МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ»

На правах рукописи

УДК 004.93:681.5.012:004.62(043.5)

Третьяков Игорь Александрович

АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ СТРУКТУРНОГО АНАЛИЗА МАССИВОВ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ НАУЧНЫХ ИССЛЕДОВАНИЙ

Специальность 05.13.06 – Автоматизация и управление технологическими процессами и производствами (по отраслям) (технические науки)

ДИССЕРТАЦИЯ

на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Данилов В.В.

Идентичность всех экземпляров
ПОДТВЕРЖДАЮ
Ученый секретарь диссертационного
совета Д 01.024.04
кандидат технических наук, доцент Т.В. Завадская

Донецк - 2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. На современном этапе развития всех сторон деятельности человека решающую роль играют процессы получения, хранения, обработки и представления информации. В частности, в области развития научных исследований первостепенное значение приобретают методы и вычислительные алгоритмы анализа, фильтрации, преобразования и классификации экспериментальных данных как средства обоснования принимаемых решений и выводов.

Значительные объемы научной информации представляются в виде экспериментальных кривых. Данные этого типа широко используются в автоматизации управления технологическими процессами в промышленности, так как являются одним из способов представления результатов в автоматизированных системах научных исследований. Таким образом представляют, например, хроматограммы в анализе физико-химических свойств веществ, электро-фонокардиограммы и электроэнцефалограммы в медицине, спектры колебаний молекул в спектроскопии.

В связи с тем, что в современном мире постоянно возрастает сложность технологических процессов и сложность новых научных теорий, результаты научных исследований в виде массивов экспериментальных данных содержат десятки и сотни тысяч компонентов. Такие массивы экспериментальных данных не содержат в явном виде информации о свойствах исследуемого процесса, а наиболее существенные свойства и характеристики исследуемого процесса оказываются недоступными для непосредственного измерения. Поэтому возникает необходимость в разработке специальных вычислительных алгоритмов и эффективных методов анализа, аппроксимации, классификации и построения точного сжатого описания экспериментальных данных. Учитывая вышесказанное, автоматизация процессов анализа массивов экспериментальных данных, их классификации и представления в виде сжатого описания является актуальной научно-технической задачей, имеющей отраслевое значение.

Степень разработанности темы исследования. Анализ литературных источников и результатов исследований, полученных рядом авторов, обеспечил наличие достаточно обширного статистического материала, что, в свою очередь, обеспечило высокую аргументированность полученных научных результатов проведенного исследования и подтвердило актуальность выбранной темы. Из анализа литературных источников следует, что в настоящее время четко определены два крупных класса методов анализа экспериментальных кривых: интегральный и структурный. В интегральном подходе сжатое описание кривой строится без предварительного выделения какого-либо её сегмента либо разделения ее на однородные фрагменты. Недостатками такого подхода являются: необходимость задания класса экспериментальных кривых, что на практике зачастую невозможно, особенно для слабо изученных процессов; необходимость достаточного уровня знаний об анализируемом процессе; сильная корреляция переменных при построении моделей, возникающая из-за плохой обусловленности системы нормальных уравнений. Методы структурного подхода представляют сжатое описание экспериментальной кривой в два основных этапа, а именно: разделение кривой на однородные участки и построение сжатого описания кривой в целом. Применение методов структурного класса в действительности является приемлемым в большинстве случаев. Таким образом для экспериментальных кривых, являющихся неоднородными на всей области определения, возможно получить разделение на такие интервалы, на каждом из которых кривая оказывается более простой, что позволяет использовать для представления сжатого описания кривой достаточно типичные локальные модели. Полученные описания участков таких кривых несут важные данные о функционировании анализируемого процесса, например, о режимах работы исследуемого объекта. В рамках структурного подхода к анализу экспериментальных кривых разработано много различных методов, однако их практическое использование часто является неэффективным, а в некоторых случаях и невозможным, что определяет необходимость дальнейших разработок.

Цели и задачи исследования. Целью работы является обоснование модернизированных методов и алгоритмов автоматизации процессов анализа массивов экспериментальных данных, их классификации и представления в виде компактных структур. Для достижения данной цели сформулированы и решены следующие задачи:

1. Разработка комбинированных вычислительных методов сегментации экспериментальных кривых.

2. Алгоритмическая реализация процедуры структурного анализа экспериментальных данных.

3. Разработка методов лингвистического описания экспериментальных кривых.

4. Исследование применимости разработанной системы структурного анализа в процессах электромагнитной совместимости радиоэлектронных средств.

Объект исследования. Объектом исследования являются массивы данных экспериментальных исследований.

Предмет исследования. Предметом исследования являются вычислительные методы автоматизации структурного анализа массивов данных.

Научная новизна полученных результатов состоит в следующем:

1. Дальнейшее развитие получил метод структурного анализа данных, в рамках лингвистического подхода к анализу экспериментальных кривых, в котором анализируемая кривая описывается в виде сжатого описания из последовательного ряда символов либо целых слов из определенного алфавита. Каждый элемент такого ряда представляет соответствующий участок либо группу участков, определенный сегментацией анализируемой экспериментальной кривой.

2. Впервые предложен вычислительный метод сегментации массивов экспериментальных данных с использованием функций сложности, отличающийся способностью осуществлять бинарную классификацию помимо сегментации, а также наличием дополнительных условий во избежание выделения ложных экстремумов.

3. Впервые предложен метод лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами, в котором через

конечное число циклов достигается устойчивая классификация и ни один вектор не переносится из одного класса в другой, и отличающийся способностью классификации по признаку минимума расстояния до эталона.

4. Впервые применены вычислительные методы сегментации и лингвистического описания для автоматизации процедуры структурного анализа экспериментальных данных к исследованию спектрограмм радиосигналов. При этом обоснована процедура составления более расширенного лингвистического описания экспериментальных кривых, позволяющая составлять это описание с учетом местоположения участков кривой на оси абсцисс.

Теоретическая значимость диссертационной работы состоит установлении закономерностей и теоретическом обосновании структурных методов анализа данных и принимаемых решений в области автоматизации процессов анализа массивов экспериментальных данных, их классификации и представления.

Практическая значимость результатов диссертационной работы заключается в классификации экспериментальных данных и представления их в виде сжатого описания с использованием вычислительных алгоритмов сегментации экспериментальных кривых на основе функции сложности, а также вычислительных алгоритмов лингвистического описания участков экспериментальных кривых для автоматизации процедуры структурного анализа массивов экспериментальных данных научных исследований при решении практических задач. Практическая ценность работы подтверждается:

а) внедрением вычислительного алгоритма сегментации массивов экспериментальных данных с использованием функций сложности и вычислительного алгоритма лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами в научно-исследовательский процесс Государственного учреждения «Донецкий физико-технический институт им. А.А. Галкина»;

б) использованием вычислительных алгоритмов сегментации экспериментальных кривых на основе функции сложности и вычислительных алгоритмов лингвистического описания участков экспериментальных кривых для

автоматизации процедуры структурного анализа массивов экспериментальных данных научных исследований при выполнении научно-исследовательских работ кафедры радиофизики и инфокоммуникационных технологий ГОУ ВПО «Донецкий национальный университет» в 2018-2019 гг. (справка №5516/01-27/6.20 от 12.12.2019 г.);

в) внедрением методики компьютерной обработки для выделения и анализа экспериментальных данных, методики сегментации и анализа экспериментальных данных и вычислительных алгоритмы сегментации на основе функции сложности для выделения информативных участков экспериментальных кривых в учебный процесс ГОУ ВПО «Донецкий национальный университет» путем использования при чтении лекций и проведении лабораторных занятий по дисциплине «Цифровая обработка сигналов» для подготовки бакалавров по направлениям подготовки 03.03.03 Радиофизика и 10.03.01 Информационная безопасность (справка №5515/01-27/6.20 от 12.12.2019 г.).

Методология и методы исследования. Для решения поставленных задач и получения результатов применялись современные научные методы исследования: систематизация, обобщение и статистический анализ экспериментальных данных, методы анализа данных, методы сравнения с эталонами, методы лингвистического описания, а также практическая апробация результатов исследований на экспериментальных данных, полученных на лабораторных макетах.

Положения, выносимые на защиту:

1. Показано, что применение структурного подхода целесообразно для анализа массивов экспериментальных данных, а применение разработанных вычислительных методов сегментации с использованием функций сложности и лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами позволяют осуществлять бинарную классификацию во избежание выделения ложных экстремумов и устойчивую классификацию по признаку минимума расстояния до эталона.

2. Установлено, что разработанные вычислительные методы и алгоритмы позволяют автоматизировать процессы структурного анализа экспериментальных

данных научных исследований, осуществлять их классификацию и представление в виде сжатого описания. При этом обоснована процедура составления расширенного лингвистического описания экспериментальных кривых, позволяющая составлять это описание с учетом местоположения участков кривой на оси абсцисс.

Степень достоверности и апробация результатов. Обоснованность и достоверность научных результатов исследования подтверждается значительным количеством наблюдений, современными методами исследования, которые соответствуют поставленным в работе целям и задачам. Научные положения и выводы, сформулированные в диссертации, подтверждаются убедительными фактическими данными, наглядно представленными в приведенных таблицах и графиках.

Полученные результаты диссертации соответствуют областям исследования паспорта научной специальности 05.13.06 – Автоматизация и управление технологическими процессами и производствами (по отраслям) (технические науки), а именно: п. 8 «Формализованные методы анализа, синтеза, исследования и оптимизация модульных структур систем сбора и обработки данных в АСУТП, АСУП, АСТПП и др.», п. 18 «Средства и методы проектирования технического, математического, лингвистического и других видов обеспечения АСУ», п. 20 «Разработка автоматизированных систем научных исследований».

Основные положения диссертационного исследования в достаточной степени апробированы на следующих научных и научно-практических конференциях:

II Международная научная конференция студентов и молодых ученых «Донецкие чтения 2017: Русский мир как цивилизационная основа научно-образовательного и культурного развития Донбасса», г. Донецк, 17-20 октября 2017 г.;

Международная научно-практическая конференция «Социально-гуманитарные и естественно-технические науки и вызовы современности», г. Ставрополь, 22 декабря 2017 г.;

Международная научно-практическая конференция "Открытые физические чтения – 2018", г. Луганск, 18-19 мая 2018 г.;

III Международная научная конференция «Донецкие чтения 2018: образование, наука, инновации, культура и вызовы современности», г. Донецк, 25 октября 2018 г.;

IV Международная научная конференция «Донецкие чтения 2019: образование, наука, инновации, культура и вызовы современности» г. Донецк, 31 октября 2019 г.;

Международная научно-практическая конференция молодых исследователей им. Д.И. Менделеева, г. Тюмень, 15 ноября 2019 г.

Личный вклад соискателя. Основные научные результаты диссертационного исследования получены соискателем самостоятельно. Личный вклад соискателя заключается в обосновании идеи исследования и ее реализации, постановке целей и задач исследования, выборе методологии и методов исследования, проведении теоретических и экспериментальных исследований, а также во внедрении результатов диссертации.

Публикации. Основные научные результаты диссертационного исследования опубликованы в 13 научных изданиях, из них: 5 - в научных изданиях, включенных в перечень ВАК ДНР, 3 – в других изданиях, 5 – в материалах научных конференций.

Структура и объем диссертации. Диссертация состоит из введения, четырех разделов с выводами по каждому из них, заключения, списка литературы из 153 наименований и 1 приложения. Полный объем диссертации составляет 150 страниц, включая 22 рисунка и 2 таблицы.

РАЗДЕЛ 1 СОВРЕМЕННОЕ СОСТОЯНИЕ И АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ КРИВЫХ

В данном разделе проанализированы методы анализа экспериментальных кривых, а именно два крупных класса: интегральный и структурный. Оба класса методов состоят из этапа сегментации и этапа описания экспериментальных кривых. Методы сегментации рассмотрены в подразделе 1.2 и в свою очередь делятся на последовательные и параллельные, каждый из которых подразделяется на четыре типа: методы нелинейной и кусочной аппроксимации, параметрические и непараметрические методы обнаружения разладок, методы сравнения с эталонами, методы выделения информативных участков. Среди методов описания экспериментальных кривых выделяются две группы: дискриминантные и лингвистические, которые рассмотрены в подразделе 1.3. В подразделе 1.4 представлены практические способы применения рассмотренных методов анализа экспериментальных кривых. В 1.5 определяется цель и задачи данного диссертационного исследования.

1.1 Методы анализа экспериментальных кривых

В настоящее время четко определены два крупных класса методов анализа экспериментальных кривых: интегральный и структурный. В интегральном подходе сжатое описание кривой строится без предварительного выделения какого-либо её сегмента либо разделения ее на однородные фрагменты. Для решения данной задачи вводится некоторый критерий приближения кривой $f(t)$ к выбранной параметрической модели. Вектора значений параметров модели, принимающие экстремальные значения, являются отыскиваемыми описаниями анализируемой кривой. Так применяются разложения по разного рода системам линейно-зависимых функций [1], канонические разложения случайных функций [2], представления по «естественным» либо базисным функциям [3, 4]. К интегральному классу так же относят методы, параметры, характеризующие экспериментальную кривую, в которых выбираются предварительно.

Отыскиваемым сжатым описанием анализируемой таким методом кривой является вектор либо набор векторов значений этих параметров, рассчитанный по её отсчётам. Например, в [5] в качестве таких параметров используются максимальное и минимальное значение площади модуля функции *f(t)*, а в [6] количество пересечений кривой *f(t)* оси абсцисс *t* и гистограмма длин интервалов между такими пересечениями. Из недостатков интегральных методов можно выделить:

-необходимость задания класса экспериментальных кривых, что зачастую невозможно, особенно для слабо изученных процессов;

-необходимость достаточного уровня знаний об анализируемом процессе;

-сильная корреляция переменных при построении моделей (полиномиальных, например), возникающая из-за плохой обусловленности системы нормальных уравнений;

-низкое быстродействие алгоритмов, реализующих оценку параметров модели.

Методы структурного подхода представляют сжатое описание экспериментальной кривой в два основных этапа, а именно: разделение кривой на однородные участки и построение сжатого описания кривой в целом. Применение методов структурного класса в действительности является оптимальным в большинстве случаев. Таким образом для экспериментальных кривых, являющимися неоднородными на всей области определения, выходит получить разделение на такие интервалы, на каждом из которых кривая оказывается более простой, что позволяет использовать для представления сжатого описания кривой достаточно типичные локальные модели. Экспериментальные кривые, такие как осциллограммы речи, хроматограммы, электрокардиограммы, шумы двигателей и т.д., являются структурными по своей природе. Полученные описания участков таких кривых несут важные данные о функционировании анализируемого процесса, например, о режимах работы исследуемого объекта.

1.2 Методы сегментации экспериментальных кривых

При решении задачи сегментации выделяют две различные группы методов: параллельные и последовательные. Методы первой группы [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27] предусматривают, что анализируемая экспериментальная кривая обрабатывается полностью. Для решения данной задачи вводится некоторый параметр оценки качества сегментации, зависящий от конкретного разделения кривой. Разделение, при котором данный параметр принимает экстремальное значение и является искомым. Например, в работах [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] для оценки качества сегментации вводится квадратичная невязка и минимаксные критерии в [18, 19, 20, 21, 22, 23]. В основном границами искомых сегментов считают моменты изменения свойств анализируемого случайного процесса. В [28, 40] используют функцию правдоподобия для решения задачи сегментации. В публикациях [54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64] применяют различные статистики совпадения распределения двух случайных выборок, а [39, 42] – квадратичный параметр. Сегментация кривой на последовательность участков простого и сложного характера поведения рассматривается в [65, 66, 67, 68]. А именно в работе [65] вводится параметр минимума квадратичной невязки, рассчитываемый на однородной(простой) части экспериментальной кривой. При анализе кривых акустических колебаний в работах [69, 70, 71, 72, 73, 74, 75, 76] применяют параметр, рассчитанный на основе деформации оси времени t.

Методы второй группы, последовательные, предусматривают, что на данный момент анализируется только определенная часть экспериментальной кривой, на которой и необходимо выделить границы соответствующих участков. Использование таких методов приводит к возникновению других типов параметров и к необходимости применять действия последовательно. Для определения разладок (моментов изменения свойств случайного процесса) в работах [86, 87, 88, 89, 90] используется функция времени в качестве параметра оценки, значение которой минимально на однородных участках экспериментальной кривой, и

значительно увеличивается при изменении свойств случайного процесса. Разладка считается определенной, когда функция достигает заданного ранее значения. В [91, 92, 93, 94, 95, 96, 97, 98] предлагаются специально построенные статистики в качестве параметров оценки, а разладка считается определенной, когда значение статистики превышает допустимое заранее заданное значение. В [99, 100, 101, 102] рассматривается способ сегментации экспериментальной кривой с помощью заранее построенных эталонных элементов. Разделение кривой на простые и сложные участки рассмотрено в [103, 104, 105, 106, 107]. Так же специфическим в своем роде является подход [108], в котором предварительно выбирают параметры, характеризующие участки экспериментальных кривых, значения которых удовлетворяют определенным заданным значениям. К нему же относятся методы кусочной аппроксимации, в которых применяются и квадратичные и минимаксные параметры [77, 78, 79, 80, 81, 82, 83, 84, 85]. Недостатком такого подхода является ориентированность на аппроксимацию одномерных сравнительно гладких экспериментальных кривых.

Обе группы рассмотренных методов имеют как преимущества, так и недостатки. Методам первой группы необходим больший объем вычислений, сравнительно со второй группой. Общим недостатком обоих групп является неопределенность в выборе количества участков.

Существуют так же работы, в которых в некотором смысле объединены методы первой и второй группы. Так, в [19, 21, 109, 110, 111, 112, 113] последовательным алгоритмом происходит первичное предварительное разделение экспериментальной кривой, которое в дальнейшем улучшается методами локальной оптимизации заданного параметра.

1.2.1 Методы нелинейной и кусочной аппроксимации

Описанные далее методы относятся к группе методов **параллельной** сегментации. Данные методы [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] основаны на предположении о том, что анализируемая экспериментальная кривая обрабатывается полностью, а для сегментации применяют параметр оценки

качества сегментации как функционал. Далее поставим задачу сегментации в общем виде. Пусть на отрезке $[t_1,t_n]$ задана экспериментальная кривая *f(t)* причем значения *f(t)* заданы в дискретные моменты $t_l, l=1\ldots n$. Необходимо:

- найти такое разделение $T=(T_0,T_1,\ldots T_r)$, $T_0<T_1<\ldots T_r$, $T_0=t_1$, $T_r=t_n$ выбранного отрезка $[t_1,t_n]$ на *r* интервалов, причем *r* неизвестно;

- построить на каждом интервале $\left[T_j,T_{j+1}\right], j=0\ldots r-1$ локальные аппроксимирующие функции $F_j(t), j=1\ldots r$, известные с точностью до параметров и удовлетворяющие заданной степени гладкости их сопряжения в узлах разделения, чтобы выбранный параметр оценки качества аппроксимации *J* принимал экстремальное значение. В качестве *J* используют либо квадратичный, либо минимаксные параметры, как было сказано ранее.

Квадратичный параметр в непрерывной форме представляет собой:

$$J(T,a)=\sum_{j=1}^{t}\int_{T_{j-1}}^{T_j}\left[y(t)-F_j\left(t,a_j\right)\right]^2 dt, \qquad (1)$$

где $a=(a_1,a_r)$, $\left\{F_j\left(t,a_j\right)\right\}$ - параметрически заданный класс функций. $F_j\left(t,a_j\right)$ принадлежит классу полиномов определенного порядка, а именно $F_j\left(t,a_j\right)=\sum_{k=0}^{m-1}a_k^j t^k$.

Множество различных разделений конечно, так, как для экспериментальной кривой *f(t)* задано конечное количество значений в дискретные моменты времени $t_l, l=1\ldots n$. Значит найти минимальное значение параметра (1) возможно методом полного подбора. Коэффициенты a_j рассчитываются из условий минимума параметра, а именно из системы уравнений $\frac{\partial J}{\partial a_j^i}; i=0\ldots m-1;\ j=i\ldots r$. Таким образом, для каждого возможного разделения *T* рассчитывается значение (1), значение которого будет минимальным при самом оптимальном разделении. Недостатком метода полного перебора являются временные затраты, которые можно существенно сократить, используя рекуррентные выражения динамического программирования [8], а именно:

$$f_0(t_Q)=\min_{t_1\le T_1\le t_Q}\left\{\int_{T_0}^{T_1}\left[f(t)-F_1(t_1\hat{a}_1)\right]^2 dt+\int_{T_0}^{T_1}\left[f(t)-F_2(t_1\hat{a}_2)\right]^2 dt\right\},\ Q=1\ldots n\ .$$

Для $N=2,3,\ldots,r-1$:

$$f_N(t_Q)=\min_{t_1\le T_N\le t_Q}\left\{f_{r-1}(T_n)+\int_{T_N}^{t_Q}\left[f(t)-F_{N+1}(t_1\hat{a}_{N+1})\right]^2 dt\right\},\ Q=1\ldots n\ ; \qquad (2)$$

$$\min_{T,a} J(T,a)=f_{r-1}(t_N),$$

где $\hat{a}$ – оценки коэффициентов, рассчитываемые на соответствующем интервале. На последнем цикле алгоритма (2) находится оптимальное значение T_{r-1}^* границы T_{r-1}. Для нахождения оптимальных границ интервалов задаётся $n=T_{r-1}^*$, $r\to r-1$ и повторяются шаги (2) заново. Например, для нахождения оптимального значения T_{r-2}^* границы T_{r-2} задаётся $n=T_{r-2}^*$, $\boldsymbol{r=r-1}$. Следовательно, за $\boldsymbol{r-1}$ повторений шагов алгоритма, определяются все оптимальные значения границ интервалов. Для определения количества интервалов, в случае если оно не задано предварительно, задается величина ошибки J^0, и r находится с помощью соотношения $f_{r-1}(t_n)\le J^0$. Рассмотренный метод полного перебора даёт возможность доставить глобальный минимум параметру (1), но всё же требует определенные временные затраты. В публикации [9] рассматривается более быстрый алгоритм, в результате работы которого удается получить локальный минимум параметру (1). А именно, задается величина максимально допустимой ошибки аппроксимации J^0 и первичное произвольное разделение $T^0=(T_0,T_1^0,\ldots,T_{N-1}^0,T_N)$. Далее каждым *k*-м шагом алгоритма строится разделение $T^k=(T_0,T_1^k,\ldots,T_{l-1}^k,T_l^k)$, а разделение на шаге $(k+1)$ находится с помощью последовательности действий:

- определяется интервал, на котором величина ошибки аппроксимации J^0 максимальна и делится на два подинтервала следующим способом: при наличии двух и более точек, в которых величина ошибки аппроксимации максимальна, границей интервала выбирается точка в середине между ними, в других случаях – интервал делится напополам. Шаг повторяется пока $J>J^0$, при $J\le J^0$ начинается выполнение следующего шага;

- находятся два смежных интервала, при соединении которых невязка (1) увеличивается минимально. При условии $J \le zJ^0$ (z – параметр алгоритма) интервалы объединяются и шаг выполняется вновь;

- осуществляется передвижение границ интервалов так, чтобы значение параметра не увеличивалось.

Алгоритм завершается, когда на каком-либо шаге k^* появляется $T^{k^*} = T^{k^*+1}$. В работах [10, 11] для определения условий экстремума параметра (1), его частные производные по всем параметрам приравнивают к нулю и получают систему уравнений:

$$\frac{\partial J}{\partial a_i^j} = \int_{T_{j-1}}^{T_j} \left[f(t) - F_j\left(t_1 a^j\right) \right] t^j dt = 0, \; i = 1 \ldots m-1, \; j = 1 \ldots r; \tag{3а}$$

$$\frac{\partial J}{\partial T_j} = e^2\left(j, T_j\right) - e^2\left(j+1, T_j\right) = 0, \; j = 1 \ldots r-1, \tag{3б}$$

где $e\left(j, T_j\right) = f\left(T_j\right) - F_j\left(T_j, a^j\right)$ и $e\left(j+1, T_j\right) = f\left(T_j\right) - F_{j+1}\left(T_j, a^{j+1}\right)$ ошибки аппроксимации на границах интервалов. Для решения системы уравнений (3) предлагается использовать итерационные алгоритмы, так как аналитическое решение получить не удается. В работе [10] заданы начальные значения $\theta_1^0, \theta_2^0, \ldots, \theta_{3r-1}^0$ параметров $\theta_1, \theta_2, \ldots, \theta_{3r-1}$, где $\theta_1 = a_1^1, \theta_2 = a_1^2, \ldots, \theta_r = a_1^r$, $\theta_{r+1} = a_2^1, \ldots, \theta_{2r} = a_2^r$, $\theta_{2r+1} = T_1, \ldots, \theta_{3r-1} = T_{r-1}$. Далее на k-м шаге алгоритма находятся значения параметров $\theta_1^k, \ldots, \theta_{3r-1}^k$, а значения $\theta_1, \ldots, \theta_{3r-1}$ на шаге $(k+1)$ находятся с помощью системы уравнений:

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial \theta_i^k} + \sum_{j=1}^{3r-1} \left(\theta_j^{k+1} - \theta_j^k \right) \frac{\partial^2 J}{\partial \theta_i^k \partial \theta_j^k} = 0, \; i = 1 \ldots 3r-1, \tag{4}$$

где $\frac{\partial J}{\partial \theta_i^k}$ – значение производной $\frac{\partial J}{\partial \theta_i}$ в точке θ_i^k. Расчёт (4) осуществляется разложением $\frac{\partial J}{\partial \theta}$ в ряд Тейлора в окрестности точки θ_i^k. Алгоритм завершается, когда на каком-либо шаге k^* появляется $\left| J\left(\theta_1^{k^*} \ldots \theta_{3r-1}^{k^*}\right) - J\left(\theta_1^{k^*-1} \ldots \theta_{3r-1}^{k^*-1}\right) \right| < z$, где z – параметр алгоритма. В работе [11] при фиксированном разделении оценки параметров a_j^i без труда рассчитываются из системы уравнений (3а). Следовательно, необходимое решение определяется из системы уравнений (3б), с

помощью итерационного алгоритма Ньютона, рекуррентные выражения которого в матричной форме представляют собой:

$$G\left(T^k\right)\left(T^{k+1}-T^k\right)=W\left(T^k\right), \tag{5}$$

где $W\left(T^k\right)$ – матрица первых производных $\left[\frac{\partial W_i\left(T^k\right)}{\partial T_j}\right]$ $(i,j=1\ldots r-1)$. Вычисление первой производной экспериментальной кривой *f(t)* из-за наличия шумов при практическом применении редко удается произвести. Для устранения этого, в [11] используют такие итерационные выражения:

$$T_j^{k+1}=T_j^k+C\left[e^2\left(j+1,T_j^k\right)-e^2\left(j,T_j^k\right)\right],\ j=1\ldots r-1, \tag{6}$$

где C – постоянная, рассчитываемая из условия $C<\inf\limits_k\left[1/\max\frac{\partial W_k}{\partial T_j^k}\right]$. Другим недостатком вышеописанных алгоритмов (3-6) является отсутствие стопроцентной вероятности получения решения, доставляющего хотя-бы локальный минимум параметру (1) при выполнении необходимых условий.

В публикациях [12, 13, 14, 15, 16, 17] рассмотрены алгоритмы построения аппроксимаций сплайнами, задача которых к поиску минимума параметра (1) при ограничениях на непрерывность производной заданного порядка аппроксимирующей функции. Данные ограничения учитываются при определении сплайна некоторой функцией $F(t,\theta)$, где θ это вектор неизвестных параметров, принимающий значения $\theta_1\ldots\theta_p$. Например, для построения полиномиального сплайна *m*-степени, с непрерывной (*m*- 1)-й производной, $F(t,\theta)$ принимает вид:

$$F(t,\theta)=\sum_{i=0}^{m}d_it^i+\sum_{j=1}^{r-1}\frac{C_j}{m!}\left(t-T_j\right)^m I_+\left(t-T_j\right), \tag{7}$$

где $I_+(x)=0$, при $x<0$ и $I_+(x)=1$, при $x\geq 0$. Таким образом, параметр (1) в дискретной форме представляет собой:

$$J(\theta)=\sum_{i=1}^{n}\left[f_i(t_i)-F(t_i,\theta)\right]^2. \tag{8}$$

Множество различных разделений конечно и найти минимальное значение параметра (8) возможно, опять же, методом полного подбора. Но в данной ситуации на выполнение этого метода потребуется еще больше временных затрат,

поскольку при фиксированном разделении для определения оценок вектора параметров необходимо решение системы алгебраических линейных уравнений. Для сокращения временных затрат на расчёты в работе [12] применяют метод случайного поиска, в котором оптимальное значение параметра (8) рассчитывается при случайно сгенерированном разделении *T*. Осуществляется сравнительно большое количество таких расчётов, среди которых искомым разделением T^* считается разделение с минимальным значением параметра $J(T^*)$. В работах [13, 14] необходимая аппроксимация находится из системы уравнений:

$$\begin{cases} \dfrac{\partial J}{\partial \theta_i} = 0, \ i = 1 \dots p, \end{cases} \tag{9}$$

для решения которой, в [13] используется модифицированный алгоритм Ньютона. А в [14] решается задача построения непрерывного полиномиального сплайна, для решения которой необходимые условия минимума (9) параметра (8) представляют собой:

$$V_j(T) = e(j, T_j) - e(j+1, T_j) = 0, \ j = 1 \dots r-1, \tag{10}$$

где $e(j, T_j) = f(T_j) - F_j(T_j, a^j)$ и $e(j+1, T_j) = f(T_j) - F_{j+1}(T_j, a^{j+1})$ ошибки аппроксимации. Необходимое решение системы (10) находится с помощью итерационного алгоритма Ньютона, рекуррентные выражения которого в матричной форме аналогичны выражениям (5) и представляют собой:

$$G_i(T^k)(T^{k+1} - T^k) = V(T^k), \tag{11}$$

где $V(T^k) = (V_1(T^k) \dots V_{r-1}(T^k))$, $G_i(T^k) = \left(\dfrac{\partial V_i}{\partial T_j} \right), i, j = 1 \dots r-1.$ Так, как матрица тридиагональна, то ее обращение требует немалого объема вычислений. Для упрощения вычислений в работе [14] предлагается эту матрицу приближенно представлять в диагональной форме, на каждом шаге алгоритма выстраивая сплайн перед расчётом границ интервалов (11). Такой сплайн должен удовлетворять двум условиям: абсолютное значение изменения тангенса угла наклона аппроксимирующих линий на границах интервалов должно быть много больше отношения ошибки аппроксимации на соответствующих границах интервалов к соответствующим их длинам. Такое отношение будет мало, если ошибки

аппроксимации на границах интервалов будут значительно меньше длин интервалов, граничащих с ними. В целом алгоритмы (10, 11) имеют такие же недостатки, как и описанные выше (3-6), а именно отсутствие стопроцентной вероятности получения решения, доставляющего хотя-бы локальный минимум параметру (1) при выполнении необходимых условий.

В работах [15, 16, 17] показаны частные случаи, для которых применим метод динамического программирования (2). А именно, в [15] строится непрерывный кусочно-линейный сплайн, концы отрезков которого лежат на заданной кривой и полностью определены выбором границ интервалов. В [16] строится предположение, что анализируемая экспериментальная кривая является выпуклой функцией, а её оптимальная кусочно-линейная аппроксимация является непрерывной. В [17] строится предположение, что количество значений кусочно-линейного сплайна на границах разделения в каждый момент времени известно и конечно.

Минимаксный параметр для оценки качества аппроксимации в работах [18, 19, 20, 21, 22, 23] представляет собой:

$$\hat{J}(T,a) = max\{e_j(T_{j-1},T_j): j = 1\ldots r\}, \tag{12}$$

где $e_j(T_{j-1},T_j) = \min\limits_{a_j} \max\limits_{T_{j-1}\le t\le T_j} |f(t) - F_j(t,a_j)|$. Для минимизации (12) применяются рекуррентные выражения метода динамического программирования:

$$f_1(t_Q) = \min\limits_{T_0\le T_1\le t_Q} max\{e(T_0,T_1), e(T_1,t_q)\}\ , Q = 1\ldots n\ .$$

Для $N = 2,3,\ldots,r-1$: (13)

$$f_N(t_Q) = \min\limits_{T_0\le T_N\le t_Q} max\{F_{N-1}(T_N), e(T_N,t_q)\}\ , Q = 1\ldots n\ .$$

В публикациях [18, 19, 20] аппроксимация определяется достаточными условиями минимума параметра (12), а именно:

$$e_1(T_0,T_1) = e_2(T_1,T_2) = \ldots = e_r(T_{r-1},T_r), \tag{14}$$

В [18] для удовлетворения условий (14) при построении аппроксимации предлагается алгоритм, основанный на:

$$e_j^k(T_{j-1}^k,T_j^k) = C_j^k(S_j^k)^m,\ j = 1\ldots r, \tag{15}$$

где $S_j^k = T_j^k - T_{j-1}^k$ – длина интервала на *k*-м шаге алгоритма, C_j^k - неизвестные коэффициенты, m – размерность полинома $F_j(t, a_j)$. Иными словами строится предположение, что ошибки аппроксимации $e_j(T_{j-1}, T_j)$, $j = 1 \ldots r$ пропорциональны длинам соответствующих интервалов в степени m. Рекуррентные выражения данного алгоритма выглядят следующим образом:

$$S_j^{k+1} = \sqrt[m]{E / C_j^k}\,, \tag{16}$$

где C_j^k, $j = 1 \ldots r$ находятся из (15), а $E = \left[\frac{(T_r - T_0)}{\sum_{j=1}^{r} \left(1 / \sqrt[m]{C_j^k}\right)} \right]^m$ – условие постоянства суммы длин интервалов. В описанном алгоритме в случае нулевой ошибки на интервале $C_j^k = 0$ вычислить S_j^k не получается. Этот недостаток предлагается устранить в [19] исключением интервалов с нулевой ошибкой из анализа, после чего происходит обработка остальных участков экспериментальных кривых. В [20] показан алгоритм схожий с (6), а именно:

$$T_j^{k+1} = T_j^k + C_j \left(e_{j+1}^k - e_j^k\right). \tag{17}$$

Общим недостатком трёх вышеописанных алгоритмов (13-17) является то, что не всегда существует решение, удовлетворяющее условиям (14) при дискретном задании экспериментальной кривой.

В [21] рассматривается алгоритм, не выполняющий условия (14), но доставляющий локальный минимум параметру (12). А именно, задается первичное произвольное разделение $T^0 = \left(T_0, T_1^0, \ldots, T_{r-1}^0, T_r\right)$. Далее каждым *k*-м шагом алгоритма строится разделение $T^k = \left(T_0, T_1^k, \ldots, T_{r-1}^k, T_r\right)$, а разделение на шаге $(k+1)$ находится с помощью последовательности действий:

- определяются величины ошибки аппроксимации $e_j\left(T_{j-1}^k, T_j^k\right)$ и $e_{j+1}^k\left(T_j^k, T_{j+1}^k\right)$ для нечетных $j = 1, 3, 5, \ldots$. Далее происходит сдвиг граничной точки T_j^k на заданное M число отсчётов к интервалу с большей ошибкой и для новой границы $\left(T_j^k\right)^H$

определяются значения величин $\left(e_j^k\right)^H$ и $\left(e_{j+1}^k\right)^H$. Величина $T_j^k = T_j^k$ при $\max\left\{\left(e_j^k\right)^H, \left(e_{j+1}^k\right)^H\right\} \geq \max\left\{e_j^k, e_{j+1}^k\right\}$ и $T_j^{k+1} = \left(T_j^k\right)^H$ в остальных случаях.

- выполняется шаг 1 для чётных $j = 2,4,6,\ldots$.

Алгоритм завершается, когда на каком-либо шаге k^* появляется $T^{k^*} = T^{k^*-1}$.

Задача построения полиномиальных сплайнов рассмотрена в [7, 22], а именно сплайнов $S_{2m-1}(t)$ степени $(2m-1)$ с непрерывной $(m-1)$-й производной и подходящих под условия: $S_{2m-1}^{(k)}(T_j) = f^{(k)}(T_j), k = 0\ldots m,\ j = 1\ldots r-1$. Параметры таких сплайнов на каждом интервале разделения полностью определяются значениями исходной экспериментальной кривой и ее производными на соответствующих границах, другими словами такие сплайны являются локальными. В таком случае для определения оптимальных границ применим метод динамического программирования, требующий, как было сказано ранее, определенные временные затраты. Поэтому в работах [7, 22] разработаны более быстрые алгоритмы, ошибки аппроксимации в которых не превышают определенной величины, но без получения решения, доставляющего хотя-бы локальный минимум параметру (12). Для произвольного разделения T сплайном $S_{2m-1}(t)$ в [7] предлагаются неравенства:

$$e_j(T_j, T_{j+1}) \leq C_m B_j(T_j, T_{j+1}),\ j = 0\ldots r-1, \tag{18}$$

где $B_j(T_j, T_{j+1}) = (T_{j+1} - T_j)\int_{T_j}^{T_{j+1}} \left| f^{(2m)}(t)\right| dt$, $C_m = \dfrac{1}{2^{2m-2}(2m-2)!}$;

$$\hat{J}(T) \leq C_m \max_{0 \leq j \leq r-1}\left\{B_j(T_j, T_{j+1})\right\} = \Psi(T). \tag{19}$$

Так же в [7] необходимое разделение T^* находят с помощью минимизации верхней границы $\Psi(T)$ по T. В работе [22] в качестве параметра $\Psi(T)$ используется:

$$E(\varphi) = \sum_{j=0}^{r-1} C_m B_j(T_j, T_{j+1}), \tag{20}$$

где $\varphi = \varphi(T)$ - некоторая функция, для которой выполняются условия:

- $\varphi(T_j) = \dfrac{(T_{j+1} - T_j)}{z}$, где z – параметр алгоритма, $j = 0\ldots r-1$;

- $\int_{T_0}^{T_r} \dfrac{1}{\varphi(t)} dt = T - T_0$.

Иными словами разделение T^*, которое доставляет минимальное значение (20), асимптотически оптимально по $\Psi(T)$. В рассмотренном алгоритме в [22] в приближении (20) используется в виде:

$$E(\varphi) \cong \tilde{E}(\varphi) = C_m z^{2m-1} \int_{T_0}^{T_r} \left| f^{2m}(t) \right|^{1/2m} dt . \qquad (21)$$

Необходимая функция $\varphi_0(t)$ определяется минимизацией $\tilde{E}(\varphi)$ по φ методом Лагранжа:

$$\varphi_0(t) = \left| f^{2m}(t) \right|^{-1/2m} \frac{1}{T_r - T_0} \int_{T_0}^{T_r} \left| f^{2m}(t) \right|^{1/2m} dt . \qquad (22)$$

Выше рассмотренные алгоритмы (18, 19, 20-22) имеют один общий недостаток, а именно необходимость расчёта $2m$-й производной анализируемой экспериментальной кривой.

Описанные далее методы относятся к группе методов **последовательной** сегментации. Данными методами [77, 78, 79, 80, 81, 82, 83, 84, 85] на момент исследования анализируется только определенная часть экспериментальной кривой, на которой и необходимо выделить границы соответствующих участков. Далее поставим задачу кусочной аппроксимации. Пусть на отрезке $t \in [T_0, T_r]$ задана экспериментальная кривая *f(t)*. Необходимо:

- найти такое разделение $T(r) = (T_0, T_1, \ldots, T_{r-1}, T_r)$, $T_0 < T_1 < \ldots < T_{r-1} < T_r$, выбранного отрезка $[t_1, t_n]$ на *r* интервалов, причем *r* неизвестно;

- построить на каждом интервале $[T_{j-1}, T_j]$, $j = 1 \ldots r$ аппроксимирующие функции $F_j(t, a)$, известные с точностью до параметров и удовлетворяющие условиям: выбранный параметр ошибки аппроксимации *E* на каждом из интервалов не превышает значение заданной ошибки E_0, а число интервалов *r* - минимально.

В качестве параметра ошибки, как и в [19] применяют либо квадратичные, либо минимаксные параметры. Квадратичные параметры, например квадратичная невязка, гораздо проще рассчитывается для разных классов функций,

встречающихся на практике. Для расчёта минимаксных параметров необходимо применять итерационные алгоритмы.

Применение квадратичных либо минимаксных параметров для экспериментальных кривых *f(t)*, заданных дискретно может привести к получению неоптимальной по числу интервалов аппроксимации, а именно к излишнему разделению участков с одинаковым характером поведения экспериментальной кривой. Для устранения данного недостатка в [21] в качестве параметра ошибки *E* применяется ее среднее значение на соответствующем участке. В данном случае E_0 задается с условием $E_0 \le J^0$, т.е. ошибка аппроксимации не превышает заданное значение J^0. В качестве J^0 применяются параметры (1, 8, 12, 19). Благодаря этому количество ложных границ сегментации уменьшается, но всё же они встречаются.

Качество кусочной аппроксимации в публикациях [77, 78, 79, 80, 81, 82] оценивается с помощью экстраполяционных и интерполяционных алгоритмов. Параметром оценки качества аппроксимации является коэффициент сжатия, который равен отношению числа отсчётов экспериментальной кривой к числу данных, представляющих ее после аппроксимации. Среди данных алгоритмов выделяется реализуемый аппаратно достаточно просто и обеспечивающий достаточно высокий коэффициент сжатия алгоритм прогнозирования нулевого порядка с плавающей апертурой [80, 82], состоящего из следующих шагов:

- заданный набор отсчётов экспериментальной кривой $y_t, t = 1,2,\ldots$ в момент времени T_j соответствует последней определенной границе интервала;

- рассчитывается значение $J_t = \left| f_{T_j - f_t} \right|$ для $t = T_j + 1, T_j + 2, \ldots$ и сравнивается с плавающей апертурой $k_t = g\left| f_t \right|$ для определения следующей границы T_{j+1}. Здесь $g < 1$ - параметр алгоритма;

- искомым $T_{j+1} = t_0$ считается момент времени t_0, на котором $J_t \ge k_t$.

Недостатком методов прогнозирования является малое значение коэффициентов сжатия для экспериментальных кривых, которые непрерывно изменяется случайным образом, или экспериментальных зашумленных высокочастотных кривых. Для анализа таких кривых применимыми являются

интерполяционные методы [77, 78, 79, 80, 81, 82]. Интерполяционные алгоритмы аналогичны описанным алгоритмам в [19], и состоят из следующих шагов:

- заданный параметр ошибки аппроксимации E в момент времени T_{j-1} соответствует последней определенной границе интервала;

- рассчитывается значение параметра ошибки аппроксимации $E(T_{j-1},t)$ для $t=T_{j-1}+1,T_{j-1}+2,\ldots$ и сравнивается с заданной величиной E_0 для определения следующей границы T_j;

- искомым $T_j=t_0-1$ считается момент времени $t=t_0$, на котором $E(T_{j-1},t)\geq E_0$.

Быстродействие данного алгоритма пропорционально скорости расчёта параметра ошибки *E* поступающих отсчётов экспериментальной кривой.

В публикации [78] параметром ошибки аппроксимации *E* выбирается модуль интеграла отклонения анализируемой экспериментальной кривой от ее аппроксимирующей функции. А именно:

$$E_1=\left|\int_{t_1}^{t_2}\left[f(t)-F(t,a)\right]^2 dt\right|, \tag{23}$$

где $F(t,a)$ – локальная функция прямой, проведенной через значения экспериментальной кривой *f(t)* в начале и в конце интервала аппроксимации. В работе [81] параметром ошибки аппроксимации *E* выбирается среднее значение (23).

Для расчёта (23) необходим расчёт двух интегралов:

$$J_1=\int_{t_1}^{t_2} f(t)dt, \tag{24}$$

$$J_2=\int_{t_1}^{t_2} F(t,a)dt=\frac{1}{2}F(t_1)F(t_2)(t_2-t_1). \tag{25}$$

В [80, 82] ошибкой аппроксимации *E* выбирается минимаксный параметр, а локальной функцией $F(t,a)$ выбирается отрезок прямой. В [77] рассматриваются ошибки аппроксимации *E* аналогичные (23-25), но с различными условиями для локальной функции $F(t,a)$. В работах [84, 85] рассматриваются эффективные алгоритмы кусочно-линейной аппроксимации выпуклых функций, практическое

применение которых ограничено необходимостью выпуклости анализируемых экспериментальных кривых.

1.2.2 Параметрические и непараметрические методы обнаружения разладок

Рассмотренные далее методы относятся к группе методов **параллельной** сегментации. Данные методы основаны на предположении о том, что анализируемая экспериментальная кривая обрабатывается полностью, а для сегментации применяют параметр оценки качества обнаружения разладки как функционал. Далее поставим задачу, границами искомых сегментов в которой, считают моменты изменения свойств (разладок) анализируемого случайного процесса. Пусть $f(t)=(f_1,\ldots,f_N)$ является реализацией некоторого случайного процесса *F(t)* в дискретные моменты $t=1,2,\ldots,N$. Необходимо:

- найти моменты времени $T_1,T_2,\ldots,T_{r-1}$, $T_0<T_1<\ldots<T_{r-1}<T_r$, $T_0=1$, $T_r=N$ разладок случайного процесса *F(t)*, причем *r* неизвестно;

- искомым считать разделение, при котором выбранный параметр оценки качества аппроксимации *J* принимает экстремальное значение. В качестве *J* используют параметрические, либо непараметрические величины.

Решение данной задачи **параметрическими методами** рассмотрено далее. В публикациях об изменениях свойств случайных процессов [26, 27] считают, что случайный процесс *F(t)*, $t=1\ldots N$ является последовательностью независимых случайных величин. Распределение у данной последовательности до момента времени T_1 представляет собой $R(F,\theta_1)$, а начиная с этого момента времени T_1 - $Q(F,\theta_2)$. Вектора параметров θ_1,θ_2 неизвестны. Для оценки качества обнаружения разладки T_1 по реализации $f(t)=(f_1,\ldots,f_N)$ используется логарифмическая функция правдоподобия:

$$J(T_1,\theta_1,\theta_2)=\sum_{j=1}^{T_1}\log x(y_j,\theta_1)+\sum_{j=T_1+1}^{T_2}\log y(y_j,\theta_2), \qquad (26)$$

где $T_2 = N$, а *x*, *y* – функции плотности вероятности до момента времени T_1 и после. Отыскиваемым моментом разладки является T_1^*, доставляющее максимальное значение параметру (26), для поиска которого применим метод полного перебора.

В случае, когда θ_1, θ_2 определены, в момент времени T_1 вид распределения не меняется и вместо (26) стоит использовать кумулятивную сумму $J'(T_1) = \sum_{j=1}^{T_1} V_j$, в которой $V_j = \log x(y_j, \theta_1) + \log x(y_j, \theta_2)$. В случае, когда *x* – функция плотности нормального распределения и изменяется только среднее значение [44], то V_j не зависит от T_1 и определяется как $V_j = y_j - \frac{(\theta_1 + \theta_2)}{2}$, где θ_1, θ_2 – средние значения до и после разладки. В случае биномиального распределения [45] $V_j = y_j \log\frac{\theta_1}{\theta_2} - (-y_j)\log\frac{(1-\theta_1)}{(1-\theta_2)}$, где θ_1, θ_2 – значение вероятностей $F(t_j) = 1$ при $t_j \in [1, T_1]$ и $t_j \in [T_1, N_1]$.

В практических случаях зачастую анализируемый случайный процесс является зависимым. Так, в публикациях [28, 29, 30, 31, 32, 33] *F(t)* рассматривают как последовательность из *r* независимых гауссовских случайных процессов, имеющих математическое ожидание $m_j(t, z_1^j)$, $j = 1 \ldots r$ и корреляционные функции $K_j(\upsilon, \tau, z_2^j)$, $j = 1 \ldots r$. Математическое ожидание, корреляционные функции и *r* известны с точностью до параметров. Для данного случая априорная плотность вероятности $\alpha(T)$ моментов времени $T_1, \ldots, T_{r-1}$ неизвестна, и предполагается $\alpha(T) = const$, поэтому параметром выбирается апостериорная плотность вероятности, совпадающая с логарифмической функцией правдоподобия:

$$J_1(T, z) = \sum_{j=1}^{r} \sum_{i,k=T_{j-1}+1}^{T_j} \chi_j(i, k, z_2^j)\left[y_i - m_j(i, z_1^j)\right]\left[y_k - m_j(k, z_2^j)\right] + \sum_{j=1}^{r} \ln\left|K_j(i, k, z_2^j)\right|, \qquad (27)$$

где $z = (z_1^j, z_2^j : j = 1 \ldots r), \left|K_j(i, k, z_2^j)\right|$ - определитель корреляционной матрицы на участке (T_{j-1}, T_j), $\chi_j(i, k, z_2^j)$ - элементы обратной матрицы K^{-1}. Сложные выражения для определения оценок параметров *z* и необходимость обращения матриц

больших размерностей приводит к большим временным затратам на максимизацию (27). Приближенные выражения для вычисления параметра в [30, 31, 32, 33] существенно снижают временные затраты для некоторых специальных типов функций корреляции.

В публикациях [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53] применяют уравнения в конечных разностях как порождающую модель экспериментальной кривой *f(t)* на участках (T_{j-1}, T_j), $j = 1 \ldots r$. А именно, авторегрессионные уравнения вида: $f + a_0^j + a_1^j f_{t-1} + \ldots + a_{P_j}^j f_{t-P_j} = b_j V_t$, где $V_t, t = 1 \ldots N$ – белый шум, $a_i^j, i = 1 \ldots P_j$ и b_j, $j = 1 \ldots r$ – неизвестные коэффициенты. Для оценки качества сегментации применяют логарифмическую функцию правдоподобия, считая $P\left(\frac{f}{T}, z\right)$ - функцией условной плотности вероятности, где $z = \left\{P_j, z_j = \left(b_j, a_j = \left(a_0^j, \ldots, a_{p_j}^j\right)\right): j = 1 \ldots r\right\}$. В работе [35] функцию условной плотности вероятности рассматривают как $P\left(\frac{f_0, \ldots, f_{P_1}}{z_1}\right) \prod_{j=1}^{r} P\left(f_{T_{j-1}+1}, \ldots, f_{T_j} / f_{T_{j-1}}, \ldots, f_{T_j - P_j}; T_j, z_j\right)$, где $T_0 = P_1$. Значение плотности вероятности первых P_1 членов зачастую крайне мало, поэтому предлагается им пренебречь, а логарифмическая функция правдоподобия принимает вид:

$$J_2\left(T, \frac{z}{a}\right) = -\sum_{j=1}^{r} \left(T_j - T_{j-1}\right) \ln b_j - \sum_{j=1}^{r} \frac{1}{2b_j^2} \left(f_t + a_0^j + a_1^j f_{t-1} + \ldots + a_{P_j}^j f_{t-t_j}\right)^2. \qquad (28)$$

В [34] предлагается учитывать плотность вероятности первых P_1 членов в виде поправки, тогда прибавка к логарифмической функции правдоподобия принимает вид:

$$\Delta J\left(\frac{z_1}{f}\right) = -\frac{1}{2} |K_1(z_1)| - \frac{1}{2} \sum_{i,j=1}^{P} \chi_{ij}(z_1) f_i f_j, \qquad (29)$$

где $|K_1(z_1)|$ - определитель корреляционной матрицы, а $\chi_{ij}(z_1)$ - элементы обратной матрицы $K^{-1}(z_1)$. Так, параметр является суммой (28) и (29) и имеет вид:

$$J_3 = J_2 + \Delta J. \qquad (30)$$

В публикациях [34, 35, 36] считают, что в анализируемом случайном процессе имеется одна разладка, а максимальное значение параметра определяется методом полного перебора по всем значениям T_1. В [34] рассмотрен вариант определения порядков P_1, P_2, в котором искомыми принимаются первые значения P_1, P_2, выполняющих условия:

$$\frac{(T_2 - T_0 - P_j)\left\{\hat{b}_j^2(T_1, P_j) - \hat{b}_j^2(T_1, P_{j+1})\right\}}{\hat{b}_j^2(T_1, P_{j+1})} \le F_j(1, T_2 - T_0 - P_j),\ j = 1,2, \qquad (31)$$

где $\hat{b}_j^2(T_1, P_j)$ - оценка дисперсии b_j^2 на участке (T_{j-1}, T_j) при порядке уравнения авторегрессии равном P_j, $F_j(1, T_2 - T_0 - P_j)$ – табличное значение распределения Фишера для заданного уровня значимости γ. В [35], для процесса, описываемого авторегрессионным уравнением первого порядка, оценки параметров определяют из необходимых условий экстремума (30). В [36] для уменьшения временных затрат на расчёты, по сравнению с (31), параметр вычисляют с некоторым шагом *h*. Для этого, после определения приближенного положения $\tilde{\tilde{T}}_1$ максимума параметра, оценивают по реализациям $f^1 = \left(f_1, \ldots, f_{\tilde{\tilde{T}}_1 - h + 1} \right)$ и $f^2 = \left(f_{\tilde{\tilde{T}}_1 + h}, \ldots, f_{T_2} \right)$ параметры z_1, z_2. Определенные оценки применяют для более точной оценки $\hat{T}_1$ с помощью расчёта параметра в точках $T_1 = \tilde{\tilde{T}}_1 - h + 1, \ldots, \tilde{\tilde{T}}_1 + h - 1$.

В публикациях [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47] в анализируемом случайном процессе предполагается существование $(r-1)(r \ge 3)$ разладок. Для оценки качества сегментации применяется логарифмическая функция правдоподобия без поправки (29). Опять таки, метод полного перебора для оптимизации такого параметра в случае $(r \ge 3)$ требует больших временных затрат на вычисления. Для сокращения вычислительных затрат в [37, 38] применяют рекуррентные выражения, имеющие вид:

$$\begin{gathered} g(1,1) = d(1,1)\ ;\ g(1, j) = g(1, j-1) + d(1, j)\ ,\ j = T_0 \ldots N\ ; \\ g(i, j) = d(1, j) + \max\{g(i-1, j-1), g(i, j-1)\}\ ,\ j = T_0 \ldots N\ ,\ i = 2 \ldots r_j\ ; \end{gathered} \qquad (32)$$

$$\max_{T} J_2\left(\frac{T}{f}\right) = g(r, N),.$$

где $d(i, j) = -\ln b_j - \left[\frac{1}{2b_j^2}\left(f_j + a_0^i + a_1^i f_{j-1} + \ldots + a_{P_i}^i f_{j-P_j}\right)^2\right]$. В публикациях [38, 39] решается задача определения количества разладок случайного процесса *F(t)* методом полного перебора приближенных значений параметров для всех $r = 2 \ldots r_{\max}$. Параметром в [38] выбирается логарифмическая функция правдоподобия (28), а в [39] некоторый информационный параметр AIC [47]:

$$AIC(r) = -2J_2\left(\frac{\tilde{\tilde{T}}}{\tilde{\tilde{z}}}\right) + 2\sum_{j=1}^{r}\left(P_j + 2\right), \qquad (33)$$

который помимо величины (28) зависит от числа параметров применяемой модели. Помимо вышеописанного в [39] рассматривается случай с неизвестными параметрами z_j, $j = 1 \ldots r$, для которого применяется метод сокращенного перебора. Метод сокращенного перебора аналогичен тому, который рассмотрен в [36], а для уточнения приближенных оценок $\tilde{\tilde{T}}_1, \ldots, \tilde{\tilde{T}}_{r-1}$ применяются рекуррентные выражения (32).

Задача нахождения моментов изменения свойств многомерного коррелированного случайного процесса $F(t) = (F_1(t), \ldots, F_s(t))$ по его реализации $f_t = \left(f_t^1, \ldots, f_t^s\right), t = 1 \ldots N$ решается в работах [40, 41, 42, 43, 44, 45, 46]. Порождающей моделью случайного процесса на участке $(T_{j-1}, T_j), j = 1 \ldots r$ выбрано авторегрессионное уравнение [40]:

$$f_t = -A_1^j f_{t-1} - A_{P_j}^j f_{t-P_j} + B_j V_t, \ j = 1 \ldots r. \qquad (34)$$

Параметром выбирается логарифмическая функция правдоподобия (28) без поправки.

Случай с известным вектором параметров $z = \left\{A_k^j, B_j : k = 1 \ldots P_j : j = 1 \cdots r\right\}$ рассмотрен в [42, 43], для которого максимизация параметра осуществляется методом, аналогичным [38]. В [44, 45, 46] значение вектора параметров *z* предварительно не задается, а задача решается в два этапа:

- определение приближённых оценок $\tilde{\tilde{z}}$ и $\tilde{\tilde{T}}_1,\ldots,\tilde{\tilde{T}}_{r-1}$;

- уточнение полученных оценок максимизацией параметра.

В [44] так же осуществляется определение приближенных оценок методом, рассмотренным в [39]. Для оценки качества нахождения разладок анализируемого случайного процесса в [39, 42, 44] используется параметр суммарной квадратичной ошибки прогноза, для минимизации которого применяется оптимизация функции правдоподобия. В [45, 46] приближенные оценки определяются с помощью более быстрого эвристического алгоритма, состоящего из следующих шагов:

- реализация *f(t)* делится на заданное число *M* одинаковых участков;

- соединение пар смежных участков, пока их количество не уменьшится до 1. Выбирается такая пара смежных участков, при соединении которых логарифмическая функция правдоподобия уменьшается минимально;

- с каждым соединением смежных участков вычисляется величина информационного параметра (33);

- выбирается шаг, на котором информационный параметр (33) минимальный, и приближенные оценки $\tilde{\tilde{T}}_1,\ldots,\tilde{\tilde{T}}_{r-1}$, соответствующие ему.

Недостатком алгоритмов, рассмотренных в [44, 45, 46], является то, что оценки разладок случайного процесса *F(t)* не могут доставить хотя-бы локальный минимум выбранному параметру при выполнении необходимых условий.

Порождающей моделью экспериментальной кривой *f(t)* на участке $(T_{j-1},T_j),\ j=1\ldots r$ в [52, 53] выбрано, как и в [40, 41, 42, 43, 44, 45, 46] авторегрессионное уравнение (34), но его параметры могут принимать только дискретное конечное множество неизвестных значений $z_k, k=1\ldots m$. Изменение параметров $z_k, k=1\ldots m$ регулирует Марковская цепь с m состояниями $h_t=1\ldots m$, заданная матрицей вероятностей переходов $Q=\{q(h_{t-1},h_t)\}$. Параметром оценки выбирается модифицированная функция апостериорной вероятности:

$$J(T)=\sum_{t=1}^{N}\left\{\frac{1}{2b_{h_t}}\left[f_t-a_0^{h_t}-a_1^{h_t}f_{t-1}-\ldots-a_p^{h_t}f_{t-p}\right]^2-\ln q(h_{t-1},h_t)\right\}. \qquad (35)$$

В [52] значения параметров $Z=\{z_k, k=1\ldots m; Q\}$ неизвестны, а для минимизации (35) используется алгоритм, сходящийся за конечное количество шагов:

- задается начальное приближение параметров Z_0;

- с помощью Z_0 находится оптимальная сегментация T^*;

- для полученной сегментации вычисляются оценки Z_1, с помощью которых находится оптимальная сегментация T^1;

- шаг 3 повторяется до вычисления Z_m и T^m соответственно.

В [53] рассмотрен случай, когда векторы параметров $z_k, k=1\ldots m$ известны, а для вычисления минимума параметра (35) применяется метод динамического программирования, рекуррентные выражения которого аналогичны (32).

Решение данной задачи **непараметрическими** методами рассмотрено в [54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64]. Для обнаружения разладок случайных процессов *F(t)* данная группа методов, в отличие от вышеописанных параметрических методов, не требует задания функций распределения случайных величин $F(t), t=1\ldots N$ до и после момента разладки. Необходимо, чтобы данные функции до момента разладки и после были отличимы, а для оценки качества обнаружения применяются двухвыборочные статистики. Например, в [55] применяют статистики Чернова-Сэвиджа, в [58] – параметр Манна-Уитни,в [63] – статистику Колмогорова-Смирнова. Отыскиваемым моментом разладки является момент времени T_1^*, доставляющий экстремум выбранному параметру. Для полученной таким способом оценки $\lim\limits_{N\to\infty} P\left\{\left|T_1^* - T_1\right| > \varepsilon\right\} = 0$, где T_1 – истинный момент разладки, т.е. оценка T_1^* состоятельна.

Рассмотренные далее методы относятся к группе методов **последовательной** сегментации. В данных методах [86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98] на момент исследования анализируется только определенная часть экспериментальной кривой, а для сегментации применяют параметр оценки качества обнаружения разладки как функционал. Далее поставим задачу обнаружения разладок. Пусть $f_t, t=1,2,\ldots$ является реализацией некоторого

случайного процесса *F(t)*, для которого происходит резкое изменение характера поведения в момент времени $t = t_p$. Необходимо:

- найти момент изменения резкого изменения характера поведения (разладки) t_p, избежав обработку ложных экспериментальных данных.

Решение данной задачи **параметрическими** методами рассмотрено в работах [86, 87, 88, 89, 90]. Метод кумулятивных сумм, предложенный Пейджем [86] и являющийся одним из наиболее встречающихся, предполагает, что случайный процесс *F(t)*, $t = 1,2,\ldots$ является последовательностью случайных величин. Функцией плотности вероятности данной последовательности известна и представляет собой $R\left(\frac{f_t}{\theta}\right)$. Распределение у данной последовательности до момента времени разладки представляет собой $\theta = \theta_1$, а после него T_1 - $\theta = \theta_1$. Вектора параметров θ_1, θ_2 так же известны. Логарифмическая функция правдоподобия определяется как:

$$g_t = \ln\left[\frac{R\left(\frac{f_t}{\theta_1}\right)}{R\left(\frac{f_t}{\theta_2}\right)}\right]. \qquad (36)$$

Математическое ожидание логарифмической функции правдоподобия (36) до момента разладки $t < t_p$ будет отрицательной величиной, а после момента разладки $t > t_p$ – положительной. Решающая функция определяется как:

$$G_0 = 0,$$

$$G_t = \begin{cases} G_{t-1} + g_t, \text{ при } G_{t-1} + g_t > 0 \\ 0, \text{ при } G_{t-1} + g_t > 0 \end{cases}. \qquad (37)$$

Значение решающей функции (37) будет крайне мало до момента разладки t_p, а после него будет расти. Разладка считается определенной, когда значение решающей функции (37) на моменте времени t_0 превышает допустимое заранее заданное значение. Недостатком вышеописанного метода является то, что его

практическое применение ограничено необходимостью наличия определенных значений параметров θ_1, θ_2 анализируемых экспериментальных кривых.

В [89] предполагается, что известно значение вектора параметров θ_1 до момента разладки t_p и известно направление изменения $d = (d_1, \ldots, d_2), |d| = 1$ данного вектора θ_1 в момент разладки. Таким образом, в качестве величины функции правдоподобия (36) применяется:

$$g_t = \sum_{j=1}^{r} d_j \frac{\partial L(t)}{\partial \theta_{1j}}. \qquad (38)$$

В (38) $\frac{\partial L(t)}{\partial \theta_{1j}}$ – частная производная приращения функции правдоподобия по параметру θ_{1j} на каждом *t*-м шаге, $\theta_1 = (\theta_{11}, \ldots, \theta_{1r})$.

В [90] значение вектора параметров θ после момента разладки t_p не известны. Предлагается метод, аналогичный методе кумулятивных сумм [86], но ограниченный при практическом применении, в связи с тем, что для некоторых значений вектора параметров θ не удается обнаружить момент разладки.

Решение задачи обнаружения момента разладки **непараметрическими** методами рассмотрено в работах [91, 92, 93, 94, 95, 96, 97, 98]. Разладка считается определенной, когда значение специально построенной статистики превышает допустимое заранее заданное значение. В работе [92] применяется статистика:

$$T_t = \frac{\left(f_t - \hat{f}_t\right)^2}{\hat{G}^2}, \qquad (39)$$

где $\hat{f} = \sum_{i=1}^{n} \hat{a}_i f_{t-i}$, а $\hat{G}^2$, $\hat{a}_i$, $i = 1, \ldots, n$ – известные оценки параметров авторегрессионного уравнения, определенные по первичному отрезку экспериментальной кривой. В обычном состоянии величина (39) имеет распределение χ^2 с одной степенью свободы. При превышении заданного порогового значения $T_t > \chi^2_\gamma(1)$ в момент времени $t = t_0$ разладка считается обнаруженной. Данный алгоритм крайне чувствителен к кратковременным выбросам, что приводит к ложным срабатываниям и является его недостатком.

В [93] уменьшение количества ложных срабатываний достигается методом прогнозирования на *m* отсчётов. Для этого применяется статистика:

$$T_t(m) = \sum_{V=t-m}^{t} \frac{\left(f_v - \hat{f}_v\right)^2}{\hat{G}^2} . \qquad (40)$$

Разладка обнаруживается, когда в момент времени $t = t_0$ величина (40) превышает заданное пороговое значение $\chi^2_\gamma(m)$.

В работе [91] в качестве статистики применяется:

$$\lambda_t = V_t^T F_t^{-1} V_t , \qquad (41)$$

где $V_t^T = (\nu_{t1}, \ldots, \nu_{tr})$ – *r*-мерный вектор частных производных $\nu_{tj} = \dfrac{\partial \ln R\left(\dfrac{f_{1,t}}{\theta_1}\right)}{\partial \theta_{1j}}, j = 1, \ldots, r$, $R\left(\dfrac{f_{1,t}}{\theta_1}\right)$ – совместная плотность распределения до момента времени *t*, $F_t = \left\| M\{\nu_{tp}\nu_{ts}\} \right\|, p, s = 1, \ldots, r$ – информационная матрица Фишера по первым *t* точкам выборки. Статистика (41) асимптотически имеет нецентральное $\chi^2(r)$ распределение с параметром нецентральности, для которого задаются два значения, характеризующие налаженное и разлаженное состояние исследуемого случайного процесса. В публикациях [96, 97] для обнаружения разладок используется информационный критерий AIC (33), а в [98] – статистика Фишера.

1.2.3 Методы сравнения с эталонами

Описанные далее методы относятся к группе методов **параллельной** сегментации. Данные методы [69, 70, 71, 72, 73, 74, 75, 76] рассмотрим на примере сегментации экспериментальных кривых акустических сигналов. Перед сегментацией таких кривых, например, осциллограмм человеческой речи, необходимо произвести построение эталонных объектов, из которых состоит речь. Применение эталонных объектов приводит задачу анализа кривой акустического сигнала к задаче построения наиболее близкого к данной кривой эталона, определяющего отыскиваемую сегментацию. Пусть $f(t) = (f_1, \ldots, f_N)$ исследуемая экспериментальная кривая акустического сигнала. Необходимо:

- разделить область определения экспериментальной кривой f на h непересекающихся участков одинаковой длины;

- определить вектор значений признаков (например, авторегрессионные коэффициенты, спектральные отсчёты), которые описывают характер поведения кривой на данных участках. В результате этого кривая f описывается набором векторов $F_n = (b_1, \ldots, b_n)$;

- определить эталонный набор, наиболее близкий к F_n, из заданного множества эталонных наборов A, где $A = \{A_m^l = (a_1^l, \ldots, a_{m_l}^l), l = 1 \ldots M\}$;

- для определения «близости» двух наборов векторов используется параметр J, оптимизация которого приводит к нахождению отыскиваемой сегментации анализируемой экспериментальной кривой;

Для анализа отдельных акустических элементов, например слов, в [76] параметр представляет собой:

$$J_1(i(x), j(x); F_n, A_m) = \frac{1}{\sum_{k=1}^{K} w(k)} \sum_{k=1}^{K} d(i(k), j(k)), \tag{42}$$

где k – вспомогательная ось, $w(k)$ – заданная весовая функция, $i(k), j(k)$ – заданные неубывающие функции, задающие числу k целое число, не превышающие n и m соответственно, $d(i(k), j(k))$ – евклидово расстояние между векторами $a_{j(k)}, b_{i(k)}$. Минимизация параметра (42) осуществляется с помощью метода динамического программирования, рекуррентные выражения которого имеют вид:

$$\begin{gathered} g_1(i(1), j(1)) = d(i(1), j(1)) w(1), \\ g_k = \min_{\{[i(k-1), j(k-1)]\}} \{g_{k-1}(i(k-1), j(k-1)) + d(i(k), j(k) w(k))\}, k = 2,3, \ldots, K, \end{gathered} \tag{43}$$

где $\{[i(k-1), j(k-1)]\}$ – множество пар $[i, j]$, для которых $i(k) - i(k-1)$ и $j(k) - j(k-1)$ удовлетворяют заданным ограничениям.

Для анализа нескольких акустических элементов, например фраз, набор эталонов строится с помощью произвольной конкатенации эталонных наборов отдельных элементов из заданного словаря эталонов. А именно, $A_m = (A^{q(1)} \oplus A^{q(2)} \oplus \ldots \oplus A^{q(x)} \oplus A^{q(r)})$, где $A^q = (a_1^q, \ldots, a_{m_q}^q)$ - эталонный набор q-го элемента,

r – неизвестное количество элементов, $q(x)\in[1,L]$, L – число элементов в словаре, $m=\sum_{x=1}^{L} m_{q(x)}$. Для определения близости применяется выражение:

$$J_2(r,q(x))=\min_{i,j} J_1(i(x),\, j(x);F_n,A_m), \tag{44}$$

где J_1 параметр оценки близости отдельных акустических элементов (42). Для минимизации параметра (44) в [70] применяется метод динамического программирование, рекуррентные выражения которого аналогичны (43).

В публикации [72] параметр (44) представлен как:

$$J_2(r,T)=\sum_{j=1}^{r}\min_{q\in[1,L]} D\left[F(T_{j-1},T_j),A^q\right], \tag{45}$$

где $D\left[F(T_{j-1},T_j),A^q\right]=\min_{i,j} J_1(i(x),\, j(x);F(T_{j-1},T_j),A^q)$, $F(T_{j-1},T_j)=(b_{T_{j+1}+2},\ldots,b_{T_j})$, $j=1\ldots r$, $T_0=0$, $T_r=n$. Для минимизации параметра (45) используются рекуррентные выражения, аналогичные (2):

$$f_0(l)=\min_{q\in[1,L]} D\left[F(T_0,l),A^q\right], l=1\ldots n.$$

Для $j=1,2,3,\ldots$:

$$f_j(l)=\min_{1\le T_j\le l}\left\{f_{j-1}(T_j)+\min_{q\in[1,L]} D\left[F(T_j,l),A^q\right]\right\}, l=1\ldots n; \tag{46}$$

$$\min_T J_2(r,T)=f_{r-1}(n).$$

В (46) для определения числа элементов используется метод полного перебора $f_{r-1}(n)$ для всех $r=1,2,\ldots,r_{\max}$.

Описанные далее методы относятся к группе методов **последовательной** сегментации. Перед сегментацией экспериментальных кривых такими методами [99, 100, 101, 102] необходимо произвести построение набора эталонных форм, из которых состоят кривые. Применение эталонных наборов приводит задачу анализа экспериментальной кривой к задаче построения наиболее близкого к данной кривой эталона, определяющего отыскиваемую сегментацию. Пусть $f(t)=(f_1,\ldots,f_N)$ исследуемая экспериментальная кривая. Необходимо:

- Задать признаки для описания участков экспериментальной кривой.

- Построить набор эталонных форм кривой.

- Задать плавающее окно, и способ его перемещения по всей экспериментальной кривой.

- Задать способ определения принадлежности анализируемого участка кривой к одному из эталонов при каждом положении такого окна

В [101] описывается метод сравнения для решения задачи управления протезом по ЭМГ-сигналам. А именно, протез выполняет известное количество функций *h*, каждой из которых соответствует некоторая форма ЭМГ-кривой. Порождающей моделью ЭМГ-сигнала выбирается авторегресионное уравнение *n* порядка $f(t)=\sum_{i=1}^{n} a_i f_{t-i} + \varepsilon_t$. Для построения набора эталонов задается множество кривых ЭМГ, полученная для *k*-го класса:

$$M_k = \left\{ f_i{}^j, i = 1,\ldots,N,\ j = 1,\ldots,L \right\}, \tag{47}$$

где *N* – число отсчётов кривой, *L* – число кривых. Множество (47) делится на два подмножества. Для экспериментальных кривых первого подмножества рассчитываются средние значения оценок коэффициентов авторегрессионного уравнения $\overline{a}_1^k,\ldots,\overline{a}_n^k$. Для экспериментальных кривых второго подмножества рассчитывается остаточная дисперсия аппроксимации данных кривых:

$$E_j^k = \frac{\left\{ \sum_{l=n+1}^{N} \left(f_l{}^j - \overline{a}_1^k f_{l-1}^j - \ldots - \overline{a}_n^k f_{l-n}^j \right) \right\}}{(N-n)},\ j = L/2 + 1, L. \tag{48}$$

Эталоном *k*-го класса является среднее значение (48), а именно $\overline{E}_k = \frac{\left[2 \sum_{j=L/2+1}^{L} E_j^k \right]}{L}$.

Участок экспериментальной кривой принадлежит к *k* классу эталонов, при условии:

$$E_k^h \le p\overline{E}^k \wedge \sum_{l=1}^{N} f_k^2 \ge A. \tag{49}$$

В (49) E_k^h – остаточная дисперсия аппроксимации кривой, $p_i, i = 1,\ldots,h$ - заданный весовой коэффициент, *A* – заданный порог, определяющий наличие сигнала.

1.2.4 Методы выделения информативных участков экспериментальных кривых

Рассмотренный далее метод частичной аппроксимации относится к группе методов **параллельной** сегментации. Информативными или аномальными участками области определения исследуемой экспериментальной кривой являются участки (сложные), характер поведения данной кривой на которых оказывается резко другим по сравнению с характером ее поведения на остальных участках (простые) области определения. Далее данный метод основывается на предположении о том, что для экспериментальной кривой, анализируемой в области простых участков, определен закон характера ее поведения. Поставим задачу метода частичной аппроксимации. Задана исследуемая экспериментальная кривая $f(t)=(f_1,\ldots,f_N)$. Необходимо:

- найти подмножество w^* сложных участков и значения параметров описания экспериментальной кривой на простых участках, для которых выбранный параметр оценки качества аппроксимации принимает экстремальное значение.

В работе [65] для анализа детерминированных процессов применяется квадратичная невязка как оценка качества аппроксимации:

$$J(w)=\sum_{t\in\frac{T}{w}} S_t^2(w),\ S_t(w)=f_t-\sum_{i=1}^{m}\hat{c}_i(w)\varphi_t^i, \tag{50}$$

где $T=(1,2,\ldots,N)$ – область определения исследуемой кривой *f*, *w* – подмножество сложных участков с неизвестной длиной *d*, $\hat{c}_i(w)$ - коэффициенты многочлена наилучшего приближения кривой f по системе N-мерных векторов $\{\varphi_t^i,\ i=1\ldots m;\ t=1\ldots N\}$ на множестве отсчетов $\frac{T}{w}$, $S_t(w)$ – невязка приближения.

В работах [65, 66] отыскиваемое подмножество w^* определяют из необходимых условий минимума параметра (50):

$$\min_{t\in w} S_t^2(w)\geq\max_{t\in\frac{T}{w}} s_t^2(w). \tag{51}$$

Для построения подмножества w^* используется алгоритм, сходящийся за конечное количество шагов:

- задается начальное произвольное подмножество w^0;

- определяется подмножество w^k;

- в качестве элементов подмножества w^{k+1} выбираются точки количеством d с максимальными значениями квадратичной невязки $S_t(w^k)$;

- шаг 3 повторяется до выполнения условия (51) на шаге k^*, а именно $w^{k^*} = w^{k^*-1}$.

Некоторые сложности при выборе мощности множества простых участков является существенным недостатком метода частичной аппроксимации при нахождении информативных участков.

Рассмотренные далее методы обнаружения информативных участков [103, 104, 105, 106, 107] относятся к группе методов **последовательной** сегментации. Данные методы не требуют предварительного задания информации об исследуемой экспериментальной кривой. Информативными или аномальными участками области определения исследуемой экспериментальной кривой являются участки (сложные), характер поведения данной кривой на которых оказывается резко другим по сравнению с характером ее поведения на остальных участках (простые) области определения. Далее данный метод основывается на предположении о том, что анализируемая кривая является последовательностью чередующихся участков со сложным и простым характером ее поведения. Поставим задачу метода обнаружения информативных участков. Задана исследуемая экспериментальная кривая $f(t) = (f_1, \ldots, f_N)$. Необходимо:

- задать функцию оценки сложности $S(t_i, t_{i+1})$ на участке $[t_i, t_{i+1}]$.

- анализируемую кривую разделить на ω_j, $j = 1, \ldots, M$ участков одинаковой длины l, расположенных с некоторым шагом δ.

- рассчитать на каждом участке функцию оценки сложности $S(\omega_j)$, $j = 1, \ldots, M$.

- обнаружить участки с экстремальными значениями $S(\omega_j)$, которые и будут считаться сложными.

В [106, 107] сложность участка оценивается качеством аппроксимации данного участка по некоторой выбранной системе базисных функций $\varphi_i(t), i=1,\ldots,n$. В [103] для этого применяется n первых членов тригонометрического ряда. Плохая аппроксимация возникает из-за появления в анализируемом процессе элементов, не входящих в выбранную систему базисных функций. Задается вектор значений экспериментальной кривой $f^j=(f_1^j,\ldots,f_l^j)$ на j-м участке и система векторов $\varphi^i=(\varphi_1^i,\ldots,\varphi_l^i), i=1,\ldots,n$, а для оценки качества аппроксимации применяется функция сложности:

$$S(w_j)=\min_{c_j^i,i=1\ldots n} S'(w_j), \tag{52}$$

где $S'(w_j)=\sqrt{\sum_{k=1}^{l}\left(f_k^j-\sum_{i=1}^{n}c_i^j\varphi_k^i\right)^l}$. Значения коэффициентов $c_i^j, i=1,\ldots,n$, доставляющих экстремальные значения (52) определяются из системы линейных уравнений:

$$\sum_{i=1}^{n}c_i^j\sum_{k=1}^{l}\varphi_k^i\varphi_k^m=\sum_{k=1}^{l}f_k^j\varphi_k^m,\ m=1,\ldots,n\ . \tag{53}$$

При применении ортонормированной системы векторов, в качестве системы базисных функций, функция сложности (52) и коэффициенты (53) рассчитываются как $S(w_j)=\sum_{k=1}^{l}(f_k^j)^2-\sum_{i=1}^{n}(c_i^j)^2$ и $c_i^j=\sum_{k=1}^{l}f_k^j\varphi_k^i, i=1,\ldots,n$ соответственно.

1.3 Методы описания экспериментальных кривых

Среди существующих методов решения задачи описания экспериментальных кривых выделяют две крупных группы: дискриминантные и лингвистические. В **дискриминантном** методе предварительно выбираются параметры, которые несут наиболее интересные для исследователя данные о характере поведения экспериментальной кривой. Величины этих параметров рассчитываются по полученному сегментацией разделению кривой и являются её отыскиваемым описанием. Такой метод применялся в работах [67, 68] для оценки состояния рудовосстановительной печи, используя в качестве параметра коэффициент зашумленности огибающей её рабочего тока. Так же данный метод применялся в

работах о системах распознавания дикторов по голосам, в которых считалось, что корреляционные связи между параметрами являются данными, несущие основную информацию о процессе и характеризующие участки экспериментальной кривой [114, 115]. В **лингвистическом** методе анализируемая кривая описывается последовательным рядом символов либо целых слов из определенного алфавита. Каждый элемент такого ряда представляет соответствующий участок либо группу участков, определенный сегментацией анализируемой экспериментальной кривой. Методы данной группы применялись в работе о системах управления конверторным процессом [106], в работе об акустической диагностике двигателя [107], а также в работах о распознавании речи [116, 117].

Выбор метода для описания экспериментальных кривых осуществляется постановкой задачи и степенью знаний об анализируемом процессе. В ряде случаев при исследовании процессов, характер и закономерности которых мало изучены целесообразно использовать лингвистические методы, позволяющие анализировать большие массивы данных, формировать различные описания экспериментальных кривых и выбирать среди них наиболее подходящие.

1.3.1 Дискриминантные методы описания экспериментальных кривых

При применении дискриминантных методов описания экспериментальных кривых необходимо предварительное задание параметров, содержащих информацию о характере поведения кривой. Алгоритмы определения значений данных параметров достаточно просто реализуемы. В работах [118, 119] применяются непосредственно значения параметров, заданных для описания участков экспериментальной кривой. Значения данных параметров сгруппированы в матрицы, размерность $(m \times r)$ которой зависит от количества параметров m и числа участков r, а j-й столбец такой матрицы и будет являться вектором значений параметров, рассчитанных на j-м участке экспериментальной кривой. В [119] для описания анализируемой экспериментальной кривой применяется матрица

энергетических спектров, а в [118] не сам спектр, а частоты и амплитуды трёх его локальных пиков.

В публикации [120] экспериментальную кривую описывают не матрицей, а вектором значений параметров, рассчитанных на всех участках кривой. В [121, 122] для описания анализируемой кривой применяются функции изменения во времени наиболее важных параметров, характеризующих ее участки и выбранных из всего множества параметров. Вышеописанные методы имеют общий недостаток, которым является высокая размерность получаемого описания экспериментальной кривой. Высокая размерность требует больших временных затрат на вычисления и приводит к усложнению дальнейшей обработки сформированного описания. Данный недостаток в [123, 124] предлагается устранить с помощью преобразования формируемых матриц в двумерное изображение, которое анализируется в дальнейшем. В [125] используется вектор средних значений параметров, характеризующих поведение экспериментальной кривой на участках. Полученное описание анализируемой кривой данным методом является значительно укороченным.

В публикациях [114, 115] используют корреляционные связи параметров, как информацию, характеризующую поведение кривой на ее участках, а в [67, 68] степень зашумленности анализируемой кривой:

$$k_{ш} = \frac{T_{в}}{T}. \qquad (54)$$

В (54) $T_{в}$ – общая длина сложных участков, обнаруженных алгоритмом сегментации, а T - длина всей анализируемой кривой. В [103] для описания кривой применяется верхний треугольник матрицы частоты смен состояний анализируемого процесса $\left[\gamma_{ij}\right]_1^k$, в котором γ_{ij} – частота перехода объекта из состояния i в состояние j, а k – количество таких состояний.

1.3.2 Лингвистические методы описания экспериментальных кривых

При применении лингвистических методов описания экспериментальных кривых, полученная сегментация кривой представляется упорядоченным набором символов, характеризующих различные характеры поведения анализируемой кривой. Каждый элемент такого набора является некоторой последовательностью символов, применяемых для распределения участков кривой различного характера поведения по классам. Для построения такого набора необходимо найти разделение множества выделенных участков на классы схожих и присвоить названия полученным классам участков символами из некоторого, предварительно заданного алфавита. Построенный таким образом набор символов будет являться искомым описанием анализируемой кривой. В публикации [126] классификация участков осуществляется по предварительно заданным правилам, являющимися логическими комбинациями параметров, характеризующих поведение кривой на участках.

Различная классификация участков осуществляется в зависимости от заданного способа представления участков анализируемой кривой. При применении для классификации отсчётов участков кривой возникают сложности, такие как несовпадающая размерность и большое количество выделенных элементов. В [66, 127] для распределения участков на классы применяются алгоритмы автоматической классификации, а вышеописанные сложности устраняются при помощи построения евклидова пространства признаков небольшой размерности.

Метод укрупнённого описания экспериментальных кривых так же относится к лингвистическим методам. Данным методом в [117, 128] строится грамматика, задающая множество последовательностей символов, которые могут быть получены при анализе исследуемых кривых. Грамматика представляет собой последовательность конечных множеств $\{S, V_N, V_T, P\}$, в которых S – первый символ. $V_N = \{S, A, \ldots, B\}$ - множество вспомогательных символов, $V_T = \{a, b, c, \ldots, d\}$ – множество основных символов, P – множество правил постановки. После

применения правил, образуется последовательность основных и дополнительных символов, которая и является укрупненным описанием анализируемой кривой. Данный метод неприменим в случае наличия искажений в последовательности символов, не учтённых грамматикой, что является его недостатком.

В [66] устраняется вышеописанный недостаток, и рассматривается метод укрупнённого описания экспериментальной кривой, способный анализировать любые последовательности символов, состоящий из следующих шагов:

- задается множество эталонных последовательностей символов (слов) $S = \left(S^1, \ldots, S^k\right)$ и эталонная последовательность $\hat{T}$ из всех возможных, получаемых произвольной конкатенацией эталонных последовательностей из *S*;

- определяется эталонная последовательность $\hat{T}^*$ наиболее близкая к анализируемой последовательности *T*;

- расстоянием между двумя произвольными последовательностями T^1 и T^2 выбирается минимальное число $m\left(T^1, T^2\right)$ трансформаций данных последовательностей, переводящих одну из них в другую;

- последовательность, из которой составлена последовательность $\hat{T}^*$, является отыскиваемым укрупнённым описанием анализируемой кривой.

Для случая с неизвестным эталонным множеством *S* применяется алгоритм, состоящий из следующих шагов:

- задается произвольное первичное множество эталонных последовательностей *S*;

- анализируемая последовательность *T* разделяется на такие участки, для которых расстояние $m\left(T, \hat{T}\right)$ минимально;

- участки, полученные разбиением, алгоритмами автоматической классификации распределяются на группы;

- составляется новое множество эталонных последовательностей S' из эталонных последовательностей, полученных на предыдущем шаге.

Алгоритм завершается, когда на каком-либо шаге появляется $S = S'$.

1.4 Применение методов анализа экспериментальных кривых

Наибольшее распространение вышеописанные методы получили в телеметрии, в медицине, в системах автоматизированного управления, в сейсмологии. В [129] представлено устройство сжатия телеметрических данных, работающее в реальном времени на борту космического корабля и способное обрабатывать информацию, поступающую одновременно по 480 каналам и обеспечивать коэффициент сжатия до 34. Данное устройство основано на алгоритме прогнозирования нулевого порядка с плавающей апертурой, имеет малые габариты и высокую эффективность. В [82] алгоритмы нелинейной и кусочной аппроксимации используются в системе «System Test and Astronaut Requirements Simulation» фирмы Lockheed. Система STARS предназначена для моделирования условий работы астронавта. В нее входит подсистема сбора данных, основанная на устройстве сжатия данных, которое выполняет сжатие информации, поступающей по каналам. Для выполнения этого применяются алгоритмы экстраполяции и интерполяции нулевого порядка. Выбор канала и алгоритма осуществляется программным устройством, совмещенным с устройством сжатия данных.

В работе [79] представлена автоматизированная система ведения архива электрокардиосигналов. Работа системы основана на непрерывном алгоритме интерполяции первого порядка. В результате электрокардиосигнал описывается последовательностью пар чисел $\Delta f_j, \Delta t_j$, $j = 1,2,\ldots$, где Δt_j – длительность j-го интервала аппроксимации, Δf_j – приращение амплитуды сигнала на данном интервале. В результате обеспечивается коэффициент сжатия 6-7. Прежде чем определять пики электрокардиограмм, необходимо выделить их из ЭКС. В [130, 131] для решения данной задачи применяется структурный алгоритм распознавания. Множество допустимых ЭКГ – сигналов задается специально построенными синтаксическими правилами, использующие кусочно-линейную аппроксимацию, для построения которой применяется алгоритм интерполяции первого порядка. В [132] анализируются электрокардиограммы. Пики ЭКГ

определяются по ее кусочно-линейному представлению, получаемого построением непрерывного полиномиального сплайна. Также алгоритмы кусочной аппроксимации применяются для оценки шумовых характеристик цифровой системы передачи информации [133]. Данная система представляет собой линию передачи и усилитель с заданной частотной характеристикой. В данной работе кусочная аппроксимация применяется для вычисления дисперсии шума на выходе усилителя. Для дисперсии известно выражение, содержащее параметры линии связи, усилителя и шума. Использование данного выражения приводит к большим временным затратам из-за расчёта сложного интеграла, которые предлагается сократить, предварительно построив кусочно-линейную аппроксимацию зависимостей, используемых в выражении дисперсии. В [94] для обнаружения отклонений в кардиотахонометрических кривых новорождённых применяются *t*-статистики.

В публикациях [19, 134] для фильтрации эмпирических кривых применяют алгоритмы кусочной аппроксимации. В [19] кусочно-линейное представление анализируемой кривой считается отыскиваемым. В [134] для фильтрации помех сперва находят все их пики по кусочно-линейному представлению анализируемой кривой, после чего пики, длина и амплитуда которых, превышают допущенные значения, принимаются шумовыми и удаляются.

Методы обнаружения разладок также применяются при решении практических задач. Например, в публикациях [135, 136] данные методы применяются в автоматизированной системе управления цементным производством. В [135] решается задача своевременного обнаружения отклонения среднего химического состава сырьевых компонентов от заданных значений в системе управления смешиванием различных сырьевых компонентов. Данная задача решается с помощью модифицированного алгоритма кумулятивных сумм изменения среднего значения случайного процесса. В [136] с помощью алгоритмов обнаружения разладок контролируются датчики, работающие в двух режимах. Переход из нормального режима в режим отказа осуществляется изменением статистических свойств сигнала.

В [137] модифицированный алгоритм кумулятивных сумм применяется для автоматического прогнозирования цунами по сейсмологическим данным. Прогнозирование осуществлялось оценкой некоторых параметров землетрясения. Данными для анализа были выбраны 29 трёхкомпонентные записи Гавайских землетрясений. Каждой записи предшествовала запись помех длительностью 4-8 минут с шагом дискретизации 1 секунда. Данные записи применялись для получения оценок параметров многомерного авторегресионного уравнения. В результате анализа доля верных обнаружений оказалась равной 83%, что на 14% выше чем обнаружение опытным интерпретатором, на таком же количестве данных. В [138, 139] проведен сравнительный анализ и обзор методов параллельной сегментации экспериментальных кривых.

Применение алгоритмов выделения информативных участков экспериментальных кривых рассмотрено в [106, 107]. А именно, в [106] данные алгоритмы применяются при анализе связи скорости обезуглероживания конверторного процесса с типом выплавляемой марки стали. Выделенные сегментацией информативные участки классифицировались на 5 классов, для каждого из которых рассчитывались матрицы расстояний между ними. Проведенный анализ подтвердил наличие зависимости выплавляемой марки стали от формы кривой скорости обезуглероживания. В [107] данный подход применялся для акустической диагностики двигателя по записям его шумов. Анализ полученных сжатых представлений экспериментальных кривых позволил построить простые и надежные правила диагностики двигателя.

Для количественной оценки содержания углерода в ванне рудовосстановительной печи [67, 68] применялся метод сравнения с эталонами, а именно степень зашумленности анализируемой кривой. Оценка избытка либо недостатка углерода в ванне оценивалась по огибающей её рабочего тока. Результаты, полученные в работе, подтверждают возможность применения степени зашумленности для оценки дисбаланса углерода.

1.5 Постановка задач диссертационного исследования

На основании анализа степени разработанности темы и проведенного аналитического обзора достоинств и недостатков существующих методов анализа экспериментальных кривых, рассмотренных в первом разделе определена цель исследования – развитие теоретических аспектов структурных методов анализа данных и практических решений автоматизации процессов анализа массивов экспериментальных данных научных исследований, их классификации и представления в виде сжатого описания. Для достижения данной цели сформулированы и решены следующие задачи:

1. Разработка комбинированных вычислительных алгоритмов сегментации экспериментальных кривых.
2. Алгоритмическая реализация процедуры структурного анализа экспериментальных данных.
3. Разработка алгоритмов лингвистического описания экспериментальных кривых.
4. Исследование применимости разработанной системы структурного анализа в процессах электромагнитной совместимости радиоэлектронных средств.

1.6 Выводы по разделу

В первом разделе проведен сравнительный анализ методов анализа экспериментальных кривых. Показано, что применение интегральных методов не целесообразно при анализе малоизученных процессов. Проанализированы существующие методы сегментации, из которых выделены две группы параллельные и последовательные. При сегментации параллельными методами экспериментальные кривые обрабатываются полностью, и вводится некоторый параметр оценки качества сегментации. При сегментации последовательными методами на данный момент анализируется только определенная часть экспериментальной кривой, на которой в дальнейшем выделяются границы соответствующих участков. Обе группы методов сегментации имеют как

преимущества, так и недостатки, учитывая которые определено, что оптимальным решением для этапа сегментации экспериментальных кривых является комбинирование последовательных и параллельных методов. Проанализированы две крупные группы методов описания экспериментальных кривых, а именно: дискриминантные и лингвистические. Лингвистические методы являются более применимыми для анализа малоизученных процессов. Таким образом, для анализа экспериментальных кривых целесообразно использовать структурный метод с использованием комбинированных методов сегментации и лингвистического метода описания.

РАЗДЕЛ 2 ЭТАП ВЫДЕЛЕНИЯ И РАСПОЗНАВАНИЯ ХАРАКТЕРНЫХ УЧАСТКОВ ЭКСПЕРИМЕНТАЛЬНЫХ КРИВЫХ

В лингвистическом анализе экспериментальных кривых выделяется последовательность реализации трех основных этапов обработки кривой: выделения и распознавания характерных участков, присвоения выделенным участкам символов некоторого алфавита, анализа полученных последовательностей символов. Из этих трех основных этапов наиболее специфичен этап выделения и распознавания участков или сегментация. Именно этому этапу посвящен данный раздел. В 2.1 представлена общая методика сегментации на основе функции сложности, в 2.2 представлен ряд комбинированных алгоритмов сегментации на основе функции сложности, в 2.3 экстраполяционные алгоритмы.

В данном разделе решена первая и частично вторая задачи диссертационного исследования, поставленных в 1.5.

2.1 Общая методика сегментации на основе функций сложности

Сегментация экспериментальных кривых или выделение участков, которые содержат информацию на кривой, является первым этапом в создании языка для описания массива кривых. Выделяется два типа кривых, которые требуют лингвистического подхода к их анализу.

К первому типу относятся кривые, описывающие несколько процессов, данные о которых содержатся на участках кривой. Стыки участков кривых интерпретируются как изменение процесса. Целью сегментации является нахождение точек, которые и представляют эти изменения. Для этого следует разделять кривую на ряд смежных участков, отличающихся формой кривой.

Для кривых второго типа предполагается, что исследуемый процесс находится, в основном, в неизменном состоянии, из которого он иногда выходит в результате кратковременных возмущений. Такие возмущения, на фоне в целом равномерного процесса, рассматриваются как участки кривой, содержащие

единицы данных. Для таких кривых сегментацией следует выделить только отдельные участки, которые считаются информативными, а остальные (неизменные) не учитываются.

Кривые обоих типов являются дискретно упорядоченными последовательностями событий, а их анализ является описанием этих событий. Таким образом, сегментация, которая делит кривую на участки - не только первый, но и самый важный этап лингвистического анализа.

В методике сегментации, рассматриваемой далее, переходы от одного события к другому в процессе, который представляет структурную форму сигнала, рассматриваются как быстрые изменения его формы. Благодаря этому, приведенное описание таких экспериментальных кривых сохраняет данные о структуре и может быть получено путем выделения участков, в которых форма существенно отличается от граничащих участков и рассматривается как пики, рывки, резкие изменения и т.д. Эти участки называются переходными или сложными, в отличие от однородных простых, у которых форма не изменяется и которые разделены сложными участками.

Введение понятия «сложности» участков при сегментации кривых двух указанных выше типов не требует предварительной информации о том, к какому типу относится кривая. Таким образом, сложные участки могут быть интерпретированы как изменения состояния процесса, как фоновые возмущения некоторого постоянного состояния, или, наконец, как кратковременные возмущения, которые переводят процесс из одного состояния в другое.

Пусть $\omega = (t_1, t_2)$ некоторый участок области определения экспериментальной кривой $f(t)$ и задана реальная функция $\phi(f, \omega)$, которая зависит от формы кривой на участке ω и является представлением о ее изменчивости в течение этого интервала. Таким образом, величина $\phi(f, \omega)$ постоянна, если кривая $f(t)$ однородна, и изменяется, если ее форма быстро изменяется в течение этого интервала. Значение этой функции может быть интерпретировано как степень изменчивости формы кривой (назовем ее, для определенности, функцией сложности) [140, 141].

Пусть в интервале ω^* находится точка соединения двух однородных участков экспериментальной кривой, которые отличаются по форме или отображают аномалию на однородном фоне, тогда, с небольшими сдвигами интервала ω фиксированной длины справа и слева от ω^*, значение функции сложности уменьшается или увеличивается в зависимости от ее конкретной формы. То есть функция $\phi(f,\omega)$ имеет локальный экстремум при $\omega=\omega^*$. Таким образом, в зависимости от выбора функции сложности, сложные участки могут быть локальными экстремумами или участками, которые являются "более сложными", чем граничащие с ними. Учитывая сказанное, методика сегментации для обработки экспериментальной кривой, должна содержать следующее:

- экспериментальная кривая $f(t)$ делится на ряд элементарных участков $\omega_j, j=1,\ldots,N$ одинаковой длины *l*, которые следуют с определенным шагом Δ вдоль оси изменения аргумента (Рисунок 1) (шаг Δ может быть равен или меньше, чем *l*);

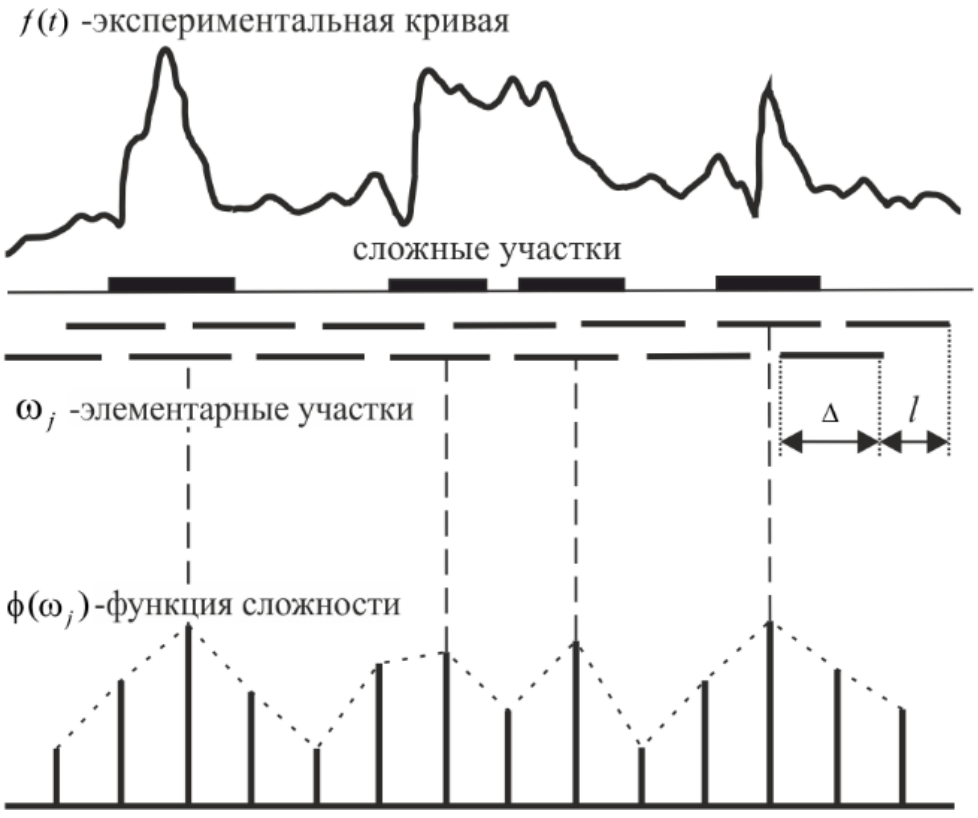

Рисунок 1 – Выделение сложных участков экспериментальной кривой

- выбирается определенная функция сложности $\phi(f,\omega)$, при этом каждый элементарный участок ω_j связан с реальной величиной в виде $\phi_j = \phi(\omega_j)$;

- выделяются сложные участки, т.е. участки с локально экстремальными значениями ϕ_j (максимальным или минимальным, в зависимости от выбранной функции сложности).

Практическое вычисление локального экстремума нуждается в дополнительном уточнении, поскольку условия $\phi_j > \phi_{j+1}$ (для случая определения максимумов функции сложности) не являются надежными и продуцируют слишком много ложных экстремумов. Чтобы этого избежать в данной работе предлагается использовать следующие дополнительные условия:

- $\phi_j > h$ – малые значения функции сложности интерпретируются как индексы, обозначающие, что состояние процесса «поддерживается» и что связанный с ним локальный экстремум является «незначительным»;

- для выявления огибающей функции сложности последовательно выбираются экстремумы $\phi_{ji}, \phi_{ji+1}, \ldots, \phi_{ji+k}$, для которых максимально $\{(j_{s+1} - j_s), i \le s \le i+k\} \le \tau$, и один локальный экстремум выбирается в точке j_i;

- используя $\phi_j > \phi_{j\pm1} > \ldots > \phi_{j\pm k}$ выбираются участки кривой, в которых процесс изменяет свое состояние плавно (используем отдельно от предыдущего).

2.1.1 Функции сложности для выявления участков кривой, наиболее отличающихся от граничащих

Пусть экспериментальная кривая будет определена как последовательность значений ее ординат $f_1, f_2, \ldots$ в виде дискретного ряда, и пусть каждый элементарный участок ω_j содержит l точек. Тогда, $f^j = (f_1^j, \ldots, f_l^j)$ будет вектором значения ординат экспериментальной кривой на участке ω_j. Введем для оценки формы кривой функцию сложности $\phi(\omega_j)$, оценивающую степень изменчивости поведения кривой на данном участке. Построим вектор k, характеристики $g^j = (g_1^j, \ldots, g_k^j)$ которого будут рассматриваться как описание формы кривой на

участке ω_j (k не обязательно равна l). Тогда функция сложности $\phi(\omega_j)$ будет простейшей функцией подобия между вектором g^j и граничащими векторами g^{j-1}, g^{j+1}, и представляет собой среднее значение произведения с граничащей величиной: $\phi(\omega_j)=\frac{1}{2}\left[\left(g^j,g^{j-1}\right)+\left(g^j,g^{j+1}\right)\right]$, где $\left(g^j,g^{j-1}\right)$ и $\left(g^j,g^{j+1}\right)$ - скалярные произведения центрированного и нормированного вектора отсчетов кривой на данном участке ω_j и соответствующего вектора на граничащих участках слева и справа от него. Обозначаем среднее значение произведения граничащих участков через величину $x^j=\frac{1}{2}\left(g^{j-1},g^{j+1}\right)$, которая характеризует степень постоянства формы кривой на данном участке. Тогда функция сложности для выделения участков, наиболее отличающихся от граничащих, принимает вид:

$$\phi(\omega_j)=(x^j,g^j)\,. \tag{55}$$

В частности, если характеристиками экспериментальной кривой являются значения ординат, то функция сложности принимает вид:

$$\phi(\omega_j)=\left(y^j,f^j\right). \tag{56}$$

где $y^j=\frac{1}{2}\left(f^{j-1}+f^{j+1}\right)$ соответственно. В этом случае необходимо сначала центрировать и нормализовать вектор f^j или выбрать в качестве его характеристик величины $g_s^j=\frac{f_s^j-\bar{f}^j}{\left\|f^j\right\|}$, где $s=1,\ldots,l$ является порядковым номером точки на данном участке, а $\bar{f}^j$ и $\left\|f^j\right\|$ равны: $\bar{f}^j=\frac{1}{l}\sum_{s=1}^{l}f_s^j$, $\left\|f^j\right\|=\sqrt{\sum_{s=1}^{l}\left(f_s^j-\bar{f}^j\right)^2}$.

Другими характеристиками, используемыми при описании участков кривой ω_j, могут быть статистические моменты, вычисленные из компонент вектора f^j:

$$g_1^j=\bar{f}^j,\quad g_2^j=\frac{1}{l}\sqrt{\sum_{s=1}^{l}\left(f_s^j-\bar{f}^j\right)^2},\ldots,\quad g_k^j=\frac{1}{l}\sqrt{\sum_{s=1}^{l}\left(f_s^j-\bar{f}^j\right)^k}\,.$$

Данный набор функций сложности (55, 56) применим в сегментации как детерминированных, так и случайных кривых, и позволяет сравнивать участки кривой различной длины. Вне зависимости от способов, используемых для определения вектора g^j, в качестве сложных участков будут выделяться участки с локально минимальными значениями $\phi(\omega_j)$. Величина $\phi(\omega_j)$ уменьшается с увеличением изменчивости формы кривой на участке ω_j и его усредненной формы на участках ω_{j-1} и ω_{j+1}. Таким образом, данная функция сложности выделяет участки, наиболее отличающиеся от граничащих.

2.1.2 Функции сложности для определения качества аппроксимации кривой

Пусть $f^j = (f_1^j, \ldots, f_l^j)$ вектор значений ординат экспериментальной кривой на элементарном участке ω_j. Задаем множество *l*-мерных ортонормированных векторов $\psi^i = (\psi_1^i, \ldots, \psi_l^i), \quad i = 1, \ldots, n$, зависящих от дискретных значений ординат некоторых элементарных функций. Для каждого элементарного участка ω_j, или для каждого вектора f^j, коэффициенты многочлена его наилучшей среднеквадратичной аппроксимации в системе векторов $\psi^i, i = 1, \ldots, n$ вычисляются как скалярные произведения: $c_{ij} = \sum_{s=1}^{l} f_s^j \psi_s^i$.

Функция сложности каждого элементарного участка ω_j равна норме остаточного рассогласования и характеризует качество аппроксимации кривой на этом участке:

$$\phi(\omega_j) = \sqrt{\sum_{s=1}^{l} \left(f_s^j - \sum_{i=1}^{n} c_{ij} \psi_s^i \right)^2} . \qquad (57)$$

Сложными участками будут являться те участки, на которых $\phi(\omega_j)$ будет иметь локально максимальное значение.

Если кривая изменяется случайным образом, то стоит предположить, что участки, где ее поведение изменяется не значительно, являются реализацией

определенного случайного стационарного процесса. Такие процессы должны быть смоделированы как совокупность стохастических уравнений авторегрессии:

$$y_t = c_0 + \sum_{i=1}^{n} c_i y_{t-i} + b\xi_t \ , \qquad (58)$$

где $c_i, i = 0,\ldots,n$ - коэффициенты авторегрессии; n является порядком авторегрессии; ξ_t представляет собой последовательность независимых нормальных случайных величин с нулевой средней и унитарной дисперсией, а коэффициент b определяет среднеквадратичное отклонение возмущающего белого шума $b\xi_t$.

Пусть экспериментальная кривая определяется на конечном интервале или на конечном множестве точек $T = \{1,\ldots,N\}$. Разделим это множество на M непересекающихся элементарных участков $\omega_j, j = 0,\ldots,M$, каждый из которых содержит l точек (в отличие от предыдущего случая, важно, что элементарные участки не перекрываются): $\bigcup_{j=1}^{M} \omega_j = T, \quad N = Ml$. Предположим также, что число точек l в каждом участке превышает порядок авторегрессии n. Тогда, на множестве элементарных участков ω_j, стохастическая функция сложности может быть представлена в виде:

$$\phi(\omega_j) = \min_c \sum_{s=n+1}^{l} \left(f_s^{\,j} - c_0 - \sum_{i=1}^{n} c_j f_{s-i}^{\,j} \right)^2 , \qquad (59)$$

где в отличии от (57) минимум берется по всевозможным $(n+1)$-мерным векторам $c = (c_0,\ldots,c_n)$. Значение этой функции сложности характеризует качество аппроксимации экспериментальной кривой по участкам с помощью авторегрессионной модели n-го порядка (58). Чем ниже это значение, тем вероятнее, что кривая на этом участке является реализацией случайного стационарного процесса. Поэтому сложными участками будут участки с локальными максимумами $\phi(\omega_j)$.

2.1.3 Функции сложности для определения экстраполяционных свойств кривой

Пусть экспериментальная кривая на элементарном участке ω_j аппроксимируется как отрезок степенного ряда $\sum_{i=0}^{n} c_i s^i$, где s является порядковым номером точки на данном участке. Тогда полученный полином распространяется вправо за границы участка (правая экстраполяция), а число точек (длина правого интервала) k_r находится таким образом, что ошибка экстраполяции не должна превышать заданное значение ε: $\sum_{s=l+1}^{k_r}\left(f_s - \sum_{i=0}^{n} c_i s^i \right)^2 \le \varepsilon$.

Длина левого интервала k_l определяется таким же образом. Значение функции сложности считается равным наименьшему из интервалов экстраполяции:

$$\phi(\omega_j) = \min(k_r, k_l). \qquad (60)$$

Сложные участки являются элементарными участками с локальными минимумами $\phi(\omega_j)$. Эта функция сложности формализует представление об изменчивости формы кривой, проверяя, сохраняет ли она свое основное направление в некоторой области текущего элементарного участка. Данная функция может быть использована и для анализа экстраполяционных свойств случайной кривой.

Пусть, как и в функции (59), участок ω_j описывается моделью с *n*-ым порядком авторегрессии (58) с коэффициентами (*c, b*). Параметры его вектора получим методом наименьших квадратов на участке ω_j. Определим для $t > jl$ величины:

$$s_t(\omega_j) = \sum_{s=l+1}^{k_r}\left[\frac{1}{b}\left(f_s - c_0 - \sum_{i=1}^{n} c_i f_{s-i} \right)^2 \right] - (t - jl).$$

В постоянной части значения $g_s(\omega_j) = \frac{1}{b}\left(f_s - c_0 - \sum_{i=1}^{n} c_i f_{s-i} \right)^2$ совпадают с квадратом стандартного белого шума ξ_s^2 и в среднем их значения близки к единице,

а значение суммы $s_t(\omega_j)$ приблизительно равно нулю. После изменения параметров (c, b), возникают систематические различия $g_s(\omega_j)-1$, что приводит к монотонному дрейфу $s_t(\omega_j)$ от нуля.

Введем пороговое значение ε и найдем минимальное значение $\phi(\omega_j)$, при котором величина $|s_t(\omega_j)|$ превышает пороговое значение $|s_t(\omega_j)| \geq \varepsilon$. В этом случае значение функции сложности будет равно:

$$\phi(\omega_j) = t_0(\omega_j), \tag{61}$$

а сложными участками будут являться участки с локальными минимумами $\phi(\omega_j)$. Данная функция (61) отличается от (60) тем, что она оценивает экстраполяционные свойства участка ω_j только в одном направлении.

2.2 Алгоритмы сегментации на основе функций сложности

Различные алгоритмы сегментации отличаются выбранными мерами оценки «сложности». Общая идея алгоритма [142], осуществляющего сегментацию кривой, основываясь на выявлении переходных участков:

- экспериментальная кривая $f(t)$ разбивается на ряд участков $\omega_j, j=1,\ldots,n$ одинаковой длины l, которые следуют с заданным шагом Δ;

- вводится функция сложности $\phi(f,\omega)$ (для оценки степени изменчивости и сложности характера поведения кривой на каждом участке);

- выделяются такие участки кривой, на которых функция сложности принимает локально экстремальные значения;

- выявленные участки принимаются за искомые переходные участки.

Два конкретных алгоритма такого типа с применением аппроксимации представлены ниже.

Алгоритм сегментации, основанный на определении качества аппроксимации участков кривой. Пусть экспериментальная кривая $f(t)$ задана набором своих ординат $f_1, f_2, \ldots$ в точках отсчета. Разобьем кривую на участки

$\omega_j, j = 1,\ldots,N$, каждый из которых представлен вектором значений ее ординат $f^j = (f_1^j,\ldots, f_l^j)$, взятых в пределах данного участка. Зададим ортонормированный базис из *l*-мерных векторов $\psi^i = (\psi_1^i,\ldots,\psi_l^i),\ i = 1,\ldots,n$, зависящих от дискретных значений ординат некоторых элементарных функций. В качестве базисной системы могут быть, например, тригонометрические функции. Размерность n этого вектора равна числу ординат кривой на участке. Введем в рассмотрение функцию [140]:

$$\phi(\omega_j) = \sqrt{\sum_{s=1}^{l}\left(f_s^j - \sum_{i=1}^{n} c_{ij}\psi_s^i \right)^2}, \qquad (62)$$

которая будет являться нормой остаточной невязки и характеризовать качество аппроксимации экспериментальной кривой на данном участке. Для каждого участка ω_j, представленного векторами $f^j = (f_1^j,\ldots, f_l^j)$ определяем набор коэффициентов многочлена его наилучшего среднеквадратичного приближения по системе векторов $\psi^i = (\psi_1^i,\ldots,\psi_l^i)$ как скалярные произведения $c_{ij} = \sum_{s=1}^{l} f_s^j \psi_s^i$. Участки с локально максимальными значениями будут идентифицироваться данным алгоритмом в качестве переходных участков [143]. Учитывая сказанное данный алгоритм состоит из следующих шагов:

- исследуемая экспериментальная кривая разбивается на участки, каждый из которых представлен вектором значений ее ординат.

- для каждого участка определяется набор коэффициентов многочлена его наилучшего среднеквадратичного приближения.

- для каждого участка определяется значение функции сложности (62).

- на кривой выявляются все участки, соответствующие локально максимальным значениям функции сложности.

- выделенные участки рассматриваются как искомые переходные участки.

- для каждого такого участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

Структурная схема вышеописанного алгоритма представлена на рисунке 2.

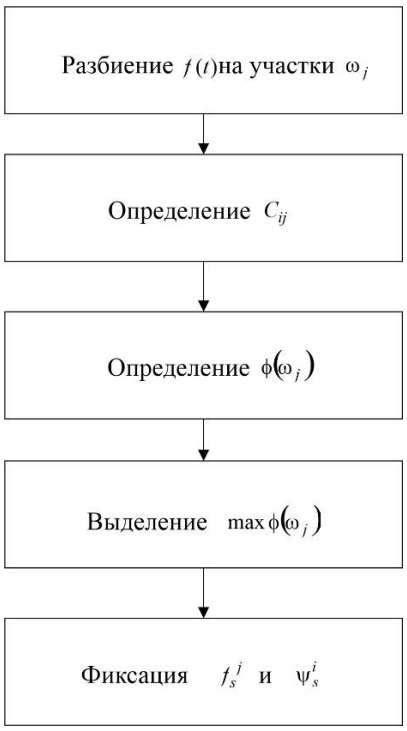

Рисунок 2 – Структурная схема алгоритма сегментации

В результате работы алгоритма полученный массив векторов в дальнейшем используется для разработки словаря описания выделенных участков.

Алгоритм сегментации, основанный на определении качества аппроксимации участков кривой авто-регрессионной моделью. Пусть экспериментальная кривая $f(t)$ изменяется случайным образом. Тогда такие участки, на которых характер ее поведения изменяется не значительно (простые, однородные), будут являться реализацией определенного случайного стационарного процесса. Такие процессы будем рассматривать как совокупность стохастических уравнений авторегрессии:

$$x_t = c_0 + \sum_{i=1}^{n} c_i x_{t-i} + b\xi_t \,, \qquad (63)$$

где c_i – коэффициент авторегрессии (i принимает значения от 0 до n); n – порядок авторегрессии; ξ_t – последовательность независимых нормальных случайных

величин с нулевым математическим ожиданием и единичной дисперсией; b – коэффициент среднеквадратичного отклонения белого шума.

И пусть экспериментальная кривая $f(t)$ имеет область определения на конечном множестве $T = \{1,\ldots,N\}$ и представлена последовательностью чисел - значений ее ординат $f_1, f_2,\cdots$ в точках отсчета, расположенных равномерно с шагом Δ на кривой. Разобьем это множество T на ряд неперекрывающихся элементарных участков $\omega_j, j=1,\ldots,M$ равной длины l. Тогда каждому участку ω_j сопоставляется вектор, компонентами которого являются упорядоченные значения ординат кривой $f^j = (f_1^j,\ldots,f_l^j)$ на этом участке и этот вектор будет являться характеристикой формы соответствующего ему участка кривой $f(t)$. Допустим также, что число точек l в каждом участке превышает порядок авторегрессии n и введем в рассмотрение стохастическую функцию сложности [140]:

$$\phi(\omega_j) = \min_c \sum_{s=n+1}^{l} \left(f_s^j - c_0 - \sum_{i=1}^{n} c_i^j f_{s-i}^j \right)^2, \qquad (64)$$

которая будет характеризовать качество аппроксимации экспериментальной кривой на участках авторегрессионной моделью порядка n (63). С помощью этой функции сложности на кривой выделяются участки, которые не являются реализацией случайного стационарного процесса. Именно они выделяются данным алгоритмом как переходные участки. Учитывая сказанное данный алгоритм состоит из следующих шагов:

- исследуемая экспериментальная кривая разбивается на участки равной длины.

- на каждом участке определяется набор коэффициентов авторегрессии.

- на кривой определяется последовательность значений функции (64).

- на кривой выделяются все участки, соответствующие локально максимальным значениям функции сложности.

- выявленные участки рассматриваются как искомые переходные участки.

- для каждого выделенного участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

Структурная схема вышеописанного алгоритма представлена на рисунке 3.

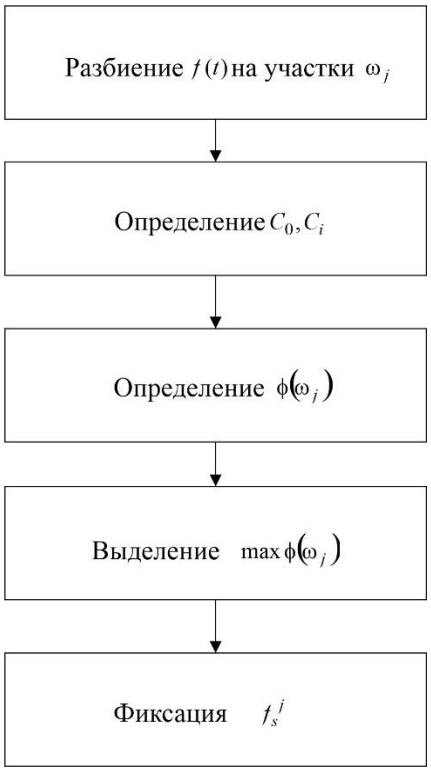

Рисунок 3 – Структурная схема алгоритма сегментации

Особенностью представленного алгоритма является то, что в результате применения этого алгоритма все информативные участки выделяются одновременно. В результате работы алгоритма полученный массив векторов в дальнейшем используется для разработки словаря описания выделенных участков.

2.2.1 Алгоритм частичной аппроксимации

Предложенные в 2.2 алгоритмы сегментации, можно назвать локальными, поскольку отнесение каждого участка кривой к простым либо к сложным основано на анализе поведения кривой только на данном элементарном участке и в его ближайшей окрестности. Данные алгоритмы сегментации используют минимум априорной информации о характере обрабатываемой кривой и могут применяться как для поиска сложных участков, соответствующих смене состояний исследуемого процесса, так и для выделения аномальных событий на общем

простом фоне. Однако в последнем случае сегментацию более естественно проводить на основе анализа всей кривой в целом и определения относительной степени изменчивости ее поведения на отдельном участке в зависимости от общего уровня изменчивости исследуемой кривой.

Пусть анализируемая экспериментальная кривая представлена в виде чередования участков простого фона, общего для всей кривой, и сложных, т. е. априори неизвестных, возмущений, локализованных на части области определения. При этом результат сегментации строится из двух моделей: модели, объясняющей простой фон, и описания расположения участков, содержащих сложные возмущения. Из содержательных соображений может быть назначена общая длина системы сложных участков.

В качестве сложных выбирается такая система участков заданной общей длины, удаление которой позволяет наилучшим образом аппроксимировать поведение кривой на оставшейся части области определения моделью из заданного класса моделей. Априорные представления о простых формах поведения кривых выразим в виде некоторого подмножества E в множестве F всех возможных кривых, определенных в области изменения аргумента T. Экспериментальная кривая $e(t)$ из E может рассматриваться как простая на всей области определения. Тогда поведение произвольной кривой $f(t) \in \frac{F}{E}$ можно приближенно описать, указав простую кривую $e(t) \in E$ достаточно близкую к $f(t)$ вне некоторой системы участков ω в интервале T и значительно отличающуюся от нее на ω. Область $\frac{T}{\omega}$ соответственно будет простой частью кривой $f(t)$, а множество ω – системой сложных участков (Рисунок 4).

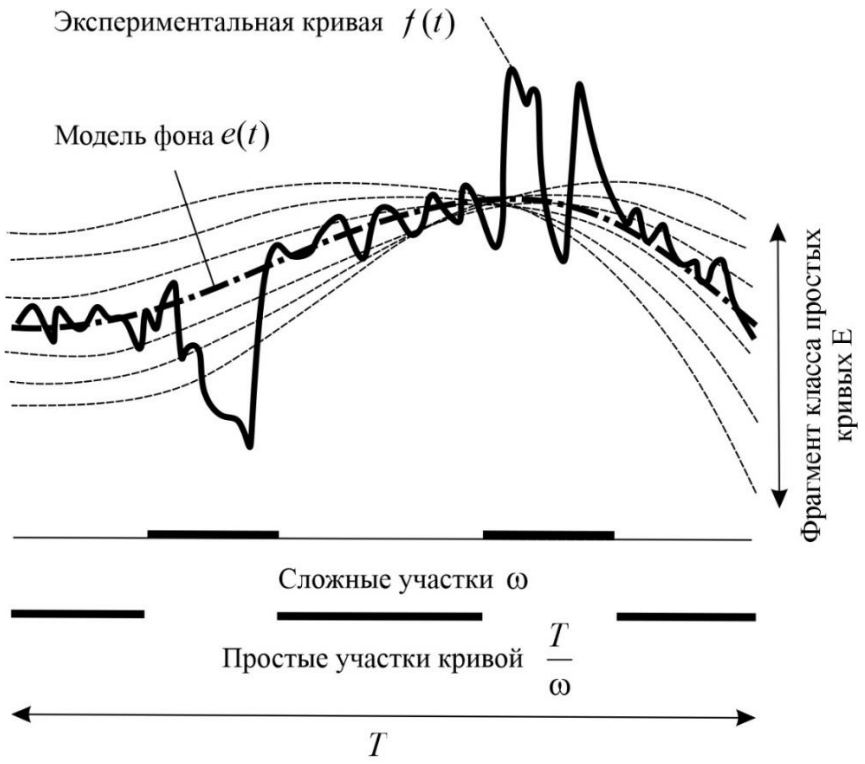

Рисунок 4 – Выделение сложных участков экспериментальной кривой алгоритмом частичной аппроксимации

В качестве способа задания класса простых кривых примем множество многочленов $e(t)=\sum_{i=1}^{m} c_i \varphi^i(t)$ по некоторой фиксированной системе линейно независимых функций $\{\varphi^i(t), i=1,...,n\}$, определенной на интервале T, выбор которой зависит от решаемой задачи. Учитывая вышесказанное, алгоритм частичной аппроксимации состоит из следующих шагов:

- обозначается через Ω_d класс всех подмножеств $\omega \subset T$, имеющих заданную длину $\mu(\omega)=d$.

- экспериментальной кривой $f(t)$ сопоставляется функция множеств $\phi(\omega)$, определенную на классе подмножеств Ω_d и численно равную интегральной квадратичной норме остаточного члена многочлена наилучшего приближения $f(t)$ по системе $\{\varphi^i(t)\}$ на множестве $\frac{T}{\omega}$, которая и будет являться функцией сложности:

$$\phi(\omega) = \int_{T/\omega} s^2(\omega, t)dt, \tag{65}$$

$$s(\omega, t) = f(t) - \sum_{i=1}^{n} c_i(\omega)\varphi^i(t),$$

где $c_i(\omega)$ - коэффициенты многочлена наилучшего приближения.

- находится для данной функции $f(t)$ среди всех подмножеств $\omega \in \Omega_d$ ее интервала определения подмножество $\omega*$, доставляющее функции сложности $\phi(\omega)$ минимальное значение.

Структурная схема вышеописанного алгоритма представлена на рисунке 5.

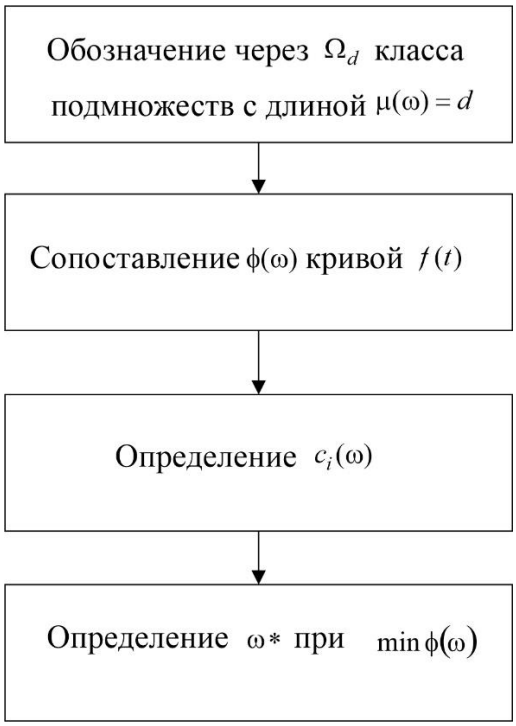

Рисунок 5 – Структурная схема алгоритма частичной аппроксимации

При дискретном задании экспериментальной кривой f_t на конечном интервале $T = \{1,..., N\}$ в качестве системы базисных функций используется система N - мерных векторов $\{\varphi_t^i, i = 1,..., n; t = 1,..., N\}$. В качестве класса подмножеств области определения Ω_d, среди которых ищется система сложных участков, рассматривается множество всех подмножеств $\omega \subset T$, состоящих из заданного числа точек d. Тогда функция сложности $\phi(\omega)$ (65), определенная на Ω_d, примет вид:

$$\phi(\omega) = \sum_{t \notin \omega} s_t^2(\omega),$$

$$s_t(\omega) = f_t - \sum_{i=1}^{n} c_i(\omega)\varphi_t^i,$$

где $c_i(\omega)$ - коэффициенты многочлена наилучшего приближения кривой f_t по системе векторов $\{\varphi_t^i, i = 1,..., n\}$ на множество отсчетов $\frac{T}{\omega}$, $s_t(\omega)$ - остаточная невязка этого приближения.

2.2.2 Алгоритм минимизации функции сложности

Для построения алгоритма минимизации функции сложности необходимо определить способ выбора очередного подмножества ω^{k+1}, удовлетворяющего неравенству $\phi(\omega^{k+1}) < \phi(\omega^k)$, на основе вычислений, проведенных только на подмножестве ω^k. Пусть $\omega \in \Omega_d$ некоторое подмножество области определения кривой T. Для любой кривой f_t, $t \in T$ и любой пары подмножеств $\boldsymbol{\omega}$, $\omega_1 \subset T$ выполняется неравенство:

$$\phi(\omega) - \phi(\omega_1) > \sum_{\substack{t \in \omega \\ \omega_1}} s_t^2(\omega) - \sum_{\substack{t \in \omega_1 \\ \omega}} s_t^2(\omega). \qquad (66)$$

По определению функции сложности $\phi(\omega) - \phi(\omega_1) = \sum_{\substack{t \in T \\ \omega}} s_t^2(\omega) - \sum_{\substack{t \in T \\ \omega_1}} s_t^2(\omega_1)$. В силу того, что $s_t(\omega_1)$ есть остаточная невязка многочлена наилучшего приближения f_t на $\frac{T}{\omega}$, значение крайней правой суммы в последнем равенстве не уменьшается, если заменить $s_t(\omega_1)$ остаточным членом другого многочлена, например $s_t(\omega)$:

$$\phi(\omega) - \phi(\omega_1) > \sum_{\substack{t \in T \\ \omega}} s_t^2(\omega) - \sum_{\substack{t \in T \\ \omega_1}} s_t^2(\omega) = \sum_{\substack{t \in \omega \\ \omega_1}} s_t^2(\omega) - \sum_{\substack{t \in \omega_1 \\ \omega}} s_t^2(\omega).$$

Для того, чтобы найти нижнюю оценку разности $\boldsymbol{\phi(\omega) - \phi(\omega_1)}$ для любого подмножества $\omega_1 \in \Omega_d$, достаточно построить многочлен наилучшего приближения кривой на $\frac{T}{\omega}$ и вычислить остаточную невязку $s_t(\omega)$ для всех $t \in T$. Искомым

является подмножество $\omega \in \Omega_d$, для которого все разности $\phi(\omega) - \phi(\omega_1) \le 0$ при любом $\omega_1 \in \Omega_d$.Такое условие может удовлетворяться лишь на подмножествах $\omega \in \Omega_d$, для которых разность в правой части неравенства (66) не положительна при любых $\omega \in \Omega_d$. Это в свою очередь имеет только в том случае, если взаимная конфигурация подмножества $\omega \in \Omega_d$ и соответствующего остаточного члена наилучшего приближения кривой в области $\frac{T}{\omega}$ удовлетворяют неравенству:

$$\min_{\omega} s_t^2(\omega) > \max_{\frac{T}{\omega}} s_t^2(\omega). \qquad (67)$$

Неравенство (67) является необходимым условием минимума функции сложности $\phi(\omega)$ на классе подмножеств Ω_d. Определим отображение класса подмножеств Ω_d сопоставив каждому подмножеству $\omega \in \Omega_d$ подмножество $\omega' \in \Omega_d$, состоящее из d точек с наибольшими значениями квадрата невязки $s_t^2(\omega)$. Введем уточнение, в случае равных значений $s_t^2(\omega)$ предпочтение отдается точкам с меньшим номером, которое делает отображение $\omega \to \omega'$ в классе подмножеств Ω_d однозначным. Подмножество ω' назовем производным по отношению к ω. Благодаря данному уточнению и неравенству (67), $\omega \to \omega'$ обладает следующими свойствами:

- $\phi(\omega') \le \phi(\omega)$, причем $\phi(\omega') < \phi(\omega)$ если ω не удовлетворяет необходимому условию минимума функции сложности на Ω_d;

- если $\omega' = \omega$, то подмножество ω удовлетворяет необходимому условию минимума $\phi(\omega)$ на Ω_d;

Пусть $\omega_2 = \omega'_1, ..., \omega_k = \omega_k = \omega'_{k-1}$ и $\omega_k = \omega_1$. Все подмножества $\omega_1, ..., \omega_{k-1}$ заведомо удовлетворяют необходимому условию минимума $\phi(\omega)$ на Ω_d, поскольку в противном случае, согласно первому свойству, $\phi(\omega_j) < \phi(\omega_{j-1})$ хотя бы для одного $j = 2, ..., k$ и при этом $\phi(\omega_k) \le ... \le \phi(\omega_1)$; отсюда следовало бы $\phi(\omega_k) \le ... \le \phi(\omega_1)$. Таким образом, имеет место неравенство

$$\min_{\omega_j} s_t^2(\omega_j) \geq \max_{\frac{T}{\omega_j}} s_t^2(\omega_j),\ j = 1,...,k. \tag{68}$$

Вместе с тем, $\phi(\omega_1) = \phi(\omega_k)$, поскольку $\omega_1 = \omega_k$. Поэтому $\phi(\omega_1) = \phi(\omega_2) = ... = \phi(\omega_k)$. Докажем, что

$$s_t^2(\omega_1) = s_t^2(\omega_2) = ... = s_t^2(\omega_{k-1}). \tag{69}$$

Действительно, согласно (68) и определению производного подмножества, $\sum_{\frac{t\in\omega_{j-1}}{\omega_{j-1}}} s_t^2(\omega_{j-1}) = \sum_{\frac{t\in\omega_{j-1}}{\omega_j}} s_t^2(\omega_{j-1}) = 0$ для любого $2 \leq j \leq k$. Следовательно $\sum_{t\in\frac{T}{\omega_j}} s_t^2(\omega_{j-1}) = \sum_{t\in\frac{T}{\omega_{j-1}}} s_t^2(\omega_{j-1}) = \phi(\omega_j - 1)$, причем аналогичное равенство имеет место также для пары подмножеств ω_{k-1}, ω_1. В то же время $\phi(\omega_{j-1}) = \phi(\omega_j)$ и таким образом:

$$\sum_{\frac{t\in T}{\omega_j}} s_t^2(\omega_{j-1}) = \phi(\omega_j) = \sum_{\frac{t\in T}{\omega_j}} s_t^2(\omega_j). \tag{70}$$

.

Тогда отображение $\omega \to \omega'$ не имеет циклов, т. е. если $\omega_2 = \omega'_1, ..., \omega_k = \omega_k = \omega'_{k-1}$ и при этом $\omega_k = \omega_1$, то $\omega_1 = \omega_2 = ... = \omega_k$.

По определению, $s_t(\omega_j)$ является остаточным членом многочлена наилучшего приближения кривой f_t, на $\frac{T}{\omega_j}$. Но, согласно равенству (70), таким же свойством на множестве $\frac{T}{\omega_j}$ обладает и $s_t(\omega_{j-1})$. Тогда в силу единственности многочлена наилучшего приближения $\hat{f}_t(\omega_j) = \sum_{i=1}^{n} c_i(\omega_j)\varphi_t^i$ остаточные члены $s_t(\omega_{j-1})$ и $s_t(\omega_j)$ совпадают для любого $2 \leq j \leq k$, откуда следует (69). Таким образом, все подмножества $\omega_1, ..., \omega_k$ образованы как производные по одному и тому же правилу из одной и той же последовательности квадратов невязок и, следовательно, совпадают.

Исходя из вышесказанного, непосредственно следует способ направленного перебора подмножеств, гарантирующий достижение через конечное число шагов некоторого $\omega* \in \Omega_d$, удовлетворяющего необходимому условию минимума

функции сложности на Ω_d. Для построения такой последовательности подмножеств достаточно, начиная с любого $\omega^1 \in \Omega_d$, выбирать в качестве ω^{k+1} подмножество, производное по отношению к ω^k, т. е. $\omega^{k+1} = (\omega^k)$. Пока для ω^k не выполняется условие (67), последовательность подмножеств не может иметь повторяющихся элементов, поскольку при этом $\phi(\omega^{k+1}) < \phi(\omega^k)$. Таким образом, в силу на некотором шаге $k*$ возникает ситуация $\omega^{k*} = \omega^{k*-1}$, при этом подмножество ω^{k*} будет удовлетворять необходимому условию минимума функции сложности.

2.2.3 Стохастический алгоритм частичной аппроксимации

В качестве порождающей модели экспериментальной кривой примем частично стационарный случайный процесс $f(t)$, совпадающий с некоторым стационарным случайным процессом, а именно его фоном $y(t)$ всюду, за исключением некоторой последовательности аномальных участков, в пределах которых закон его формирования неизвестен. Такая модель соответствует представлению об источнике кривой как об объекте, находящемся преимущественно в неизменном состоянии и генерирующем при этом некоторый стационарный случайный процесс. Отдельные возмущения неизвестного характера выводят его из основного состояния и вызывают появление локальных изменений генерируемых им колебаний.

Для сегментации экспериментальной кривых, рассматриваемых как отрезки реализации частично стационарного случайного процесса, необходимо определение системы участков внутри интервала наблюдения, содержащих отдельные возмущения, и в оценке закона формирования стационарного фона кривой, несущего информацию об основном состоянии изучаемого объекта. В качестве класса моделей фона примем множество авторегрессий n -го порядка (63). Разобьем область определения экспериментальной кривой T на M

непересекающихся элементарных участков (интервалов) σ_q, $q=1,\dots$, M, каждый из которых состоит из l точек: $T=\bigcup_{q=1}^{M}\sigma_q$, $N=Ml$.

Под классом будем подмножеств Ω_d будем понимать множество всевозможных комбинаций из d элементарных интервалов, $d<M$: $\Omega_d=\left\{\omega\subset T;\omega=\bigcup_{j=1}^{d}\sigma_{qj}\right\}$. Таким образом, длина аномального подмножества d понимается как число входящих в него элементарных интервалов σ_q.

Поскольку каждый элементарный интервал σ_q неделим, то минимальным элементом информации о поведении экспериментальной кривой является вектор ее отсчетов на σ_q, Для удобства наложения введем двойную нумерацию отсчетов кривой в области определения T. Каждый отсчет будем обозначать f_{qr} где первый индекс $q=1,\dots,M$ указывает на номер элементарного интервала σ_q, а второй, $r=1,\dots,l$, равен порядковому номеру данного отсчета в его пределах. Дополнительно предположим, что число точек l в каждом элементарном интервале больше порядка авторегрессии n.

Если подмножество ω содержит все аномальные участки экспериментальной кривой f_{qr} то ее отсчеты в области $\frac{T}{\omega}$ формируются согласно стохастическому уравнению авторегрессии (63). В противном случае ее поведение на $\frac{T}{\omega}$ не соответствует этому уравнению. Для оценки согласованности поведения кривой в области $\frac{T}{\omega}$ с принятым классом моделей фона примем функцию:

$$\phi(\omega)=\sum_{\sigma_q\subset\omega}\sum_{r=n+1}^{l}s_{qr}^2(\omega), \qquad (71)$$

где $s_{qr}(\omega)=f_{qr}-c_0(\omega)-\sum_{i=1}^{n}c_i(\omega)f_{q,r-i}$ функция невязки уравнения авторегрессии, подсчитанная при значениях коэффициентов, равных оценкам, полученным по

методу наименьших квадратов для области $\frac{T}{\omega}$. Функцию (71) назовем стохастической функцией сложности. Тогда задачу выделения аномалий на фоне стационарного случайного процесса авторегрессии можно сформулировать так: для данной кривой f_{qr}, $q=1,\ldots$, M, $r=1,\ldots$, l среди всех подмножеств ее области определения T принадлежащих классу Ω_d, найти подмножество ω^*, доставляющее стохастической функции сложности минимальное значение.

В отличии от алгоритма частичной аппроксимации: множество многочленов по заданной системе базисных функций заменяется классом стохастических разностных уравнений авторегрессии (63), остаточным членом $s_t(\omega)$ наилучшего приближения кривой многочленом на подмножестве значений аргумента $\frac{T}{\omega}$ является вектор невязок кривой и модели фона $s_{qr}(\omega)$, а функции сложности соответствует стохастическая функция сложности. Этим аналогия не ограничивается. Задача поиска подмножества, доставляющего минимальное значение стохастической функции сложности для данной экспериментальной кривой, полностью сводится к задаче многочленной частичной аппроксимации по специально выбранной системе базисных функций.

Вместо двойной индексации отсчетов экспериментальной кривой f_{qr} и каждый отсчет помечается одним индексом f_t, равным его порядковому номеру в интервале T: $t=(q-1)l+r$. Выделяется в каждом элементарном интервале q_r, его часть $\sigma_q \subset \sigma_q$, не включающую первые n точек $\hat{\sigma}_q = \{t \in T; t=(q-1)l+r; r=n+1,\ldots l\}$.

Соответственно в области определения кривой T, которая является объединением всех элементарных интервалов σ_q, выделим подмножество $\hat{T} = \bigcup_{q=1}^{M} \hat{\sigma}_q \subset T$. Обозначим множество всех подмножеств области $\hat{T}$, составленных из интервалов $\hat{\sigma}_q$ как $\hat{\Omega}_d = \left\{ \hat{\omega} \subset \hat{T}; \hat{\omega} = \bigcup_{j=1}^{d} \hat{\sigma}_{qj} \right\}$. Между элементами систем подмножеств

Ω_d и $\hat{\Omega}_d$ устанавливается взаимно однозначное соответствие, а именно, если $\omega \in \Omega_d$ состоит из элементарных интервалов σ_{qj}, $j = 1,..., d$, то ему соответствует подмножество $\hat{\omega}$, объединяющее интервалы $\hat{\sigma}_{qj}$ с теми же индексами q_j, $j = 1,..., d$, и наоборот. При этом $\hat{\omega} \subset \omega$, $\hat{T}/\hat{\omega} \subset T/\omega$.

На множестве $\hat{T}$ определим $n+1$ функций $\varphi_t^0 = 1$, $\varphi_t^1 = f_{t-1}, \ldots, \varphi_t^n = f_{t-n}$, $t \in T$. Тогда, если рассматривать систему функций $\{\varphi_t^i, i = 0,..., n\}$ как базис, то вектор оценок коэффициентов авторегрессии $c(\omega)$ для данного подмножества $\omega \in \Omega_d$, полученный по методу наименьших квадратов, равен вектору коэффициентов многочлена наилучшего приближения кривой f_t, $t \in \hat{T}$ по системе функций $\{\varphi_t^i, i = 0,..., n\}$ на множестве $\frac{\hat{T}}{\hat{\omega}}$, $\hat{\omega} \in \Omega_d$, определяемому из условия минимума остаточной суммы квадратов $\sum_{t \in \hat{T}/\hat{\omega}} \left(f_t - \sum_{i=0}^{n} c_i \varphi_t^i \right)^2 = \sum_{t \in \hat{T}/\hat{\omega}} s_t^2(c)$. Стохастическая функция сложности $\phi(\omega)$, определенная согласно (57), равна квадрату нормы остаточного члена многочлена наилучшего приближения кривой f_t на подмножестве $\frac{\hat{T}}{\hat{\omega}}$, а именно $\phi(\omega) = \hat{\phi}(\hat{\omega}) = \sum_{t \in \hat{T}/\hat{\omega}} s_t^2(\hat{\omega})$.

Для построения последовательности подмножеств $\omega^{k+1} = (\omega^k)'$ необходимо на каждом шаге находить оценки коэффициентов авторегрессии $c(\omega^k)$ из условия минимума остаточной суммы квадратов на подмножестве $\frac{T}{\omega^k}$. Приравнивание нулю частных производных $\frac{\partial}{\partial c_j} \sum_{\sigma_q \subset \omega^k} \sum_{r=n+1}^{t} \left(f_{qr} - c_0 - \sum_{i=1}^{n} c_i f_{q,r-1} \right)^2 = 0$, $j = 0,..., n$, приводит к системе линейных алгебраических уравнений относительно компонент вектора $c(\omega^k)$:

$$a(\omega^k) \cdot c(\omega^k) = h(\omega^k), \tag{72}$$

где матрица $a(\omega^k)$ размерности $(n+1) \times (n+1)$ и вектор $h(\omega^k)$ размерности $n+1$ имеют следующую структуру:

$$a(\omega^k)=\left\|\begin{matrix}(M-d)(l-n) & \sum\limits_{\sigma_q\subset\omega^k}\sum\limits_{r=n+1}^{l} f_{q,r-i} \\ \sum\limits_{\sigma_q\subset\omega^k}\sum\limits_{\substack{r=n+1\\ i=0}}^{l} f_{q,r-j} & \sum\limits_{\sigma_q\subset\omega^k}\sum\limits_{\substack{r=n+1\\ i=1,\dots n}}^{l} f_{q,r-i}\,f_{q,r-j}\end{matrix}\right\|\begin{matrix} j=0 \\ j=1,\dots,n\end{matrix}, \qquad (73)$$

$$h(\omega^k)=\left\|\begin{matrix}\sum\limits_{\sigma_q\subset\omega^k}\sum\limits_{r=n+1}^{l} f_{q,r-i} \\ \sum\limits_{\sigma_q\subset\omega^k}\sum\limits_{r=n+1}^{l} f_{q,r}\,f_{q,r-j}\end{matrix}\right\|\begin{matrix} j=0 \\ j=1,\dots,n\end{matrix}. \qquad (74)$$

Элементы матрицы $a(\omega^k)$ и вектора $k(\omega^k)$ являются суммами величин, соответствующих элементарным интервалам σ_q, не входящим в подмножество ω^k. Эти величины в свою очередь представляют собой суммы отсчетов экспериментальной кривой и их произведений внутри каждого элементарного интервала. Вычисляются суммы заранее для всех элементарных интервалов области определения кривой $\sigma_q, q=1,\dots,M$, в виде матриц a_q векторов h_q имеющих следующую структуру:

$$a_q=\left\|\begin{matrix}l-n & \sum\limits_{r=n+1}^{l} f_{q,r-i} \\ \sum\limits_{\substack{r=n+1\\ i=0}}^{l} f_{q,r-j} & \sum\limits_{\substack{r=n+1\\ i=1,\dots,n}}^{l} f_{q,r-i}\,f_{q,r-j}\end{matrix}\right\|\begin{matrix} j=0 \\ j=1,\dots,n\end{matrix},$$

$$h_q=\left\|\begin{matrix}\sum\limits_{r=n+1}^{l} f_{qr} \\ \sum\limits_{r=n+1}^{l} f_{qr}\,f_{q,r-1}\end{matrix}\right\|\begin{matrix} j=0 \\ j=1,\dots,n\end{matrix}.$$

Такой набор матриц и векторов содержит всю информацию, необходимую для формирования матриц $a(\omega^k)$ и векторов $h(\omega^k)$ для любого подмножества $\omega^k\in\Omega_d$, а именно $a(\omega^k)=\sum\limits_{\sigma_q\subset\omega^k} a_q$, $h(\omega^k)=\sum\limits_{\sigma_q\subset\omega^k} h_q$.

Вычисление $a(\omega^k)$ и $h(\omega^k)$ таким способом позволяет значительно сократить время на формирование системы уравнений (72), поскольку при непосредственном расчете по формулам (73) и (74) пришлось бы на каждом шаге работы алгоритма заново вычислять суммы, входящие в a_q и h_q. Вместе с числом $p_q=\sum\limits_{r=n+1}^{l} f_{qr}^2$ матрица

a_q и вектор h_q полностью описывают характер кривой на элементарном участке при данном значении его длины l и порядке авторегрессии n, поэтому сама кривая, которая обычно имеет десятки тысяч отсчетов, может не храниться в оперативной памяти. Если учесть, что для устойчивости работы алгоритма элементарный участок должен содержать как минимум несколько отсчетов, сокращение использования памяти оказывается существенным. Суммарная невязка в пределах элементарного участка $u_q(\omega^k)$ может быть вычислена через элементы a_q , h_q и p_q:

$$u_q(\omega^k) = p_q + \sum_{i=0}^{n}\sum_{j=0}^{n} c_i(\omega^k)c_j(\omega^k)a_{ij}^q - 2\sum_{i=0}^{m} c_i(\omega^k)h_i^q .$$

При этом оценка параметра шума $b(\omega^k)$ уравнения авторегрессии (63) определяется остаточной суммой квадратов в пределах $\frac{T}{\omega^k}$, а именно:

$$b^2(\omega^k) = \frac{1}{(M-d)(l-n)} \sum_{\sigma_q \subset \omega^k} u_q(\omega^k) .$$

Алгоритм сегментации, реализующий вышеописанные процедуры, выделяет подмножество области определения экспериментальной кривой, состоящее из заданного числа ее элементарных интервалов. В то же время на практике имеющейся априорной информации о характере изучаемого объекта, когда его поведение имеет стохастическую природу, далеко не всегда достаточно для разумного выбора величины d. В связи с этим был разработан эвристический способ автоматической оценки этой величины. Способ основан на анализе формы логарифмической функции правдоподобия относительно коэффициентов авторегрессии в ее экстремальной точке.

Пусть для некоторого d с помощью алгоритма частичной аппроксимации найдено подмножество $\omega_d^* \in \Omega_d$, содержащее все аномальные точки экспериментальной кривой, так что в области $\frac{T}{\omega_d^*}$ она является реализацией нормального стационарного случайного процесса авторегрессии. И пусть $\theta(\omega_d^*) = \left[c_0(\omega_d^*),\cdots,c_n(\omega_d^*),b(\omega_d^*)\right]$ - соответствующие оценки параметров стохастического

уравнения фона. Если считать подмножество ω_d^* постоянным, то математическое ожидание матрицы вторых производных логарифмической функции правдоподобия $L(\theta, \omega_d^*)$ в точке $\theta(\omega_d^*)$

$$\left.\frac{\partial^2 L(\theta, \omega_d^*)}{\partial c_i \partial c_j}\right|\theta = \theta(\omega_d^*), i, j = 0,..., n, \qquad (75)$$

называется информационной матрицей Фишера. Эта матрица обратна по отношению к ковариационной матрице оценок $c_0(\omega_d^*),..., c_n(\omega_d^*)$. Оценки этих ковариаций можно получить, обращая матрицу вторых производных (75), отображающую кривизну логарифмической функции правдоподобия относительно $c_0,..., c_n$ в экстремальной точке.

Элементы матрицы вторых производных логарифмической функции правдоподобия для данной экспериментальной кривой полностью совладают с элементами матрицы

$$\frac{1}{b^2(\omega_d^*)} a(\omega_d^*),$$

вычисленной на последнем шаге работы алгоритма, согласно (73) по формуле $a(\omega_d^*) = \sum_{\sigma_q \subset \omega_d^*} a_q$. Тогда след обратной матрицы $S(d) = b^2(\omega_d^*) Spa^{-1}(\omega_d^*)$ связан обратной зависимостью с обобщенной кривизной логарифмической функции правдоподобия в экстремальной точке и является мерой «безразличия» значений $L(\theta, \omega_d^*)$ к значениям коэффициентов авторегрессии $c_0,..., c_n$. Если предположить, что ω_d^* содержит все аномальные участки кривой, то $S(d)$ имеет смысл суммы выборочных дисперсий компонент вектора $c(\omega_d^*)$.

2.3 Экстраполяционные алгоритмы сегментации

Экстраполяционные методы прогнозирования широко распространены и исследованы. Все они основываются на исследовании динамических рядов - массивов наблюдений, полученных последовательно во времени. Широко

применяется метод математической экстраполяции, метод подбора функций, матричный метод. Все они с математической точки зрения предполагают распространение некоторого закона изменения функции из области ее определения в области, находящейся вне ее. Вводится некоторая функция, представляющая собой математико-статистическую модель исследуемого события. В зависимости от характера и специфики поведения кривой, функция может быть линейного, гиперболического, логарифмического, экспоненциального и других типов. Линейная, например, применяется для исследования событий, равномерно изменяющихся во времени. В лингвистическом подходе роль такой функции может выполнять функция сложности.

В 2.2 использован ряд алгоритмов сегментации переходных участков экспериментальных кривых с применением функции сложности. На основе этого представим общую идею алгоритма [144], прогнозирующего поведение кривой за границами ее определенных участков:

-определенная область экспериментальной кривой $f(t)$ разбивается на ряд элементарных участков $\omega_j, j=1,\ldots,n$ одинаковой длины l (Рисунок 6);

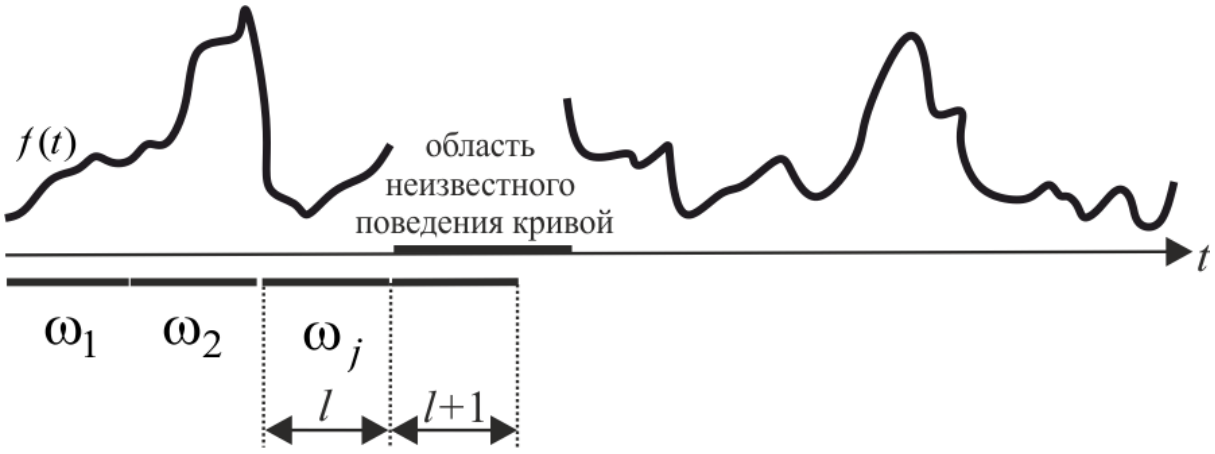

Рисунок 6 – Экстраполяция участков экспериментальных кривых

-вводится функция сложности $\phi(f,\omega)$ для оценки экстраполяционных свойств поведения кривой на граничащих участках (l+1);

-задается величина порога ошибки экстраполяции;

-выделяются такие участки кривой, на которых функция сложности принимает локально экстремальные значения;

-выделенные участки принимаются за искомые участки, по которым стоит прогнозировать поведение кривой на данных участках.

Учитывая вышесказанное, представим два конкретных алгоритма такого типа [145] далее.

Алгоритм прогнозного типа для экстраполяции кривых стохастического характера. Пусть экспериментальная кривая $f(t)$ изменяется случайным образом. Тогда такие участки, на которых характер ее поведения изменяется не значительно (простые, однородные), будут являться реализацией определенного случайного стационарного процесса. Такие процессы будем рассматривать как совокупность стохастических уравнений авторегрессии:

$$x_t = c_0 + \sum_{i=1}^{n} c_i x_{t-i} + b\xi_t \ ,$$

где c_i – коэффициент авторегрессии (i принимает значения от 0 до n), n – порядок авторегрессии, ξ_t – последовательность независимых нормальных случайных величин с нулевым математическим ожиданием и единичной дисперсией, b – коэффициент среднеквадратичного отклонение белого шума.

Представим эту кривую $f(t)$ как последовательность значений ее ординат $f_1, f_2, \ldots$, расположенных равномерно с шагом Δ на кривой и разобьем ее на ряд элементарных участков $\omega_j, j = 1,\ldots,N$ равной длины l, каждый из которых описывается моделью n-го порядка авторегрессии (63) с коэффициентами c_i и b. Коэффициенты рассчитаем методом наименьших квадратов на каждом участке ω_j и при дополнительном условии $t > jl$ рассчитаем величину: $s_t(\omega_j) = \sum_{s=l+1}^{t} \left[g_s(\omega_j) \right] - (t - jl)$, где $g_s(\omega_j) = \frac{1}{b}\left(f_s^j - c_0 - \sum_{i=1}^{n} c_i^j f_{s-i}^j \right)^2$. Если данный участок будет стационарного поведения, то значение $g_s(\omega_j)$ будет совпадать с квадратом белого шума ξ_s^2 и будет стремиться к единице, соответственно значение

$s_t(\omega_j)$ будет стремиться к нулю. При изменении коэффициентов c_i и b возникают систематические невязки $g_s(\omega_j)-1$, соответственно $s_t(\omega_j)$ станет отлично от нуля в ту или иную сторону. Введем в рассмотрение функцию сложности [140]

$$\phi(\omega_j)=t_0(\omega_j), \tag{76}$$

и зададим пороговое значение ошибки экстраполяции ε. Вычислим такое минимальное значение $t_0(\omega_j)$, при котором модуль величины $|s_t(\omega_j)|$ превысит заданное пороговое значение ε. С помощью данной функции сложности будут анализироваться экстраполяционные свойства участков ω_j в одном направлении и прогнозировать поведение кривой за границами участка следует по участкам с локальными минимумами $\phi(\omega_j)$. Учитывая сказанное данный алгоритм состоит из следующих шагов:

- исследуемая экспериментальная кривая разбивается на участки равной длины;

- на каждом участке определяется набор коэффициентов уравнения авторегрессии (63);

- на каждом участке вычисляется величина $s_t(\omega_j)$;

- задается величина порога ошибки экстраполяции;

- на кривой определяется последовательность значений функции (76);

- на кривой выделяются все участки, соответствующие локально минимальным значениям функции сложности;

- выделенные участки рассматриваются как искомые участки для прогнозирования;

- для каждого выделенного участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

Данный алгоритм будет анализировать экстраполяционные свойства кривой стохастического характера поведения только в одном направлении, а именно вправо за границы определенного участка.

Алгоритм формального типа для экстраполяции экспериментальных кривых с применением аппроксимации отрезком степенного ряда. Пусть

исследуемая кривая $f(t)$ представлена последовательностью значений ее ординат $f_1, f_2, \ldots,$ расположенных равномерно с шагом Δ на кривой и разбита на ряд элементарных участков $\omega_j, j = 1, \ldots, N$ равной длины l. И каждый участок аппроксимируется отрезком степенного ряда $\sum_{i=0}^{N} c_i^j s^i$, где s является порядковым номером точки на данном участке. Данный полином продолжается вправо (правая экстраполяция) и влево (левая экстраполяция) за границы участка. Введем в рассмотрение функцию сложности [140]

$$\phi(\omega_j) = \min(k_r, k_l), \tag{77}$$

формализующую представление об изменчивости формы кривой, проверяя, сохраняет ли она свое основное направление в некоторой области текущего элементарного участка. Зададим пороговое значение ε и определим длины левого k_l и правого k_r интервалов экстраполяции так, чтобы они не превышали его: $\sum_{s=l-1}^{k_l} \left(f_s^j - \sum_{i=0}^{n} c_i^j s^i \right)^2 \le \varepsilon$ и $\sum_{s=l+1}^{k_r} \left(f_s^j - \sum_{i=0}^{n} c_i^j s^i \right)^2 \le \varepsilon$. Учитывая сказанное данный алгоритм состоит из следующих шагов:

- исследуемая экспериментальная кривая разбивается на участки равной длины;

- исходные векторы значений ординат кривой, соответствующие участкам, нормируются;

- на каждом участке определяется величина аппроксимирующего полинома;

- на кривой определяется последовательность значений функции (77);

- задается величина порога ошибки экстраполяции;

- на кривой выделяются все участки, соответствующие локально минимальным значениям функции сложности;

- выделенные участки рассматриваются как искомые участки для прогнозирования;

- для каждого такого участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

Особенностью данного алгоритма является то, что он будет прогнозировать поведение кривой в обеих направлениях, а именно и вправо за границы определенного участка.

2.4 Выводы по разделу

Во втором разделе решалась задача разработки комбинированных вычислительных методов сегментации экспериментальных кривых и алгоритмической реализации процедуры структурного анализа экспериментальных кривых. Для методики сегментации представлены дополнительные уточнения во избежание выделения ложных экстремумов. Представлены функции сложности, которые применимы при сегментации детерминированных и случайных кривых различной длины. Особенностью алгоритма с определением качества аппроксимации участков экспериментальной кривой авто-регрессионной моделью является то, что в результате применения этого алгоритма все информативные участки выделяются одновременно. Представлен алгоритм частичной аппроксимации, для применения которого в отличие от эвристических алгоритмов сегментации, достаточно выбрать длину системы выделяемых участков d. Отличие стохастического варианта алгоритма частичной аппроксимации состоит в том, что минимум функции сложности $\hat{\phi}(\hat{\omega})$ необходимо найти в классе подмножеств, состоящих из заранее нарезанных интервалов области определения, а не среди любых подмножеств заданной длины.

РАЗДЕЛ 3 ЭТАП ЛИНГВИСТИЧЕСКОГО ОПИСАНИЯ УЧАСТКОВ ЭКСПЕРИМЕНТАЛЬНЫХ КРИВЫХ

Данный раздел посвящен одному из трех основных этапов обработки экспериментальных кривых, а именно этапу лингвистического описания участков кривых. В 3.1 рассмотрена реализация данного этапа. В 3.1.1 и 3.1.2 представлены алгоритмы, осуществляющие выбор степени отличия для анализа участков различной длины и построение опорных участков. В 3.2 рассмотрена реализация формирования языка описания экспериментальных кривых. В 3.2.1 представлена процедура сопоставления участков каждого класса к последовательности символов, а в 3.2.2 и 3.2.3 представлены алгоритмы для ее реализации. В 3.2.4 предложены модификации представленных в 3.2.2 и 3.2.3 алгоритмов с применением метода потенциальных функций.

В данном разделе завершено решение второй задачи и решена третья задача диссертационного исследования, поставленных в 1.5.

3.1 Классификация участков по векторам признаков

В результате этапа сегментации, рассмотренного во втором разделе, исследуемая экспериментальная кривая представляет собой набор интервалов с чередованием простых (однородных) и сложных (переходных) участков. Далее следует этап присвоения выделенным участкам символов, соответствующих различным характерам поведения экспериментальной кривой. Реализация этого этапа анализа возможна несколькими способами:

1.Для анализа и обработки выбираются только сложные участки, которые характеризуют изменение состояния исследуемого процесса, фоновые возмущения некоторого постоянного состояния, или переход исследуемого процесса из одного состояния в другое.

2.Для анализа и обработки выбираются только простые участки, которые характеризуют фоновое состояние исследуемого процесса.

3.Анализируются и обрабатываются как сложные, так и простые участки.

Во всех данных случаях для составления лингвистического описания кривой необходим процесс присвоения символов каждому из анализируемых участков экспериментальной кривой. В первом случае выбранные участки рассматриваются отдельно. Во втором случае каждый простой участок показывает отдельное состояние процесса, тогда наборы простых участков рассматриваются отдельно, либо фоновые аномалии одного и того же процесса, тогда наборы простых участков рассматриваются как целое. В третьем же случае наборы простых и сложных участков не объединяются, а исследуются отдельно.

Набор присваиваемых символов представляет собой алфавит, в котором компоненты являются кодовыми обозначениями поведения кривой на каждом участке. Для формирования такого алфавита необходимо применять алгоритмы автоматической классификации, которые будут осуществлять распределение массивов векторов на классы, количество которых определяется самим алфавитом, и устанавливать критерии, по которым каждый новый вектор будет распределен в тот или иной класс, иными словами - присваивать им конкретные символы.

Классификация данных участков зависит от того, чем можно охарактеризовать анализируемые участки, т.е. их векторами признаков. Простейший вектор признаков это набор ординат экспериментальной кривой на анализируемом участке $f^j = (f_1^j, \ldots, f_l^j)$, но данный вариант может использоваться только в исключительном случае, когда вектор признаков, характеризующий исследуемые участки, имеет равную размерность на каждом участке. Например, при использовании алгоритмов сегментации с применением аппроксимации [143], переходные участки могут быть идентифицированы с перекрытием, т.е. шаг Δ меньше длины участка l. Следовательно, возникает задача составления набора векторов признаков для дальнейшего анализа участков разной длины l (Рисунок 7).

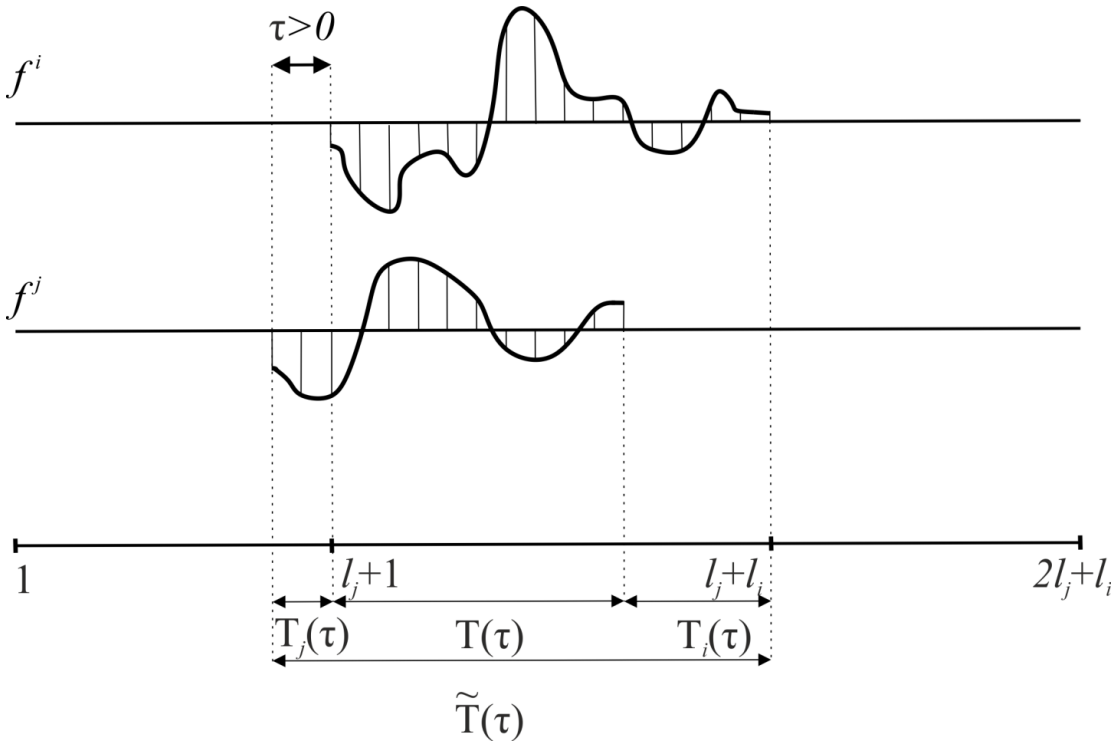

Рисунок 7 – Пример сравнения исследуемых участков, имеющих разную длину

На рисунке 7 показана возможность сравнения исследуемых участков разной длины, на котором:

$|\tau|$ – заданный фиксированный сдвиг, который должен быть больше, чем $\max(l_i, l_j)$;

$\tilde{T} = \{1,...,2l_j + l_i\}$ – множество индексов;

$\tilde{T}(\tau)$ – подмножество индексов в $\tilde{T}$, которым соответствует как минимум один индекс векторов f_i или f_j;

$T_i(\tau)$ – подмножество индексов в $\tilde{T}$, находящиеся в области определения векторов f_i и неопределенные значения f_j;

$T_j(\tau)$ – подмножество индексов в $\tilde{T}$, находящиеся в области определения векторов f_j и неопределенные значения f_i;

$T(\tau)$ – подмножество, компенсирующее $T_i(\tau)$ и $T_j(\tau)$ до множества $\tilde{T}(\tau)$.

Таким образом, длину участка *l*, выделенного алгоритмами сегментации, необходимо учитывать, как одну из важных его характеристик. Существенное отличие длин участков свидетельствует о том, что такие участки необходимо группировать в разные классы. Исходя из этого, следует, что степень отличия участков необходимо сориентировать на сравнение участков с относительно малыми отличиями по длине.

Введем некоторую степень $r(f^i, f^j)$, характеризующую расстояние для любых значений f^i, f^j и учитывающую отличие их длин. Используя данную метрику возможно классифицировать исследуемые участки в метрическом пространстве. В связи с тем, что в современном мире постоянно возрастает сложность технологических процессов, сложность новых научных теорий и результаты научных исследований в виде экспериментальных данных содержат десятки и сотни тысяч компонентов, возникает задача перехода от метрического пространства в координатное. Рассмотрим относительно малое количество из всех участков в метрическом пространстве, для которых построено достаточно большое количество k опорных (искусственных) участков $h^1,\ldots,h^k$, эвристически характеризующих их форму. Тогда вектор признаков $g^i=(g_1^i,\ldots,g_k^i)$ характеризует форму любого участка f^i. В качестве вектора признака принимается вектор его расстояний до опорных участков $g_p^i=r(f^i,\ldots,h_p)$. Такой вектор признаков g^i учитывает зависимость выбранной степени расстояния в пространстве форм исследуемых участков.

Исходя из вышесказанного, набор опорных участков $\{h^1,\ldots,h^k\}$ позволяет осуществить преобразование участков различной длины из метрического пространства X в *k*-мерное пространство Y, в котором будут применяться алгоритмы автоматической классификации. Поэтому важным этапом является определение способа задания метрики $r(f^i, f^j)$ для участков с различной длиной и алгоритмизация построения опорных участков $\{h^1,\ldots,h^k\}$ во введенном метрическом пространстве X.

3.1.1 Выбор степени отличия для участков различной длины

Степень отличия должна определять «похожими» участки, соответствующие однородным событиям. Такие однородные события приводят к появлению на экспериментальной кривой изменений приблизительно одинаковой формы. Сравнение форм двух участков необходимо производить методом их наложения друг на друга. При наложении необходимо учитывать время от начала данного события. В связи с этим, вводится некоторый сдвиг τ, позволяющий компенсировать неточности этапа сегментации. При этом компоненте f_t^i вектора f^i будет соответствовать компонента $f_{t+\tau}^j$ вектора f^j. Учитывая сказанное, алгоритм выбора степени отличия [146] для участков различной длины должен состоять из следующих шагов:

- выбор двух участков $f^i = (f_1^i, \ldots, f_{l_i}^i)$ и $f^j = (f_1^j, \ldots, f_{l_j}^j)$, где *li* и *lj* их длины;

- центрирование векторов: $\overline{f}^i = \frac{1}{l_i}\sum_{t=1}^{l_i} f_t^i = 0$ и $\overline{f}^j = \frac{1}{l_j}\sum_{t=1}^{l_j} f_t^j = 0$;

- выбор фиксированного сдвига $\tau = \max(l_i, l_j)$;

- выбор множества индексов $\tilde{T} = \{1,\ldots,2l_j + l_i\}$ (Рисунок 7) и сопоставление из него: вектору f^i индексы $s = l_j + 1,\ldots, l_j + l_i$; вектору f^j индексы $s = l_j + 1 - \tau,\ldots,2l_j - \tau$;

- примем степень расстояния $p(f^i, f^j, \tau)$ между векторами f^i и f^j за степень их отличия на всем множестве $\tilde{T}(\tau)$.

Очевидно, что различным значениям сдвига τ соответствуют различные значения степени отличия векторов f^i и f^j. Определять к одному ли классу относятся два исследуемых участка экспериментальной кривой нужно в таком их взаимном расположении, в котором они максимально совпадают по форме. Поэтому в качестве степени отличия участков разной длины следует принять величину $r(f^i, f^j) = \min_{\tau} p(f^i, f^j, \tau)$.

3.1.2 Алгоритм построения опорных участков векторной параметризации

Целью построения участков векторной параметризации является формирование векторного пространства признаков, в рамках которого и будет производиться распределение на классы участков анализируемых экспериментальных кривых. Исходя из этого, в массиве таких опорных участков необходимо наличие отображения разнообразия форм выделенных участков. Выберем часть множества векторов, отображающих разнообразие их форм $\{f^1,\ldots, f^n\}$, $f^j \in X$, и выберем количество опорных участков k формируемого пространства векторов ($k>>n$). Тогда $\{h^i,\ i=1,\ldots,k\}$ - множество участков в пространстве X, рассматриваемых как опорные, а критерий, выражающий разнообразие форм векторов множества $\{f^1,\ldots, f^n\}$ и опорных участков в X будет выбираться из условия минимума $L(h^1,\ldots,h^k)$:

$$L(h^1,\ldots,h^k)=\sum_{j=1}^{n}\min_i r\left(h^i, f^j\right), \qquad (78)$$

Любой массив $\{h^1,\ldots,h^k\}$ зависит от разделения множества $\{f^1,\ldots, f^n\}$ на k непересекающихся классов $A^1,\ldots, A^k$, в которых $f^j \in A^i$, при условии $r\left(h^i, f^j\right)=\min$ из всех опорных участков. Следовательно, задача построения опорных участков переходит в задачу разделения множества $\{f^1,\ldots, f^n\}$ на k классов, в которой (78) будет параметром качества такого разделения, но с необходимостью его минимизации. Учитывая сказанное, алгоритм построения опорных участков [146] векторной параметризации, показывающих разнообразие форм векторов $\{f^1,\ldots, f^n\}$ должен состоять из следующих шагов:

- выбор нескольких k векторов из множества $\{f^1,\ldots, f^n\}$, обозначение их $\{h^i,\ i=1,\ldots,k\}$;

- сопоставление каждому вектору h^i подмножества $A^i \subset \{f^1,\ldots, f^n\}$;

- расчёт расстояния $r\left(h^i, f^{k+1}\right)$ от вектора f^{k+1} до всех векторов h^i и определение ближайший, обозначив его h^{i^*};

- внесение вектора f^{k+1} в множество A^{i^*} и определение вектора $(h^{i^*})'(A^{i^*})$;

- расчёт расстояния $r(h^i, f^n)$ от вектора f^n.до всех h^{i*} и определение $i*$ с самым минимальным расстоянием;

- определение класса A^{i_n} для вектора f^n, перенос его в A^{i*} и пересчет эталонов изменившихся классов согласно правилу $h^{i_n} = (h^{i_n})'(A^{i_n})$, $h^{i*} = (h^{i*})'(A^{i*})$;

- векторы $h^1,\ldots,h^k$, полученные в результате минимизации, принимаются опорными участками векторной параметризации или же эталонами их классов.

Через конечное число циклов работы данного алгоритма будет достигнута устойчивая классификация, а именно для любого $f^j \in A^i$ расстояние r до эталона своего класса не будет превышать расстояние до других эталонов, что свидетельствует о том, что ни один вектор не будет перенесен из одного класса в другой.

3.2 Формирование языка описания экспериментальных кривых

В результате применения алгоритмов сегментации, предложенных в разделе 2 и классификации выделенных фрагментов, предложенных в 3.1, экспериментальная кривая оказывается представленной в виде упорядоченной последовательности символов конечного алфавита. Такую последовательность можно рассматривать как некоторый текст на неизвестном языке, в данном случае языке, специально приспособленном для описания обрабатываемых экспериментальных кривых. При этом каждую цепочку можно рассматривать как некоторую фразу на данном языке. С этой точки зрения задачу формирования языка описания данного массива экспериментальных кривых можно сформулировать следующим образом:

- задано некоторое количество текстов в виде множества упорядоченных последовательностей символов;

- необходимо сформулировать алфавит более крупных лингвистических единиц, чем отдельные символы, т. е. словарь слов, каждое из которых есть и том или ином смысле устойчивая цепочка символов;

- необходимо представить каждый из заданных текстов в виде одного или

нескольких слов этого словаря.

Пусть имеется всего один достаточно длинный текст $T = \langle a_1 \dots a_N \rangle$. Каждая упорядоченная пара индексов (i, j), $i \le j$, $1 \le i$, $j \le N$ вырезает из T некоторый отрезок последовательности, получающийся стиранием в T символов с индексами, меньшими i и большими j: $c(i, j) = \langle a_i, a_{i+1} \dots a_j \rangle$. Каждому отрезку соответствует образ — упорядоченная последовательность его символов, у которых отброшены индексы, учитывающие местоположение этого отрезка в тексте T. Очевидно, что в тексте может встретиться несколько отрезков, соответствующих одному образу. Необходимы найти такие наборы образов, из которых можно составить заданный текст T. Один из таких наборов и есть искомый словарь.

Разбиение D текста T на непересекающиеся отрезки задает некоторый словарь $M(D)$. Формирование словаря можно рассматривать как выделение набора макрособытий в развитии исследуемого процесса. Эти макрособытия отражены в символическом представлении кривых в виде стабильных цепочек символов. Например, в задаче распознавания устной речи такого рода слова можно отождествлять с фонемами. В этом случае выделение фонем в сигнале речи является не процессом сегментации этого сигнала, а процессом его лингвистического анализа. Такой анализ требует предварительной обработки непрерывного речевого сигнала, которая преобразует его в первичную последовательность символов, которые кодируют значения дифференциальных признаков или каких-либо других простейших структурных единиц информации речевого сигнала [147].

Пусть сформирован словарь M. Обозначим через L, множество всех конечных цепочек символов из алфавита A. Рассмотрим подмножество $\hat{L}(M) \subset L$ всех таких цепочек, каждая из которых представляет собой последовательность слов из словаря M. Подмножество цепочек $\hat{L}(M)$ является языком, полностью определяемым словарем M. Цепочка T принадлежит языку $\hat{L}(M)$, т.е. является его правильной цепочкой, тогда и только тогда, когда существует такое ее разбиение D, что образ каждого отрезка из D является в точности словом из M. При этом

всякое слово также является правильной цепочкой языка $\hat{L}(M)$. Язык, в котором любая последовательность слов из M является правильной цепочкой, будет являться языком с морфологической грамматикой. Для того чтобы морфологическую грамматику можно было использовать для анализа текстов, т.е. других экспериментальных кривых, необходимо иметь процедуру преобразования символьного текста в упорядоченный набор слов, или же процедуру разбиения текста на слова из словаря. Иными словами, необходима процедура распознавания правильности этого текста на языке $\hat{L}(M)$.

Экспериментальные кривые можно описать правильными текстами вполне строго. Задачи обработки кривых нацелены на автоматизацию анализа данных научных экспериментов, в которых смена состояний исследуемого процесса не имеет жесткой логической закономерности. Даже если алгоритмы сегментации и формирования алфавита используются для обработки высокоструктурированных сигналов (например, речевых), то и в этом случае алгоритмы сегментации и автоматической классификации вносят в формируемые ими последовательности символов существенный элемент случайности.

Морфологическую грамматику следует рассматривать лишь как модель, более или менее точно аппроксимирующую тексты, порождаемые экспериментальными кривыми. Чтобы ее использовать в такой роли, требуется дополнить механизм порождения правильных текстов над словарем некоторым искажающим механизмом, позволяющим порождать цепочки «не особо» отличающиеся от правильных. Для этого прежде всего необходимо задать некоторую меру сходства двух произвольных цепочек символов алфавита A. Она должна отражать степень искажении при переходе от одной из цепочек к другой. Роль искажающего механизма необходимо возложить на трансформационную грамматику, содержащую некоторое множество элементарных трансформаций, т.е. единичных искажений. Тогда в качестве меры сходства между двумя цепочками естественно принять минимальное число элементарных трансформаций, необходимых для перехода от одной цепочки к другой.

Назовем язык $\hat{L}(M)$ ядром нечеткого языка £,определяемого парой $\langle \hat{L}(M), Q \rangle$,

где Q — некоторая трансформационная грамматика. Пусть Q определяет меру отличия $r(\hat{T},T)$ произвольной цепочки $T \in L$ от ядерной цепочки $\hat{T} \in \hat{L}(M)$. Под степенью несоответствия ρ (T, £) данной цепочки T нечеткому языку £ понимается величина ρ (T, £)= $\min_{\hat{T}\in L(M)} r(\hat{T},T)$. Тогда задача анализа текста T может представляет собой нахождение для данной цепочки T последовательность слов $\langle m_1, m_2 \ldots, m_k \rangle = \hat{T}$ из словаря M такую, чтобы мера отличия T от ядерной цепочки $\hat{T}$ имела минимальное значение на множестве всех возможных ядерных цепочек, что является естественным аналогом задачи нахождения жесткого разбиения текста на слова. Полученная в результате ее решения последовательность слов принимается в качестве окончательного описания рассматриваемой экспериментальной кривой.

Введем две элементарные трансформации текста, а именно стирание и дописывание одного символа из алфавита A. Каждой паре цепочек $\langle T_1, T_2 \rangle$ сопоставим сеть $G(T_1, T_2)$, которая строится следующим образом:

- составление прямоугольной решетки, столбцы которой соответствуют символам первой цепочки T_1, а строки — символам второй цепочки T_2. Каждая клетка этой решетки является квадратом с ребром единичной длины;

- клеткам сопоставляются упорядоченные пары индексов, указывающие место соответствующих ей символов в цепочках: (i, j)-й клетке соответствует i -й по порядку символ в T_1 и j -й по порядку символ в T_2;

- выделение всех такие клеток, для каждой из которых символ, соответствующий ее строке, совладает с символом, соответствующим ее столбцу и проведение в каждой такой клетке дополнительного диагонального ребра, связывающего ее левую верхнюю вершину с правой нижней вершиной;

- все вертикальные и дополнительные диагональные ребра, связывающие соседние вершины, ориентируем стрелкой вниз, а горизонтальные стрелкой слева направо.

Множество X всех вершин такой решетки является множеством вершин графа искомой сети $G(T_1, T_2)$. Число таких вершин равно $q = (l_1 + 1)(l_2 + 1)$, где l_1 —

число символов в T_1, а l_2 — число символов в T_2, а множество Y всех построенных и ориентированных ребер — множество дуг этого графа. Истоком x_1 сети служит самая левая верхняя вершина, а стоком x_q — самая правая нижняя вершина.

В качестве меры сходства цепочек T_1 и T_2 принимается минимум из длин всех путей из множества $W(T_1, T_2)$:

$$r(T_1T_2) = \min l(S), \quad S \in W(T_1T_2). \tag{79}$$

Пусть теперь задано ядро $\hat{L}(M)$ и наблюдается цепочка $T \in L$ конечной длины, полученная при обработке некоторой экспериментальной кривой. Такая цепочка $\hat{T}$ является правильным описанием данной экспериментальной кривой на языке $\hat{L}(M)$, аппроксимирующем исходный текст T, если эта цепочка удовлетворяет условию:

$$r(\hat{T}, T) = \min r(T', T), \quad T' \in \hat{L}(M). \tag{80}$$

Ядро $\hat{L}(M)$ вместе со способом нахождения правильной аппроксимации произвольного текста конечной длины над алфавитом A назовем трансформационной морфологической грамматикой. Из описанного выше следует, что задача формирования такой грамматики сводится к формированию словаря M. Для этого необходимо осуществление следующих шагов:

- нахождение такого разбиения D текста T на отрезки, которое минимизирует расстояние (79) от $\hat{T}$ до T;

- разбиение множества отрезков D на классы похожих в смысле меры (79);

- сопоставление каждому классу похожих отрезков одной эталонной цепочки, ближайшей ко всем элементам класса. Набор этих эталонных цепочек и является искомым словарем M.

При фиксированном словаре M первый из этих шагов осуществляется с помощью описанного выше правила (80). Поскольку мера сходства двух произвольных цепочек определена (78), то на множестве отрезков D одним из алгоритмов автоматической классификации легко осуществляется второй шаг.

Третий шаг требует составления другого алгоритма формирования цепочки, ближайшей в смысле (79) к заданному множеству цепочек. При наличии такого алгоритма общую процедуру можно реализовать как повторяющийся процесс поочередного решения указанных трех шагов.

3.2.1 Сопоставление участков каждого класса к последовательности символов

Пусть $\{T_1,\ldots,T_n\}$ — некоторый массив цепочек. Разбором R этого массива является прямоугольная таблица, состоящая из n строк и некоторого числа k столбцов, каждая клетка которой либо пуста, либо содержит один символ из алфавита A, причем символы в j й строке, расположенные в порядке возрастания номеров столбцов, составляют цепочку T_j, а цепочка, составленная из символов в клетках одного столбца, является повторением одного символа. Запись двух цепочек T_{j_1} и T_{j_2} в некоторых двух строках разбора указывает способ перехода от одной цепочки к другой с помощью последовательности элементарных трансформаций. Их число равно числу столбцов, в которых есть символ в одной из этих строк и нет в другой. Пусть R - некоторый разбор, k - количество столбцов в нем, n - количество строк или количество цепочек в массиве. Каждый столбец разбора содержит некоторое число повторений одного и того же символа. Цепочку, образованную символами столбцов, обозначим как $T(R)$ и будем называть объединяющей цепочкой разбора R. Для каждого i-го столбца рассчитаем количество символов в нем $n_i(R)$ и частоту символа $p_i(R)=n_i(R)/n$ в i-м столбце. В цепочке $T(R)$ выделим символы, для которых $p_i(R)\geq 0{,}5$. Пусть это символы с порядковыми номерами $i_1,\ldots,i_a$. Цепочку, образованную этими символами, обозначим как $\tilde{T}(R)=\langle a_{i_1}\ldots a_{i_s}\rangle$ и будем считать собственной цепочкой разбора R. Качество разбора будем оценивать критерием, называемым неплотностью разбора:

$$J(R)=\sum_{i=1}^{k}\min[p_i(R),1-p_i(R)]. \qquad (81)$$

Нахождение цепочки, доставляющей минимальное значение функции

$$\phi(T)=\sum_{j=1}^{n} r(T,T_j) \qquad (82)$$

для данного массива цепочек, сводится к нахождению его разбора с минимальной неплотностью разбора (81).

Точное решение задачи нахождения разбора предъявленного массива цепочек с минимальной неплотностью разбора $J(R)$ найти непросто. В то же время объединяющая цепочка $T(R_n)=\langle a_1,\ldots,a_k\rangle$ вместе с частотами входящих в нее символов $p_1^n,\ldots,p_k^n$ соответствующая некоторому разбору R_n массива цепочек $\{T_1,\ldots,T_n\}$, дает значительную информацию о всем разборе. Это позволяет указать простой в вычислительном отношении способ включения в разбор новой цепочки T_{n+1} без изменения первых n строк таким образом, чтобы увеличение неплотности $J(R_{n+1})-J(R_n)$ было минимальным. Пусть $T(R_n)$ и $p_1^n,\ldots,p_k^n$ — объединяющая цепочка и вектор частот входящих в нее символов, соответствующие некоторому разбору R_n массива из n цепочек. Пусть T_{n+1} — новая цепочка, которую надо включить в разбор, построив тем самым новый разбор R_{n+1} массива из $n+1$ цепочек и соответствующую ему объединяющую цепочку $T(R_{n+1})$ с частотами символов p_i^{n+1}.

Построим сеть $G[T(R_n),T_{n+1}]$ трансформации $T(R_n)$ в T_{n+1}. Всякий путь на такой сети, ведущий из левой верхней в правую нижнюю вершину, определяет следующий способ дописывания $(n+1)$-й строки к разбору R_n: прохождение по горизонтальной дуге в клетке сети, расположенной под i-м символом цепочки $T(R_n)$, соответствует пустой клетке в новой строке, прохождение по диагональной дуге в клетке на пересечении i-го столбца и j-й строки — записыванию в i-ю клетку новой строки разбора j-го символа цепочки T_{n+1}, а прохождение по дуге в j-й строке на границе i-й и $(i+1)$-й клеток означает появление в разборе нового столбца, все клетки которого с номерами $1,\ldots,n$ пусты, а в $(n+1)$-й клетке записан

j-й символ цепочки T_{n+1}. И пусть выбран некоторый путь и тем самым способ дописывания строки. Обозначим через $\hat{\omega}$ подмножество номеров столбцов разбора R_{n+1}, соответствующих непустым клеткам $(n+1)$-й строки. Из общего определения неплотности разбора (81) следует, что $J(R_{n+1})$ равно:

$$J(R_{n+1}) = \sum_{i \in \hat{\omega}} \min\left(\frac{np_i^n + 1}{n+1}, \frac{n - np_i^n}{n+1} \right).$$

Построим новую сеть $G'[T(R_n), T_{n+1}]$,отличающуюся от $G[T(R_n), T_{n+1}]$ только системой весов на дугах того же графа: припишем всем вертикальным дугам сети одинаковые веса $1/(n+1)$, горизонтальным дугам в i-м столбце – веса, равные:

$$\min\left(\frac{np_i^n}{n+1}, \frac{n+1-np_i^n}{n+1} \right),$$

а диагональным дугам в j-м столбце:

$$\min\left(\frac{np_i^n + 1}{n+1}, \frac{n - np_i^n}{n+1} \right).$$

Длина всякого пути из левого верхнего в правый нижний узел новой сети в точности равна приращению неплотности разбора для соответствующего способа дописывания строки. Отсюда следует, что для выбора оптимального способа дописывания строки достаточно найти кратчайший путь. Частоты символов объединяющей цепочки $T(R_{n+1})$ нового разбора находятся следующим образом:

- для элементов, соответствующих столбцам предыдущего разбора R_n:

$$p_i^{n+1} = \frac{np_i^n + 1}{n+1};$$

- если в этом столбце есть символ цепочки T_{n+1}:

$$p_i^{n+1} = \frac{np_i^n}{n+1};$$

- если в столбце нет символа из : для элементов, соответствующих вновь образованным столбцам:

$$p_i^{n+1} = \frac{1}{n+1}.$$

Указанный метод оптимального дописывания новой $(n+1)$-й строки к уже

имеющемуся разбору массива из n цепочек является основой для построения алгоритмов для решения следующих двух шагов:

- накопление хорошего разбора заданного массива цепочек;

- улучшение уже имеющегося разбора путем оптимального изменения одной строки.

3.2.2 Алгоритм накопления разбора заданной последовательности символов

Алгоритм, осуществляющий накопление хорошего разбора заданной последовательности символов должен состоять из следующих шагов:

- задается таблица R_1, содержащая одну строку, в которой записана цепочка $T_1 . R_1$ и будет являться наилучшим разбором массива $\{T_1\}$, состоящего из единственной цепочки;

- к разбору R_1 оптимальным образом достраивается строка, содержащую цепочку T_2;

- полученный разбор обозначается как R_2;

- разбор R_2 достраивается до R_3 оптимальным добавлением цепочки T_3 и т. д., пока не будет получен разбор R_n;

- собственная цепочка $\tilde{T}(R_n)$ разбора R_n принимается в качестве приближенного решения задачи поиска цепочки с минимальным значением $\phi(T)$ (82).

Достоинством данного алгоритма является то, что он не требует наличия большой памяти, поскольку на каждом шаге достаточно хранить лишь объединяющую цепочку очередного разбора и частоты входящих в нее символов. Так же применение данного алгоритма находит хорошее начальное приближение для итерационной процедуры минимизации неплотности разбора, подробно рассмотренной в 3.2.3.

Так, как на очередном шаге алгоритма разбор, построенный на предыдущих шагах, остается без изменения, то уже после включения в него третьей по счету

цепочки значение $J(R_3)$ не будет, вообще говоря, минимальным и результат может зависеть от порядка предъявления цепочек, что является его недостатком. Данный алгоритм находит наилучшую цепочку, если массив состоит из не очень отличающихся цепочек. Таким образом, алгоритм хорошо работает именно в той ситуации, для которой он предназначен.

3.2.3 Алгоритм улучшения разбора заданной последовательности символов

Алгоритм, осуществляющий улучшение уже имеющегося разбора путем оптимального изменения одной строки должен состоять из следующих шагов:

- имеется R^0 – разбор всего заданного массива из n цепочек, построенный, например, в результате работы алгоритма, представленного в 3.2.2. $T(R^0)$ и $p_1^0,\dots p_{k_0}^0$ – объединяющая цепочка и частоты входящих в нее символов данного разбора;

- выбирается одна строка разбора с номером j;

- через $\hat{\omega}_j$ обозначается множество номеров столбцов, в которых клетки j-ой строки не пусты;

- из разбора вычеркиваются такие строки;

- получается некоторый разбор R_j^0 массива цепочек $\{T_1,\dots,T_{j-1},T_{j+1},\dots,T_n\}$ с той же объединяющей цепочкой $T(R_j^0)=T(R^0)$, частоты символов которой равны (частоты некоторых символов могут оказаться нулевыми):

- при $i\in\hat{\omega}_j$:

$$p_i(j)=\frac{np_i^0-1}{n-1};$$

- при $i\notin\hat{\omega}_j$:

$$p_i(j)=\frac{np_i^0}{n-1};$$

- из $T(R_j^0)$ вычеркиваются нулевые символы и соответствующие им пустые столбцы из R_j^0;

- строится разбор R^1 путем оптимального дописывания ранее вычеркнутой цепочки T_j к разбору R_j^0. Ему соответствует объединяющая цепочка $T\left(R_j^0\right)$ с частотами $p_1^1,\ldots,p_{k_1}^1$.

Для разбора R^1 выполняется неравенство $J\left(R_0\right)-J\left(R^1\right)=\left[J\left(R^0\right)-J\left(R_j^0\right)\right]-\left[J\left(R^1\right)-J\left(R_j^0\right)\right]\geq 0$, причем уменьшение неплотности при переходе от R^0 и R^1 максимально среди всех R^1, отличающихся от R^0 одной j- ой строкой.

3.2.4 Алгоритмы разбора на основе метода потенциальных функций

Два алгоритма, представленные в 3.2.2 и 3.2.3 конечны и позволяют построить следующий алгоритм минимизации неплотности разбора из n цепочек:

- алгоритмом из 3.2.2 строится исходный разбор R заданного массива цепочек;

- присваивается $q:=0;\ j:=1;$

- алгоритм из 3.2.3 применяется к j-й строке разбора R и если строка изменилась, то присваивается $q:=q+1;$

- при $j<n$, то $j:=j+1$ и повторяется предыдущих шаг, в другом случае переход к следующему шагу;

- при $q:=0$, процедура закончена, при $q\neq 0$ переход ко второму шагу.

С помощью данного алгоритма за конечное число циклов находится такой разбор R^* предъявленного массива цепочек, который нельзя улучшить изменением никакой отдельно взятой строки. Данный алгоритм не гарантирует отсутствие другого разбора, отличающегося от R^* более чем одной строкой, для которого значение неплотности меньше, однако алгоритм находит достаточно глубокий минимум $J(R)$.

Описанный выше алгоритм предлагается модифицировать в алгоритм классификации множества фрагментов некоторой последовательности символов, который одновременно находит разбиение этого множества на классы похожих, и

каждый класс снабжает представляющим его эталоном. Данный алгоритм можно модифицировать, основываясь на алгоритмах индексации и объединения метода потенциальных функций [148]. Такой алгоритм на основе алгоритма индексации должен состоять из следующих шагов:

- имеется построенное множество разборов $\{R_1^t,\ldots,R_q^t\}$ на t-м шаге, где q - число слов в формируемом словаре;

- обозначается совокупность цепочек этих разборов через $\{\hat{T}_1^t,\ldots,\hat{T}_q^t\}$;

- рассматривается новый отрезок T^{t+1} и находится такую цепочку $\hat{T}_s^t$, расстояние которой до T^{t+1} по $G'(\hat{T}_s^t,T^{t+1})$ минимально $r(\hat{T}_s^t,T^{t+1})=\min\limits_k r(\hat{T}_k^t,T^{t+1})$;

- к соответствующему разбору R_s^t присоединяется T^{t+1} способом, описанным выше в алгоритме 3.2.2, а именно: $\hat{T}_s^t$ преобразуется в $\hat{T}_s^{t+1}$ и соответствующим образом пересчитываются частоты символов в старом разборе;

- получаем разбор R_s^{t+1}, а остальные разборы остаются без изменения $R_k^{t+1}=R_k^t, k\neq s$ и вместе с ними, неизменными остаются и соответствующие объединяющие цепочки и частоты их символов;

- переход к следующему отрезку.

Для составления алгоритма на основе алгоритма объединения можно использовать шаг перестройки $T(R_n)$ в $T(R_{n+1})$ алгоритма 3.2.2 для построения по объединяющим цепочкам $T(R^1)$ и $T(R_2)$ двух разборов R^1 и R^2, состоящих соответственно из n^1 и n^2 строк, объединяющей цепочки $T(R^{12})$ наилучшего набора этих разборов R^{12}. Для этого необходимо изменить правила расчета весов сети $G'[T(R^1),T(R_2^2)]$. А именно:

- диагональной дуге в клетке (i,j) приписывается вес:

$$\min\left(\frac{n^1p_i^1+n^2p_j^2}{n^1+n^2},\frac{n^1+n^2-n^1p_i^1-n^2p_j^2}{n^1+n^2}\right);$$

- горизонтальным дугам в i-м столбце приписываются веса:

$$\min\left(\frac{n^1 p_i^1}{n^1+n^2}, \frac{n^1+n^2-n^1 p_i^1}{n^1+n^2}\right);$$

- вертикальным дугам в j-й строка приписываются веса:

$$\min\left(\frac{n^2 p_i^2}{n^1+n^2}, \frac{n^1+n^2-n^2 p_i^2}{n^1+n^2}\right).$$

В таком случае частоты символов объединяющей цепочки $T\left(R^{12}\right)$ нового разбора вычисляются следующим образом:

- для элементов, соответствующих общим столбцам i и j разборов R^1 и R^2

$$p = \frac{n^1 p_i^1 + n^2 p_j^2}{n^1+n^2};$$

- для i-го столбца R^1, которому не соответствует никакой столбец R^2:

$$p = \frac{n^1 p_i^1}{n^1+n^2};$$

- для аналогичных столбцов R^2:

$$p = \frac{n^2 p_j^2}{n^1+n^2}.$$

Матрица $\left\|J\left(R^{ij}\right)\right\|$ неплотностей разборов всех возможных пар цепочек определяется из D. В соответствии с модификацией алгоритма объединение легко строится дерево набора наиболее плотных разборов последовательной перестройкой их объединяющих цепочек и рекуррентным пересчетом матрицы неплотностей разборов их пар.

3.3 Выводы по разделу

В третьем разделе решалась задача алгоритмической реализации процедуры структурного анализа экспериментальных кривых и задача разработки методов лингвистического описания экспериментальных кривых. Полученный в результате классификации алфавит символов будет являться классификатором необычных явлений в ходе анализируемого процесса. При применении алгоритма построения опорных участков через конечное число циклов работы будет достигнута

устойчивая классификация, а именно для любого $f^j \in A^i$ расстояние r до эталона своего класса не будет превышать расстояние до других эталонов, что свидетельствует о том, что ни один вектор не будет перенесен из одного класса в другой. Разбиение полученного текста на непересекающиеся отрезки задает некоторый словарь, формирование которого можно рассматривать как выделение набора макрособытий в развитии исследуемого процесса. Эти макрособытия отражены в символическом представлении кривых в виде стабильных цепочек символов. Экспериментальные кривые можно описать правильными текстами вполне строго. Обработка кривых нацелена на автоматизацию анализа данных научных экспериментов, в которых смена состояний исследуемого процесса не имеет жесткой логической закономерности. Морфологическую грамматику следует рассматривать лишь как модель, более или менее точно аппроксимирующую тексты, порождаемые экспериментальными кривыми. Для этого задается некоторая мера сходства двух произвольных цепочек символов алфавита, которая отражает степень искажении при переходе от одной из цепочек к другой, а роль искажающего механизма необходимо возложить на трансформационную грамматику, содержащую некоторое множество элементарных трансформаций, т.е. единичных искажений. Тогда в качестве меры сходства между двумя цепочками принимается минимальное число элементарных трансформаций, необходимых для перехода от одной цепочки к другой.

Указанный в 3.2.1 метод дописывания новой $(n+1)$-й строки к уже имеющемуся разбору массива из n цепочек является основой для построения алгоритмов для накопления хорошего разбора заданного массива цепочек и улучшения уже имеющегося разбора путем оптимального изменения одной строки. Достоинством алгоритма накопления разбора, представленного в 3.2.2, является то, что он не требует наличия большой памяти, поскольку на каждом шаге достаточно хранить лишь объединяющую цепочку очередного разбора и частоты входящих в нее символов. Так же применение данного алгоритма находит хорошее начальное приближение для итерационной процедуры минимизации неплотности разбора. С помощью алгоритма, представленного в 3.2.3 за конечное число циклов находится

такой разбор R^* предъявленного массива цепочек, который нельзя улучшить изменением никакой отдельно взятой строки. Предложены модификации описанных в 3.2.2 и 3.2.3 алгоритмов, которые одновременно находят разбиение множества фрагментов на классы похожих, и каждый класс снабжает представляющим его эталоном.

РАЗДЕЛ 4 ПРИМЕНИМОСТЬ СИСТЕМЫ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ПРИ РЕШЕНИИ ПРАКТИЧЕСКИХ ЗАДАЧ

Данный раздел посвящен исследованию практической применимости разработанной автоматизированной системы структурного анализа экспериментальных кривых и последнему этапу обработки кривых - анализу полученных последовательностей символов. В 4.1 исследована применимость разработанных алгоритмов сегментации к выделению переходных участков на кривой акустических колебаний, в 4.2 исследована применимость алгоритмов лингвистического описания или присвоения символов участкам исследуемых кривых. В 4.3 исследована применимость разработанной системы лингвистического анализа к анализу экспериментальных данных ЭКГ, а в 4.4 - к анализу спектрограмм радиочастот FM диапазона.

В данном разделе решена четвертая задача диссертационного исследования, поставленная в 1.5, а именно – исследование применимости разработанной системы структурного анализа в процессах электромагнитной совместимости радиоэлектронных средств.

4.1 Исследование применимости алгоритмов сегментации к выделению переходных участков кривой акустических колебаний

Для исследования применимости алгоритмов сегментации рассмотрим общую методику сегментации на основе функций сложности, предложенную в 2.1, а именно:

- экспериментальная кривая $f(t)$ делится на ряд элементарных участков $\omega_j, j=1,\ldots,N$ одинаковой длины l, которые следуют с определенным шагом Δ вдоль оси изменения аргумента;

- выбирается определенная функция сложности $\phi(f,\omega)$, при этом каждый элементарный участок ω_j связан с реальной величиной в виде $\phi_j = \phi(\omega_j)$;

- выделяются сложные участки, т.е. участки с локально экстремальными значениями ϕ_j (максимальным или минимальным, в зависимости от выбранной функции сложности).

В зависимости от выбора функции сложности [149], сложные участки являются локальными экстремумами. Для выделения таких участков применим два различных алгоритма сегментации, предложенных в 2.2:

Первый, основанный на оценке подобия граничащих участков кривой:

- исследуемая экспериментальная кривая разбивается на участки одинаковой длины;

- исходные векторы значений ординат кривой, соответствующие участкам, нормируются;

- на кривой определяется последовательность значений функции сложности (62);

- на кривой выделяются все участки, соответствующие локально минимальным значениям функции сложности;

- выделенные участки рассматриваются как искомые переходные участки;

- для каждого такого участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

Второй, основанный на исключении переходных участков кривой:

- исследуемая экспериментальная кривая разбивается на участки, каждый из которых представлен вектором значений ее ординат;

- для каждого участка определяется набор из первых $2m+1$ коэффициентов Фурье;

- для каждого участка определяется значение функции сложности, характеризующее качество аппроксимации кривой на участке (64);

- выделяются участки, соответствующие условию $\phi(\omega_{j\pm1}) > \phi(\omega_j)$;

- выделенные участки рассматриваются как искомые переходные участки;

- для каждого такого участка фиксируется его позиция на кривой и соответствующий вектор, характеризующий его форму.

В качестве экспериментальных данных для проведения сравнительного исследования работы двух вышеуказанных алгоритмов была использована кривая акустических колебаний в металлическом цилиндре [150], возбуждаемых ударным методом, полученная в ГОУ ВПО «Донецкий национальный университет» на лабораторном макете. Удар и съем осуществлялся по оси вращения цилиндра. Исследуемая экспериментальная кривая (Рисунок 8) содержала 5000 точек отсчета.

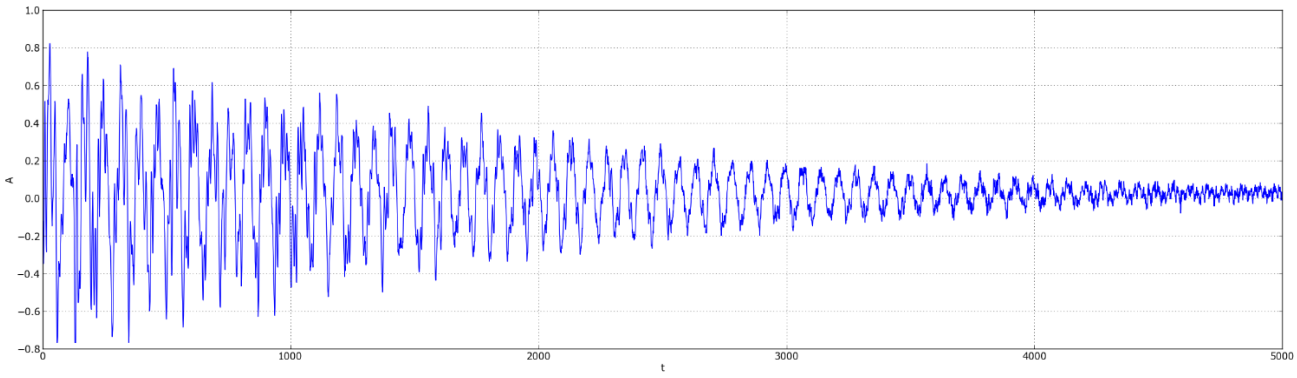

Рисунок 8 – Экспериментальная кривая акустических колебаний в металлическом цилиндре

Два алгоритма сегментации были реализованы на высокоуровневом языке программирования общего назначения Python и применены к анализу выделения переходных участков кривой акустических колебаний. Исследование содержало два эксперимента.

Первый эксперимент. Исследуемая кривая разбивается на 100 непересекающихся участков, каждый в 50 точек отсчета. Последовательность действий первого алгоритма сегментации, основанного на оценке подобия граничащих участков, описана далее. Сперва предварительно разбитый массив участков нормировался. Нормировка заключалась в делении каждого вектора значений на величину $\left\|g^j\right\| = \sqrt{\sum_{i=1}^{l} \left| f_i^j \right|^2}$, вычисленную для данного участка. Величина $\left\|g^j\right\|$ также имеет название метрика L2, норма или евклидова норма и является геометрическим расстоянием между двумя точками в многомерном пространстве. Затем, для каждого участка вычислялось значение функции сложности

$\phi(\omega_j)=\frac{1}{2}\left(g^j, g^{j-1}+g^{j+1}\right)$, являющейся средним значением скалярного произведения вектора g^j с граничащими предварительно нормированными векторами этой же кривой g^{j-1} и g^{j+1}. Значения функции сложности для первого и последнего участка, граничащих всего с одним участком, вычислялись как скалярные произведения: $\phi(\omega_1)=\left(g^1, g^{j+1}\right)$ и $\phi(\omega_{end})=\left(g^{end-1}, g^{end}\right)$ соответственно. Далее алгоритмом выделялись участки, соответствующие локальным минимумам, а именно участки, значения функции сложности $\phi(\omega_j)$ которых принимали значения менее -0.4

Второй алгоритм сегментации, основанный на исключении переходных участков, аналогично предыдущему, сперва нормировал предварительно разбитый массив участков. Нормировка так же заключалась в делении каждого вектора значений на евклидову норму данного участка. Затем для каждого участка вычислялся набор коэффициентов Фурье. Далее для каждого участка вычислялось значение функции сложности $\phi(\omega_j)=\sqrt{\sum_{i=1}^{n}\left[g_i^j-\left(\frac{C_0^j}{2}+\sum_{s=1}^{m}C_{s1}^j\cos(2\pi s\frac{i}{n})+\sum_{s=1}^{m}C_{s2}^j\sin(2\pi s\frac{i}{n})\right)\right]^2}$, являющейся характеристикой качества аппроксимации кривой на участке рядом Фурье, где n-количество точек на участке, m принимает значения от 1 до n, а s от 1 до m. После этого алгоритм выделял участки, значения функции сложности $\phi(\omega_j)$ которых отличались от граничащих на заданную величину. Графики функций сложности, полученные в результате работы двух алгоритмов представлены на рисунках 9 и 10.

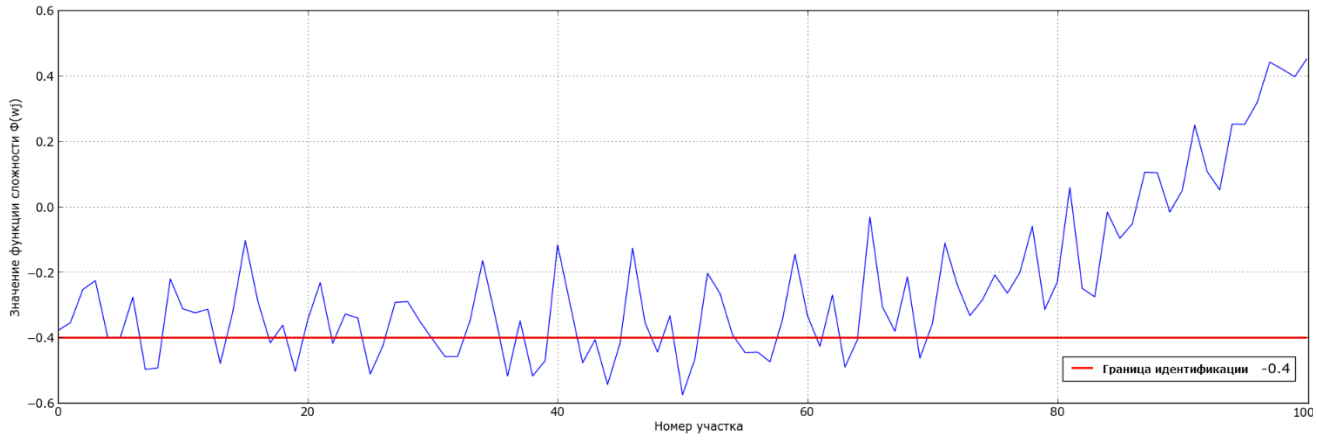

Рисунок 9 – График функций сложности при применении первого алгоритма

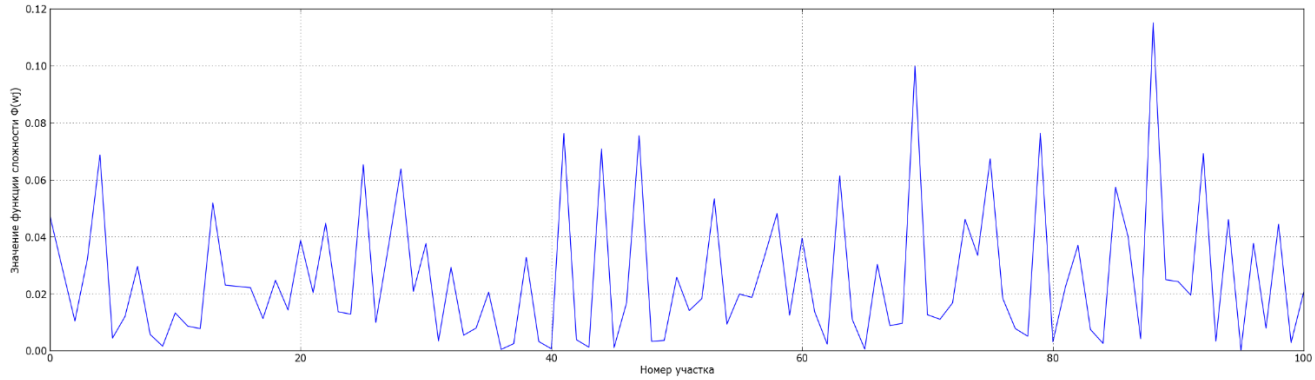

Рисунок 10 – График функций сложности при применении второго алгоритма

Первым алгоритмом выделено 29 переходных участков, вторым - 31. Следовательно, данный эксперимент показывает сходства в результате применения предложенных алгоритмов, несмотря явные отличия способов вычисления функции сложности. Примеры выделенных участков показаны на рисунке 11.

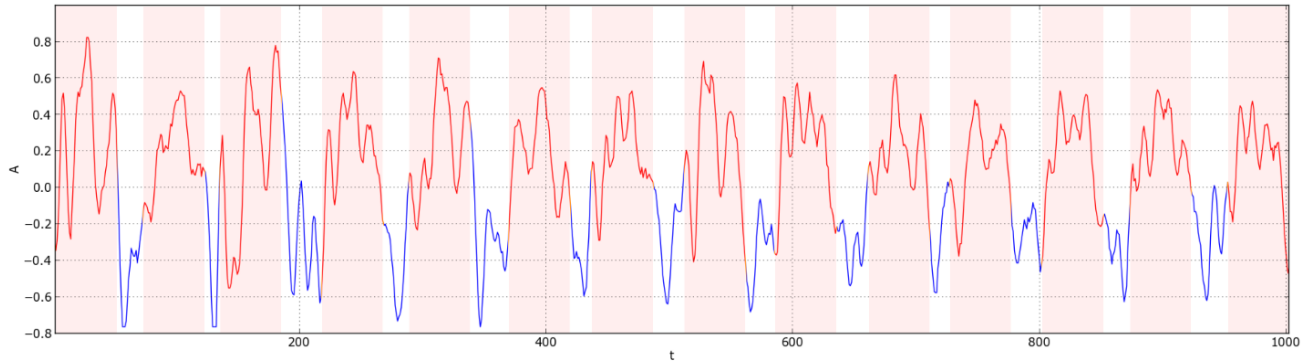

Рисунок 11 – Примеры выделенных участков в первом эксперименте

Второй эксперимент. Исследуемая кривая разбивается на 500 участков, каждый в 10 точек отсчета, с шагом в 5 точек. Таким образом граничащие участки имели перекрытие размером в половину длины участка.

Первым алгоритмом выделено 114, вторым - 119 переходных участков. Данный эксперимент так же показал наличие сходства в результате применения предложенных алгоритмов. Примеры выделенных участков показаны на рисунке 12.

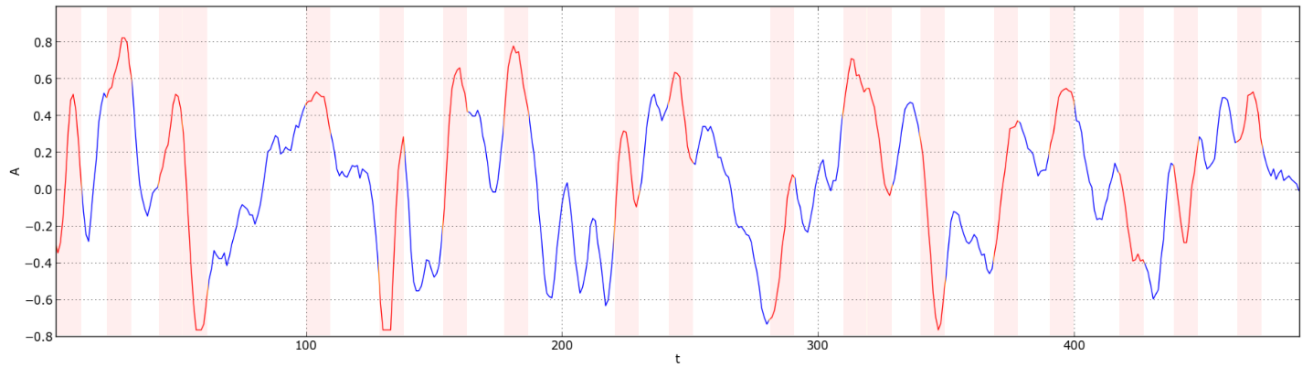

Рисунок 12 – Примеры выделенных участков во втором эксперименте

Проведенные эксперименты показали, что выделенные участки непосредственно и «на глаз» являются более сложными, чем остальные. Исходя из этого, можно заключить что предложенные алгоритмы с применением функции сложности целесообразно применять к задачам сегментации переходных участков экспериментальных кривых акустических колебаний. Первый эксперимент показал сходства в результаты работы обеих алгоритмов. Второй эксперимент так же подтвердил наличие сходства, а также показал, что разбиение на большее количество участков и применение перекрытия приводит к увеличению точности выделения.

4.2 Исследование применимости алгоритмов лингвистического описания участков экспериментальных кривых

В процессе формирования лингвистического описания, в текст могут вноситься искажения. Такие искажения могут быть следующих видов:

- некоторая локальная особенность на кривой, соответствующая элементарному событию в составе исследуемого процесса может быть не выделена алгоритмам сегментации, что соответствует потере символа в тексте;

- алгоритмом сегментации может быть случайно выделен фрагмент кривой, который не соответствует никакому событию, что равносильно появлению в цепочке символов лишнего символа;

- в процессе присваивания символов выделенным фрагментам кривой возможны искажения, в результате которых фрагменту, соответствующему в действительности одному событию, будет приписан символ другого события.

Такие искажения могут иметь место в силу большой вариабельности формы кривой на выделенных фрагментах. Они соответствуют замене символа в цепочке на другой символ. Такую замену удобно рассматривать как результат последовательного применения двух первых трансформаций — стирания ошибочного символа и дописывания вслед за ним правильного. Далее на примерах рассмотрены построение трансформационной грамматики для учета специфики структуры экспериментальных кривых и метод построения цепочки ближайшей к заданному множеству.

4.2.1 Построение трансформационной грамматики

Рассмотрим на примере элементарные трансформации текста, такие как стирание и дописывание одного символа из алфавита A, представленные в 3.2. Пусть $T_1 = \langle pqpp \rangle$ и $T_2 = \langle ppqp \rangle$. Каждой паре цепочек $\langle T_1, T_2 \rangle$ сопоставим сеть $G(T_1, T_2)$.

Каждый путь S, ведущий из истока в сток сети $G(T_1, T_2)$, порождает цепочку элементарных трансформаций, переводящую T_1 в T_2. При этом движение по левой вертикали клетки (i, j) означает, что между последним символом уже трансформированной части цепочки T_1 и первым символом еще не трансформированной ее части (в данном случае это символ, стоящий в T_1 на i-м месте) вставляется символ, стоящий в T_2 на j-м месте. Движение по верхней горизонтали этой клетки означает, что в T_1 требуется удалить символ, стоящий на i-м месте. Движение по диагонали означает, что символ, стоящий в T_1 на i-м месте, остается в выстраиваемой цепочке без изменения.

Зададим на множестве дуг сети $G(T_1, T_2)$ систему весов, а именно: каждой вертикальной и горизонтальной дуге припишем вес 1, а каждой диагональной дуге припишем вес 0. Длину пути $l(S)$ определим как сумму весов всех дуг, лежащих

на этом пути. В качестве меры сходства цепочек T_1 и T_2 примем минимум из длин всех путей из множества (79). Сеть $G(T_1, T_2)$ соответствующая паре цепочек $T_1 = \langle pqpp \rangle$ и $T_2 = \langle ppqp \rangle$ представлена на рисунке 13.

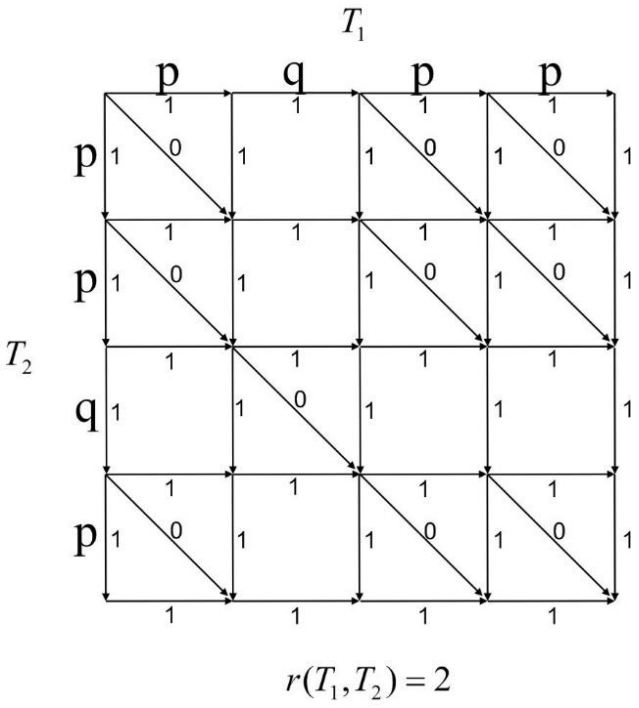

Рисунок 13 – Сеть трансформации $G(T_1, T_2)$ цепочки $\langle pqpp \rangle$ в цепочку $\langle ppqp \rangle$

Теперь зададим ядро $\hat{L}(M)$ и рассмотрим цепочку $T \in L$ конечной длины, полученную при обработке исследуемой экспериментальной кривой. Правильным описанием данной экспериментальной кривой на языке $\hat{L}(M)$, аппроксимирующем исходный текст T, является цепочка удовлетворяет условию (80). Идея нахождения правильного описания для данного текста полностью раскрывается на примере, изображенном на рисунке 14.

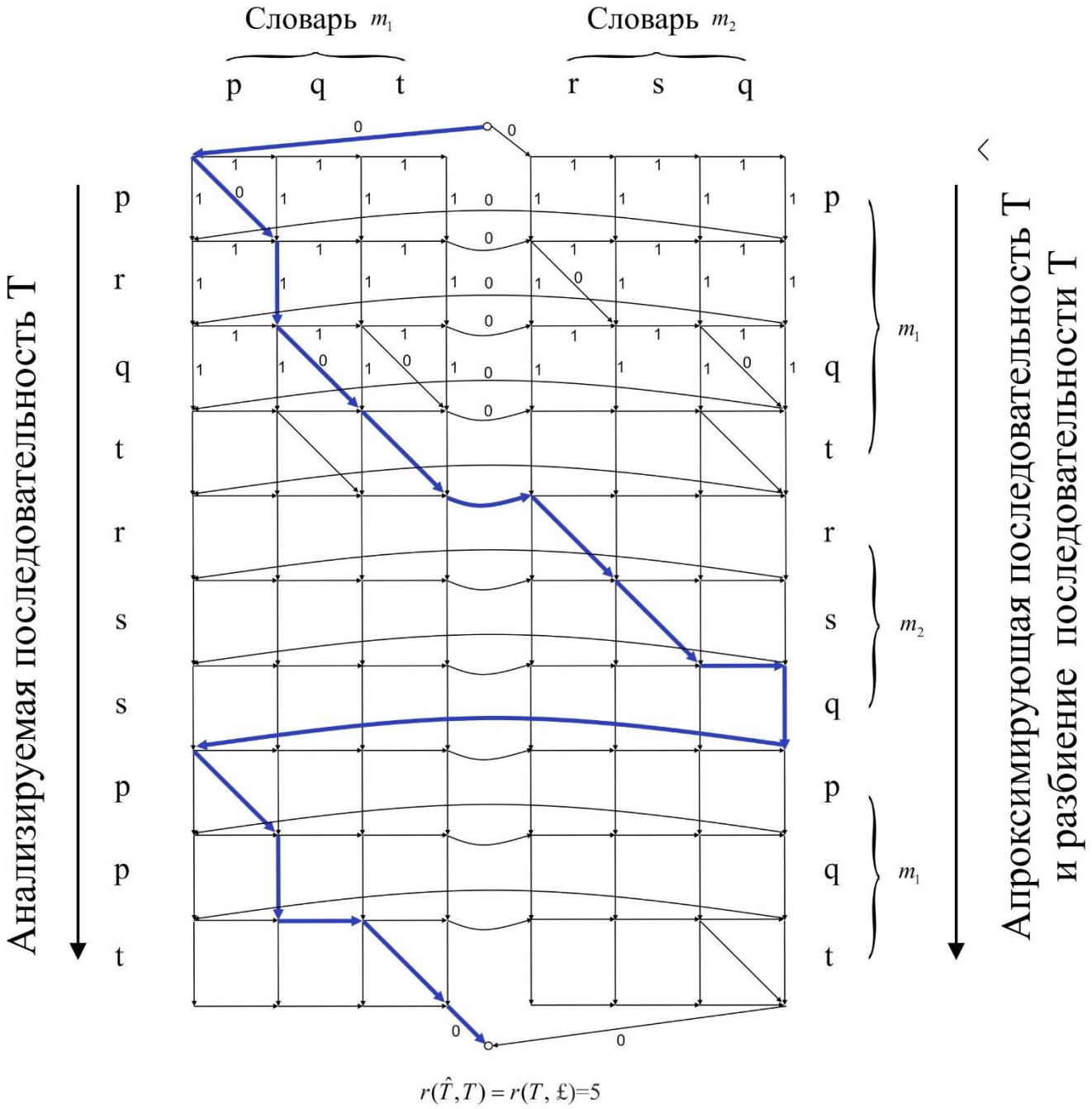

Рисунок 14 – Разбиение цепочки в соответствии с заданным словарем

Пусть словарь M состоит из двух слов: $m_1 = \langle pqt \rangle$, $m_2 = \langle rsq \rangle$, $A = \{p, q, r, s, t\}$. Для анализа предъявлена цепочка $\langle prqtrsppt \rangle$, состоящая из 10 символов. Построим две прямоугольные решетки для каждого слова отдельно точно так же, как это делалось при построении сети $G(T_1, T_2)$. При этом вершины клеток понимаются как вершины графа, а ребра, ориентированные соответствующим образом как его дуги.

Каждой дуге припишем вес, равный единице. Такой граф, построенный на основе двух решеток, является несвязным. Добавим в него еще ряд дуг, имеющих нулевые веса, а именно:

- диагональные дуги в клетках, сверху и слева от которых стоят одинаковые символы;

- дуги, соединяющие вершины крайнего правого вертикального ряда каждого слова с вершинами крайнего левого вертикального ряда других слов, расположенных в том же горизонтальном ряду (кроме верхнего и нижнего рядов);

- дуги, ведущие из специальной входной вершины (истока) в верхние вершины левых вертикальных рядов каждого из слов, а также из нижних вершин правых рядов в выходную вершину (сток).

Входную и выходную вершины полученного графа связывают все те и только те пути, по которым предъявленная цепочка T может быть получена в соответствии с правильными последовательностям трансформаций из всевозможных последовательностей слов данного словаря, а длина пути равна числу необходимых при этом элементарных трансформаций и, следовательно, значению меры различия соответствующей правильной цепочки $\hat{T}$ ядра языка $\hat{L}(M)$ и анализируемой цепочки T. Кратчайший путь дает морфологический анализ предъявленной цепочки T в виде последовательности слов, а длина этого пути дает степень ее несоответствия $r(\hat{T},T)$ правильному описанию $\hat{T}$.

4.2.2 Метод построения цепочки, ближайшей к заданному множеству

Рассмотрим на примере метод построения цепочки, ближайшей к заданному множеству цепочек, представленный в 3.2.1. Пусть $\{T_1,\ldots,T_n\}$ — некоторый массив цепочек. Разбором R этого массива называется прямоугольная таблица, состоящая из n строк и некоторого числа k столбцов, каждая клетка которой либо пуста, либо содержит один символ из алфавита A причем символы в j й строке, расположенные в порядке возрастания номеров столбцов, составляют цепочку T_j, а цепочка, составленная из символов в клетках одного столбца, является

повторением одного символа. Например, для массива из трех цепочек $\{\langle qrp\rangle,\langle qpr\rangle,\langle sqpr\rangle\}$ могут быть записаны, в частности, два разбора, показанные на рисунке 15.

	q	r	p	
	q		p	r
s	q		p	r

m_1

			q	r	p
	q	p		r	
s	q	p		r	

m_2

Рисунок 15 – Разбиение цепочки в соответствии с заданным словарем

Запись двух цепочек T_{j_1} и T_{j_2} в некоторых двух строках разбора указывает способ перехода от одной цепочки к другой с помощью последовательности элементарных трансформаций. Их число равно числу столбцов, в которых есть символ в одной из этих строк и нет в другой. Пусть R – некоторый разбор, k – число столбцов в нем, n – число строк (число цепочек в массиве). Каждый столбец разбора содержит некоторое число повторений одного и того же символа. Цепочка, образованную символами столбцов, обозначим как $T(R)$ и будет объединяющей цепочкой разбора R.

Для каждого i-го столбца подсчитаем число символов в нем $n_i(R)$ и величину $p_i(R)=n_i(R)/n$, называемую частотой символа в i-м столбце. В цепочке $T(R)$ выделим символы, для которых $p_i(R)\geq 0{,}5$. Цепочку, образованную этими символами, обозначим как $\tilde{T}(R)=\langle a_{i_1}\ldots a_{i_s}\rangle$ и будем считать собственной цепочкой разбора R. Для приведенных на рисунке 15 двух примеров разбора:

$$T(R_1)=\langle sqrpr\rangle,\ T(R_2)=\langle sqpqrp\rangle,$$

$$\tilde{T}(R_1)=\langle qp\rangle,\ \tilde{T}(R_2)=\langle qpr\rangle.$$

Рассмотрим оптимальное дописывание нового столбца к построенному разбору, показанного на рисунке 16.

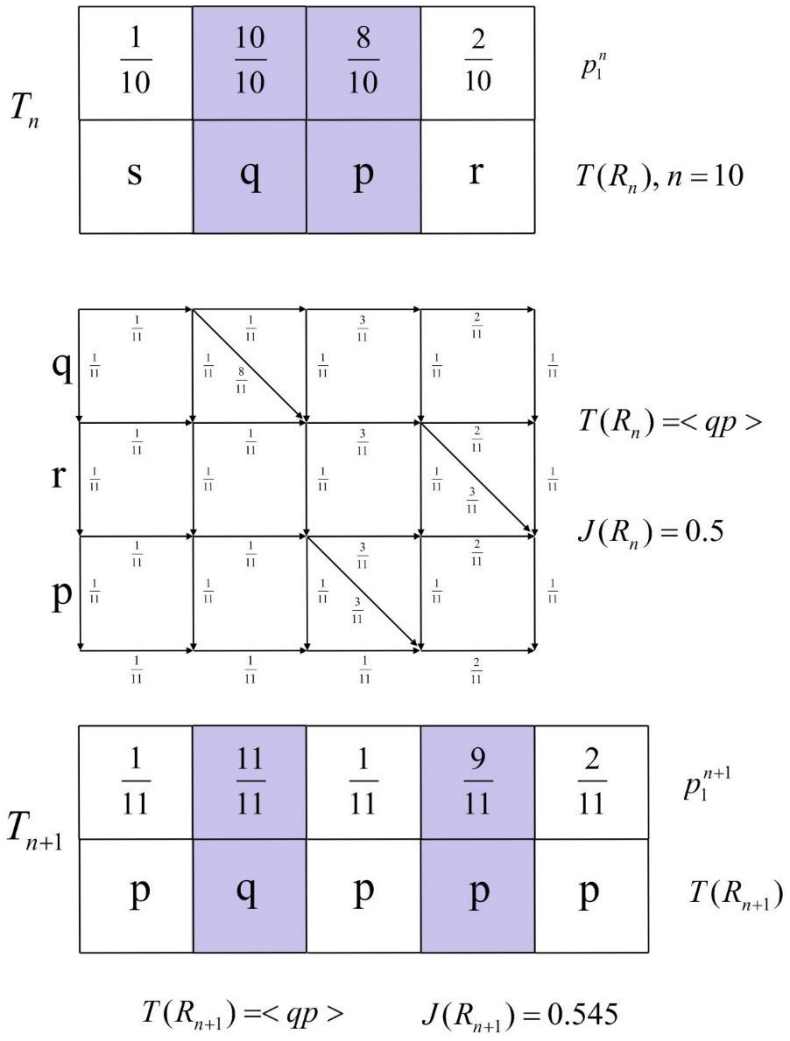

Рисунок 16 – Дописывание новой строки к построенному разбору

На рисунке 16 в закрашенных столбцах обозначены символы с частотами $p_i \geq 0{,}5$, образующие собственные цепочки $\tilde{T}(R_n)$ и $\tilde{T}(R_{n+1})$. Для понимания правил вычисления весов дуг сети эти веса, а также частоты символов объединяющих цепочек $T(R_n)$ и $T(R_{n+1})$ представлены в виде правильных дробей.

4.3 Исследование применимости автоматизированной системы лингвистического анализа к анализу экспериментальных данных ЭКГ

Предлагаемые алгоритмы были реализованы на высокоуровневом языке программирования общего назначения Python и применены к задаче присвоения символов анализируемым участкам экспериментальных кривых [151, 152]. В качестве экспериментальных данных были выбраны записи ЭКГ, с использованием одной из схем усиления биопотенциала из лабораторного сеанса длительностью 16 минут на частоте дискретизации 100 Гц, находящиеся в свободном доступе на электронном ресурсе [153]. Экспериментальные данные ЭКГ содержали более 95000 точек отсчета. Сегмент экспериментальной кривой, построенной по этим данным изображен на рисунке 17.

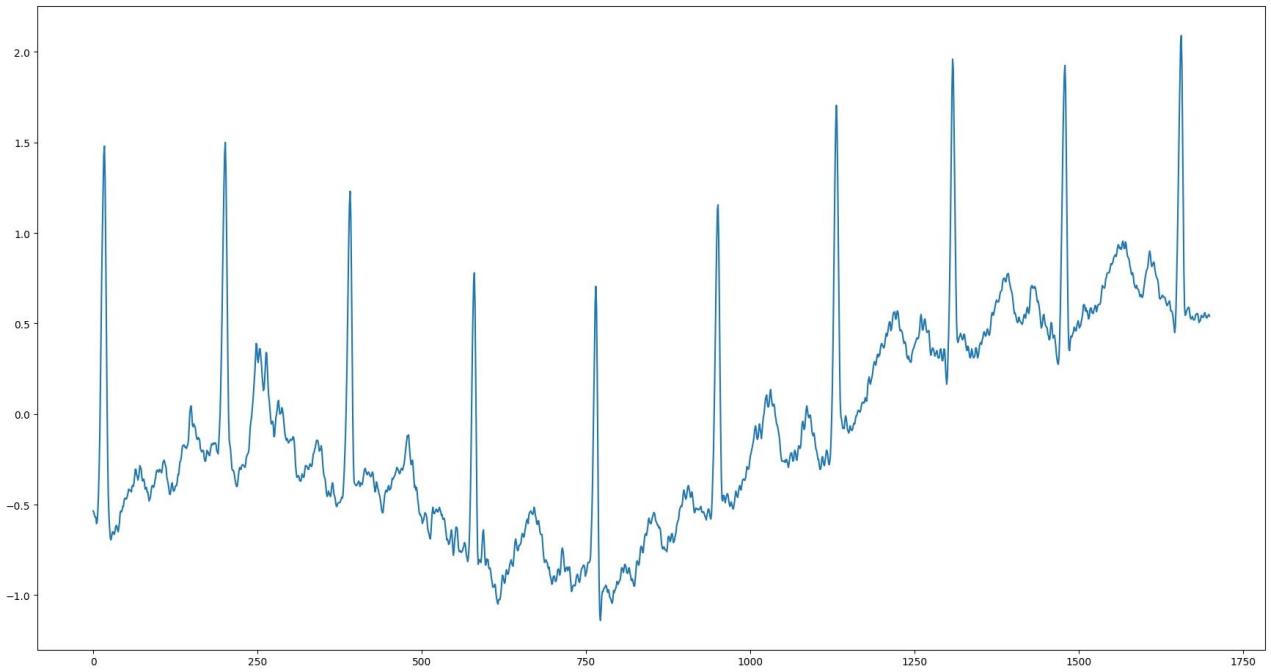

Рисунок 17 – Сегмент кривой ЭКГ

Предложенным алгоритмом экспериментальная кривая была классифицирована на 5 классов. Классам были присвоены соответствующие символы P, Q, R, S, T. После анализа всех точек отсчета экспериментальной кривой в первый класс было определено 378 участков, во второй - 420, в третий - 480, в

четвертый -452, и наконец в пятый -370. Сегменту экспериментальной кривой, построенной на участке от 0 до 500 точек, изображенному на рисунке 18, соответствует цепочка символов: PQRSTPQRST.

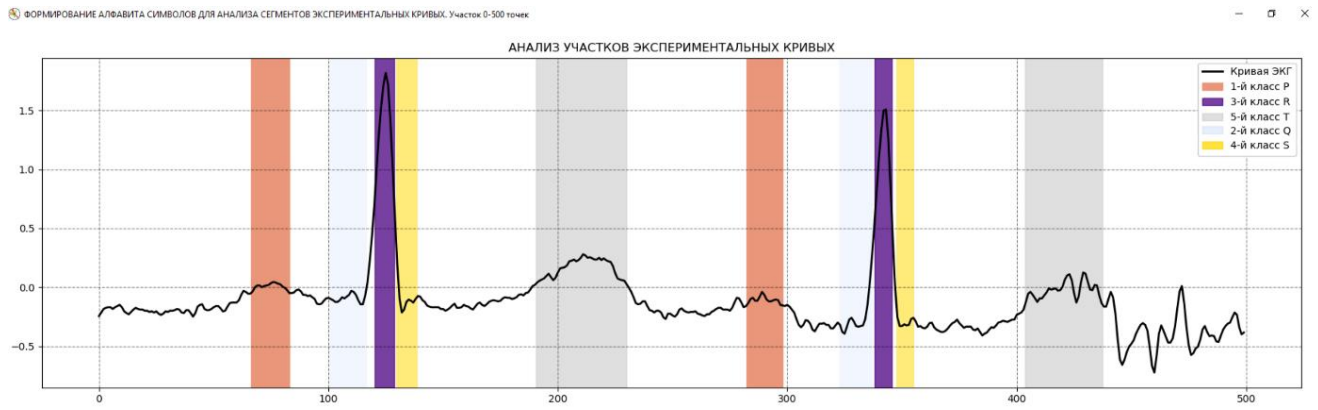

Рисунок 18 – Присвоение символов алфавита сегментам кривых

Полученная классификация находит значительную схожесть с реальной расшифрованной электрокардиограммой (Рисунок 19).

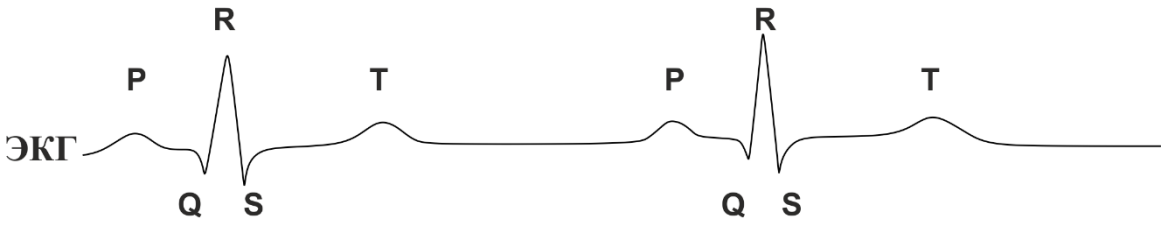

Рисунок 19 – Электрокардиограмма человека

4.4 Исследование применимости автоматизированной системы лингвистического анализа к исследованию спектрограмм радиочастот FM диапазона

Для исследования разработанной системы лингвистического анализа был проведен эксперимент. Исходной информацией для эксперимента послужили экспериментальные кривые спектрограмм радиочастот FM-диапазона, полученные

в ГОУ ВПО «Донецкий национальный университет» на лабораторном макете. Экспериментальные данные были сняты с помощью SDR приемника на базе RTL2832 и R820T. На рисунке 20 представлен характер поведения экспериментальных кривых при различных состояниях.

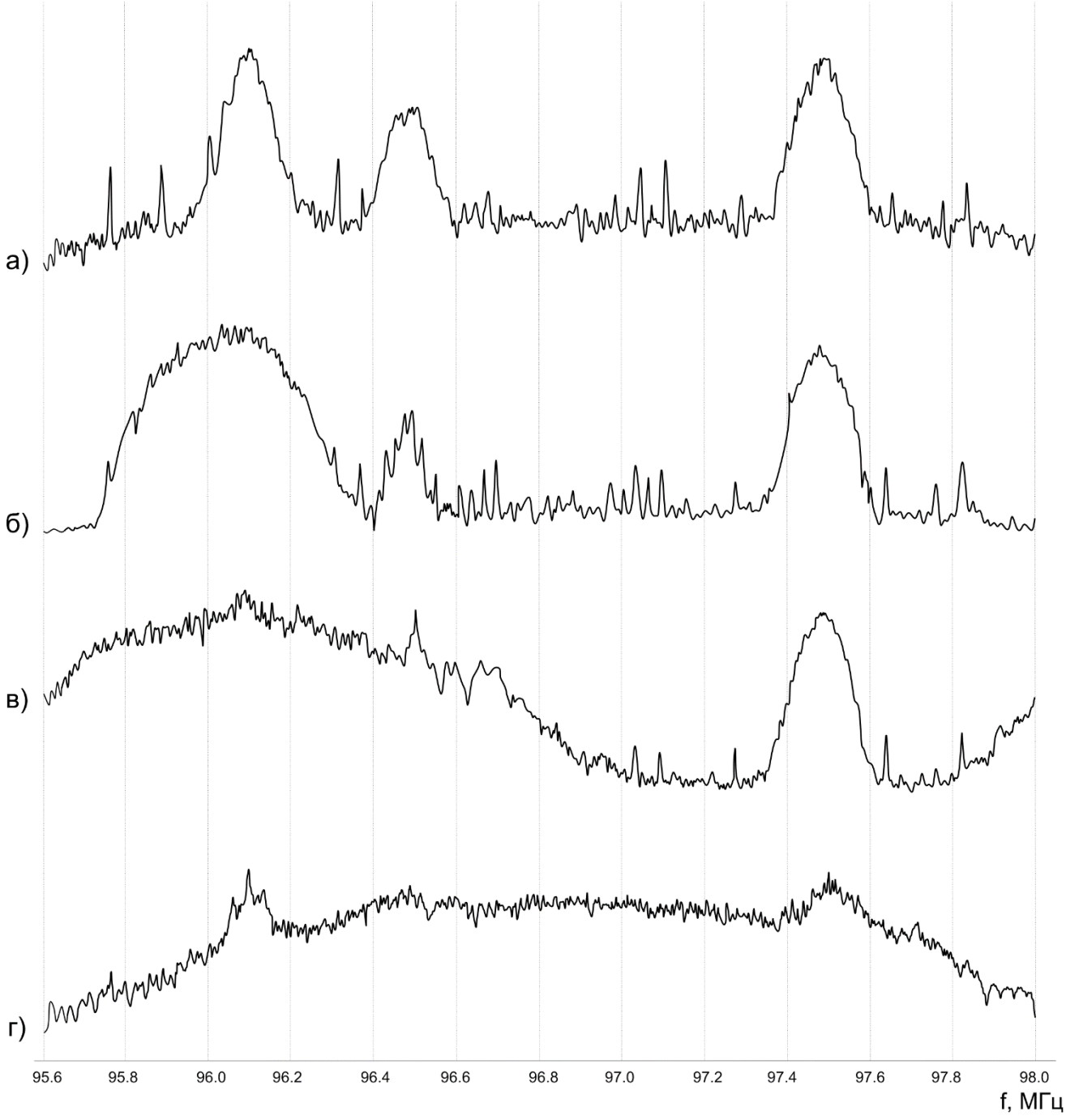

Рисунок 20 – Экспериментальные кривые спектрограмм радиочастот: а) без подавления, б) с подавлением 1-го сигнала, в) с подавлением 2-х сигналов, г) с подавлением 3-х сигналов

Исследуемый диапазон ограничен шириной полосы приёмника от 0,25 МГц до 3,2 МГц. Для удобства выбрана частота дискретизации RTL приемника 2432 MSPS (~2.400 МГц). Регистрировалась зависимость частоты от относительной мощности сигнала dBFS в диапазоне частот 95.6-98 МГц. Данный диапазон выбран в связи с большим количеством источников сигналов (радиостанций). Каждая кривая представлялась набором 2400 значений ординат. Эти числа соответствовали точкам отсчета, шагом в 1 КГц. В исследовании применялось частотное подавление сигналов. Подавление осуществлялось лабораторным широкополосным генератором качающейся частоты Кривые регистрировались при 4 состояниях: без подавления, с подавлением 1-го одного сигнала, 2-х сигналов и 3-х сигналов соответственно. Всего было зарегистрировано и проанализировано 60 таких кривых (по 15 кривых для каждого состояния).

Сегментация анализируемых кривых осуществлялась алгоритмами на основе функции сложности, представленные в 2.2. Каждая кривая разбивалась на равные участки длиной в 200 КГц, взятые с шагом следования вдоль кривой в 100 КГц. Таким образом имелось перекрытие в половину длины участка. Алгоритмом определялись участки, для которых функция сложности принимала локально минимальные значения. Количество выделенных участков, соответствующих сложному поведению кривой, для различных кривых составляло от 4 до 12. Итого на всех 60 кривых было выделено 452 таких участка, которые затем были разбиты алгоритмом классификации на 3 класса. На рисунке 21 представлены образцы эталонов каждого класса.

1 класс 2 класс 3 класс

Рисунок 21 – Образцы эталонов классов

Каждый из этих эталонов представляет собой участок кривой, ординаты которого являются среднеарифметическим значением ординат всех участков соответствующего класса. Из рисунка 21 можно отметить, что участки 1-го класса характеризуют левый фронт сигнала, участки 2-го класса – центральную часть сигнала и участки 3-го класса – правый фронт сигнала. Обозначим участки следующим образом: 1 класс – L (left), 2 класс – C (central), 3 класс – R (right). Таким образом сформирован алфавит языка описания исследуемых экспериментальных кривых, состоящий из 3х символов $A = \{L, C, R\}$.

Лингвистические описания исследуемых кривых представлены в таблице 1. Для удобства анализа, полученные описания кривых сортированы на группы, соответствующие различным состояниям, а именно: 1 группа - без подавления, 2 группа - с подавлением 1-го сигнала, 3 группа - с подавлением 2-х сигналов, 4 группа - с подавлением 3-х сигналов. Символы в описаниях расположены в соответствии с последовательностью их естественного следования вдоль кривых.

Таблица 1 – Лингвистические описания исследуемых кривых

1 группа	2 группа	3 группа	4 группа
L C C L C	L L C R C C	C C C C C C C R C	L L C C C C C C C C C C R
L C C L C	L L C R C C	C C C C C C C R C	L L L C C C C C C C C C R
C C L C	L L C R C C	C C C C C C C R C	L C C C C C C C C C C R
L C C L C	L C R C C	C C C C C C C R R	L C C C C C C C C C C R
L C C L C	L C R C C	C C C C C R L C	L C C C C C C C C C C R
L C R C L C	L L C R C C	C C C C C R C	L C C C C C C C C C R R
L C R C L C	L L C R C L C	C C C C C C C R R	L L C C C C C C C C R R
L C C L C	L L C R C L C	C C C C C C C R C	L C C C C C C C C C R R
L C C L C	L C R C L C	C C C C C C C R C	L C C C C C C C C C C R
L C C C	L L C R C C	C C C C C R L C	L C C C C C C C C C C R
L C C L C	L C C R C C	C C C C C C C R C	L C C C C C C C C C C R
L C C L C	L L C R C C	C C C C C C C R C	L C C C C C C C C C C R
L C C L C R	L L C R C C R	C C C C C R C R	L C C C C C C C C C C R
L C C L C	L L C R C C	C C C C C C C R C	L C C C C C C C C C R
L C C L C	L L C R C C	C C C C C C C R C	L C C C C C C C C C C R

Из анализа таблицы 1 следует, что лингвистические описания, соответствующие одной группе очень близки, а описания, взятые из разных групп,

заметно отличаются. Можно отметить, что все лингвистические описания третьей группы начинаются с символа C, а описания четвертой группы заканчиваются на R. В описаниях третьей группы практически не встречается символ L, в описаниях третьей и четвертой групп большую часть символов составляет C. Так же описания первой группы не превышают пяти символов, а описания четвертой группы состоят из не менее чем десяти символов.

Таким образом, можно заключить что приведенные описания экспериментальных кривых, полученные разработанной автоматизированной системой лингвистического анализа, представляют короткие и надежные правила для анализа спектрограмм радиочастот.

4.4.1 Расширенное лингвистическое описание спектрограмм радиочастот FM диапазона

Для более расширенного лингвистического описания экспериментальных кривых предлагается составлять это описание с учетом местоположения участков кривой на оси абсцисс. Такая информация не использовалась при описании в 4.4, однако эта информация может оказаться необходимой для более глубокого анализа представленной кривой. Например, при анализе спектрограмм важно учитывать, на какой из частот находится пик амплитуды. Кроме этого имеется большой класс экспериментальных кривых, для анализа которых необходима информация о характере поведения как на отдельных информативных участках, так и информация о положении этих участков на оси аргумента.

Для составления такого расширенного лингвистического описания, учитывающего положение участков на исследуемой кривой, на примере спектрограмм радиочастот FM диапазона воспользуемся следующей процедурой. Область определения кривой разбивается на l зон (в нашем случае l=12), одинаковых по длине (200 КГц). Алфавит языка описания кривых дополняется символом b – background, обозначающим не информативные фоновые участки. Таким образом алфавит состоит из четырех символов $A = \{L, C, R, b\}$. Исходя из

этого, каждому участку будет присвоен символ, либо характеризующий его поведение на кривой, либо характеризующий фоновое и не информативное возмущение.

Полученное расширенное лингвистическое описание кривой будет отражать не только классы выделенных на ней элементарных событий, но и их фазы. Таким образом, цепочка символов $\langle L\,C\,C\,L\,C\rangle$ будет представлена в виде $\langle b\,L\,C\,b\,C\,b\,b\,b\,L\,C\,b\,b\rangle$. При этом расширенные описания разных экспериментальных кривых с одинаковой длиной будут иметь одинаковое количество символов.

При определении автоматизированной системой лингвистического анализа конечного подмножества $\hat{L}$ в множестве всех таких цепочек L трансформационная грамматика сможет состоять из одной элементарной трансформации, а именно в замене одного символа другим. Минимальное количество элементарных трансформаций, переводящих цепочку символов T_1 в T_2, является мерой отличия этих цепочек друг от друга. Рассмотрим метрику на множестве всех цепочек длины l – числу несовпадающих символов на одинаковых местоположениях:

$$r(T_1,T_2)=\sum_{i=1}^{l}\left|a_i^1,a_i^2\right|, \tag{83}$$

где $\left|a_i^1,a_i^2\right|=\begin{cases}0, & a_i^1=a_i^2\\ 1, & a_i^1\neq a_i^2\end{cases}$. Так образуются символьные цепочки фиксированной длины, которые можно сравнивать посимвольно, например, как сравнивают числовые векторы одинаковой размерности. Для ядра языка, состоящего из одной символьной цепочки $\hat{T}$ степень принадлежности произвольной символьной цепочки T к этому языку определяется монотонной убывающей функцией расстояния в метрическом пространстве L от этой цепочки T до эталонной цепочки $\hat{T}$. Пусть имеется выборка экспериментальных кривых одной группы, представленная расширенными описаниями $T_j=\langle a_1\ldots a_l\rangle$, $j=1,\ldots,n$, где n - количество кривых в данной выборке. Для того чтобы найти символьную цепочку $\hat{T}=\langle\hat{a}_1\ldots\hat{a}_l\rangle$, для которой $\phi(T)=\sum_{j=1}^{n}r(T,T_j)=\min$ необходимо подставить критерий $\phi(T)$ в (83) и изменить порядок суммирования: $\phi(T)=\sum_{i=1}^{l}\sum_{j=1}^{n}\left|\hat{a}_i,a_i^j\right|$. Минимум критерия

будет обеспечен когда $\hat{a}_i$ будет символом из расширенного алфавита A, который чаще других оказывается на *i*-м месте в символьных цепочках T_j, $j = 1,\ldots,n$.

Исходной информацией для эксперимента послужили те же экспериментальные кривые спектрограмм радиочастот, что и в 4.4. В таблице 2 представлены расширенные лингвистические описания этих спектрограмм.

Таблица 2 – Расширенные лингвистические описания исследуемых кривых

Группы	Описания кривых	Эталоны групп и расстояние $r(\hat{T},T_j)$			
		1	2	3	4
		bLCbCbbbLCbb	LLCRCbbbbCbb	CCCCCCRbbCbb	bLCCCCCCCCCR
1	2	3	4	5	6
1 группа	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C b b	0	6	10	14
	b b C b C b b b L C b b	2	10	10	18
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C b b	0	6	10	14
	b L C R C b b b L C b b	4	8	10	14
	b L C R C b b b L C b b	4	8	10	14
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b b C b b	2	6	10	14
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C R b	2	6	10	12
	b L C b C b b b L C b b	0	6	10	14
	b L C b C b b b L C b b	0	6	10	14
2 группа	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C b b	6	0	10	14
	b L C R C b b b b C b b	4	2	10	14
	b L C R C b b b b C b b	4	2	10	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b L C b b	8	4	12	14
	L L C R C b b b L C b b	8	4	12	14
	b L C R C b b b L C b b	4	2	10	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C R b	8	2	8	14
	L L C R C b b b b C b b	6	0	10	14
	L L C R C b b b b C b b	6	0	10	14

Продолжение таблицы 2

1	2	3	4	5	6
3 группа	C C C C C C R b b C b b	10	8	0	10
	C C C C C C R b b C b b	10	8	0	10
	C C C C C C R b b C b b	10	8	0	10
	C C C C C C R b b R b b	10	10	2	12
	b C C C C C R b L C b b	10	12	2	12
	b C C C C C R b b C b b	10	10	2	12
	C C C C C C R b b R b b	10	10	2	12
	C C C C C C R b b C b b	10	8	0	10
	C C C C C C R b b C b b	10	8	0	10
	b C C C C C R b L C b b	10	12	2	12
	C C C C C C R b b C b b	10	8	0	10
	C C C C C C R b b C b b	10	8	0	10
	b C C C C C R b b C R b	10	12	4	12
	C C C C C C R b b C b b	10	8	2	10
	C C C C C C R b b C b b	10	8	0	10
4 группа	L L C C C C C C C C C R	18	14	12	2
	L L L C C C C C C C C R	20	16	14	4
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R R	14	14	10	2
	b L L C C C C C C C R R	18	14	12	4
	b L C C C C C C C C C R R	14	14	10	2
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R	14	14	10	0
	b L C C C C C C C C C R b	12	12	8	2
	b L C C C C C C C C C R	14	14	10	0

Положение символов описаний кривых в таблице 2 соответствуют естественным положениям участков на этих кривых. Каждой группе соответствует усредненная расширенная цепочка символов (эталон группы) полученные по $\phi(T)=\sum_{j=1}^{n} r(T,T_j)=\min$, а справа для каждой кривой указаны расстояния $r(\hat{T},T_j)$ от соответствующей ей расширенной цепочки до эталонов всех четырех групп, вычисленные по формуле (83), отражающие минимальное число элементарных трансформаций, переводящих эталон в эту цепочку. Применение к каждой группе

цепочек из таблицы 2 алгоритма поиска разбора с минимальной неплотностью (81) позволяет получить следующие эталоны групп: 1 группа – bLCbCbbbLCbb, 2 группа – LLCRCbbbbCbb, 3 группа – CCCCCCRbbCbb, 4 группа – bLCCCCCCCCCR, что подтверждает правильность вычисления эталонов. Для удобства восприятия расширенных лингвистических описаний и наглядности, на рисунке 22 на кривых из разных групп отмечены символы, присвоенные каждому анализируемому участку.

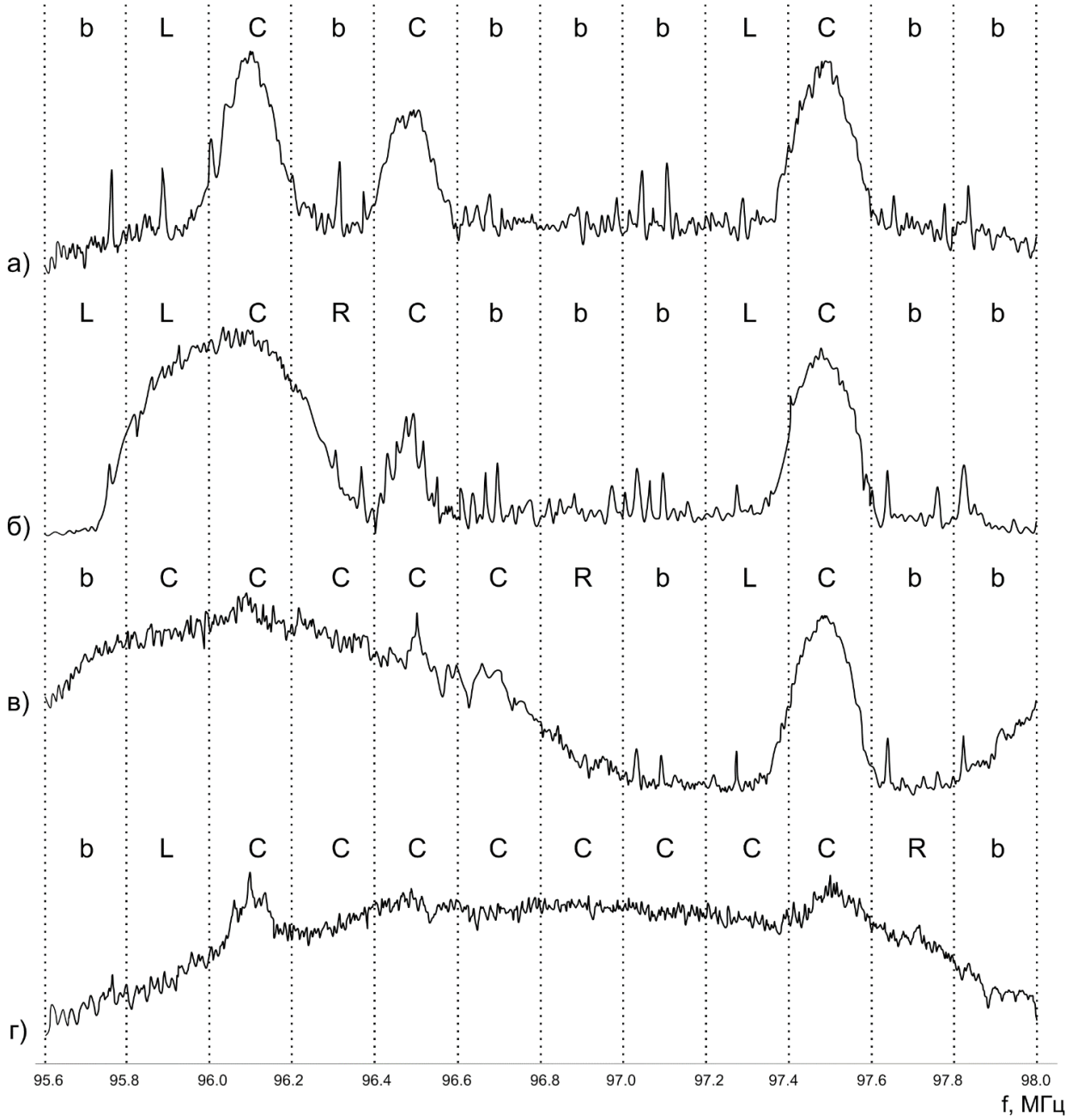

Рисунок 22 – Экспериментальные кривые спектрограмм радиочастот и их расширенные описания: а) без подавления, б) с подавлением 1-го сигнала, в) с подавлением 2-х сигналов, г) с подавлением 3-х сигналов

Применение эталонов позволяет безошибочно отнести каждую цепочку символов к своей группе по признаку минимума расстояния до эталона. Более детальное описание исследуемых кривых можно получить изменяя длину участков и шаг их следования на этапе сегментации, что в случае с анализом приведенных спектрограмм не является необходимым. Полученные расширенные описания достаточно точно описывают характер поведения исследуемых кривых.

4.5 Выводы по разделу

В четвертом разделе решалась задача исследования применимости разработанной системы структурного анализа в процессах электромагнитной совместимости радиоэлектронных средств. Для ее реализации проведен ряд экспериментов. Эксперименты проводились как отдельно по каждому из этапов обработки кривой (выделения и распознавания характерных участков, присвоения выделенным участкам символов некоторого алфавита, анализа полученных последовательностей символов), так и автоматизированной системы лингвистического анализа в целом.

Проведенные эксперименты алгоритмов сегментации показали, что выделенные участки непосредственно и «на глаз» являются более сложными, чем остальные и можно заключить, что предложенные алгоритмы с применением функции сложности целесообразно применять к задачам сегментации переходных участков экспериментальных кривых акустических колебаний. Первый эксперимент показал сходства в результаты работы обеих алгоритмов. Второй эксперимент так же подтвердил наличие сходства, а также показал, что разбиение на большее количество участков и применение перекрытия приводит к увеличению точности выделения.

При построении трансформационной грамматики входную и выходную вершины полученного графа связывают только те пути, по которым предъявленная цепочка может быть получена в соответствии с правильными последовательностями трансформаций из всевозможных последовательностей слов данного словаря, а длина пути равна числу необходимых при этом

элементарных трансформаций и, следовательно, значению меры различия соответствующей правильной цепочки ядра языка и анализируемой цепочки. Кратчайший путь дает морфологический анализ предъявленной цепочки в виде последовательности слов, а длина этого пути дает степень ее несоответствия правильному описанию.

В результате применения системы лингвистического анализа к анализу данных электрокардиограмм можно сделать вывод, что полученная классификация находит значительную схожесть с реальной расшифрованной электрокардиограммой.

В результате применения системы лингвистического анализа к анализу спектрограмм радиочастот можно заключить, что приведенные описания экспериментальных кривых представляют короткие и надежные правила для анализа спектрограмм радиочастот. Применение эталонов позволяет безошибочно отнести каждую цепочку символов к своей группе по признаку минимума расстояния до эталона. Полученные расширенные описания достаточно точно описывают характер поведения исследуемых кривых.

ЗАКЛЮЧЕНИЕ

Диссертационная работа является завершенным научным исследованием, в котором получено решение важной научно-технической задачи, состоящей в обосновании модернизированных методов и алгоритмов автоматизации процессов анализа массивов экспериментальных данных, их классификации и представления в виде компактных структур, что позволяет использовать фактические результаты работы в практической, научно-исследовательской и преподавательской деятельности. Основные научные результаты и выводы, полученные при выполнении работы, состоят в следующем:

1. Дальнейшее развитие получил метод структурного анализа данных, в рамках лингвистического подхода к анализу экспериментальных кривых, в котором анализируемая кривая описывается в виде сжатого описания из последовательного ряда символов либо целых слов из определенного алфавита.

2. Разработаны вычислительные методы сегментации экспериментальных кривых с использованием функций сложности, способные осуществлять бинарную классификацию помимо сегментации, и содержащие дополнительные условия во избежание выделения ложных экстремумов.

3. Осуществлена алгоритмическая реализация процедуры структурного анализа данных, точнее лингвистического подхода к анализу экспериментальных кривых, в котором анализируемая кривая описывается последовательным рядом символов либо целых слов из определенного алфавита.

4. Разработаны методы лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами, в котором через конечное число циклов достигается устойчивая классификация и ни один вектор не переносится из одного класса в другой, благодаря классификации по признаку минимума расстояния до эталона.

5. Обоснована применимость разработанной системы структурного анализа к выделению информативных участков кривой акустических колебаний, к анализу

экспериментальных данных электрокардиограмм, к анализу спектрограмм радиочастот.

6. Обоснована процедура составления более расширенного лингвистического описания экспериментальных кривых, позволяющая составлять это описание с учетом местоположения участков кривой на оси абсцисс.

СПИСОК ЛИТЕРАТУРЫ

1. Чернышев, А. Б. Теория информационных процессов и систем [Текст] : учебное пособие / А. Б. Чернышев // - Ставрополь : Северо-Кавказский федеральный университет, 2015. - 169 с.
2. Яглом, А. М. Корреляционная теория стационарных случайных функций [Текст] : с примерами из метеорологии / А. М. Яглом. - Изд. 2-е. - Москва : URSS : ЛЕНАНД, 2018. - 279 с.
3. Демидович, Б. П. Численные методы анализа [Текст] : приближение функций, дифференциальные и интегральные уравнения / Б. П. Демидович, И. А. Марон, Э. З. Шувалова / под ред. Б. П. Демидовича. - Изд. 5-е стер. - Санкт-Петербург и др. : Лань, 2010. - 400 с.
4. Ганичев, А. В. Структурное распознавание образов [Текст] : монография / А. В. Ганичев, А. В. Ганичева/ Минобрнауки России, Федеральное государственное бюджетное образовательное учреждение высшего образования "Тверской государственный технический университет" (ТвГТУ). - Тверь : ТвГТУ, 2018. - 107 с.
5. Сальников, И. И. Анализ пространственно-временных параметров удаленных объектов в информационных технических системах [Текст] / И. И. Сальников. - Москва : Физматлит, 2011. - 252 с.
6. Трофимов, Д. М. Современные методы и алгоритмы обработки и анализа комплекса космической, геолого-геофизической и геохимической информации для прогноза углеводородного потенциала неизученных участков недр [Текст] / Д. М. Трофимов, В. Н. Евдокименков, М. К. Шуваева. - Москва : Физматлит, 2012. - 319 с.
7. Стечкин, С. Б. Сплайны в вычислительной математике [Текст] / С. Б. Стечкин, Ю. Н. Субботин. М.: Наука, 1976, - 157 с.
8. Bellman, R. On the approximation of curves by line segments using dynamic programing [Текст] / R. Bellman. – Communication of the ACM, 1961, v.4. - 284 p.

9. Pavlidis, T. Segmentation of plane curves / T. Pavlidis, L. Horowitz // IEE Trans. Computers. – 1974. - V. 23. -P. 860-870.
10. Stone, H. Approximation of curves by line segments / H. Stone // Mathematics of computation. – 1961 - V. 15. -P. 40-47.
11. Pavlidis, T. Optimal piecewise polynomial approximation of functions of one and two variables / T. Pavlidis // IEEE Trans. Computers. – 1975. - V.25. P. 98-102.
12. Canroni, A. Optimal curve fitting with piecewise linear functions / A. Canroni // IEEE Trans. Computers. – 1971. - V.20. P. 59-67.
13. Gallant, A.R. Fitting segmented polynomial regression models whose joints points have to be estimated / A. R. Gallant, W. A. Fuller // J. Amer. Statis. Assoc. – 1973. - V.68 - P. 144-147.
14. Pavlidis, T. Polynomial approximations by Newtons method / T. Pavlidis // IEEE Trans, Computers. – 1977. - V.26. -P. 801-807.
15. Gluss, B. Futher remarks on line segment curve – fitting using dynamic programming / B. Gluss// Communication of the ACM. – 1962. - V. 5. - P. 441-445.
16. Hawkins, D. On the choice of segments in piecewise approximation. / D. Hawkins // J. of the Inst. Of Mathematics and its Application. – 1972. -V. 9. - P. 250-256.
17. Эсбенсен, Ким Анализ многомерных данных : избр. главы [Текст]/ Ким Эсбенсен ; пер. с англ. С. В. Кучерявского ; под ред. О. Е. Родионовой. - Черноголовка : изд-во ИПХВ РАН, 2005. - 157 с.
18. Lawson, C. L. Characteristic properties of the segmented rational minimax approximation problem / C. L. Lawson // Numerishe Mathematik. – 1964. - V. 6. - P.293-301.
19. Pavlidis, T. Linguistic analysis of waveforms / T. Pavlidis // In Software engineering. N.-Y: Academic. – 1971. - P. 203-225.
20. Pavlidis, T. Uniform piecewise polynomial approximation with variable joints / T. Pavlidis, A. Maika // – J. of the Approx. Theory. – 1974. - V. 12. - P. 61-69.
21. Pavlidis, T. Waveform segmentation through functional approximation/ T. Pavlidis // IEEE Trans. Computers. – 1973. - V. 22. - P. 689-697.

22. Гребенков, А. И. О выборе узлов при аппроксимации функции сплайнами / А. И. Гребенков // – ЖВМ и МФ. – 1976. -Т. 16. - № 1. - С. 219-224.
23. Методы сплайн-функций. Российская конференция, посвящённая 80-летию со дня рождения Ю. С. Завьялова (Новосибирск, 31 января –2 февраля 2011 г.): Тез. докладов / ИМ СО РАН. Новосибирск, 2011. - 113 с.
24. Дикусар, Н. Д. Оптимизация решения в задачах кусочно-полиноминальной аппроксимации [Текст] / Н. Д. Дикусар. - Дубна : Издательский отд. Объединенного ин-та ядерных исслед., 2017. - 13 с.
25. Дикусар, Н. Д. Кусочно-полиномиальная аппроксимация шестого порядка с автоматическим обнаружением узлов [Текст] / Н. Д. Дикусар. - Дубна, Московская обл. : Объединенный ин-т ядерных исслед., 2012. - 15 с.
26. Шевчук, В. П. Моделирование метрологических характеристик интеллектуальных измерительных приборов и систем [Текст] / В. П. Шевчук. - Москва : Физматлит, 2011. - 319 с.
27. Коралов, Л. Б. Теория вероятностей и случайные процессы [Электронный ресурс] : электронное издание / Л. Б. Коралов, Я. Г. Синай ; пер. с англ. Э. В. Переходцевой ; под ред. Б. М. Гуревича ; Независимый Московский ун-т. - Москва : Изд-во МЦНМО, 2014.
28. Телькснис, Л. А. Определение наиболее вероятностного времени измерения характера случайного процесса / Л. А. Телькснис, В. Ю. Черняускас // Нелинейные и оптимальные системы. М.: Наука. – 1971. - Т. I. - С. 223-229.
29. Козинов, И. А. Модифицированный алгоритм обнаружения разладки случайного процесса и его применение при обработке многоспектральных данных / И. А. Козинов, Г. Н. Мальцев // Информационно-управляющие системы. - 2012. - №3 (58). - С. 9-17.
30. Буркатовская, Ю. Б. Оценивание параметров и обнаружение момента их изменения для обобщенного авторегрессионного процесса с условной неоднородностью / Ю. Б. Буркатовская, С. Э. Воробейчиков, Е. Е. Сергеева // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. - 2012. - №1 (18). - С. 48-57.

31. Суворов, И. С. Методика структурного обучения динамических байесовских сетей на основе статистических данных / И. С. Суворов, П. И. Бидюк // ММС. - 2010. - №4. С. 110-118.
32. Мартынов, В. В. Автоматизация процедуры обнаружения разладки процесса функционирования сложных технологических объектов / В. В. Мартынов, П. В. Мартынов // Вестник СГТУ. - 2011. - №2 (60). - С. 219-224.
33. Карташов, В. Я. Обнаружение структурно-параметрических изменений в стохастических системах в реальном масштабе времени алгоритмами непрерывных дробей и структурного анализа / В. Я. Карташов, М. А. Новосельцева // УБС. - 2011. - №34. - С. 62-91.
34. Телькснис, Л. А. Определение изменений свойств случайных процессов при неполных априорных данных / Л. А. Телькснис // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит. ССР. – 1975. - Вып. 12. - С. 9-27.
35. Монтвилас, А. М. Определение изменения свойств авторегресионной последовательности при неизвестных параметрах / А. М. Монтвилас // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит. ССР. – 1973. - Вып. 7. - С. 21-39.
36. Монтвилас, А. М. Определение изменений состояний стохастической системы в начале интервала наблюдения / А. М. Монтвилас // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1976. - Вып. 15. - С. 104-115.
37. Липейка, А. Определение моментов изменения свойств авторегрессионной последовательности / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1976. -Вып. 24. - С. 27-53.
38. Липейка, А. Об определении моментов изменения свойств авторегрессионной последовательности / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1976. - Вып. 39. - С. 9-25.

39. Липейка, А. Определение моментов изменения свойств авторегрессионных последовательностей с неизвестными параметрами / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1982. - Вып. 54. - С. 9-28.
40. Крицкий, О. Л. Информационная матрица Фишера для многомерного метода динамических условных корреляций dcc-mgarch(1,1) / О. Л. Крицкий // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. - 2009. - №4 (9). С. - 67-82.
41. Лапко, А. В. Анализ свойств смеси непараметрических оценок плотности вероятности многомерной случайной величины / А. В. Лапко, В. А. Лапко // Сибирский журнал науки и технологий. - 2010. - №2. - С. 32-35.
42. Липейка, А. Определение моментов изменения свойств многомерной авторегрессионной последовательности / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1981. - Вып. 51. - С. 9-33.
43. Липейка, А. Оценка моментов времени изменения параметров многомерных авторегрессионных последовательностей / А. Липейка // Y Всесоюз. совещ. по статистическим методам в процессах управления. Москва-Алма-Ата: Наука. – 1981. - С. 47-49.
44. Липейка, А. Определение моментов изменения свойств многомерных авторегрессионных последовательностей с неизвестными параметрами / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1982. - Вып. 59. - С. 25-35.
45. Липейка, А. Об определении моментов изменения свойств многомерных авторегрессионных последовательностей с неизвестными параметрами / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. -1984. - Вып. 65. - С. 102-106.
46. Липейка, А. Определение моментов изменения свойств многомерных авторегрессионных случайных последовательностей с неизвестными параметрами при длинных реализациях / А. Липейка, В. Малинаускас //

Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1984. - Вып. 65. -С. 121-126.

47. Kitafava, G. A procedure for the modelling of non-stationary time series / G. Kitafava, H. Akaike // Annals of Institute of Statistical Mathematics. – 1978. - V. 30, part B. - P. 351-363.
48. Гришко, А. К. Анализ временных рядов и методов обработки измерительной информации на основе регрессионных и авторегрессионных моделей / А. К. Гришко, В. А. Корж, В. А. Канайкин, А. С. Подсякин // НиКа. - 2012. - №2. - С. 104-105.
49. Кутузов, В. М. Многосегментный авторегрессионный алгоритм обработки сложномодулированных сигналов. Характеристики обнаружения скоростных целей / В. М. Кутузов, К. А. Мазуров // Известия вузов России. Радиоэлектроника. - 2009. - №4. - С. 43-50.
50. Клименченко, П. В. Авторегрессионный алгоритм Берга для обнаружения целей и определения их скоростей на фоне пассивных помех, основанный на спектральных и статистических различиях целей и помех / П. В. Клименченко, В. Н. Жураковский // Радиостроение. - 2017. - №4. - С. 1-15.
51. Глушко, Е. В. Особенности применения метода SSA для обнаружения разладки во временных рядах / Е. В. Глушко, И. С. Синева // T-Comm. - 2010. - №10. - С. 25-34.
52. Яковлев, В. Г. Оптимальная сегментация экспериментальных кривых / В. Г. Яковлев // YIII Всесоюзное совещание по проблемам управления. Таллин: Таллинский политехнический ин-т. – 1980. - №. 2. - С. 358-360.
53. Буробин, Н. Алгоритм определения моментов многократного изменения свойств случайного процесса на основе метода динамического программирования / Н. Буробин, В. В. Моттль, И. Б. Мучник // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1984. - Вып. 65. - С. 49-56.
54. Буркатовская, Ю. Б. Асимптотические свойства процедур оценивания параметров и обнаружения разладки обобщенного авторегрессионного

процесса с условной неоднородностью / Ю. Б. Буркатовская, С. Э. Воробейчиков, Е. Е. Сергеева // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. - 2012. - №2 (19). - С. 59-71.

55. Асатрян, Д. Г. Непараметрическое оценивание момента разладки случайной последовательности / Д. Г. Асатрян, И. А. Сафарян // ДАН Арм.ССР. – 1982. Т. XXY. - №4.- С.160-164.
56. Воробейчиков, С. Э. Обнаружение момента разладки процесса авторегрессии первого порядка / С. Э. Воробейчиков, Т. В. Кабанова // Вестн. Том. гос. ун-та. - 2003. - №280. - С.170-174.
57. Лапко, А. В Непараметрические методы обнаружения закономерностей в условиях малых выборок / А. В. Лапко, М. А. Шарков, В. А. Лапко // Приборостроение. - 2008. - №8. - С. 62-67.
58. Дарховский, Б. С. Непараметрический метод для апостеприрного обнаружения момента разладки последовательности независимых случайных величин / Б. С. Дарховский // Теория вероятностей и ее применения. – 1976. - № 1. - С. 180-184.
59. Устинов, Ф. А. Задача о скорейшем обнаружении смены режима для процессов Леви / Ф. А. Устинов // Вестник Московского университета. Серия 1. Математика. Механика. - 2009. - №2. - С. 69-71.
60. Сергеева, Е. Е. Гарантированная оценка параметров и обнаружение момента разладки GARCH(1,1)-процесса / Е. Е. Сергеева, С. Э. Воробейчиков // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. - 2011. - №3 (16). - С. 31-42.
61. Мазалов, В. В. Байесовская модель в задаче наилучшего выбора с «Разладкой» / В. В. Мазалов, Е. Е. Ивашко // Вестник СПбГУ. Серия 10. Прикладная математика. Информатика. Процессы управления. - 2009. - №4. - С. 142-151.
62. Микка, К. В. Статистическое моделирование процедуры обнаружения разладки по среднему значению в гауссовской последовательности независимых случайных величин / К. В. Микка // Известия вузов. Северо-Кавказский регион. Серия: Естественные науки. - 2004. - №3. - С. 17-21.

63. Назаров, Н. В. Непараметрическое обнаружение и оценивание момента появления статистической неоднородности в реализации случайного процесса / Н. В. Назаров, Е. П. Пономарев // Радиотехника и электроника. – 1978. - № 10. - С. 2230-2233.
64. Сидоров, Ю. Е. Исследование непараметрического обнаружителя радиосигналов / Ю. Е. Сидоров, А. В. Шумилов // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика, телекоммуникации и управление. - 2008. - №6 (69). - С. 78-85.
65. Моттль, В. В. Метод частичной аппроксимации в задаче выделения информативных участков экспериментальных кривых / В. В. Моттль // – Автоматика и телемеханика. – 1977. - № 1. - С. 97-108.
66. Моттль, В. В. Лингвистический анализ экспериментальных кривых / В. В. Моттль, И. Б. Мучник // ТИИЭР. – 1979. - Т. 67. - № 5. - С. 12-38.
67. Богданов, Е. А. Математическая модель колебаний рабочего тока рудовосстановительной печи / Е. А. Богданов, А. Н. Кузнецов, А. С. Микулинский, В. В. Моттль // Электротехническая промышленность, серия Электрометрия. -1977. - Вып. 3(175). - С. 14-15.
68. Богданов, Е.А. Форма колебаний рабочего тока как носитель нформации о содержании углерода в шахте/ Е. А. Богданов, А. Н. Кузнецов, А. С. Микулинский, В. В. Моттль // Электротехническая промышленность, серия Электрометрия. – 1977. - Вып. 7(179). - С. 5-7.
69. Хеин, Мин Зо Алгоритмы динамического программирования в распознавании речи / Хеин Мин Зо // Auditorium. - 2017. - №4 (16). – С. 55-60.
70. Винцюк, Т. К. Поэлементное распознавание непрерывной речи, составленной из слов заданного словаря / Т. К. Винцюк // Кибернетика, Киев. – 1971. - № 2. - С. 133-143.
71. Сажок, Н. Н. Речевые информационные технологии и системы / Н. Н. Сажок // Управляющие системы и машины. - 2017. -№ 2. -С. 38-45.

72. Sakoe, H. Two –level DP – Matching- A dynamic programming – based pattern matching algorithm for connected word recognition / H. Sakoe // IEEE Trans. ASSP. – 1979. - V. 27. - N 6. - P.578-595.
73. Гладышев, К. К. Система автоматического распознавания речевых команд / К. К. Гладышев, Е. А. Шульгин // Приборостроение. - 2009. - №3. - С 17-21.
74. Овчинников, В. Г. К алгоритмам динамического программирования оптимальных процессов / В. Г. Овчинников // Вестн. Сам. гос. техн. ун-та. Сер.: Физ.-мат. науки. - 2012. - №3 (28). - С. 215-218.
75. Гусев, А. С. Прогнозирование остаточного ресурса по результатам диагностирования натурных конструкций и при непрерывном отслеживании их технического состояния / А. С. Гусев, С. А. Стародубцева, В. И. Щербаков // Известия МГТУ. - 2014. - №1 (19). - С. 100-104.
76. Rabiner, L. R. Perfomance trade offs in dynamic time warping algorithms for isolated word recognition / L. R. Rabiner, A. E. Rosenberg, C. Myers // IEEE Trans. ASSP/ - 1980. - V. 28. - N 6. - P. 623-635.
77. Ольховский, Ю. Б. Сжатие данных при телеизмерениях [Текст] / Ю. Б. Ольховский, О. Н. Новоселов, А. П. Мановцев. М.: Сов.радио, 1971, -143 с.
78. Валужис, К. К. Адаптивная дискретизация электробиологических сигналов / К. К. Валужис, С. Л. Корсакас, А. Г. Рашимас, Р. И. Цитварас // Вычислительная техника. Материалы XXII республиканской научной-технической конф. Лит.ССР, Каунас. – 1972. - Т. 3. - С. 507-511.
79. Янушкевичус, С. Структуры и принципы построения автоматизированного архива ЭКС / З. Янушкевичус, Г. Витенштейнас, А. Валужис, С. Корсакас // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1974. - № 9. - С. 175-213.
80. Эндрюс, К. Адаптивное сжатие данных / К. Эндрюс, Д. Девис, Г. Шварц // ТИИЭР. – 1967. - № 3. - С. 25-38.
81. Валужис, К. К. О точности кусочно-линейной аппроксимации ЭКС / К. К. Валужис, А. П. Рашимас, Р. И. Цитварас // Техническая кибернетика. Каунас: Каунасский политехнический ин-т. – 1972. - Т. 4. - С. 237-240.

82. Кортман, С. М. Сокращение избыточности как практический метод сжатия данных / С. М. Кортман // – ТИИЭР. – 1967. - № 3. - С. 8-20.
83. Кузьмин, О. В. Параллельные алгоритмы вычисления локальных минимумов целочисленных решеток / О. В. Кузьмин, В. С. Усатюк // Программные продукты и системы. - 2015. - №1 (109). - С. 55-62.
84. Phillips, G. M. Algorithms for piecewise straight line approximation / G. M. Phillips // – Comput. Journal. – 1968. - V. 2. - P. 211-212.
85. Vanderwalle, I. On the calculation of the piecewise linear approximation to deiscrete function / I. Vanderwalle // IEEE Trans. Computers. – 1975. - V. 24. - P. 843-846.
86. Красильников, Э. М. Непрерывное распознавание базовых жестов в реальном времени с применением скрытых марковских моделей / Э. М. Красильников // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. - 2013. - №3. - С. 46-52.
87. Габибов, И. А. Диагностика состояния магистрального газопровода с использованием последовательных методов обнаружения разладок / И. А. Габибов // Sciences of Europe. - 2018. - №35-1 (35). - С. 30-35.
88. Медведев, А. В. Теория непараметрических систем. Управление-i / А. В. Медведев // Сибирский журнал науки и технологий. 2013. №2 (48). С. 57-63.
89. Никифоров, И. В. Применение кумулятивных сумм для обнаружения изменения характеристик случайного процесса / И. В. Никифоров// – Автоматика и телемеханика. – 1979. - № 2. - С. 48-58.
90. Бородкин, Л. И. Алгоритм обнаружения моментов времени изменения параметров уравнения случайного процесса / Л. И. Бородкин, В. В. Моттль // Автоматика и телемеханика. – 1976. - № 6. - С. 23-29.
91. Никифоров, И. В. Модификация и исследование процедуры кумулятивных сумм / И. В. Никифоров // Автоматика и телемеханика. - 1980. - № 9. - С. 74-80.
92. Jones, R. H. A method for detecting change in a time series applied to newborn ECG / R. H. Jones, D. H. Growell, L. E. Kapunial // Electroenceph. Clin. Neurophysiol. – 1969. - V. 27. - P. 87-93.

93. Jones, R. H. Change detection model for serially correlated multivariate data / R. H. Jones, D. H. Growell, L. E. Kapunial // Biometrics. – 1970. - V. 26. - N 2/ - P. 269-281.
94. Jones, R. H. An adaptive method for testing for change in digitized cardiotachometer data / R. H. Jones, D. H. Growell, L. E. Kapunial // IEEE Trans. On Biomed. End. – 1971. - V. 18. - N 5. - P. 360-363.
95. Клячкин, В. Н. Алгоритмы обнаружения нарушений при многомерном статистическом контроле технологического процесса / В. Н. Клячкин, Ю. А. Кравцов, И. А. Охотников // Вестник УлГТУ. - 2014. - №1 (65). - С. 48-51.
96. Kitagawa, G. A procedure for the modelling of non-stationary time series analysis / G. Kitagawa, H. Akaike // Annal Inst. Statist. Math. – 1978. - V. 30, part B. - P. 351-363.
97. Ozaki, T. On the fitting on non-stationary autoregressive models is time series analysis / T. Ozaki, H. Tong / In Proc. Of the 8 th Hawaii Internat. Conf. on System Science. Hawaii: Western Periodical C. – 1975. - P. 224-227.
98. Каминаскас, В. А. Обнаружение изменения параметров процесса авторегрессии. / В. А. Каминаскас, Д. А. Шиените // Тр. АН Лит.ССР. – 1975. - Серия Б, Т. 4(89). - С 143-147.
99. Липейка, А. Классификация авторегрессионных последовательностей со скачкообразно меняющимися параметрами / А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1978. - Вып. 30. - С. 9-28.
100. Гринявичус, К. Автоматическое определение стадий сна / К. Гринявичус, В. Лесене, А. Липейка // Статистические проблемы управления. Вильнюс: Институт физики и математики АН Лит.ССР. – 1981. - Вып. 51. - С. 83-99.
101. Grauper, D. A microprocessor system for multifunctional control of upper limb prostheses via ECG signal identification / D. Grauper, J. Magnussen, A. Beex // Proc. The 4 th International Joint Conf. on Pattern Recognition. Kyoto, Japan: IAPR. – 1978. - P. 1113-1117.

102. Немирко, А.П. Алгоритм оперативного распознавания опасных аритмий / А. П. Немирко, Л. А. Менило // Вопросы овтоматизации съема и обработки биомедицинской информации. Л.: ЛЭТИ. – 1981. - Вып. 283. - С. 71-74.

103. Мучник, Р. Б. Алгоритмы выделения и анализа характерных участков и их взаимного расположения / Р. Б. Мучник // Техническая кибернетика. Материалы литовской респуб. XX науч.-техн. конф. Каунас: Каунасский политехнический институт. – 1970. - С. 196-204.

104. Мучник, Р. Б. Сокращение описания временных сигналов. – В кн.: Актуальные вопросы технической кибернетики / Р. Б. Мучник // М.: Наука. – 1972. - С. 208-210.

105. Мучник, И. Б. Формирование языка для описания экспериментальных кривых / И. Б. Мучник, Р. Б. Мучник // Автоматика и телемеханика. – 1973. - № 5. - С. 86-98.

106. Бейтельман, Л. С. Применение структурного анализа кривых к задаче исследования конверторного процесса / Л. С. Бейтельман, И. Б. Мучник, Р. Б. Мучник, Р. А. Симсарьян // Изв. Вузов. Черная металлургия. – 1971. - № 12. - С. 149-155.

107. Гуров, О.Б. Исследование применимости лингвистических методов к задачам акустической диагностики / О. Б. Гуров, И. Б. Мучник, Р. Б. Мучник, Ю. В. Тимофеев, Р. Б. Шапуро // Диагностика и обслуживание машин. Тр. Сибирского института механизации и электрификации сельского хозяйства. Новосибирск: Западно-сибирское книжное изд-во. – 1972. - Вып. 8. - Ч. 2. - С. 145-151.

108. Яковлев, В. Г. Алгоритм выделения всплесков на физиологических кривых / В. Г. Яковлев // Автоматика и телемеханика. – 1977. - № 12. - С. 94-105.

109. Никифоров, И. В. Последовательное обнаружение изменения свойств временных рядов [Текст] / И. В. Никифоров. – М.: Наука, 1983, - 244 с.

110. Дорофеюк, Ю. А. Структурно-классификационные методы анализа и прогнозирования в системах управления / Ю. А. Дорофеюк // ТВИМ. - 2008. - №1 (12). - С. 166-170.

111. Огнев, И. В. Распознавание речи методами скрытых марковских моделей в ассоциативной осцилляторной среде / И. В. Огнев, П. А. Парамонов // Известия ВУЗов. Поволжский регион. Технические науки. - 2013. - №3 (27). - С. 115-126.

112. Дорофеюк, А. А. Методология экспертно-классификационного анализа в задачах управления и обработки сложноорганизованных данных (история и перспективы развития) / А. А. Дорофеюк // Проблемы управления. - 2009. - №3.1. - С. 19-28.

113. Дорофеюк, Ю. А. Структурная идентификация сложных объектов управления на базе методов кусочной аппроксимации / Ю. А. Дорофеюк // УБС. - 2010. - №30. - С. 79-88.

114. Grenier, Y. Speaker identification from linear prediction / Y. Grenier // In: Proc. the 4 th International Joint Conf. on Pattern Recognition. Kyoto, Japan: IAPR. – 1978. - P. 1019-1021.

115. Li, K. P. Talker differences as they appear In correlation matrices of continuous speech spectra / K. P. Li, G. W. Hugnes // JASA. – 1974. - V. 55. - P. 833-837.

116. Bellisant, C. A system for segmentation and phonemic labeling of speech / C. Bellisant // In Proc. the 4 th International Joint Conf. on Pattern Recognition. Kyoto, Japan: IAPR. – 1978. - P. 1003-1005.

117. Kashyap, R. L. Recognition on spoken words and phrases in multitalker environment using syntactic methods / R. L. Kashyap, M. C. Mittal // IEEE Trans. Comput. – 1978. - V. 27. – N 5. - P. 442-450.

118. Tanaka, K. A standart category pattern making method with application to phoneme recognition / K. Tanaka // In: Proc. the 4 th International Joint Conf. on Pattern Recognition. Kyoto, Japan: IAPR. – 1978. - P. 1030-1033.

119. Pruzansky, S. Pattern – matching procedure for automatic talker recognition / S. Pruzansky // JASA. – 1963. -V. 35. - P. 354-358.

120. Матвеев, Ю. Н. Анализ возможности применения методов машинного обучения на основе многообразий в задачах распознавания дикторов / Ю. Н. Матвеев, А. К. Шулипа // Приборостроение. - 2014. - №2. - С. 70-76.

121. Савченко, В. В. Метод фонетического декодирования слов в информационной метрике Кульбака лейблера для систем автоматического анализа и распознавания речи с повышенным быстродействием / В. В. Савченко, А. В. Савченко // Информационно-управляющие системы. 2013. №2 (63). С. 7-12.

122. Бочаров, И. В. Распознавание речевых сигналов на основе метода спектрального оценивания / И. В. Бочаров, Д. Ю. Акатьев // Исследовано в России. - 2003. - С. 1537-1546.

123. Желтов, П. В. Алгоритмы идентификации фонем и формирования слова в системах распознавания речи на основе вейвлет-преобразования / П. В. Желтов, В. И. Семенов, А. И. Трофимова, А. К. Шурбин // Вестник ЧГУ. - 2014. - №2. - С. 98-102.

124. Савченко, Л. В. Алгоритм пофонемного распознавания устной речи на основе метода нечеткого фонетического кодирования-декодирования слов / Л. В. Савченко // Информационно-управляющие системы. - 2014. - №1 (68). - С. 23-31.

125. Грачев, А. М. Статистические подходы к автоматическому распознаванию речи / А.М. Грачев // Вестник ННГУ. 2015. №2-2. С. 376-379.

126. Лапко, А. В. Непараметрический алгоритм автоматической классификации статистических данных / А. В. Лапко, В. А. Лапко, А. Н. Хлопов // Приборостроение. - 2011. - №4. - С. 73-79.

127. Кошеков, К. Т. Методика и алгоритм автоматической классификации сигналов / К. Т. Кошеков // Вестник СГТУ. - 2008. - №1. -С. 107-113.

128. Крицкий, С. П. Реализация оптимизирующих преобразований программ с помощью структурных предикативных грамматик /С. П. Крицкий, Б. Ю. Тапкинов // Известия вузов. Северо-Кавказский регион. Серия: Естественные науки. 2006. №1. С. 3-14.

129. Massey, H. N. An experimental telemetry data compressor / H. N. Massey // In: Proc. of the National Telemetring Conf. N.-Y. – 1965. - P. 25-28.

130. Максимов, А. В. Структура алгоритмического и программного обеспечения микропроцессорной системы сбора и обработки ЭКГ-сигналов /А. В. Максимов // Известия ЮФУ. Технические науки. - 2004. - №2. - С. 127-132.

131. Тупиков, В. А. Лингвистические методы в задачах распознавания изображений / В. А. Тупиков, В. А. Павлова, С. Н. Крюков, М. В. Созинова, П. К. Шульженко // Известия ЮФУ. Технические науки. - 2015. - №1 (162). - С. 142-151.

132. Менлитдинов, А. С. Алгоритм анализа сердечных аритмий с использованием лингвистического и секвенциального анализа и алгоритма кластеризации COBWEB / А. С. Менлитдинов, М. А. Барков, А. В. Коробейников // Интеллектуальные системы в производстве. -2013. -№1 (21). –С. 131-136.

133. Митянок, В. В. Метод аппроксимации для определения числовых характеристик некоторых низкочастотных звуков человеческой речи / В. В. Митянок // Техническая акустика. - 2008. - №8. - С. 87-97.

134. Enrich, R.W. Representation of random waveforms by relational trees / R.W. Enrich, J.P. Foith // IEEE Trans. Comput. – 1976. - V. 25. - N 7. - P. 725-736.

135. Никифоров, И. В. Применение имитационного моделирования для разработки АСУ ТП непрерывного производства / И. В. Никифоров // Кибернетические проблемы АСУ технологическими процессами. М.: Знание. – 1978. –С. 99-106.

136. Никифоров, И. В. Статистический метод обнаружения момента времени изменения свойств измерительных средств / И. В. Никифоров // Статистические методы в теории измерений. Л.: ЛПИ. – 1978. - С. 49.

137. Королев, В. Ю. Прогнозирование параметров цунами у побережий Сахалинской и Камчатской областей методом оценивания вероятностей катастроф в неоднородных потоках экстремальных событий / В. Ю. Королев, Е. В. Арефьева, Р. А. Лазовский // Научные и образовательные проблемы гражданской защиты. - 2015. - №3 (26). - С. 73-77.

138. Третьяков, И. А. Методы параллельной сегментации экспериментальных кривых / И. А. Третьяков // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2018. – № 4. – С. 36-41.

139. Третьяков, И. А. Сравнительный анализ методов параллельной сегментации экспериментальных кривых / И. А. Третьяков // Донецкие чтения 2019: образование, наука, инновации, культура и вызовы современности: Материалы IV Международной научной конференции (Донецк, 31 октября 2019 г.). – Том 1: Физико-математические и технические науки. Часть 2 / под общей редакцией проф. С.В. Беспаловой. – Донецк: Изд-во ДонНУ, 2019. – С. 176-178.

140. Третьяков, И. А. Функции сложности для выделения и распознавания характерных участков экспериментальных кривых / И. А. Третьяков, В. В. Данилов // Вестник Донецкого национального университета. Серия А: Естественные науки. – 2017. – № 2. – С. 101-107.

141. Третьяков, И. А. Функции сложности для выделения и распознавания характерных участков экспериментальных кривых / И. А. Третьяков, А. В. Шалаев, В. В. Данилов // Донецкие чтения 2017: Русский мир как цивилизационная основа научно-образовательного и культурного развития Донбасса: Материалы Международной научной конференции студентов и молодых ученых (Донецк, 17-20 октября 2017 г.). – Том 1: Физико-математические и технические науки / под общей редакцией проф. С.В. Беспаловой. – Донецк: Изд-во ДонНУ, 2017. – С. 206-207.

142. Третьяков, И. А. Алгоритмы идентификации переходных участков экспериментальных кривых / И. А. Третьяков // Социально-гуманитарные и естественно-технические науки и вызовы современности: материалы международной научно-практической конференции (Ставрополь, 21 декабря 2017 г.) / С. Е. Шиянов, А. П. Федоровский (отв.ред.) – Ставрополь: АНО ВО СКСИ, 2017. – С. 824-828.

143. Данилов, В. В. Алгоритмы идентификации переходных участков экспериментальных кривых с применением аппроксимации / В. В. Данилов, И. А. Третьяков, А. В. Шалаев, Я. И. Рушечников // Сборник научных трудов Донецкого института железнодорожного транспорта. – 2018. – № 48. – С. 19-23.

144. Данилов, В. В. Алгоритмы экстраполяции участков экспериментальных кривых / В. В. Данилов, И. А. Третьяков, Я. И. Рушечников, А. В. Шалаев // Сборник научных трудов Донецкого института железнодорожного транспорта. – 2018. – № 50. – С. 10-15.

145. Третьяков, И. А. Алгоритмы прогнозного и формального типов экстраполяции участков экспериментальных кривых / И. А. Третьяков // Донецкие чтения 2018: образование, наука, инновации, культура и вызовы современности: Материалы III Международной научной конференции (Донецк, 25 октября 2018 г.). – Том 1: Физико-математические и технические науки / под общей редакцией проф. С.В. Беспаловой. – Донецк: Изд-во ДонНУ, 2018. – С. 162-164.

146. Данилов, В. В. Алгоритмизация присвоения символов анализируемым участкам экспериментальных кривых / В. В. Данилов, И. А. Третьяков, Я. И. Рушечников // Сборник научных трудов Донецкого института железнодорожного транспорта. – 2018. – № 51. – С. 15-22.

147. Алимурадов, А. К. Обзор и классификация методов обработки речевых сигналов в системах распознавания речи / А. К. Алимурадов, П. П. Чураков // Измерение. Мониторинг. Управление. Контроль. - 2015. - №2 (12). – С. 27-35.

148. Богоносцева, Т. А. (2013). Метод потенциальных функций в распознавании образов / Т.А. Богоносцева // Труды Международного симпозиума «Надежность и качество». – 2013. - Т. 1. - С. 154-155.

149. Третьяков, И. А. Исследование применимости функций сложности к задачам идентификации переходных участков экспериментальных кривых акустических колебаний / И. А. Третьяков, А. В. Шалаев, Я. И. Рушечников, В. В. Данилов // Вестник Луганского национального университета имени Владимира Даля. – 2018. – № 5(11). – С. 332-335.

150. Данилов, В. В. Идентификация переходных участков кривой акустических колебаний / В. В. Данилов, И. А. Третьяков, А. В. Шалаев, Я. И. Рушечников // Сборник научных трудов Донецкого института железнодорожного транспорта. – 2018. – № 49. – С. 10-16.

151. Третьяков, И. А. Исследование алгоритмов лингвистического описания участков экспериментальных кривых / И. А. Третьяков // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2019. – № 3. – С. 26-30.

152. Третьяков, И. А. Реализация алгоритмов формирования алфавита символов для анализа экспериментальных кривых / И. А. Третьяков // Материалы международной научно-практической конференции молодых исследователей им. Д. И. Менделеева (г. Тюмень, 15 ноября 2019 г.): сборник статей / отв. ред. А. Н. Халин. – Тюмень: ТИУ, 2020. – С. 28-31.

153. Data from ECG recording in today’s class [Электронный ресурс]. — Режим доступа: URL: https://bioelectromagnetism.wordpress.com/2012/11/28/data-from-ecg-recording-in-todays-class/ (дата обращения: 01.09.2019).

ПРИЛОЖЕНИЕ А

Документы, подтверждающие внедрение результатов диссертации

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ»
ул. Университетская, 24, г. Донецк, 83001, тел: приемная (062) 302-07-22,
справочная служба (062) 302-06-00, факс: (062) 302-07-49
e-mail: rector@donnu.ru Идентификационный код 02070803

12.12.2019 №5516/01-23/6.20
на № _____ от _____

Диссертационный совет Д 01.024.04 при ГОУ ВПО «Донецкий национальный технический университет» и ГОУ ВПО «Донецкий национальный университет»

СПРАВКА
о внедрении результатов исследований диссертационной работы
Третьякова Игоря Александровича
на тему «Автоматизированная система лингвистического анализа экспериментальных кривых», представленную на соискание ученой степени кандидата технических наук по специальности
05.13.06 – Автоматизация и управление технологическими процессами и производствами (по отраслям)

Результаты диссертационных исследований Третьякова Игоря Александровича, а именно: автоматизированная система научных исследований для структурного анализа экспериментальных данных; вычислительный алгоритм сегментации, способный выделять только отдельные участки, которые считаются информативными; вычислительный алгоритм лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами, в котором через конечное число циклов достигается устойчивая классификация были использованы при выполнении двух этапов научно-исследовательской работы Г-18/39 №0118D000013 «Моделирование защищенных инфокоммуникационных систем» в 2018-2019 гг.

Проректор по научной и инновационной деятельности ГОУ ВПО «ДонНУ», д-р техн. наук, проф. В.И. Сторожев

Соответствует оригиналу
Ученый секретарь Д 01.024.04
Т.В. Завадская

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ»
ул. Университетская, 24, г. Донецк, 83001, тел: приемная (062) 302-07-22,
справочная служба (062) 302-06-00, факс: (062) 302-07-49
e-mail: rector@donnu.ru Идентификационный код 02070803

12.12.2019 №5515/01-17/620
на № _____от _____

Диссертационный совет Д 01.024.04 при ГОУ ВПО «Донецкий национальный технический университет» и ГОУ ВПО «Донецкий национальный университет»

СПРАВКА
о внедрении результатов исследований диссертационной работы
Третьякова Игоря Александровича
на тему «Автоматизированная система лингвистического анализа экспериментальных кривых», представленную на соискание ученой степени кандидата технических наук по специальности
05.13.06 – Автоматизация и управление технологическими процессами и производствами (по отраслям)

Основные этапы обработки экспериментальных данных структурными методами, методика компьютерной обработки для выделения и анализа экспериментальных данных, методика сегментации и анализа экспериментальных данных и вычислительные алгоритмы сегментации на основе функции сложности для выделения информативных участков экспериментальных кривых, предложенные в диссертационном исследовании Третьякова Игоря Александровича, внедрены в учебный процесс ГОУ ВПО «Донецкий национальный университет» путем использования при чтении лекций и выполнении лабораторных работ по дисциплине «Цифровая обработка сигналов» для подготовки бакалавров по направлениям подготовки 03.03.03 Радиофизика и 10.03.01 Информационная безопасность, что отражено в рабочих программах вышеуказанных учебных дисциплин.

Проректор по научно-методической и учебной работе ГОУ ВПО «ДонНУ»,
д-р пед. наук, проф. Е.И. Скафа

Соответствует оригиналу
Ученый секретарь Д 01.024.04
Т.В. Завадская

ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ
«ДОНЕЦКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
ИМ. А.А. ГАЛКИНА»
ул. Розы Люксембург, 72, г. Донецк, 83114, тел./факс: (62) 311-52-27, факс: (62) 342-90-18,
E-mail: scsecr@donfti.ru web: http://www.donfti.ru ИКЮЛ 05420497

СПРАВКА
о внедрении результатов исследований диссертационной работы Третьякова Игоря Александровича
на тему «Автоматизация процедуры структурного анализа экспериментальных данных научных исследований», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.06 – Автоматизация и управление технологическими процессами и производствами (по отраслям)

Настоящим подтверждается, что результаты диссертационных исследований Третьякова Игоря Александровича были приняты в виде рекомендаций к использованию в научно-исследовательском процессе ГУ ДонФТИ, а именно:

- вычислительный алгоритм сегментации массивов экспериментальных данных с использованием функций сложности, способный осуществлять бинарную классификацию и выделять информативные участки;
- вычислительный алгоритм лингвистического описания участков экспериментальных кривых на основе сравнения с эталонами, способный осуществлять классификацию по признаку минимума расстояния до эталона.

ГУ ДонФТИ никаких финансовых обязательств перед соискателем не имеет.

Заместитель директора ГУ ДонФТИ по научной работе, канд. физ.-мат. наук, доцент А.В. Головчан

Соответствует оригиналу
Ученый секретарь Д 01.024.04
Т.В. Завадская