электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

**УДК 004.4**

# Автоматизированная информационная система идентификации фаз неизвестного вещества по его штрих-диаграмме

***Зинченко А.М.,*** *студент*
*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,*
*кафедра «Программное обеспечение ЭВМ и информационные технологии»*

*Научный руководитель: Романова Т.Н., к. ф.-м. н., доцент*
*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана*

*Научный руководитель: Винтайкин Б.Е., д. ф.-м. н, профессор*
*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана*
*irudakov@bmstu.ru*

По результатам последнего конкурса алгоритмов автоматизированной идентификации Round Robin Search-Match [4], проводимого всемирным обществом кристаллографии можно сделать вывод о том, что на сегодняшний день большинство существующих алгоритмов автоматизированной идентификации фаз однозначно определяют фазы монофазных кристаллов или фазы с наибольшей концентрацией многофазных кристаллов. Достоверность результатов идентификации снижается с повышением числа фаз, входящих в состав образца и уменьшением их концентрации. В основном это связано с плохим соответствием эталонных спектров экспериментальной штрих-диаграмме [2]. Поэтому в настоящий момент остаётся актуальной проблема определения малой фазы вещества, то есть фазы с малой концентрацией, в условиях плохого соответствия дифрактограммы образца дифрактограммам эталонных спектров.

Несоответствие штрих-диаграмм стандарта и образца проявляется в отклонениях от истинных значений как положений штрихов, так и их интенсивностей [5]. Основными факторами, из-за которых появляется отклонение положений штрихов, являются смещение нуля детектора при съёмке образца на дифрактометре и неточность измерений пика дифракционной линии. Смещение нуля детектора влияет на значение систематической погрешности, а неточность измерений – на величину случайной погрешности положения штрихов. На интенсивность штрихов оказывают влияние размеры зёрен поликристалла, а также наличие текстуры в кристалле.

Стандартным алгоритмом проведения автоматизированной идентификации фаз по штрих-диаграмме является двухэтапная процедура поиска, известная как процедура Ханавальта [3]:

1. На первом этапе проводится предварительный отбор стандартов из базы данных, штрих-диаграммы которых удовлетворяют просто мере близости. В качестве простой меры близости обычно рассматривают число совпавших по положению штрихов стандарта и образца из первых N штрихов, обладающих наибольшей интенсивностью.
2. На втором этапе для каждого из прошедших первый этап стандартов рассчитывается ресурсоёмкая мера близости, имеющая сложный функциональный вид от вида двух сравниваемых штрих-диаграмм. Обычно данную меру рассчитывают на восстановленных по штрихам дифракционных линиях, а в качестве моделирующих данные линии функций берут гауссиан или лоренциан. Сама мера рассчитывается как отношение суммарной площади пересечения восстановленных дифракционных линий стандарта с восстановленными линиями образца к суммарной площади линий стандарта.

На вход алгоритму идентификации фаз поступают штрих-диаграммы стандартов и образца, причем положения штрихов равны межплоскостным расстояниям в элементарных ячейках кристаллов, а высоты штрихов – относительным интенсивностям отраженного электромагнитного излучения при съёмки образца.

На первом этапе для определения совпавших линий используется скользящее по штрих-диаграмме образца окно постоянной ширины. Если какие-либо 2 штриха стандарта и образца находятся в рамках одного окна, то они считаются совпавшими [5]:

$$x_0 - \Delta x \leq x^s \leq x_0 + \Delta x, \quad (1)$$

где $x_0 -$ положение линии образца (Å),

$x_s -$ положение линии стандарта (Å),

$\Delta x -$ ширина скользящего окна (Å)

Для повышения достоверности результатов идентификации в данной работе предлагается следующая модификация меры близости.

Во-первых, необходимо отметить важный момент, касающийся выбора величины максимального отклонения $\Delta x$. В физическом смысле $\Delta x$ должна быть численно равна абсолютной погрешности при измерении межплоскостных расстояний стандартов и образца. Однако использование постоянного значения $\Delta x$ на всём интервале некорректно

по следующей причине. Дело в том, что погрешность при измерении положения дифракционной линии постоянна и равна $d(\theta)$. Рассчитаем погрешность при определении межплоскостного расстояниях [5]:

$$d(d) = \frac{\lambda}{2}\frac{1}{\sin^2\theta}\cos\theta\, d(\theta)$$

$$d(d) = \frac{\lambda}{2}\frac{1}{\sin\theta}ctg\theta\, d(\theta) \quad (2)$$

Из последней формулы вытекает, что погрешность при определении межплоскостного расстояния нелинейно зависит от погрешности измерения угла. По этой причине при использовании окна сравнения $\Delta x$ постоянной ширины неизбежно будет возникать ситуация, когда с одного края штрих-диаграммы ширина $\Delta x$ соответствует физической погрешности, а с другого края существенно больше или меньше соответствующей погрешности. Поскольку недопустимо отбрасывание истинной фазы на предварительном этапе из-за того, что при сравнении с ней использовалось слишком узкое окно, то на практике используют окно максимальной постоянной ширины. Использование же скользящего окна динамически изменяющейся ширины даст оптимальный результат.

Этап предварительного отбора в отличие от этапа уточнения в качестве результата сравнения двух линий возвращает логическое значение: линии могут совпасть или не совпасть. В связи с тем, что часто при выборе скользящего окна большого размера в список стандартов первого этапа попадают стандарты с весьма схожими штрих-диаграммами и с абсолютно одинаковой оценкой первого этапа, целесообразным является придание данным стандартам несколько отличающихся оценок, поскольку их штрих-диаграммы не являются абсолютно одинаковыми. Так, например, если максимальное отклонение одной из штрих-диаграмм стандартов от штрих-диаграммы образца составляет половину ширины скользящего окна, а другой – всю ширину скользящего окна, то уже на этапе предварительно отбора можно дать похожие, но несколько отличающиеся оценки данным стандартам. Таким образом, стандарты будут упорядочены по степени близости и данную оценку можно будет учесть и на этапе уточнения.

Другая причина, по которой стоит рассчитывать меру совпадения уже на предварительном этапе – наличие систематического отклонения штрихов по положениям. Наличие систематического отклонения означает, что все штрихи образца будут сдвинуты относительно штрихов истинной фазы на некоторое расстояние. Таким образом, если

использовать ширину окна, достаточную для учёта систематической погрешности, то, анализируя ширину диапазона отклонений по всем штрихам, можно будет разделить стандарты на совпадающие и на случайно похожие. У совпадающих стандартов диапазон возможных отклонений будет малым, поскольку систематическая погрешность имеет постоянную величину. У случайно похожих стандартов будет большая вероятность того, что диапазон отклонений по положениям будет широк – достаточно хотя бы 2-х сильнейших линий для того, чтобы диапазон стал широким.

Введём новую меру близости $M(S_i, O_i)$, рассчитываемую на предварительном этапе:

$$M(S_i, O_i) = \begin{cases} 1 - \dfrac{|d_i^o - d_i^s|}{\Delta d K_1} - \dfrac{|I_i^o - I_i^s|}{K_2}, \text{если } |d_i^o - d_i^s| < \Delta d \\ 0, \text{иначе} \end{cases}$$

$$M(S,O) = \frac{\sum_{i=1}^{n_{совп}} (S_i, O_i)}{n_{совп}} + \tag{3}$$

$$+ \left( \min\left( \frac{|d_k^o - d_k^s|}{\Delta d} \right) - \max\left( \frac{|d_k^o - d_k^s|}{\Delta d} \right) \right), k = \overline{1, n_{\text{совп}}}$$

где S, O – штрих-диаграммы стандарта и образца

$S_i, O_i$ –штрихи стандарта и образца,

$d_i^s, d_i^o$ – межплоскостные расстояния стандарта и образца,

$I_i^s, I_i^o$ – интенсивности штрихов стандарта и образца,

$n_{\text{совп}}$ – количество совпавших линий образца и стандарта,

$\Delta d$ – ширина скользящего окна,

$K_1, K_2$ – коэффициенты, определяющие вклад сравнения двух штрихов в значение меры близости штрих-диаграмм

Значения коэффициентов $K_1, K_2$ вычисляются в рамках разработанной программы из соответствующих экспериментов, которые позволяют определить, при каких значениях данных коэффициентов кластерная функция предоставит наиболее точное решение задачи идентификации.

Последнее слагаемое в формуле (3) определяет ширину всего диапазона отклонений положений штрихов стандарта от образца.

На уточняющем этапе предлагается учитывать различные отклонения штрихов следующим образом. Для прошедших этап предварительного отбора стандартов необходимо рассчитать математическое ожидание разности положений штрихов образца и стандарта для всех совпадающих штрихов. На данную величину необходимо сдвинуть штрих-диаграмму образца для того, чтобы избавиться от систематической ошибки. При этом стандарты, являющиеся истинными фазами, получат большее суммарное приращение меры близости, чем случайно совпавшие:

$$D = M\left[\frac{|d_k^o - d_k^s|}{\Delta d}\right] k = \overline{1, n_{\text{совп}}} \tag{4}$$

где $d_i^s, d_i^o$ – межплоскостные расстояния стандарта и образца,

$n_{\text{совп}}$ – количество совпавших линий образца и стандарта,

$\Delta d$ – ширина скользящего окна,

$M$ –математическое ожидание разности положений штрихов на штрих-диаграммах стандарта и образца.

Вместо расчёта простой суммарной площади пересечения восстановленных по штрихам дифракционных линий стандарта и образца предлагается использовать полный информационный вес линий и учитывать его при подсчёте площадей под дифракционными линиями. В качестве информационного веса линии предлагается использовать величину, обратную произведению плотностей распределения дифракционных линий по положениям и по интенсивностям в заданной точке:

$$\omega_N(\theta) \sim \frac{1}{N(\theta)} \qquad \omega_M(I) \sim \frac{1}{M(I)} \tag{5}$$

где $\omega_M(I)$ – информативность штриха с интенсивностью $I$

$M(I)$ – плотность распределения штрихов по их интенсивностям $I$

$\omega_N(\theta)$ –информативность штриха с положением $\theta$,

$N(\theta)$ –плотность распределения штрихов по их положениям $\theta$

Тогда мера близости, рассчитываемая на уточняющем этапе, будет равна:

$$F = \frac{\sum_{i=1}^{n_{совп}} \Delta S_i \omega^N(x_i) \omega^M(y_i^s)}{\sum_{k=1}^{n_s} S_k \omega^N(x_k) \omega^M(y_k^s)} \tag{6}$$

где $\Delta S_i$ –площадь пересечения реконструированных линий, построенных на паре совпадающих $i$-х штрихов штрих-диаграмм стандарта и образца,

$S_k$ −площадь реконструированной линии, построенной на $k$-м штрихе стандарта,

$n_{\text{совп}}$ − количество совпавших линий образца и стандарта,

$n_s$ −количество линий стандарта,

$x_i, x_k$ −положение $i$-й и $k$-й линий стандарта,

$y_k^s, y_i^s$ −интенсивность $i$-й и $k$-й линий стандарта,

$\omega^M(y_i^s)$ − информативность штриха стандарта с интенсивностью $y_i^s$,

$\omega^N(x_i^s)$ − информативность штриха стандарта с положением $x_i^s$

Таким образом, модификация, предлагаемая в данном исследовании, состоит из следующих действий:

1. Количество сильнейших линий, участвующих в предварительном этапе, рассчитывается исходя из имеющихся стандартов в базе данных, а не задается константой
2. На предварительном и уточняющем этапах используется скользящее окно с шириной, изменяющейся в зависимости от сравниваемой линии.
3. На предварительном этапе помимо факта совпадения или несовпадения двух линий используется простая оценка степени совпадения двух линий, если имеется факт совпадения
4. На уточняющем этапе используется мера близости, сравнивающая реконструированные по штрихам дифракционные линии, учитывающая систематическую погрешность положений линий, а в качестве схемы оценки веса используется полный информационный вес линии.

## Структура системы

В рамках работы была разработана АИС, позволяющая проводить автоматизированную идентификацию фаз неизвестного вещества по его штрих-диаграмме с использованием изложенных выше модификаций стандартной процедуры Ханавальта и модифицированной меры близости [1].

Ниже перечислены подсистемы, составляющие программный продукт, и решаемые ими задачи.

1. Подсистема “Импорт из cif файлов” - подсистема импортирования информации из cif файлов в хранилище стандартов, используемое следующими модулями системы.

2. Подсистема “Генератор штрих-диаграммы образца” – подсистема, содержащая алгоритмы создания штрих-диаграммы смеси кристаллических веществ, использующая штрих-диаграммы кристаллов из хранилища.
3. Подсистема “Идентификация фаз” – подсистема, содержащая рассмотренные в аналитической части стратегии идентификации фаз, меры близости.
4. Подсистема “Проведение эксперимента” – подсистема, проводящая многократный запуск процедуры идентификации с заданной мерой близости на различных тестовых выборках.
5. “База данных кристаллических веществ” – база данных, содержащая информацию о кристаллах с известной структурой элементарной ячейки, их штрих-диаграммы, химический состав и другую информацию

Данные подсистемы и их связи показаны на диаграмме потоков данных на рис. 1.

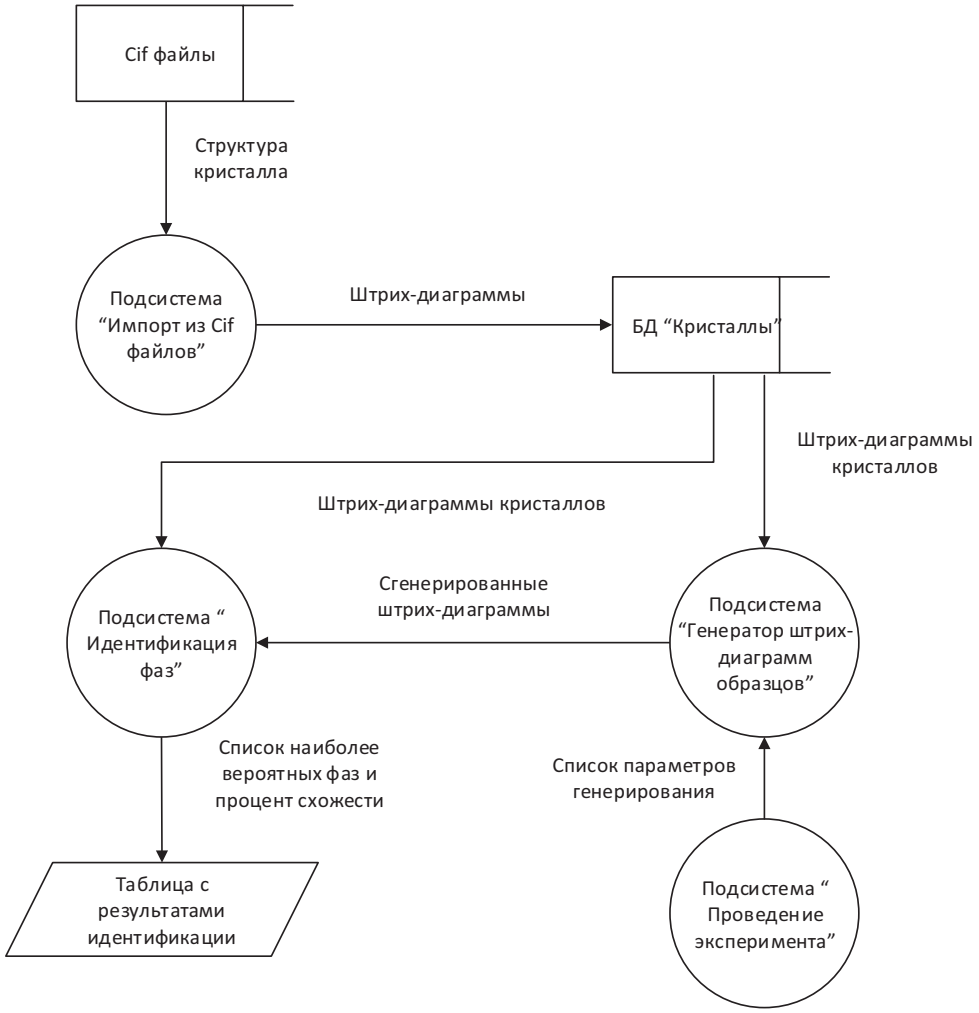


Рис. 1. Диаграмма потоков данных для разрабатываемой системы

**Результаты проведения экспериментов по измерению достоверности результатов идентификации**

Каждый из экспериментов по установлению влияния систематической погрешности положений штрихов образца и параметров модифицированной меры близости на достоверность результатов идентификации был проведён по следующей схеме:

1. В ходе каждого эксперимента был определён один варьируемый параметр, шаг изменения его значения и граничные значения. Остальные параметры генерирования штрих-диаграммы и меры близости имели постоянные значения.

2. Для каждого значения варьируемого параметра была сгенерирована выборка штрих-диаграмм объёмом 50 штрих-диаграмм. Данный объём был установлен в ходе экспериментов, как достаточный для оценки точности идентификации, но в тоже время, не требующий больших временных ресурсов (больше 10 минут).

3. Для каждой из штрих-диаграмм выборки была запущена процедура идентификации неизвестных фаз.

4. Оценка результатов идентификации была проведена с помощью выставления баллов: если в результате идентификации одной штрих-диаграммы все неизвестные фазы были определены правильно, то есть занимали места в результирующей таблице фаз начиная с первого, и между ними не было других фаз, то данному результату начислялся 1 балл. В иных случаях баллов не начислялось. Таким образом удачным считался эксперимент, в результате которого пользователю не требовалось принимать решение о том, какая из представленных фаз входит в исследуемое вещество, а достаточно было определиться только с числом фаз.

5. Таким образом, оценка точности идентификации для фиксированных параметров имела значение от 0 до 50.

В экспериментах проводилось как сравнение точностей, которую предоставляли различные меры близости, так и установление зависимости точности разработанной меры от параметров генерирования штрих-диаграмм

Первый эксперимент – установление влияния систематической ошибки в расчетах положений штрихов на достоверность результатов идентификации. Параметры генерирования – образец, состоящий из одной фазы. Погрешность определения положения штриха $\delta\theta$ = 0,025°, количество случайно выбранных отсутствующих фаз – 5. Границы варьирования систематической ошибки измерения положения штриха – от 0.1° до 2°. Результаты эксперимента приведены на рис.2.

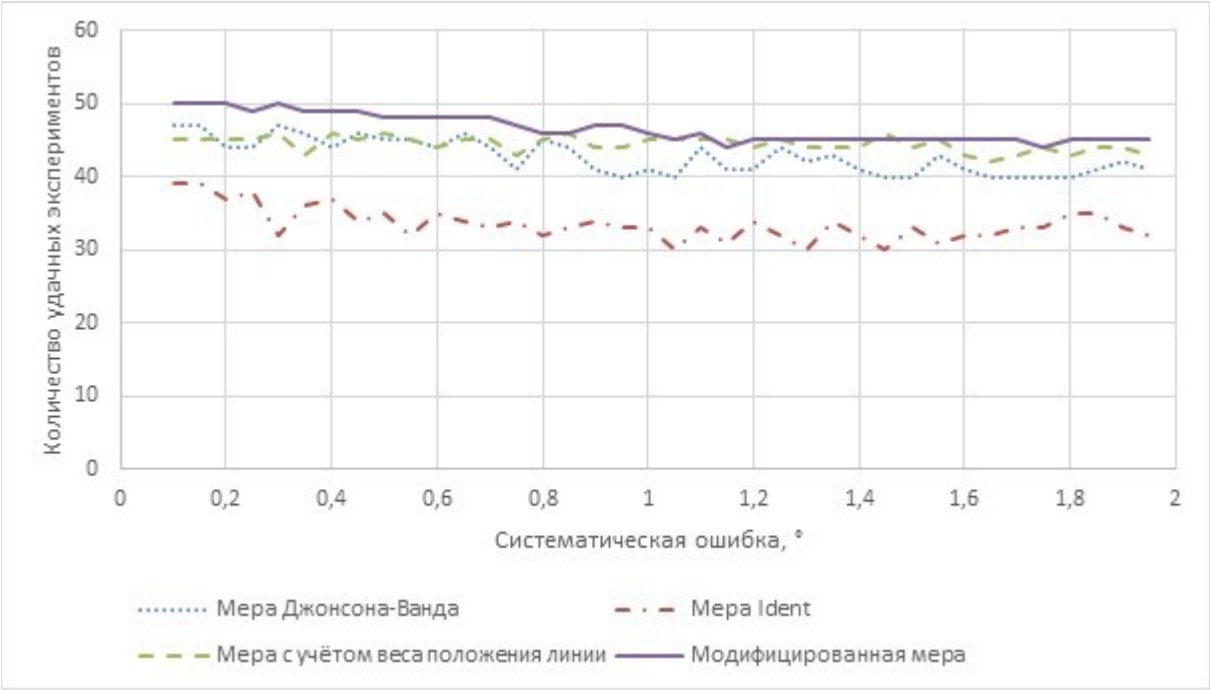


Рис. 2. Зависимость точности идентификации монофазного образца от систематической ошибки измерения положений штрихов

Из эксперимента следует два вывода: во-первых, размер систематической ошибки даже в 2° оказывает слабое влияние на точность определения фазы монофазного образца с помощью каждой из мер; во-вторых, модифицированная мера даёт несколько лучший результат, чем другие меры: для каждого значения систематической ошибки данная мера позволяет определить на 2-3 вещества точнее, чем другие меры. Поэтому можно сделать вывод о том, что для идентификации монофазного образца с большой систематической ошибки модифицированная мера не даёт ощутимого преимущества в точности.

Для того чтобы определить, влияет ли каким-нибудь образом состав и концентрация фаз в смеси, был проведён аналогичный эксперимент, но с образцами, состоящими из двух фаз одинаковой концентрации (50% и 50%). Результаты эксперимента представлены на рис.3.

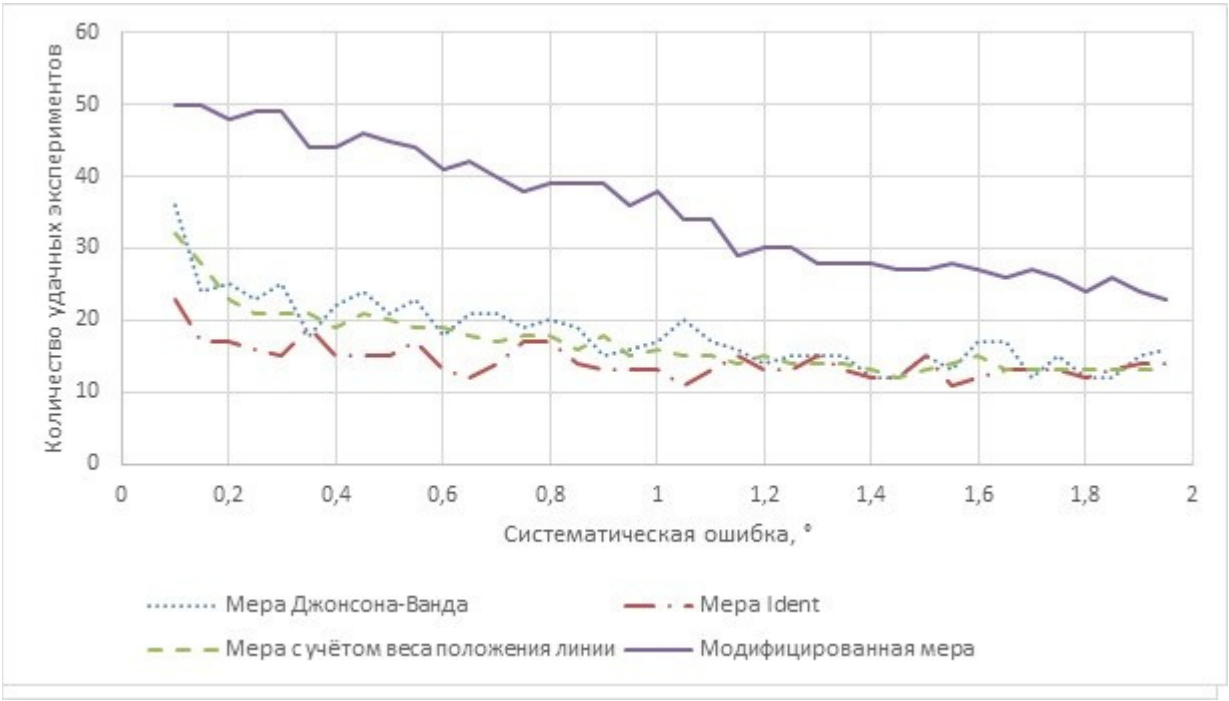


Рис. 3. Зависимость точности идентификации двухфазного образца с фазами одинаковой концентрации от систематической ошибки измерения положений штрихов

Результаты эксперимента позволяют сделать вывод о том, что рост систематической ошибки при идентификации двухфазного образца стал сильнее понижать точность идентификации при использовании модифицированной меры. На точность идентификации двухфазного образца рост с помощью остальных мер систематическая ошибка оказала меньшее влияние. Однако при минимальной систематической ошибке (до 0,4°) модифицированная мера существенно превосходит по точности остальные меры - при ошибке в 0,1° правильно идентифицированы все 50 штрих-диаграмм, что на 20 правильно идентифицированных штрих-диаграмм больше, чем у остальных мер. Можно заметить, что при максимальном значении систематической ошибки все меры близости имеют примерно одинаковую точность идентификации.

На основании двух проделанных экспериментов можно выдвинуть гипотезу о том, что увеличение числа фаз образца и соответственно уменьшение их концентраций увеличивают падение достоверности идентификации модифицированной мерой при одном и том же приросте систематической ошибки. Для того чтобы проверить данную гипотезу, был проведён третий эксперимент, аналогичный первым двум, однако вместо двухфазных образцов были сгенерированы 4-х фазные образцы, 3 фазы которых имели равные концентрации (30%) а одна фаза была фазой малой концентрации (10%). Результаты эксперимента представлены на рис.4.

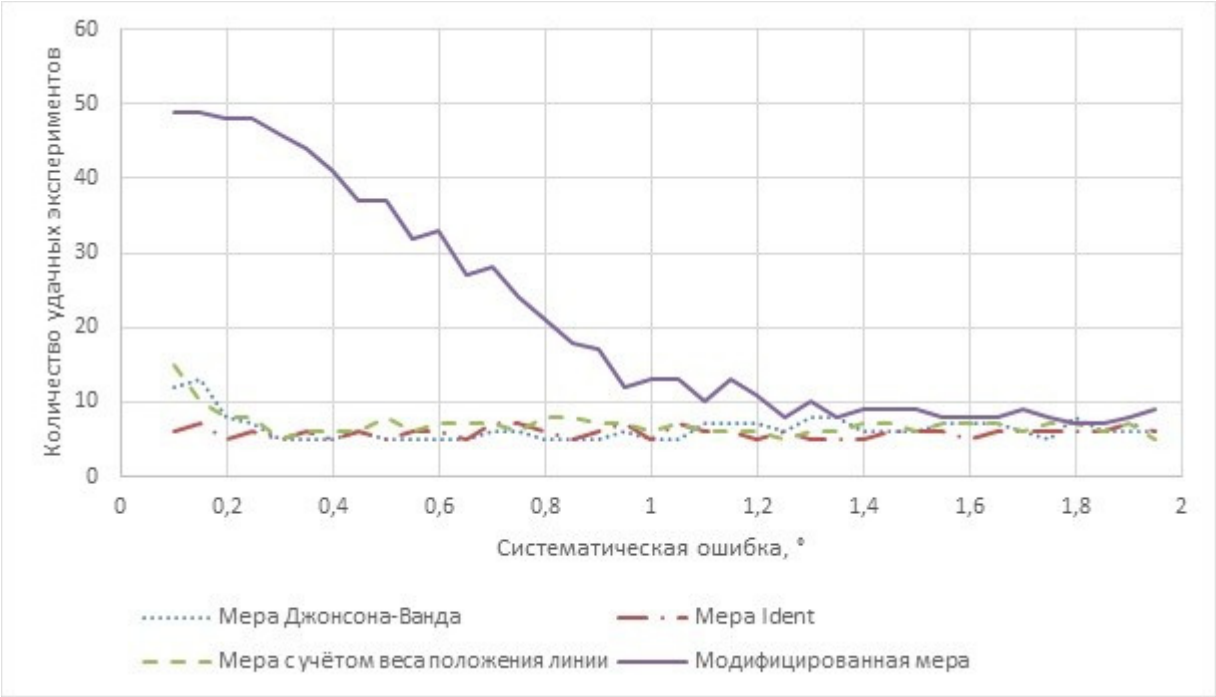


Рис. 4. Зависимость точности идентификации 4-фазного образца с одной фазой малой концентрации от систематической ошибки измерения положений штрихов

В результате проведенного эксперимента гипотеза подтвердилась. Идентификация наличия малой фазы в образце на фоне фаз больших концентраций всегда представляет затруднение, поскольку обычно на дифрактограмме линии фазы малой концентрации перекрываются остальными линиями. Именно из-за того, что фаза малой концентрации не оказалась в группе первых 4-х стандартов с наибольшим значением, все меры близости кроме модифицированной, показали стабильную низкую точность (в среднем, 10 правильно идентифицированных штрих-диаграмм из 50, или 20%). Достоверность идентификации модифицированной мерой стала спадать еще быстрее, чем во втором эксперименте, и уже при систематической ошибке в 1° составила всего 24% правильно идентифицированных штрих-диаграмм (12 из 50).

Таким образом, на основании первых 3-х экспериментов можно сделать следующий вывод: использование модифицированной меры близости позволяет достичь высокой степени точности идентификации как монофазных образцов, так и образцов многофазных с фазой малой концентрации, но при условии малой систематической ошибки. Точность идентификации падает со 100% правильно идентифицированных штрих-диаграмм при систематической ошибке 0,1 до 40% при систематической ошибке в 0,8°. Второй вывод: точность идентификации штрих-диаграммы модифицированной

мерой близости не зависит от состава и концентрации фаз многофазного образца при условии малой систематической ошибки (до 0.2°).

## Заключение

Разработана усовершенствованная методика оценки меры близости штрих диаграмм исследуемого многофазного вещества с эталонными штрих-диаграммами. Главными ее особенностями являются использование ширины окна, меняющейся в зависимости от положения линий по углу, отделение случайно хорошо совпавших стандартов на предварительном этапе, учёт систематической ошибки положений штрихов.

Проведены численные эксперименты влияния систематических ошибок экспериментальных данных на вероятность правильной идентификации фаз. Получено, что предложенный метод превосходит существующие по стандартным параметрам.

## Список литературы


1. Романова Т.Н., Винтайкин Б.Е., Зинченко А.М. Симулятор рентгеновского дифрактометра // Вестник МГТУ им. Н. Э. Баумана. Сер. Приборостроение. 2013. №.18. С. 34–46
2. Якимов И.С. Регуляризация методов бесстандартного рентгенофазового анализа // Журнал структурной химии. 2011. № 2. С. 329–335
3. Винтайкин Б. Е. Физика твёрдого тела: учебное пособие. М.: Издательство МГТУ имени Н.Э. Баумана, 2006. 360 с.
4. Le Meins J.-M., Granswick L.M.D., Le Beil A. Protein powder diffraction // Powder Diffraction. 2003. Vol. 18. P. 106–113.
5. Нахмансон М.С. Фекличев В.Г. Диагностика состава материалов рентгенодифракционными и спектральными методами. Л.: Машиностроение. Ленинградское отделение, 1990. 357 с.
6. Миркин Л.И. Справочник по рентгеноструктурному анализа поликристаллов. Л.: Изд-во физ.-мат. литературы, 1961. 863 с.