УДК 004.056:004.93

# Процедура формирования грамматики для описания спектрограмм технических каналов утечки информации

И. А. Третьяков
ГОУ ВПО «Донецкий национальный университет», г. Донецк
i.tretiakov@mail.ru

***Аннотация***
*В данной работе подробно рассмотрена реализация этапа присвоения сегментированным участкам спектрограмм символов некоторого алфавита, соответствующим определенным типам поведения в структурно-лингвистическом подходе анализа данных к задаче обнаружения технических каналов утечки информации. Решена задача построения трансформационной грамматики. Реализован метод построения цепочки, ближайшей к заданному множеству. Получены лингвистические описания исследуемых спектрограмм.*

## *Введение*

В настоящее время актуальной научно-технической задачей в области информационной безопасности является задача обнаружения технических каналов утечки информации и побочных электромагнитных излучений и наводок [1-4].

В данной работе рассмотрено решение этой задачи в рамках структурно-лингвистического подхода [5, 6] к анализу экспериментальных данных.

Структурный анализ предполагает последовательность реализации трех основных этапов обработки спектрограмм:

-выделения и распознавания характерных участков (сегментация);

-присвоения выделенным участкам символов некоторого алфавита, соответствующих определенным типам поведения кривой (формирование грамматики);

-анализа полученных последовательностей символов.

Данная статья посвящена реализации второго этапа - формированию грамматики для описания спектрограмм технических каналов утечки информации. В рамках этого этапа 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 будут присваивать им конкретные символы.

## *Постановка задачи формирования грамматики*

В качестве массива экспериментальных данных для реализации процедуры формирования грамматики послужили спектрограммы технических каналов утечки информации, полученные с помощью программно-определяемой радиосистемы (SDR) (рис. 1) на базе RTL2832 и R820T в ГОУ ВПО «Донецкий национальный университет».

*Рисунок 1 – Внешний вид программно-определяемой радиосистемы*

SDR представляют набор программных и специальных аппаратных средств, которые позволяют решать круг задач анализа взаимодействия радиоизлучений в широком диапазоне частот. Некоторые SDR системы могут использоваться для решения задач мониторинга спектра радиочастот [7, 8]. Данный класс устройств, благодаря возможностям программного управления, реализует функции физического уровня, что обеспечивает возможность обработки различных типов сигналов без изменения аппаратной части принимающего устройства. На рис. 2 показан процесс выявления технических каналов утечки информации и образец получаемых спектрограмм.

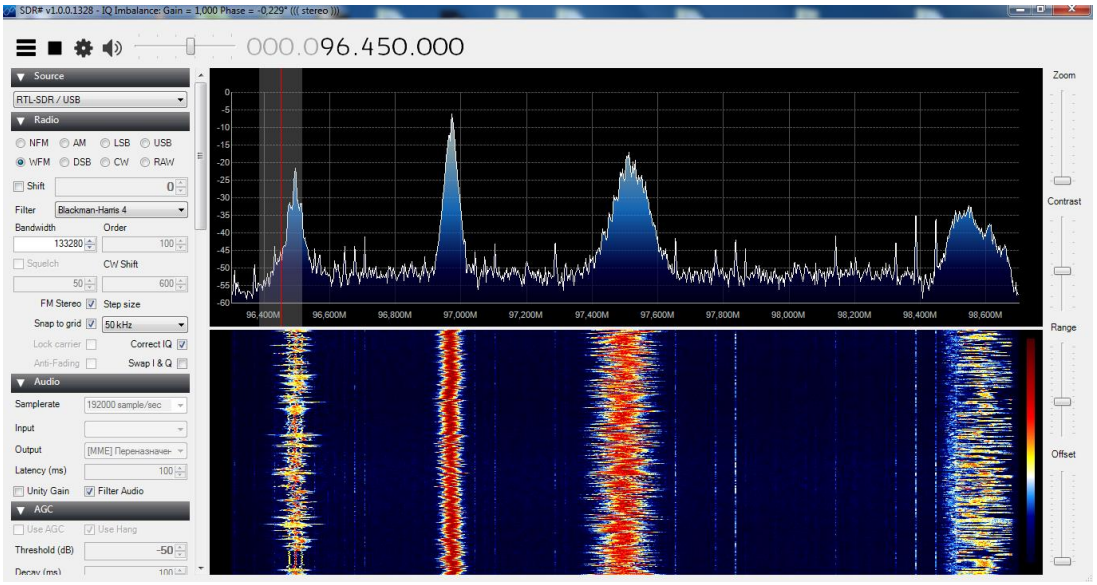


*Рисунок 2 – Визуализация спектра радиообстановки*

В результате выполнения первого этапа, а именно применения алгоритмов сегментации и классификации выделенных фрагментов [9] анализируемая спектрограмма 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 спектрограмм. При этом каждую цепочку можно рассматривать как некоторую фразу на данном языке.

Для описания получаемых спектрограмм зададим следующий алфавит: передний фронт обозначим символом L (left), центральную часть - C (central), задний фронт - R (right). Не информативные (фоновые) участки будут обозначаться символом b (background). Таким образом сформирован алфавит языка описания исследуемых спектрограмм, состоящий из 4х символов $A=\{L, C, R, b\}$.

С такой точки зрения задачу формирования языка описания спектрограмм технических каналов утечки информации 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 $T=\langle a_1 \ldots a_N \rangle$. 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 $j$: $c(i, j)=\langle a_i, a_{i+1} \ldots a_j \rangle$. 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. Необходимо найти такие наборы образов, из которых можно составить заданный текст $T$. Один из таких наборов и есть искомый словарь.

### *Формирование грамматики*

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 спектрограмм 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.

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 данных, 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 $\hat{L}(M)$.

Исследуемые спектрограммы можно описать правильными текстами вполне строго. Задачи их обработки нацелены на автоматизацию анализа таких данных, 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.

Таким образом, такую морфологическую грамматику следует рассматривать лишь как модель, более или менее точно аппроксимирующую тексты, порождаемые экспериментальными данными. Чтобы ее использовать в полной мере, требуется дополнить механизм порождения правильных текстов над словарем некоторым искажающим механизмом, позволяющим порождать цепочки, мало отличающиеся от правильных. Для этого прежде всего необходимо задать некоторую меру сходства двух произвольных цепочек символов алфавита $A$. Она должна отражать степень искажения 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 $\xi$ ,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 $\rho(T, \xi)$ данной цепочки $T$ нечеткому языку $\xi$ понимается величина:

$$\rho(T, \xi) = \min_{\hat{T} \in L(M)} r(\hat{T}, T). \qquad (1)$$

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 анализируемой спектрограммы.

### *Построение трансформационной грамматики*

Рассмотрим на примере элементарные трансформации текста, такие как стирание и дописывание одного символа из алфавита $A$. Пусть $T_1 = \langle LCLL \rangle$ и $T_2 = \langle LLCL \rangle$. 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 (1). Сеть $G(T_1, T_2)$, соответствующая паре цепочек $T_1 = \langle LCLL \rangle$ и

$T_2 = \langle LLCL \rangle$, представлена на рис. 3.

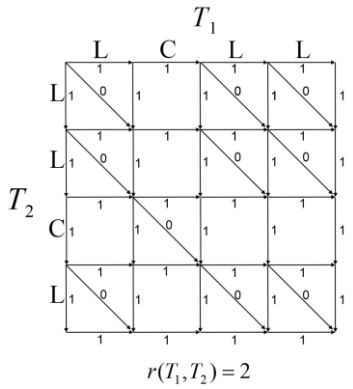


*Рисунок 3 – Сеть трансформации цепочек*

Теперь зададим ядро $\hat{L}(M)$ и рассмотрим цепочку $T \in L$ конечной длины, полученную при обработке исследуемой спектрограммы. Правильным описанием данной спектрограммы на языке $\hat{L}(M)$, аппроксимирующем исходный текст $T$, является цепочка удовлетворяет условию:

$$r(\hat{T}, T) = \min_{T' \in \hat{L}(M)} r(T', T). \qquad (2)$$

Идея нахождения правильного описания для данного текста полностью раскрывается на примере, изображенном на рис. 4, и описывается далее.

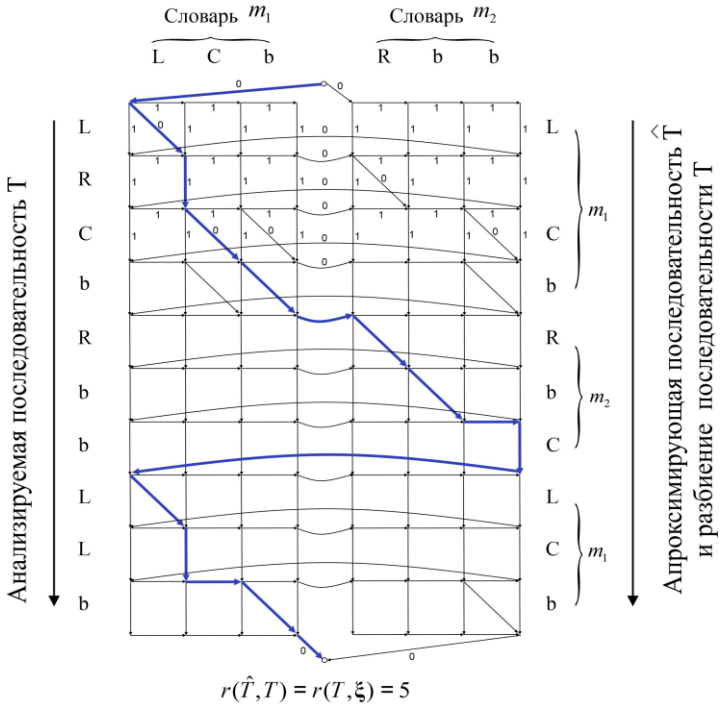


*Рисунок 4 – Разбиение цепочки в соответствии с заданным словарем*

Пусть словарь $M$ состоит из двух слов: $m_1 = \langle LCb \rangle$, $m_2 = \langle Rbb \rangle$, $A = \{L, C, R, b\}$. Для анализа предъявлена цепочка $\langle LRCbRbbLLb \rangle$, 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 (2) правильному описанию $\hat{T}$.

### *Метод построения цепочки, ближайшей к заданному множеству*

Рассмотрим на примере реализацию метода построения цепочки, ближайшей к заданному множеству цепочек. 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 $\{\langle CRL\rangle,\langle CLR\rangle,\langle bCLR\rangle\}$ могут быть записаны, в частности, два разбора, показанные на рисунке 5.

| | | | | |
|---|---|---|---|---|
| | C | R | L | |
| | C | | L | R |
| b | C | | L | R |

$m_1$

| | | | | | |
|---|---|---|---|---|---|
| | | | C | R | L |
| | C | L | | R | |
| b | C | L | | R | |

$m_2$

*Рисунок 5 – 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 $p_i(R)\ge 0{,}5$. 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 5 двух примеров разбора:

$$T(R_1)=\langle bCRLR\rangle,\ T(R_2)=\langle bCLCRL\rangle,$$
$$\tilde{T}(R_1)=\langle CL\rangle,\ \tilde{T}(R_2)=\langle CLR\rangle.$$

Рассмотрим оптимальное дописывание нового столбца к построенному разбору, показанного на рисунке 6.

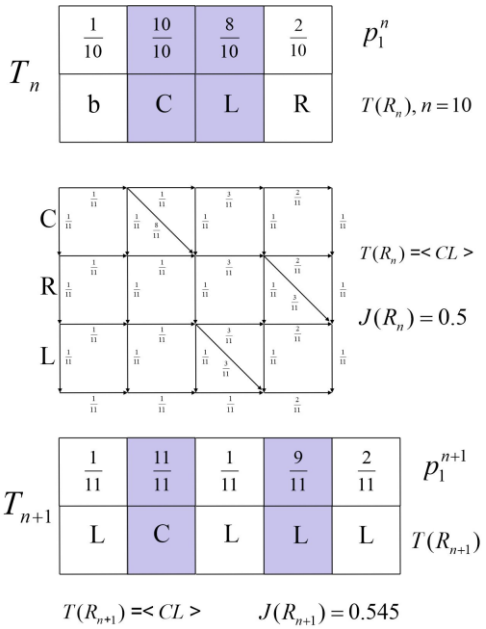


*Рисунок 6 – Дописывание новой строки к построенному разбору*

На рисунке 6 в закрашенных столбцах обозначены символы с частотами $p_i\ge 0{,}5$, 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.

### *Выводы*

В результате применения структурно-лингвистического подхода к задаче обнаружения технических каналов утечки информации и побочных электромагнитных излучений и

наводок можно заключить, что полученные лингвистические описания исследуемых спектрограмм представляют короткие и надежные правила для их анализа и позволяют в автоматизированном режиме выявлять изменения этих сигналов.

### *Литература*

1. Фаустов, И. С. Радиоконтроль служебных параметров сигналов Bluetooth / И. С. Фаустов, А. Б. Токарев, В. А. Сладких, В. А. Козьмин, И. Б. Крыжко // Системы управления, связи и безопасности. – 2021. – №3. – С. 135-151.

2. Ашихмин, А. В. Способ однопозиционного местоопределения источников радиоизлучения с использованием бортового радиопеленгатора беспилотного летательного аппарата вертолетного типа / А. В. Ашихмин, А. Д. Виноградов, А. М. Рембовский, В. А. Сладких // Системы управления, связи и безопасности. – 2021. – № 4. – С. 40-57.

3. Третьяков, И. А. Спектральный анализ радиосигналов в реальном времени на основе применения эхо-эффекта / И. А. Третьяков, В. В. Данилов // Вестник Астраханского государственного технического университета. Серия: управление, вычислительная техника и информатика. 2022. № 1. С. 53-59.

4. Третьяков, И. А. Исследование спектрограмм радиочастот методами лингвистического анализа / И. А. Третьяков, В. В. Данилов // Вестник Астраханского государственного технического университета. Серия: управление, вычислительная техника и информатика. – 2020. – № 3. – С. 26-33.

5. Моттль, В.В. Лингвистический анализ экспериментальных кривых / В. В. Моттль, И. Б. Мучник // ТИИЭР. – 1979. – Т.69. – №5. – С. 12-39.

6. Чистова, Г. К. Методы и процедуры построения лингвистической системы обнаружения и распознавания нарушителя / Г. К. Чистова, В. И. Волчихин // Вестник МГТУ им. Н.Э. Баумана. Серия «Приборостроение». – 2004. – № 3. – С. 96-114.

7. Рушечников, Я. И. Информационная технология радиомониторинга на основе программно-определяемой радиосистемы / Я. И. Рушечников, В. В. Данилов // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2020. – № 1. – С. 31-36.

8. Рушечников, Я. И. Информационная технология автоматизированной локализации источника излучения / Я. И. Рушечников, В. В. Данилов, С. В. Борщевский // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2020. – № 4. – С. 26-34.

9. 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 ДОНИЖТ. – 2018. – № 48. – С. 19-23.

***Третьяков И.А. Процедура формирования грамматики для описания спектрограмм технических каналов утечки информации.*** *В данной работе подробно рассмотрена реализация этапа присвоения сегментированным участкам спектрограмм символов некоторого алфавита, соответствующим определенным типам поведения в структурно-лингвистическом подходе анализа данных к задаче обнаружения технических каналов утечки информации. Решена задача построения трансформационной грамматики. Реализован метод построения цепочки, ближайшей к заданному множеству. Получены лингвистические описания исследуемых спектрограмм.*

***Ключевые слова:*** *структурно-лингвистический подход, программно-определяемая радиосистема, анализ данных, формирование грамматики, радиомониторинг*

***Tretiakov I.A. The procedure for forming a grammar for describing spectrograms of technical channels of information leakage.*** *In this paper, the implementation of the stage of assigning segmented sections of spectrograms of symbols of a certain alphabet corresponding to certain types of behavior in the structural-linguistic approach of data analysis to the task of detecting technical channels of information leakage is considered in detail. The problem of constructing a transformational grammar is solved. The method of constructing a chain closest to a given set is implemented. Linguistic descriptions of the studied spectrograms are obtained.*

***Key words:*** *structural-linguistic approach, software-defined radio system, data analysis, grammar formation, radio monitoring.*

*Статья поступила в редакцию 19.05.2022*
*Рекомендована к публикации*