

Разработка программного комплекса спектральный анализ спектров веществ для низкочастотного ЯМР спектрометра низкого разрешения

Аннотация

В данном проекте представлен комплекс программного обеспечения, предназначенный для проведения анализа спектров, полученных с помощью низкочастотных ЯМР спектрометров, в том числе низкого разрешения. Комплекс является комплементарным к проекту кафедры ЯМР СПбГУ, представляющим из себя доступный ЯМР спектрометр, работающий со слабыми полями. (не финальная версия)

ОГЛАВЛЕНИЕ

1. Введение	2
2. Обработка	3
2.1 Цель обработки	3
2.2 Предобработка	4
2.3 Корректировка	5
2.3.1 Первичная корректировка фазы.....	5
2.3.2 Разработанная система корректировки фазы	6
2.3.3 Свёртка	7
2.3.4 Компенсация опорной линии	8
2.4 Анализ	10
3. Определение структуры сканируемой молекулы	12
4. Реализация и заключение	14
5. Литература (неактуально).....	15

1. Введение

Ядерный магнитный резонанс (ЯМР) является одним из самых мощных бесконтактных, неразрушающих методов для исследования и анализа вещества, а также практически самым тончайшим инструментом измерения напряжённости магнитных полей. Ни одно химическое производство не обходится без ЯМР-спектрометра. С появлением ЯМР химики получили уникальный метод, который позволяет не только определять химический состав вещества, но и геометрию самой молекулы с точными расстояниями между отдельными атомами в молекуле. Таким образом, ЯМР-спектр каждого соединения уникален.

На данный момент рынок сложных ЯМР спектрометров хорошо освоен несколькими крупными предприятиями, но большинство импортных ЯМР-спектрометров стоят сотни тысяч евро из-за сложных сверхпроводящих магнитных систем, входящих в их конструкцию. Кроме того, они громоздки и имеют исключительно стационарное исполнение. Таким образом, доступ к качественному анализу имеют только крупные предприятия и организации. Владельцам малых предприятий приходится довольствоваться примитивным анализом образцов или тратить время и деньги на проведение анализов своих образцов сторонними организациями. Вместе с тем, очень часто бывает необходимо оперативно провести анализ закупаемого для производства сырья или проконтролировать качество своего продукта. В таком случае доступный ЯМР-спектрометр просто необходимо иметь в арсенале производства.

Решением данной проблемы может стать получивший развитие в последнее десятилетие ЯМР в слабых полях, в частности в земном магнитном поле. Магнитом для такого прибора является геомагнитное поле — бесплатное, с высокой однородностью, что крайне необходимо для ЯМР-спектроскопии. В основном используют так называемые протонные ЯМР-спектры, т.е. регистрируют сигнал от ядер водорода. Недостатком ЯМР в земном поле является отсутствие так называемого химического сдвига. Этот эффект порождается электронными оболочками ядер. Он ослабляет магнитное поле вокруг ядер, вызывает смещение спектральных линий относительно частоты резонанса и делает ЯМР-спектры более информативными. Земное магнитное поле недостаточно сильное, из-за чего химический сдвиг становится меньше ширины спектральной линии. Тем не менее существует взаимодействие магнитных моментов ядер друг с другом через электронные оболочки, которые не зависят от напряженности внешнего магнитного поля. Это взаимодействие называют косвенным диполь-дипольным (или спин-спиновым, КССВ), в слабом магнитном поле оно вызывает характерное симметричное расщепление спектральной линии протонов и позволяет так же по спектру различать протонсодержащие жидкости, если они содержат в молекулах фтор или фосфор. Дело в том, что только эти элементы имеют 100% природное содержание магнитных изотопов ^{19}F и ^{31}P и фиксировать расщепление линии протонов на этих ядрах очень просто. Такие изотопы магнитных ядер, как, например,

кремний ^{29}Si и углерод ^{13}C , имеют всего несколько процентов природного содержания и требуют для ЯМР спектроскопии либо искусственного обогащения, либо так называемого накопления сигнала, когда эксперимент повторяется многократно и ЯМР-сигналы этих экспериментов складывают, при этом шумы в спектре ослабевают в корень из числа экспериментов. Благодаря этому можно регистрировать очень слабые линии от взаимодействия протонов с ядрами ^{13}C , природное содержание которых около 1%. Это значит, что имеется возможность различать большую часть потребности химической промышленности — продукты органического синтеза, органические жидкости.

Однако, на этом пути имеются некоторые трудности. Во-первых, земное поле нестабильно и малые его флуктуации вызывают сдвиги частоты ЯМР на несколько герц, что делает невозможным традиционное накопление сигнала. Во-вторых, расщепленные линии очень малы и их трудно наблюдать на «склонах» основной линии протонов. В этой статье будет показано, как можно решить эти проблемы с помощью программно-математических методов.

Примечание: далее вся терминология и названия процессов подразумеваются применимо к низкочастотным системам, если не указано иного.

2. Обработка

2.1 Цель обработки

Обработка ЯМР спектра для слабого поля требует реализации ряда техник обработки сигналов для получения информативного спектра. Весь процесс работы со спектром может быть разбит на три этапа: предобработка, корректировка и анализ. Последний может быть представлен различными алгоритмами с разными целями. В рамках данной статьи целью анализа является нахождение констант КССВ (также именуемых J-coupling) и дальнейшее распознавание сканируемого вещества. Они представляют из себя расстояния между сателлитами (парой пиков, расположенных симметричных относительно некоего третьего) относительно главной спектральной линии. Главной спектральной линией является пик с наибольшей амплитудой для всего спектра, расположенный на частоте настройки прибора (в данном случае производится настройка на ЯМР-частоту протонов).

2.2 Предобработка

Входные данные представляют собой набор из 20-100 сигналов, записанных подряд. Каждый сигнал является словарём, где каждой дискретной отметке времени t соответствует дискретное значение V - амплитуда сигнала в момент времени t .

Первым этапом предобработки является аподизация сигнала, заключающаяся в домножении каждого сигнала на некую спадающую кривую. В данном случае использовалась часть косинусоиды, такой, что при $t = 0$ её значение равно $2/3$, а в последней точке сигнала - нулю. Этот процесс позволяет усилить сигнал в середине и уменьшить шумы в конце, а также избавиться от осцилляций в спектре при последующем дополнении сигнала нулями из-за резкого обрыва данных перед нулевыми значениями. Для вышеописанных условий, данная операция может быть представлена следующим образом:

$$V_k = V_k \cdot \cos \left(x \frac{\frac{\pi}{2} + \arccos \left(\frac{2}{3} \right)}{t} - \arccos \left(\frac{2}{3} \right) \right).$$

Для увеличения разрешения спектра, аподизированный сигнал искусственно расширяется путём добавления N нулевых значений через промежутки промежутки времени, обратные частоте дискретизации изначального сигнала. Для достижения значительного эффекта, N должно быть не менее, чем в 3 раза больше количества дискретных отметок расширяемого сигнала.

На данном этапе, ввиду сильной флуктуации Земного поля (см. рис. 1), в котором проводится запись, необходимо провести ряд дополнительных преобразований, компенсирующих искажения, порождаемые вариациями магнитного поля в процессе записи. Основным из таких преобразований является квадратурное детектирование, выполняемое по следующему алгоритму:

1. происходит разложение каждого сигнала на составляющие его гармонические колебания, что делается путём применения быстрого преобразования Фурье;
2. находится положение главной спектральной линии на оси частот F_0 ;
3. каждое значения сигнала преобразовывается следующим образом

$$V_k = V_k \cdot [\cos(2\pi F_0 t_k) + i \sin(2\pi F_0 t_k)].$$

Далее, к каждому сигналу вновь применяется преобразование Фурье, и все полученные спектры складываются почастотно, что позволяет избавиться от

значительной части шума (амплитуда шума уменьшается в \sqrt{n} раз при сложении n спектров). Также, так, как в проекте используются спектры с единственной главной спектральной линией, для удобства последующей обработки, она принимается за начало координат по оси частот.

В записи последующих формул будет применяться следующая нотация: A_k или $A(k)$ - амплитуда спектра на частоте k , F - частота, $Re(x)$ и $Im(x)$ - действительная и мнимая части числа соответственно.

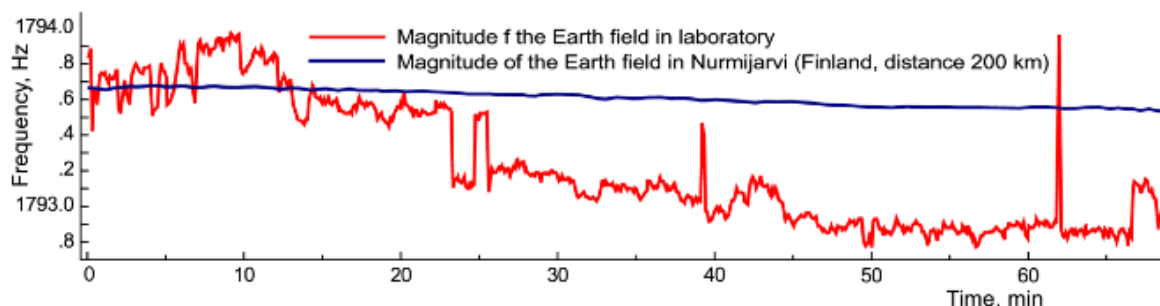


Рис. 1. Изменения земного поля в лаборатории (красный) и данные с вариометрической станции в Нурмиярви, Финляндия (синий) в один и тот же промежуток времени.

2.3 Корректировка

Получив первичный "сырой" спектр, остаётся ряд проблем, которые необходимо решить. Наиболее важными являются сдвиг фазы и искажение опорной линии. Разработанные алгоритмы будут представлены на примере уксусной кислоты и 2,2,2-Трифторэтанола.

2.3.1 Первичная корректировка фазы

Ввиду применения квадратурного детектирования, а так же из-за особенности процесса записи изначального сигнала, фаза получившегося спектра смещена, что ощутимо искажает вид спектра и в большинстве случаев может сделать невозможным дальнейший его анализ. Для корректировки фазового смещения обычно используют линейное смещение фазы: $A_k = A_k \cdot e^{(p_0 + p_1 \frac{j}{N})i}$, где i - мнимая единица, j - порядковый номер преобразуемой дискретной точки A_k , N - количество дискретных величин, p_0 и p_1 - углы смещения фазы 0 и 1 порядков соответственно. Для подбора и оценки оптимальности пары p_0 и p_1 для сдвига фазы конкретного спектра используется функция потерь АСМЕ.

2.3.2 Разработанная система корректировки фазы

Метод описанный в предыдущем пункте разрабатывался для спектров в сильном поле, ввиду чего результат его применения не полностью выравнивает фазу. Для решения данной проблемы был разработан собственный метод корректировки фазы. За основу берётся изменённая формула квадратурного детектирования: $A_k = A_k \cdot [\cos(2\pi\gamma F_k + \delta) + i \sin(2\pi\gamma F_k + \delta)]$, где γ и δ - настраиваемые входные параметры. Они подбираются с помощью метода Нелдера — Мида, минимизируя функцию потерь. Функция потерь принимает на вход массив значений оцениваемого спектра (после применения функции смещения фазы), а результатом её выполнения является скаляр, представляющий собой "оценку" рассматриваемого преобразования. Для решения поставленной задачи была разработана собственная функция потерь,

выглядящая следующим образом: $\sigma(A) = \sqrt{\frac{1}{n} \sum_{i=1}^N [\arg(A_i) - \overline{\arg(A)}]^2}$, однако,

так, как извлечение квадратного корня, равно как и деление на константу, не меняет монотонность конечной функции, в целях оптимизации вычислений

формула была упрощена до вида $\sigma(A) = \sum_{i=1}^N |\arg(A_i) - \overline{\arg(A)}|$. Фактически,

целью подбора γ и δ становится снижение дисперсии фаз для всего спектра. Дальнейшая работа со спектром будет производится исключительно с его модулем. Разница между изначальным и скорректированным спектром представлена на рис. 2.

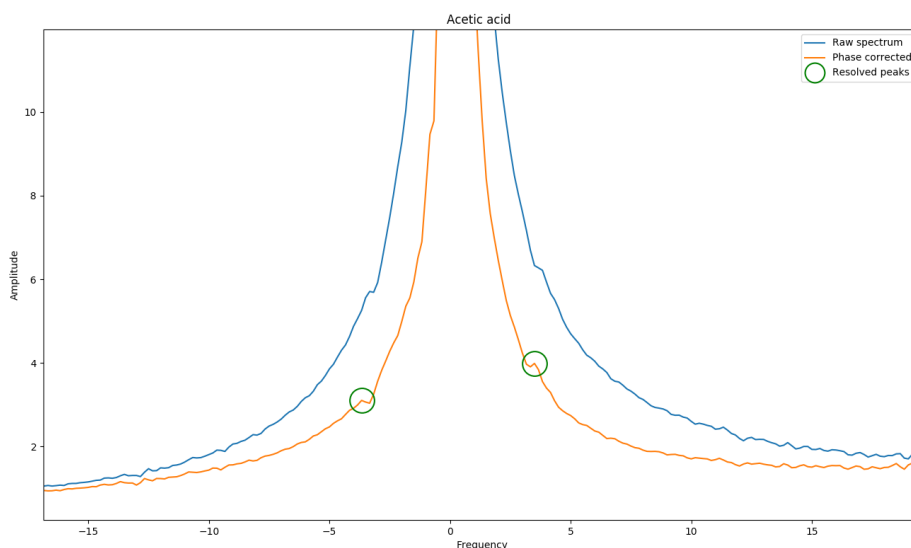


Рис. 2 — Выделение сателлитов после корректировки фазы

2.3.3 Свёртка

В условиях реальных экспериментов существует множество шумов, которые даже в случае малой амплитуды значительно затрудняют обработку данных. В частности, разработанный алгоритм компенсации опорной линии (см. пункт 2.4) крайне чувствителен к дополнительным экстремумам на пиках предполагаемых спутников. Для выделения несущей информации пиков был разработан метод, сглаживания для ЯМР спектров. Идея заключается в нахождении такой функции, что её использование в роли ядра свёртки для обрабатываемого спектра сглаживает пики спутников, делая их пригодными для дальнейшей обработки.

Так, как в идеальном случае каждый не шумовой пик представлен кривой распределения Коши, следовательно, наилучшей функцией-свёрткой будет лоренциана. Известно, что ширина пика прямо пропорциональна ширине спектральной линии. Коэффициент пропорциональности был экспериментально выведен как $\mu \approx \frac{1}{2}$ (среднее отношение ширины главной спектральной линии на уровне половины её амплитуды к ширине пика у основания). Таким образом, функция свёртки определяется как $g(x) = \frac{\mu^2}{\mu^2 + x^2}$. Далее, применяется Фурье-свёртка $A = A * g$, что значительно сглаживает пики. Пример разницы до и после применения свёртки представлен на рис. 3.

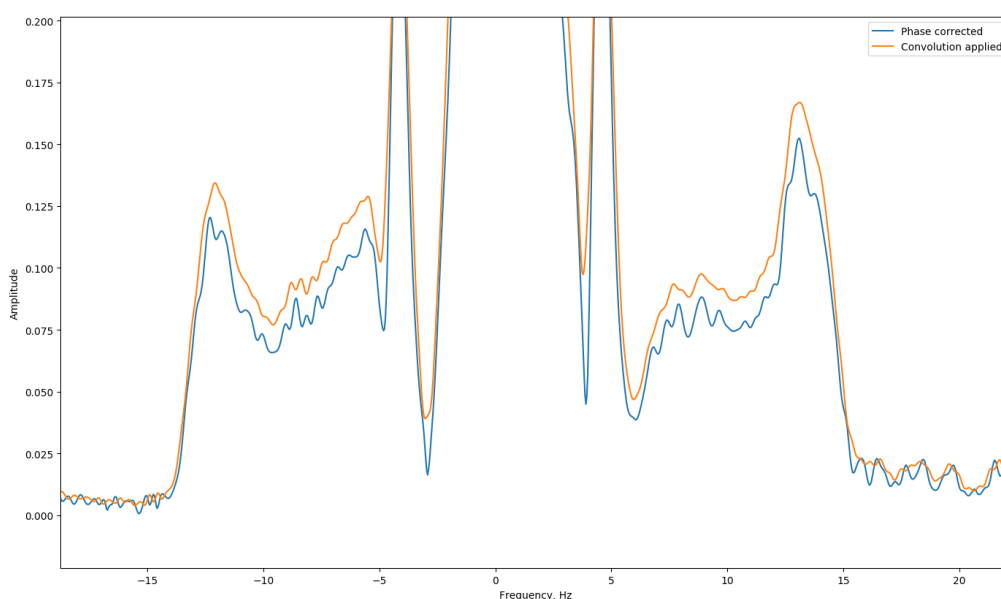


Рис. 3 — Применение свёртки

2.3.4 Компенсация опорной линии

Для начала, будет удобно представить спектр, как сумму трёх функций $A(k) = A_{pure}(k) + N(k) + B(k)$, где A_{pure} - чистый (идеальный) спектр, N - шум, а B - опорная линия (рис. 4). Следовательно, для её компенсации достаточно найти значение опорной линии в каждой точке спектра и вычесть её.

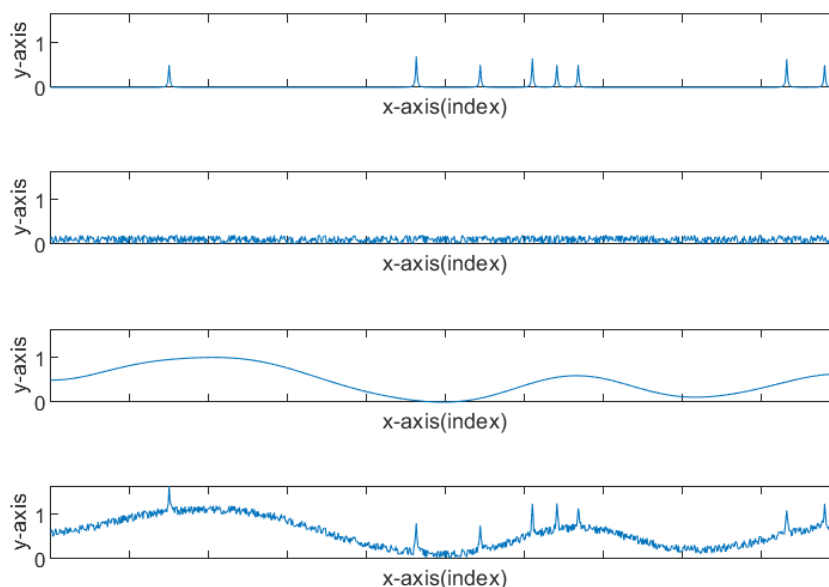


Рис. 4 — Разложение спектра. Сверху вниз: чистый спектр, шум, опорная линия, результирующий спектр

Для расчёта опорной линии существует несколько алгоритмов, основанных на Вейвлет-преобразовании и на методах машинного обучения. Были опробованы оба, однако, Вейвлет-компенсация была нестабильна, и зачастую выделяла неверные пики, что не позволило использовать данный метод. Подходы, основанные на машинном обучении предполагают создание искусственных спектров из частей описанных в вышеуказанной формуле. Такое решение показывает удовлетворяющие результаты в рамках оригинальной статьи, однако добиться подобной точности для низкопольных спектров не удалось.

Путём множественных экспериментов был разработан собственный подход, отчасти вдохновленный методом нахождения "конверта" сигнала. Идея заключается в поиске отрицательных пиков на всём спектре и линейной интерполяции по всей оси частот, используя найденные пики как опорные точки. Далее, получившийся сплайн вычитается из основного спектра. Поиск

отрицательных пиков производится путём нахождения множества корней λ уравнения $\frac{\partial A}{\partial F} = 0$ с условием, что $\begin{cases} \lim_{\epsilon \rightarrow +0} A(\lambda_i - \epsilon) < 0 \\ \lim_{\epsilon \rightarrow +0} A(\lambda_i + \epsilon) > 0 \end{cases}$

Вычитая из спектра полученную опорную линию, заметно упрощается задача по последующему нахождению пар пиков, расстояния между которыми являются константами КССВ. Результат выполнения алгоритма представлен на рис. 5.

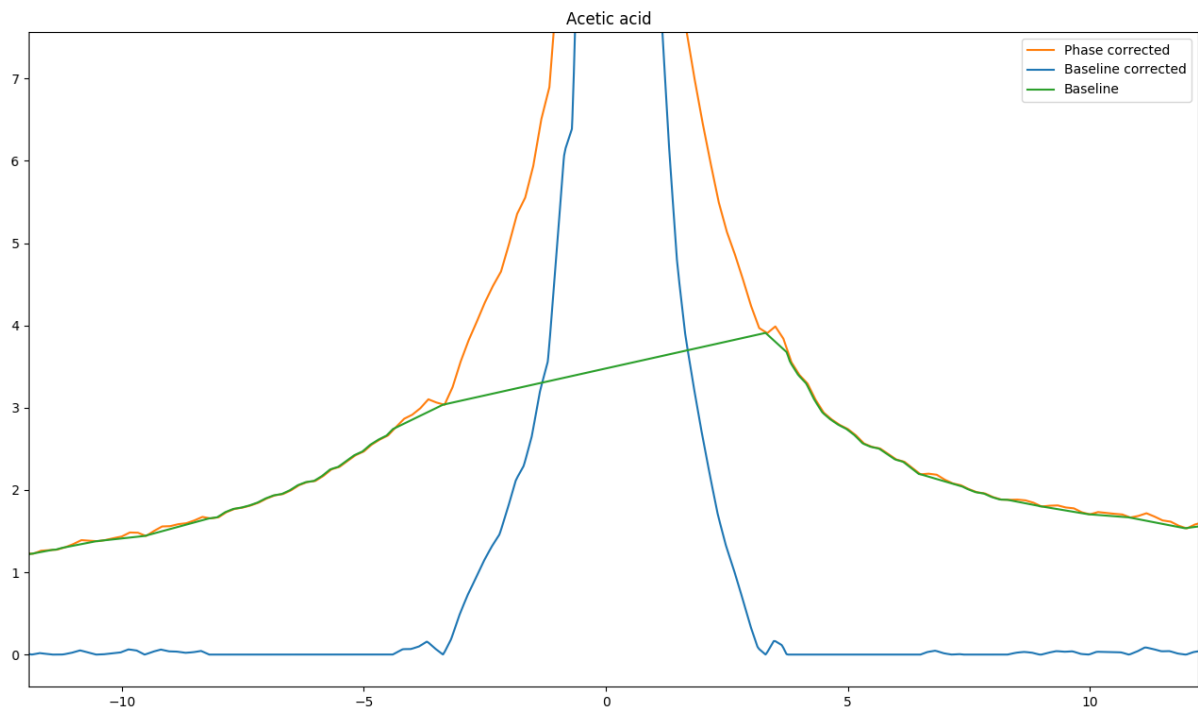


Рис. 5 — Компенсация опорной линии

Дополнительно, для упрощения анализа удаляется спектральная линия, так как она не несёт в себе полезной (в рамках исследования) информации, но усложняет дальнейшие алгоритмы. Так как спектральная линия представлена пиком наибольшей амплитуды с вершиной на частоте $F_0 = 0$ Гц и не имеет внутренних пиков, достаточно вновь найти множество корней λ уравнения $\frac{\partial A}{\partial F} = 0$ и выбрать из них пару корней λ_1, λ_2 с наименьшими абсолютными значениями, но имеющих разные знаки. Далее, приравнять все промежуточные точки спектра к нулю: $A(\rho) = 0, \rho \in [\lambda_1; \lambda_2]$. В результате остаются только несущие информацию сателлиты и пики, порождённые шумом (рис. 6).

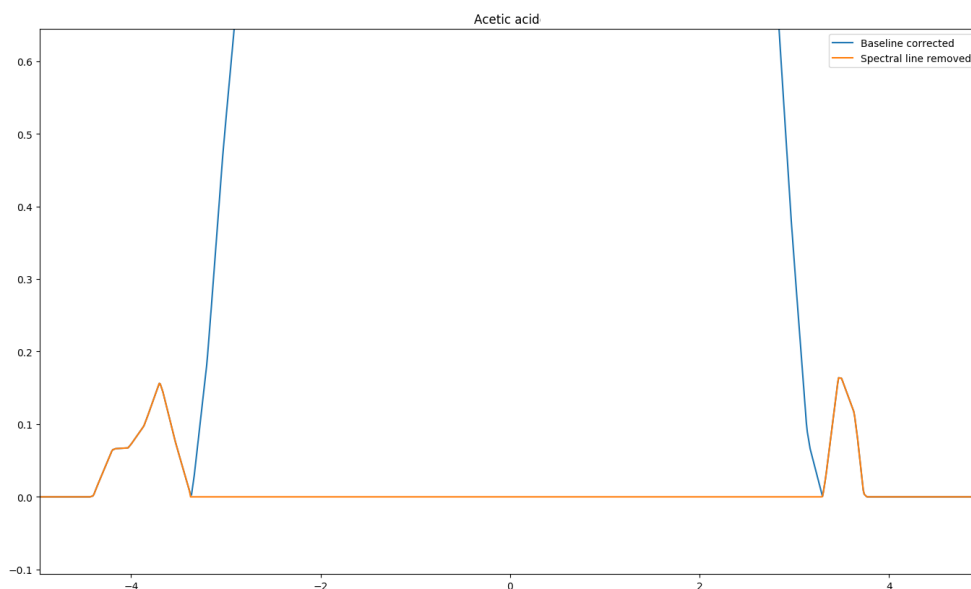


Рис. 6 — Удаление главной спектральной линии

2.4 Анализ

Получив удобную для работы репрезентацию спектра, можно приступить к обработке данных, цель которой, в рамках проекта, сводится к измерению расстояний между пиками спутников. Однако, для большинства спектров на этом этапе возникают две проблемы:

1. шум хоть и уменьшен на предыдущих шагах, но всё ещё сравним по амплитуде с спутниками;
2. следствием корректировки фазы является частичный сдвиг оси частот, к тому же он по большей части неоднороден, что исключает возможность простого выравнивания.

Для решения первой проблемы было решено ввести весовую функцию $w(k)$, на которую домножается получившийся спектр. Подход основывается на знании, что искомые пики:

1. симметричны относительно нуля (с погрешностью);
2. чем меньше сумма модулей положений пары пиков - тем выше их амплитуда;
3. чем больше сумма модулей положений пары пиков - тем больше будет расстояние между ними и последующими (по направлению удаления от нуля).

Из первого и второго пунктов следует, что $w(k)$ должна быть симметрична относительно нуля и затухать отдаляясь от него. Эмпирическим методом была

получена следующая весовая функция, имеющая максимум в точке (0, 1) и равная нулю в крайних точках спектра: $w(k) = \max[0, \cos \frac{k\pi}{F_{\max}} \cdot (1 - \frac{|k|}{F_{\max}})]$. Данная функция, а также результат её применения представлены на рис. 7.

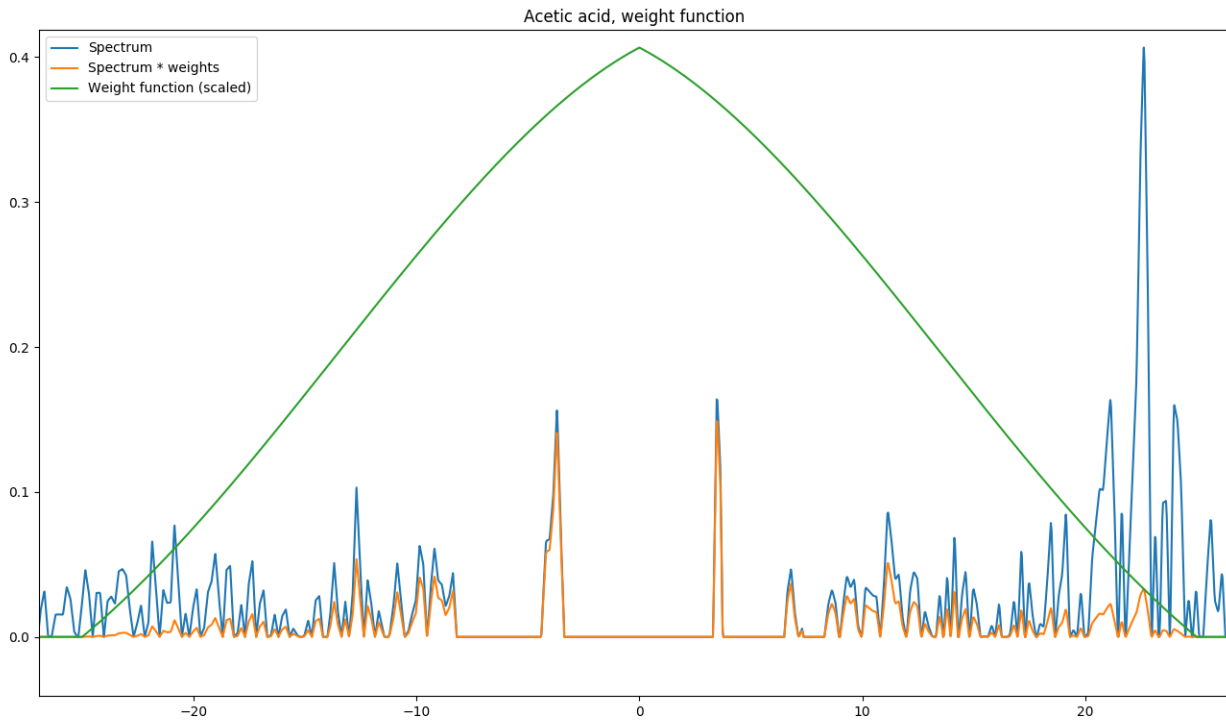


Рис. 7 — Применение весовой функции

Для непосредственного поиска и выборки пиков следует определить четыре константы: N - число первично находимых пиков, S - максимальная допустимая разница модулей точек двух предположительно симметричных пиков, P - минимальное линейное расстояние между двумя пиками одного знака для двух пар (защита от шума), W - размер окна поиска для выравнивания координат пиков на оригинальном и обработанном спектре. Оптимальные значения

указанных констант были статистически подобраны как
$$\begin{cases} N = 8 \\ S = P = 1.5 \text{ Hz.} \\ W = 2.7 \text{ Hz} \end{cases}$$

На обработанном спектре необходимо найти все положительные пики, отсортировать их по возрастанию модуля соответствующих им частот и взять N первых, затем, для каждого из полученных пиков выполнить следующий алгоритм:

1. пусть, f - частота рассматриваемого пика;

2. взять выборку ϑ массива значений изначального спектра A_0 так, что $\vartheta = [A_0(m_1), \dots, A_0(m_n)]$, $m \in [f - W; f + W]$;
3. для производной (градиента) ϑ' найти частоту, соответствующую точке, где модуль ϑ' минимален;
4. полученная частота перезаписывает собой частоту рассматриваемого пика.

Далее, необходимо отсеять все пики, не имеющие пары, такой что её частота противоположна по знаку и разница модулей их частот не превышает S . Расстояния между оставшимися парами пиков будут являться константами КССВ.

3. Определение структуры сканируемой молекулы

Практической задачей исследования является определение структуры молекулы основываясь на полученных данных. Для этого можно выделить два подхода: аналитический и накопительный. Первый заключается в математическом подборе молекулы, создающей наиболее схожий с рассматриваемым спектр, основываясь на имеющихся данных и форме спектра. Второй подход основывается на выделении ряда признаков и отличительных черт спектра в числовом виде, с последующим сравнением их с заранее сформированной базой известных веществ. Он представлен в двух версиях в зависимости от условий.

Первый из них сводится к накоплению базы наборов скалярных параметров для известных молекул и наиболее оптимален для случаев с большим разнообразием непохожих (с точки зрения спектров) молекул. Для последующего распознавания выделяются следующие параметры: количество найденных пар пиков, массив частот, состоящий из отсортированных по модулю местоположений найденных J-coupling пиков, а также массив отношений их амплитуд к амплитуде спектральной линии. Полученный массив записывается в базу данных (в проекте используется SQLite3) и ассоциируется с названием молекулы. Далее, после накопления достаточной базы молекул, обучается модель на методе опорных векторов SVC, используемая для последующего распознавания неизвестной молекулы.

Другой подход может быть применён в случае плохого качества записей или при большом количестве похожих молекул. Необходимо выделить такое окно-маску вокруг найденного пика, что в его крайних точках будут находиться локальные минимумы. Далее, с помощью метода наименьших квадратов найти k - уровень совпадения рассматриваемого пика с лоренцианой, описанной в пункте 2.2.3. В данном методе в базу записываются следующие параметры для каждого пика: точки начала и конца выделенного окна, k и отношение максимальной амплитуды внутри окна к амплитуде спектральной линии. Данный набор параметров также ассоциируется с некоторой молекулой. При добавлении данных другой записи аналогичной молекулы, следует перезаписать вышеуказанные данные как среднее между ними и новыми данными. Для определения заранее неизвестной молекулы, необходимо перебрать все ранее записанные маски и выбрать как результат распознавания такую молекулу, что Евклидово расстояние между данными в базе и данными, полученными после применения маски - минимально.

4. Реализация и заключение

В процессе работы над проектом был использован подход двух ветвей разработки: исследовательско-экспериментальная среда и финальное приложение, в которое попадают уже оптимизированные и протестированные алгоритмы.

Первая была реализована на языке программирования Python (v3.6) и использовала технологию Jupyter Lab для упрощения работы и экспериментов. Основными библиотеками стали: NumPy для работы с векторами, SciPy для удобства анализа данных, plotly для построения графиков во время работы. В данной среде проводилась непосредственно разработка, оптимизация и тестирование всех алгоритмов.

Конечное приложения написано на языке C++20 с использованием библиотек MetalAPI+DearImGUI для реализации графического интерфейса и Eigen в роли аналога NumPy. Все алгоритмы полученные в исследовательской среде были переписаны и по возможности векторизованы для ускорения анализа. Помимо этого, часть вычислений была перенаправлена на GPU т.к. большинство операций применяются ко всему спектру, а значит представляют из себя набор несложных независимых подпрограмм, выполняемых множество раз.

Результатом проекта является программный комплекс для обработки и анализа ЯМР спектров, в частности, с функционалом, позволяющим идентифицировать молекулу по собранной базе. Также, процессе были разработаны собственные, оптимальные для низкочастотных ЯМР спектров алгоритмы накопления сигнала, коррекции фазы и коррекции опорной линии.

5. Литература (неактуально)

1. Исследование и разработка научно-технических решений в области проведения сортировочных операций в режиме реального времени, с объектами, имеющими сложные характеристики, с использованием высокоэффективных робототехнических средств автоматизации. Электронный ресурс. Режим доступа: http://fcpir.ru/upload/iblock/94b/corebofs000080000ldo38um1mqn7hvc_annotation.pdf
2. D. Jirak, S. Wermter Potentials and Limitations of Deep Neural Networks for Cognitive Robots. Электронный ресурс. Режим доступа: <https://arxiv.org/pdf/1805.00777.pdf>
3. Detection of ArUco Markers. Электронный ресурс. Режим доступа: https://docs.opencv.org/trunk/d5/dae/tutorial_aruco_detection.html, – Проверено 08.01.2020
4. Цветовая модель HSV. Электронный ресурс. Режим доступа: https://en.wikipedia.org/wiki/HSL_and_HSV – Проверено 08.01.2020.
5. Ле Мань Ха Свёрточная нейронная сеть для решения задачи классификации // ТРУДЫ МФТИ. 2016. Том 8, № 3. 2016. — с.91-97
6. ImageNet. Электронный ресурс. Режим доступа: <http://image-net.org/index> – Проверено 08.01.2020.
7. Матрица ошибок. Электронный ресурс. [https://learnmachinelearning.wikia.org/ru/wiki/ Матрица_ошибок_\(Confusion_matrix\)](https://learnmachinelearning.wikia.org/ru/wiki/Матрица_ошибок_(Confusion_matrix)) – Проверено 08.01.2020.
8. R. Berwick, An Idiot's guide to Support vector machines (SVMs) – с.5-28.
9. SCRA Robot Kinematics. Электронный ресурс. Режим доступа: <http://www.deltatau.com/Common/technotes/SCARA%20Robot%20Kinematics.pdf> – Проверено 08.01.2020.
10. Вапник В.Н. The nature of statistical learning theory – с.138-176