинационных разностей неприменим.

## 2. Количественные критерии для идентификации линий

Необходимо пояснить, что достоверная идентификация линий может быть сделана только по результатам расчетов *ab initio*. Однако расчеты *ab initio* центров и интенсивностей КВ-линий многоатомных молекул в настоящее время недостаточно точны для идентификации линий путем простого их сопоставления с наблюдаемыми характеристиками. В свою очередь, расчеты, использующие эмпирически определяемые функции потенциальной энергии, дипольного момента молекулы, в общем случае также не могут быть напрямую использованы для идентификации, по крайней мере для линий, образованных переходами на высоковозбужденные КВ-состояния, когда ошибки предсказания положений линий могут достигать нескольких десятков обратных сантиметров. Поэтому необходимо компенсировать недостаток точной информации, применяя, одновременно с расчетом, другие приемы отнесения.

Такие способы идентификации основываются на поиске характерных особенностей в спектре, например серий дублетов с отношением интенсивностей 1:3 (для молекул типа воды), на поиске подполос (в спектрах молекул типа симметричного волчка), сильных *Q*-ветвей для линейных молекул [4].

При рассмотрении задачи идентификации необходимо сформулировать общие критерии отнесения линий.

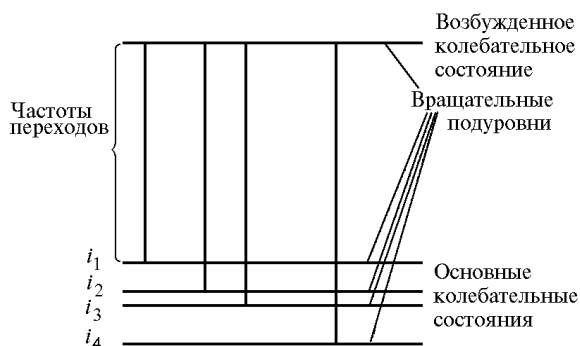


Рис. 1. Пример выполнения комбинационного правила

Согласно комбинационному правилу Ридберга–Рица (применительно к КВ-спектрам многоатомных молекул) в спектре имеются группы линий, образованных переходами с нескольких исходных состояний на одно конечное. На рис. 1 в качестве примера приведены вращательные подуровни энергии нижнего и верхнего колебательных состояний.

Таким образом, согласно комбинационному правилу необходимо найти в спектре систему линий, центры которых попарно удовлетворяют соотношению

$$\left| \nu_{n_1}^e - \nu_{n_2}^e \right| - \left| \nu_{i_1}^c - \nu_{i_2}^c \right| \leq \Delta\nu_{n_1} + \Delta\nu_{n_2}, \quad (1)$$

где  $\nu_{n_1}^e, \nu_{n_2}^e, \nu_{i_1}^c, \nu_{i_2}^c, \nu_{n_2}^c$  – измеренные и рассчитанные центры линий;  $n_1$  и  $n_2$  – номера линий;  $\Delta\nu_n$  – ошибки измерений частот;  $i_1, i_2$  – квантовые числа начальных уровней переходов  $f \leftarrow i_1$  и  $f \leftarrow i_2$  соответственно. Здесь предполагается, что энергия начальных уровней (для определенности мы будем считать, что это основное колебательное состояние) определена с точностью более высокой, чем  $\Delta\nu_n$ . Отметим, что разности (1) не зависят от ошибок расчета верхних уровней.

В [6] были предложены некоторые общие количественные критерии, которые можно применить для идентификации КВ-линий. Задача состоит в определении параметров  $\chi$  эффективного гамильтониана, параметров  $\mu$  эффективного оператора дипольного момента таких, чтобы для каждой рассчитанной линии с вычисленной интенсивностью  $S_{if}^c > S_{гр}$  в спектре находилась единственная линия такая, что

$$\left| S_{if}^c - S_n^e \right| < \Delta S_{if}^c - \Delta S_n^e, \quad (2)$$

$$\left| \nu_{if}^c - \nu_n^c \right| < \Delta\nu_{if}^c - \Delta\nu_n^c. \quad (3)$$

Параметр  $S_{гр}$  определяется порогом чувствительности прибора.

Данный критерий предполагает идентификацию линий, исходя из набора точных спектроскопических постоянных, и является по сути «естественным требованием» объяснения идентификации в рамках единой теоретической модели.

Процесс нахождения спектроскопических постоянных  $\chi, \mu$  является итерационным. Сначала с помощью начального приближения  $\chi = \chi_0, \mu = \mu_0$  идентифицируются наиболее сильные линии и параметры  $\chi, \mu$  уточняются путем решения обратной задачи. Далее по уточненным параметрам ищутся другие, более слабые, линии, вновь найденные линии также используются для дальнейшего уточнения параметров и т.д. Использование результатов обратных задач является в большинстве случаев вполне надежным способом проверки правильности отнесения линий, поскольку неверное отнесение линий приводит к резкому увеличению стандартного отклонения при подгонке параметров и может быть легко обнаружено.

### 3. Задачи распознавания образов (ЗРО)

Методы теории распознавания образов уже применялись для решения задач газоанализа и физики молекул [10–12], например для определения структуры молекул путем анализа ИК-спектров поглощения, электронно-колебательных спектров, спектров ЭПР, ЯМР и масс-спектров. Необходимо отметить, что описываемые ниже методы и результаты теории распознавания являются классическими и широко применяются для решения статистических задач кибернетики. Описание необходимо для того, чтобы дать представление о реализации метода в виде конкретных алгоритмов.

Пусть идентифицируемые объекты (в нашем случае это могут быть отдельные линии, группы линий, образующие характерные структуры в спектре) имеют  $N$  характеристик, тогда каждый объект будет иметь свой образ в  $N$ -мерном пространстве описаний  $X$ .

Пусть  $U$  – пространство решений  $U = \{u_1, u_2, u_3, \dots, u_m\}$ , а  $\Lambda$  – пространство ситуаций  $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_k\}$ . Формально решить ЗРО означает построить отображение

$$u(x) : X \rightarrow U \quad (4)$$

– наилучшее в некотором смысле. Критериями качества могут быть, например, минимум байесового риска, минимальность различия между  $u(x)$  и отображением  $u^*(x)$ , создаваемым учителем, в смысле некоторого введенного расстояния между двумя функциями, минимум числа ошибок, минимум вероятности ошибки и т.д.

В зависимости от используемого критерия качества могут быть реализованы различные методы (однако при некоторых условиях разные методы генерируют одинаковые классы процедур). Можно указать, по крайней мере, два различных подхода к решению ЗРО: это вероятностный подход (оценка совместных плотностей вероятности пространства описаний  $X$  и пространства ситуаций  $\Lambda$ ) и детерминистский подход (сведение ЗРО к задаче построения разделяющих поверхностей в пространстве описаний), однако оба эти подхода приводят к одному классу процедур [13].

Из данных выше определений следует, что построение решающего правила эквивалентно построению разделяющих поверхностей в признаковом пространстве, которые отделяют один класс объектов от другого. Для простоты будем рассматривать случай двух классов, а разделяющую поверхность обозначим как  $f(x)$ .

Пусть  $f^*(x)$  – искомая функция, определяющая разделяющую поверхность. В процессе обучения будем искать  $f(x)$ , ближайшую (в некотором смысле) к  $f^*(x)$ . Для этого нужно ввести соответствующий функционал расстояния. Достаточно широкий класс функционалов можно записать в виде

$$J = M_x\{Q(f(x), f^*(x))\}, \quad (5)$$

где  $M_x\{f(x)\} = \int f(x) p(x) dx$  – матожидание;  $Q(x, y)$  – некая функция (например,

$J = M_x\{(f(x) - f^*(x))^2\}$  будет соответствовать среднеквадратичному расстоянию и т.д.) Пусть  $f^*(x)$  представлена рядом

$$f(x) = \sum_{i=0}^{\infty} c_i \psi_i(x) = f(x, \mathbf{c}), \quad (6)$$

где  $\psi_i(x)$  – полный набор ортогональных функций. Тогда ЗРО сводится к отысканию коэффициентов  $c_i$  в разложении (6). Эта задача решается с помощью известного метода стохастической аппроксимации Роббинса–Монро [13, 14]:

$$\mathbf{c}[n+1] = \mathbf{c}[n] + H[n+1] \nabla_c Q\{f(x[n+1], \mathbf{c}[n]), f^*(x[n+1])\}, \quad (7)$$

где  $\mathbf{c}[n]$  – вектор-столбец параметров  $c_i$ , полученных на  $n$ -м (предыдущем) шаге обучения. Элементы диагональной матрицы  $H[n]$  задают величину шагов итерационной процедуры,  $x[n+1]$  – вектор-столбец признаков  $n+1$  элемента обучающей последовательности,  $\nabla_c$  – вектор-градиент по параметрам  $\mathbf{c}$ . В [13–15] приводятся условия и доказательства сходимости процедуры (7).

Если  $Q(f(x, \mathbf{c}), f^*(x)) = (f(x, \mathbf{c}) - f^*(x))^2$ , то (7) можно представить в виде

$$\mathbf{c}[n+1] = \mathbf{c}[n] + \gamma^T[n+1] \left( \sum_i c_i[n] \psi_i(x[n+1]) - f^*(x[n+1]) \right) \Psi(x[n+1]), \quad (8)$$

где  $\Psi(x[n+1])$  обозначает вектор-столбец, составленный из базисных функций  $\psi_i(x)$ , вычисленных в точке  $x[n+1]$ ;  $T$  – знак транспонирования.

Равенство (7) определяет простой и достаточно общий способ построения разделяющей поверхности в процессе обучения. Очевидно, что этот способ может быть легко применен и для случая нескольких классов распознаваемых объектов. Вследствие простоты и удобства алгоритмизации процедуры метода стохастической аппроксимации являются весьма популярными среди специалистов по автоматике. Они дают возможность на любом этапе  $n$  использовать разделяющую функцию (6) и при необходимости поправлять ее параметры, применяя (7).

Одним из методов, применяющихся в задачах обучения для построения разделяющей поверхности (а также для оценки плотностей распределения), является метод потенциальных функций (МПФ). Процедуры этого метода входят в класс процедур экстремизации Роббинса–Монро [14].

В так называемой *машинной* реализации МПФ приближение  $f(x)$  функции  $f^*(x)$  на шаге  $N+1$  берется в виде

$$f[n+1](x) = f[n](x) + \gamma[n+1] K(x[n+1], x), \quad (9)$$

где потенциальная функция  $K(x, y)$  удовлетворяет следующим условиям:

$$K(x, y) \geq 0, \quad K(x, y) \leq K(x, x), \quad \forall x, y, \quad (10)$$

$$K(x, y) < \infty, \quad K(x, y) = K(y, x),$$

$\gamma [n]$  – числа, которые определяют степень изменения разделяющей поверхности в процессе обучения, их можно называть «дозами обучения». «Дозы обучения» необходимо выбирать по следующему правилу:

$$\gamma [n] = \begin{cases} > 0, & \text{если } f[n](x) \leq 0, f^*(x) > 0, \\ < 0, & \text{если } f[n](x) > 0, f^*(x) \leq 0, \\ 0, & \text{в другом случае;} \end{cases} \quad (11)$$

$$\gamma [n] = r(f(x), f^*(x)),$$

и  $r(x, y)$  – некоторая функция такая, что  $|r(x, y)| \leq A |x - y| + B$ ,  $A$  и  $B$  – константы.

Из равенства (9) следует, что функцию  $f[n](x)$  ( $f(x)$  на  $n$ -м шаге обучения) можно записать в виде

$$f[n](x) = \sum_{i=1}^n \gamma [i] K(x[i], x). \quad (12)$$

Возьмем  $K(x, y)$  в виде

$$K(x, y) = \sum_{i=1}^M \lambda_i^2 \psi_i(x) \psi_i(y), \quad (13)$$

где  $\psi_i(x)$ ,  $i = \overline{1, \infty}$  – ортонормированный набор функций. Можно доказать, что такая функция удовлетворяет перечисленным выше условиям (10) [13]. Обозначим  $\varphi_i(x) = \lambda_i \psi_i(x)$ , тогда  $f(x) = \sum_{i=1}^M c_i \varphi_i(x)$ . Процедура

$$\mathbf{c}[n+1] = \mathbf{c}[n] + \gamma [n+1] \Phi(x[n+1]), \quad (14)$$

где  $\gamma [i]$  – числовая последовательность, удовлетворяющая условиям (11), называется *перцептронной* реализацией МПФ. Процедуры МПФ экстремизируют функционал

$$J = M_x \{f(x) [\text{sign } f(x) - \text{sign } f^*(x)]\}. \quad (15)$$

Необходимо заметить также, что описанные выше методы работают лишь при соблюдении гипотезы о компактности образов [15]. Приведенное в [15] понятие о компактности образов, если говорить кратко, означает, что искомая разделяющая поверхность должна быть достаточно гладкой – поскольку только в этом случае можно обойтись сравнительно небольшим числом слагаемых в разложении (6).

Одной из реализаций МПФ является перцептрон Розенблатта [15]. В терминах метода потенциальных функций перцептрон Розенблатта использует базисные функции вида

$$\varphi_i(x) = \lambda_i \psi_i(x) = \text{sign} \left[ \theta \left( \sum_j a_j^i x_j \right) \right], \quad (16)$$

где  $\theta(x)$  – некоторая функция;  $a_j^i$  – числовые коэффициенты. В работе [13] показана ортогональность такой системы функций. Алгоритм обучения второго слоя перцептрона Розенблатта  $u_k = \sum_i \mu_i^k \varphi_i(x)$ ,  $k = \overline{1, K}$ , где  $K$  – число классов, в векторной форме выглядит следующим образом:

$$\boldsymbol{\mu}^k[n+1] = \boldsymbol{\mu}^k[n] + T_k[n+1] \Phi(x[n+1]). \quad (17)$$

Здесь  $\mu^k[n]$  – вектор коэффициентов  $\mu^k$ , полученный после  $n$ -го шага обучения;  $\Phi(x)$  – вектор базисных функций в точке  $x$ , и числа  $T_k$  выбираются согласно правилу

$$T_k = \begin{cases} 1, & \text{если } u_k(x) \leq 0, \quad x \in U_k \\ -1, & \text{если } u_k(x) > 0, \quad x \notin U_k \\ 0. & \end{cases} \quad (18)$$

Подробное описание метода приведено также в [15].

#### 4. Применение методов распознавания образов для анализа спектров

Как уже было отмечено, при любом способе идентификации необходимым этапом являются сравнение и оценка «похожести» полученных при вычислении и «наблюдаемых» в спектре характерных особенностей. Обычно в спектре можно найти несколько наборов линий, «похожих» на искомый, и каждый из вариантов анализируется с помощью некоторых дополнительных критериев. Единственный правильный набор выбирается спектроскопистом по ряду признаков – определенных сочетаний характеристик линий (примерами могут быть соотношение интенсивностей, отклонение измеренных центров и интенсивностей от расчетных, точность выполнения комбинационного правила в сравнении с точностью измерений и т.д.). Такая задача является типичной задачей распознавания, поэтому естественно применить для автоматизации процесса анализа спектров уже хорошо разработанные методы и алгоритмы теории распознавания образов и математической статистики. Применение методов распознавания позволяет ввести количественные критерии «аналогичности» распознаваемых спектральных элементов определенным теоретическим представлениям и, следовательно, обосновать интерпретацию линий.

Для применения теории распознавания необходимо конкретизировать объекты, подлежащие распознаванию (т.е. характерные особенности), их признаки. Анализ показывает, что для идентификации линий необходимо использовать одновременно комбинационное правило Ридберга–Рица [4] – единственное точное правило, не зависящее от модельных представлений (гамильтониана молекулы, параметров потенциальной функции, дипольного момента и т.д.), и рассчитанные характеристики линий (их положение и интенсивность), что обеспечивает достаточную надежность и общность правил распознавания. В то же время другие особенности, которые могут присутствовать в спектрах, – дублеты, наиболее сильные линии, подполосы и т.д., зависят от типа молекулы, вращательных постоянных, случайных резонансов и т.д.

Задача распознавания сводится, таким образом, к определению «похожести» рассчитанных и «наблюдаемых» характеристик в данной совокупности линий, приближенно удовлетворяющих комбинационному правилу. Очевидно, что в этом случае желательно использовать алгоритмы с «обучением» распознаванию, чтобы получить достаточно гибкую систему.

Каждая линия анализируемого набора имеет определенные характеристики, например отклонение от расчета по частоте, интенсивности, точности выполнения комбинационного правила т.д. В процессе построения системы были выявлены следующие характеристики, которые целесообразно использовать как признаки:

- 1) отклонение измеренной частоты линии от расчетной

$$x_0 = \ln(1 + |v^e - v^c| k_0^0) k_0^1; \quad (19)$$

- 2) отклонение измеренной интенсивности линии от расчетной

$$x_1 = \ln(1 + |s^e - s^c| k_1^0) k_1^1; \quad (20)$$

- 3) точность выполнения комбинационного правила для данной линии

$$x_2 = \ln(1 + |E^c - \bar{E}^c| k_2^0) k_2^1; \quad (21)$$

- 4) отклонение измеренной интенсивности линии от вычисленной средней интенсивности (по всему набору)

$$x_3 = \ln(1 + |s^e - \bar{s}^c| k_3^0) k_3^1; \quad (22)$$

- 5) средняя пропорциональность измеренных и расчетных интенсивностей линий

$$x_4 = 1 - K\{s^c, s^c\}, \quad (23)$$

где  $K\{s^c, s^c\}$  – коэффициент корреляции между вычисленными и найденными интенсивностями по проверяемому набору:

$$K\{x, y\} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}};$$

6) отношение суммарной интенсивности всех найденных линий к вычисленной общей интенсивности линий

$$x_5 = \ln \left( 1 + \ln \left( \left| \frac{\sum S^e}{\sum S^c} \right| \right) k_4^0 \right) k_4^1, \quad (24)$$

7) точность соблюдения комбинационного правила для всей группы

$$x_6 = \sqrt{\frac{1}{n-1} \sum_i (E_i^e - \bar{E}^e)^2} \times 1000. \quad (25)$$

Константы  $k_j^i$  подбираются так, чтобы диапазоны изменения признаков были сосредоточены на отрезке  $[-10; 10]$ , содержащем начало координат. Это обеспечивает компактность изображения классов в признаковом пространстве. Логарифмирование в (19)–(25) применялось для сужения диапазонов возможных значений величин  $x_n$ .

Обучение и проверка информативности признаков проводились на примере спектра поглощения высокого разрешения  $H_2O$  в области 1,4 мкм, идентификация линий в этой области известна [16]. Отладка признаков производилась поочередно на примере сильной полосы  $\nu_1 + \nu_3$ : все признаки, кроме одного, опускались и варьированием формулы добивались, чтобы классификатор указывал правильные варианты.

Величины  $\gamma[n]$  в (11) были взяты в виде

$$\gamma[n+1] = \begin{cases} +r & \text{если } f(x, c) \leq 0 \text{ и линия правильная} \\ 0 & \\ -r & \text{если } f(x, c) > 0 \text{ и линия неправильная} \end{cases} \quad (26)$$

$$r[n] = 0,2 f(x[n+1]) + 1,0 \cdot 10^{-7}. \quad (27)$$

(После отправки в редакцию данной статьи были разработаны выражения вместо (26) и (27), дающие более качественную и устойчивую работу.)

Базисные функции на пространстве векторов  $X = R_7$  есть произведения вида

$$\Psi_n(x) = \prod_{i_1+i_2+\dots+i_{10} \leq 3}^{s=1,7} \Psi_{i_s}(x_s), \quad (28)$$

где  $\Psi_i(x) = N_i \exp\{-x^2/2\} H_i(x)$ ;  $H_i(x)$  – полиномы Эрмита;  $N_i = \sqrt{2^{i+1}/i!}$  – нормировочные множители. Величины  $\lambda_i$  в (13) взяты равными 1.

В экспертной системе, кроме описанной выше реализации, был опробован также перцептрон Розенблатта. При этом в качестве пространства описаний  $X$  взято пространство векторов  $\chi$ , имеющих вид

$$\chi = (\chi_1^1, \chi_2^1, \dots, \chi_{20}^1, \chi_1^2, \chi_2^2, \dots, \chi_{20}^2, \dots, \chi_1^6, \chi_2^6, \dots, \chi_{20}^6), \quad (29)$$

компоненты  $\chi_j^i, j = \overline{1, 20}$  вычисляются следующим образом:

$$\chi_j^i(x) = \begin{cases} 1 & \text{если } x_i \in [-10 + j; -10 + j + 1], \\ 0 & \text{если } x_i \notin [-10 + j; -10 + j + 1]; \end{cases} \quad (30)$$

величины  $x_i$  вычисляются по формулам (19)–(25).  
В терминах метода МПФ в данном случае базисные функции взяты

$$\varphi_k(x) = \text{sign} \left[ \sum_{j=1}^{20} \sum_{i=1}^6 a_{ij}^k \Theta_{[-10+j; -10+j+1]}(x_i) \right], \quad (31)$$

где  $k = \overline{1, 120}$ ;  $\Theta_{[a; b]}(x)$  – индикатор отрезка  $[a; b]$  равен 1, если  $x \in [a; b]$ , и равен 0 в остальных случаях;  $a_{ij}^k = \delta_{k(20i+j)}$ ;  $\delta_{ij}$  – символ Кронекера.

Распознавание переходов происходит следующим образом. Сначала производится поиск всех групп линий, удовлетворяющих комбинационному правилу (1). Найденные варианты обрабатываются классификатором, который находит и отмечает правильные, включается режим просмотра вариантов и на экран выводится наилучший из них (который содержит линию с наибольшим показанием функции  $f(x)$ ).

### 5. Экспертная система для идентификации колебательно-вращательных линий

С использованием описанных методов была разработана экспертная система, позволяющая проводить идентификацию спектров в автоматическом и в полуавтоматическом режимах. Блок-схема системы приведена на рис. 2.

Система содержит блоки расчета уровней энергии, волновых функций, центров линий и интенсивностей, блок распознавания и обучения, специализированные базы данных, блоки редактирования и управления базами. Специальный блок «инициализации» содержит программы подготовки спектров к работе, проверки и подготовки необходимых рабочих файлов.

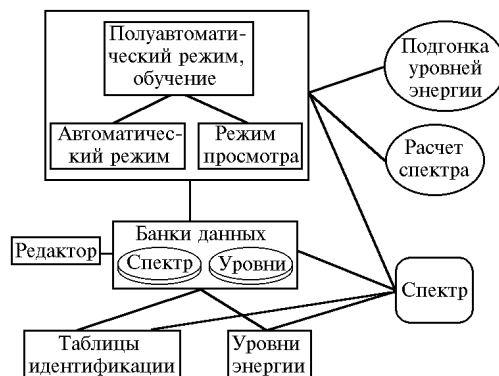


Рис. 2. Блок-схема экспертной системы для идентификации КВ-линий

Блоки распознавания – обучаемые, в ходе работы легко дообучать систему для конкретного спектра или обучать заново, благодаря встроенным средствам диалога между системой и пользователем.

Система может работать с молекулами различного типа (в данном варианте реализованы два случая –  $C_s$  и  $C_{2v}$ ). Для расчета уровней энергии, частот и интенсивностей линий может использоваться эффективный вращательный гамильтониан Уотсона [17], аппроксиманты Падэ–Бореля [18] и метод производящих функций [19]. Максимальный размер обрабатываемого спектра – 32000 линий.

Исходными данными для работы системы являются параметры гамильтониана для верхних состояний (начальное приближение), параметры основного состояния, параметры дипольного момента и обрабатываемый спектр.

Результаты накапливаются в специализированных банках данных для полученных «экспериментальных» верхних уровней и для спектра с идентификацией. Имеется возможность получения готовых таблиц «экспериментальных» уровней энергии, файлов, содержащих центры, интенсивности и квантовую идентификацию линий.



Работа с системой производится следующим образом. В начале работы, с помощью программы инициализации, пользователь устанавливает имена файлов со спектром (списком частот и интенсивностей линий), параметрами модели для верхних и для нижних состояний, параметрами дипольного момента, имена банков данных, устанавливает тип молекулы, используемую модель, режимы поиска и указывает некоторые из условий эксперимента. После этого можно войти в режим работы со спектром. Сначала выбирается колебательно-вращательное состояние, с которым нужно работать. Затем пользователь может включить поиск линий с использованием вычисленных или экспериментальных уровней энергии нижнего состояния.

Если найден хотя бы один вариант комбинационной разности, то включается режим просмотра вариантов и на экран выводится наилучший из них (если система обучена) в виде списка квантовых чисел и уровней энергии нижнего состояния, расчетных и экспериментальных частот и интенсивностей и экспериментальных верхних уровней энергии, получающихся при сложении энергии нижнего состояния и измеренного центра линии. В этом режиме можно посмотреть, исправить, добавить или убрать идентификацию найденных линий, занести или стереть в банке данных для уровней энергии значение верхнего экспериментального уровня энергии, усредненного по всем помеченным пользователем линиям. Каждая линия может иметь до трех идентификаций.

Если ни один вариант не найден, то включается режим просмотра ближайших линий, где можно идентифицировать одиночные линии. В этом режиме имеются такие же возможности по работе с банками данных, как и в режиме просмотра комбинационных разностей. В обоих режимах можно посмотреть весь спектр подряд. Режим сплошного просмотра спектра также позволяет работать с банками данных. После возврата из режимов просмотра можно выбрать другое колебательно-вращательное состояние и продолжить работу или вызвать подпрограмму инициализации (о которой говорилось выше) и изменить режимы поиска, посмотреть содержимое банков данных с результатами, вывести их содержимое в текстовый файл в виде таблиц. Можно также включить автоматическую обработку, начиная с текущего состояния и до какого-либо другого состояния, с различными режимами для нахождения одиночных линий (пропустить, обработать или обратиться к пользователю).

Данная система применялась для идентификации линий в спектрах поглощения HDO, D<sub>2</sub>O, H<sub>2</sub>O, H<sub>2</sub><sup>18</sup>O.

Автор выражает благодарность Л.Н. Синице, А.Д. Быкову, О.В. Науменко и Т.М. Петровой за помощь в работе, ценные замечания и предоставление ряда программ вычисления параметров линий.

Работа была выполнена при поддержке Российского фонда фундаментальных исследований (грант N 96-07-89321).

1. Зуев В.Е., Макушкин Ю.С., Пономарев Ю.Н. Современные проблемы атмосферной оптики. Т. 3. Спектроскопия атмосферы. Л.: Гидрометеиздат, 1987. 247 с.
2. Bösenberg J. Measurements of the pressure shifts of water vapor absorption lines by simultaneous photoacoustic spectroscopy // *Appl. Opt.* 1985. V. 24. N 21. P. 3531–3534.
3. Grossman B.E., Browell E.V. Water-Vapor Line Broadening and Shifting by Air, Nitrogen, Oxygen, and Argon in the 720-nm Wavelength Region // *J. Mol. Spectrosc.* 1989. V. 138. N 3. P. 562–595.
4. Герцберг Г. Колебательно-вращательные спектры многоатомных молекул. М.: Изд-во иностр. лит-ры, 1949. 403 с.
5. Jensen P. High Resolution FT-IR Spectroscopy with a PC // BRUKER FT-IR application note 35.
6. Pine A.C., Dresselhaus G., Palm B., Davies W., Glought S.A. Analysis of 4  $\mu\text{m}$   $\nu_1 + \nu_2$  Combination Bands of SO<sub>2</sub> // *J. Mol. Spectrosc.* 1977. V. 67. N 1–3. P. 386–415.
7. Kozin I.N., Klee S., and Jensen P. The Far-Infrared Fourier Transform Spectrum of H<sub>2</sub>Se // *J. Mol. Spectrosc.* 1993. V. 157. P. 409–422.
8. Науменко О.В. Методы и некоторые результаты решения обратной спектроскопической задачи для высоковозбужденных колебательных состояний молекул типа асимметричного волчка: Дис. канд. физ.-мат. наук. Томск, 1986.
9. Urban S. and Behrend J. A Software Package for Assignment of Vibration-Rotation Bands of Asymmetric Top Molecules // *Proc. Forteen Colloquium on High Resolution Molecular Spectroscopy.* 1995. P. B30.
10. Эляшберг М.Е., Грибов Л.А., Серов В.В. Молекулярный спектральный анализ и ЭВМ. М.: Наука, 1980. 305 с.
11. Катаев М.Ю., Мицель А.А. Идентификация состава газовых смесей по спектрам поглощения // *Автометрия.* 1985. N 4. С. 15–20.
12. Катаев М.Ю., Мицель А.А., Тарасова С.Р. Выбор информативных спектральных участков для решения задач газоанализа с помощью ОАД // *Оптика атмосферы.* 1990. Т. 3. N 8. С. 832–841.
13. Айзерман М.А., Браверман Э.И., Розоноэр Л.И. Метод потенциальных функций в задачах обучения машин. М.: Наука, 1970. 850 с.

14. Невельсон М. Б., Хасьминский Р. З. Стохастическая аппроксимация и рекуррентное оценивание. М.: Наука, 1972. 295 с.
15. Левин Л. Л. Введение в теорию распознавания образов. Томск: Изд-во ТГУ, 1982. 250 с.
16. Robert A. Toth, Extensive measurements of H<sub>2</sub><sup>16</sup>O line frequencies and strengths: 5750 to 7965 cm<sup>-1</sup> // Appl. Optics. 1994. V. 33. N 21. P. 4851–4867.
17. Watson J. K. G. Determination of centrifugal distortion coefficients of asymmetric top molecules // J. Chem. Phys. V. 46. P. 1935–1948.
18. Polyanski O. L. One-Dimensional Approximation of Effective Rotational Hamiltonian of the Ground State of the Water Molecule // J. Mol. Spectrosc. 1985. V. 112. P. 79.
19. Starikov V. I., Tashkun S. A., and Tyuterev V. I. G. Description of Vibration-Rotation Energies of Nonrigid Triatomic Molecules Using Generating Function Method // J. Mol. Spectrosc. 1992. V. 151. P. 130–147.

Институт оптики атмосферы СО РАН,  
Томск

Поступила в редакцию  
3 марта 1997 г.

**A. P. Shcherbakov. Pattern Recognition Method in Line Assignment of Vibrational-Rotational Spectra.**

The pattern recognition methods are used for the line assignment of vibrational-rotational spectra. The method is based on the recognition of sets of lines which satisfy the Ridberg–Rits combinational rule and have line positions and line strengths most similar to calculated ones when using the estimated rotational and centrifugal distortion constants and dipole transitions moments. The specialized computer Expert System is created for analysis of asymmetric rotor-type molecular spectra, allowing automated line assignment. The Expert System uses models such as Watson-type effective rotational Hamiltonian, Pade–Borel approximations, and Generation functions method, the resonance interactions are explicitly taken into account. To build the decision rule, the Potential Function Method is used which allows one to determine the decision rule using some sequence of teacher's answers.