

МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

Вестник СибГУТИ Том 18, № 1, 2024

Выпускается ежеквартально, выходит с 2007 г.

Учредитель –

федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»

Председатель редакционного совета А. Н. Фионов, д.т.н., проф.

Редакционный совет:

С. С. Абрамов, д.т.н., доц.	В. И. Носов, д.т.н., проф.
В. Б. Барахнин, д.т.н., доц.	С. В. Поршневу, д.т.н., проф.
В. М. Белов, д.т.н., проф.	А. С. Родионов, д.т.н., доц.
В. Н. Васюков, д.т.н., проф.	А. И. Романенко, д.ф.-м.н., проф.
В. Ю. Васильев, д.х.н., доц.	Б. Я. Рябко, д.т.н., проф.
Н. И. Горлов, д.т.н., проф.	И. И. Рябцев, д.ф.-м.н., чл.-корр. РАН
Н. Л. Казначеева, д.э.н., доц.	Э. Сименс, д.т.н.
В. С. Канев, д.т.н., проф.	О. В. Стукач, д.т.н., доц.
А. И. Карпович, д.э.н., проф.	В. К. Трофимов, д.т.н., проф.
М. Г. Курносов, д.т.н., проф.	А. И. Фалько, д.т.н., проф.
В. В. Лебедев, д.т.н., проф.	С. В. Федоренко, д.т.н., доц.
А. В. Лихачёв, д.т.н.	Б. Г. Хаиров, д.э.н., доц.
С. Н. Мамоиленко, д.т.н., доц.	А. Г. Черевко, к.ф.-м.н., доц.
О. Г. Мелентьев, д.т.н., проф.	С. В. Шидловский, д.т.н.
Р. В. Мещеряков, д.т.н., проф.	Ю. И. Шокин, д.ф.-м.н., акад. РАН
И. Г. Неизвестный, д.ф.-м.н., чл.-корр. РАН	В. П. Шувалов, д.т.н., проф.
С. Н. Новиков, д.т.н., доц.	

Редакция:

А. Н. Фионов (главный редактор), М. Ю. Галкина (заведующая редакцией),
Н. А. Двуреченская (технический и литературный редактор, компьютерная вёрстка),
Т. А. Алфёрова (лингвист-корректор).

Адрес редакции и издателя

630102, г. Новосибирск, ул. Кирова, д. 86

e-mail: vestnik@sibguti.ru

Информация о журнале доступна в сети Internet по адресу <https://vestnik.sibsutis.ru>.

Журнал включён в Перечень ВАК российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёных степеней доктора и кандидата наук.

Журнал зарегистрирован Федеральной службой по надзору за соблюдением законодательства в сфере массовых коммуникаций и охране культурного наследия, свидетельство о регистрации ПИ № ФС77-25835 от 29.09.2006. ISSN 1998-6920. Подписной индекс в каталоге «Урал-Пресс» – 82519.

Дата выхода в свет 29.03.2024. Отпускная (редакционная) цена 750 руб.

Отпечатано в типографии Альфа по адресу: 630015, г. Новосибирск, ул. Шишкина, д. 3.

Бумага офсетная, формат А4. Тираж 300 экз.

© СибГУТИ, 2024 г.

СОДЕРЖАНИЕ

А. Д. Петров, Е. А. Харченко Морфологический метод обнаружения аномальных состояний сервера	3
С. А. Рудин, К. В. Павский, А. Л. Ревун, А. В. Двуреченский Моделирование гетероэпитаксиального роста Ge на структурированной подложке Si(100) методом Монте-Карло и распараллеливание вычислений	16
К. Е. Крамаренко, А. Ю. Поляков, А. В. Ефимов Система анализа оптимизации доступа к общей памяти для реализации стандарта PMIx	29
В. А. Павский, К. В. Павский Стохастическая модель функционирования вычислительных систем с накопителем при групповом обслуживании	40
В. М. Деревяшкин, А. О. Виркунин, Н. И. Горлов, А. Е. Малашенко, А. И. Сидоренко Волоконно-оптическая система морского базирования для сбора и передачи данных	48
Е. П. Белова Использование формантных характеристик гласных и согласных звуков русского языка для аутентификации пользователей	59
И. В. Перцев, Е. И. Ситняковская Лингвистическое обеспечение информационных систем и процессов	70
М. С. Шушнов, Т. В. Шушнова, А. М. Оноприенко, Н. О. Абросимова Вопросы нелинейной акустики в проектировании громкоговорителей	78
Г. А. Фокин, И. В. Гришин Оценка влияния процедуры пространственного сглаживания корреляционной матрицы на эффективность метода 2D-MUSIC	91
А. А. Калачиков Анализ характеристик алгоритмов прекодирования сигналов в системе MU-MIMO с оцениванием канала	109
В. В. Подколзин, А. Н. Полетайкин, М. Ю. Галкина Математический аппарат для оценивания некоторых показателей качества учебного вебинара посредством машинного обучения	121

Морфологический метод обнаружения аномальных состояний сервера

А. Д. Петров^{1,2}, Е. А. Харченко¹

¹ Московский политехнический университет

² ООО «Безопасная информационная зона»

Аннотация: В работе предложен вычислительно простой алгоритм выявления выбросов и аномалий на основе морфологического анализа внутренней структуры многомерных данных. Важным преимуществом метода является возможность одновременной работы как с качественными, так и с количественными признаками. От аналогов его также отличает простота представления и интерпретации результатов. Доверительная область значений изучаемых объектов аппроксимируется объединением доверительных областей значений качественно однородных объектов (кластеров). Принадлежность объектов одному кластеру обуславливается характерными для предметной области причинно-следственными связями между признаками. В основе метода лежит построение конечного вероятностного пространства, каждый элемент которого (двоичный вектор) однозначно ставится в соответствие объектам выборки. На основании неравенства Чебышёва за выбросы принимаются малоомощные кластеры. За аномалии принимаются объекты, не принадлежащие совокупной доверительной области. Проработаны основанные на расстоянии Хэмминга механизмы сравнения: 1) кластера и кластера; 2) кластера и объекта; 3) объекта и объекта. Для демонстрации действенности метода разработан программный модуль для обнаружения аномальных состояний сервера на базе операционной системы семейства Linux. Он также может быть использован в качестве вспомогательного в профессиональных системах обнаружения вторжений.

Ключевые слова: многомерные данные, выбросы, аномалии, кластеризация, сегментация, машинное обучение без учителя, системы обнаружения вторжений.

Для цитирования: Петров А. Д., Харченко Е. А. Морфологический метод обнаружения аномальных состояний сервера // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 3–15. <https://doi.org/10.55648/1998-6920-2024-18-1-3-15>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Петров А. Д., Харченко Е. А., 2024

Статья поступила в редакцию 05.07.2023;
принята к публикации 12.08.2023.

1. Введение

Фокус внимания в области кибербезопасности неуклонно смещается в сторону разработки специализированных средств защиты информационных систем от целевых атак и сложных угроз, в том числе и комплексных таргетированных угроз [1]. В отличие от типового вредоносного программного обеспечения, целевые атаки осуществляются под активным контролем и удалённым управлением мотивированных и квалифицированных злоумышленников. Их сложно выявить, т.к. они всегда многоэтапные, при этом каждый отдельный шаг злоумышленника

в защищаемой системе может выглядеть легитимно. Наибольшую опасность из-за масштабности последствий представляют целевые атаки на объекты критической информационной инфраструктуры.

Большинство специализированных средств обнаружения целевых атак являются пассивными, поскольку не должны нарушать непрерывность производственных процессов: недопустимо прерывание работы основного программного обеспечения (в результате значительного потребления ресурсов защищаемой системы) или оперативная блокировка процессов (в результате ложных срабатываний). Такие решения не следят за поведением приложений или процессов в режиме реального времени, а изучают оставленные в системных журналах следы (логи) программ, задействуя при этом стандартные интерфейсы эксплуатируемой операционной системы.

Существуют два принципиально различных подхода к выявлению несанкционированной и вредоносной активности в компьютерной сети или на отдельном хосте: 1) на основании сигнатур (ранее известных способов проникновения) – проверяемые данные сравниваются с известными образцами сигнатур атаки, и в случае их совпадения создаётся оповещение безопасности; 2) на основании аномалий (новых угроз) – активность в сети или на хосте сравнивается с моделью корректного, доверенного поведения контролируемых элементов и фиксирует отклонения от неё.

Сложность самой природы предметной области обуславливает применение в целях детектирования аномальных состояний серверного оборудования методов машинного обучения. Наибольшую точность показывают модели, построенные на нейронных сетях [2], использование которых не лишено очевидных недостатков: невозможность интерпретации результатов и ресурсоёмкость вычислительных процессов. При использовании же базовых алгоритмов машинного обучения [3, 4, 5, 6] не исследуется или искажается внутренняя структура исходных данных. Приведём пример.

Обучающие выборки (датасеты) являются многомерными и в общем случае имеют неоднородную структуру. Но зачастую неразмеченные данные или данные одного класса рассматриваются как объекты одной сущности, каждый атрибут которой имеет нормальное распределение (его значения усредняются). Согласно первичному определению, за аномалии принимаются объекты, лежащие за границами доверительной области. Очевидно, что основным недостаток такого подхода – повышение порога детектирования аномалий (рис. 1, слева).

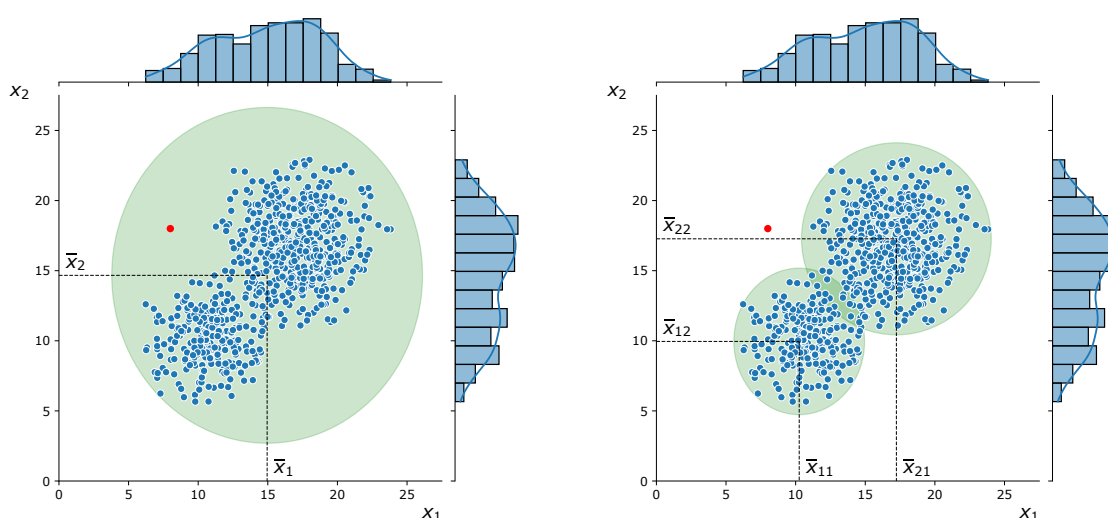


Рис. 1. Аномалия на фоне несегментированных (слева) и сегментированных (справа) данных

Кроме того, признаки датасета часто рассматриваются как независимые, что не может объективно отражать специфику предметной области. Так, у аномального объекта каждый признак

по отдельности вполне может попадать в свою область допустимых значений, в то время как в совокупности значения всех признаков не будут допустимыми (рис. 1, справа). Концепция современных решений мониторинга информационных процессов состоит в том, что каждое событие изучается не обособленно, а в контексте остальных событий.

В настоящей работе предложен один универсальный метод выявления выбросов и детектирования аномалий, учитывающий морфологию многомерных данных. Он прост для понимания и использования, т.к. построен на базовых понятиях линейной алгебры, теории вероятностей и статистики, что немаловажно для решения практических задач информационной безопасности. Также в рамках работы выработана реляционная модель состояния сервера и способ наполнения её данными.

Для подтверждения действенности метода и адекватности модели приводятся результаты работы программного модуля, предназначенного для пассивного обнаружения аномальных и подозрительных состояний сервера на базе операционной системы семейства Linux. Выбор операционной системы продиктован последними изменениями законодательства, согласно которым госорганам и госзаказчикам на критической инфраструктуре с 1 января 2025 года запрещается использовать иностранное программное обеспечение [7]. Наиболее распространёнными отечественными операционными системами в настоящее время являются Astra Linux и ALT Linux.

2. Морфологический метод выявления аномалий в многомерных данных

Предлагаемый метод выявления аномалий извлечён из метода принятия управленческих решений, освещённого в работах [8, 9], и по сути представляет собой метод кластеризации многомерных данных. В его основе лежит построение конечного вероятностного пространства L , множеством элементарных событий которого является линейное пространство двоичных векторов, в котором определены следующие операции:

- 1) сложение векторов – операция над двумя векторами, результатом которой является вектор, каждая составляющая которого равна сумме по модулю два одноимённых составляющих исходных векторов:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} (a_1 + b_1) \bmod 2 \\ (a_2 + b_2) \bmod 2 \\ \vdots \\ (a_k + b_k) \bmod 2 \end{pmatrix};$$

- 2) умножение вектора на скаляр – операция, результатом которой является вектор, составляющие которого равны конъюнкции скаляра и одноимённой составляющей исходного вектора:

$$\lambda \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} \lambda \& a_1 \\ \lambda \& a_2 \\ \vdots \\ \lambda \& a_k \end{pmatrix}.$$

Рассмотрим этапы построения пространства L .

Область допустимых значений X_j каждого показателя x_j разбивается на непересекающиеся подмножества, каждому из которых $X_{\alpha j}$ взаимно однозначно соответствует своя составляющая (свой бит) произвольного вектора пространства L (рис. 2):

$$X_j = X_{1j} \cup X_{2j} \cup \dots \cup X_{k_j j},$$

причём $X_{\alpha j} \cap X_{\beta j} \neq 0$ тогда и только тогда, когда $X_{\alpha j} = X_{\beta j}$. Здесь k_j – число качественно однородных подмножеств области допустимых значений показателя x_j . Размерность пространства L равна:

$$k = k_1 + k_2 + \dots + k_n,$$

где n – число показателей.

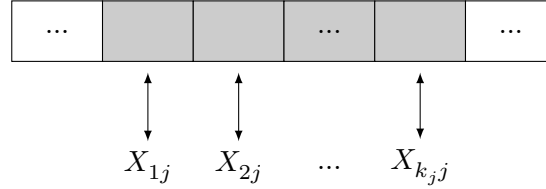


Рис. 2. Группа компонент вектора $\bar{y}_i \in L$, соответствующая показателю x_j

В случае, когда показатель x_j является качественным (измеренным в номинальной или ранговой шкале), т.е.

$$x_j \in \{x_{1j}, x_{2j}, \dots, x_{k_j j}\},$$

где $x_{\alpha j}$ – α -ое значение показателя x_j , в рассмотрение вводятся одноэлементные непересекающиеся множества

$$X_{\alpha j} = \{x_{\alpha j}\}.$$

Каждая область $X_{\alpha j}$ представляется единственным входящим в неё элементом.

В случае, когда показатель x_j является количественным (измеренным в шкале интервалов или отношений), строится сглаженная статистическая плотность распределения $\varphi(x_j)$ значений случайной величины x_j , в общем случае она имеет многовершинный характер (рис. 3).

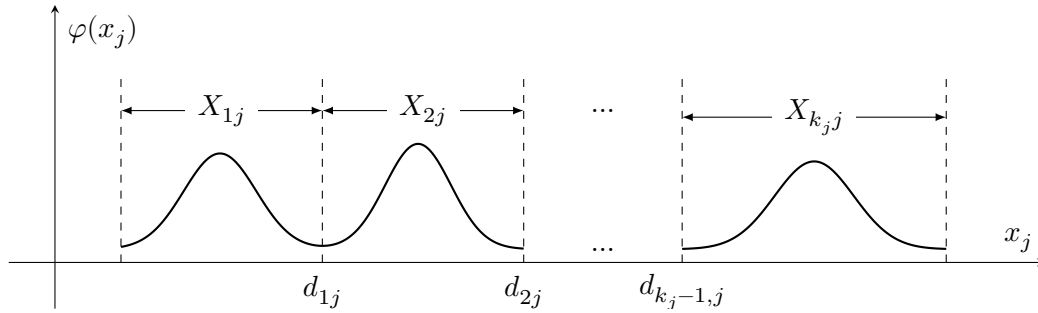


Рис. 3. Область допустимых значений X_j количественного показателя x_j

Точки минимумов $d_{\alpha j}$ плотности распределения $\varphi(x_j)$ случайной величины x_j разбивают область допустимых значений X_j количественного показателя x_j на непересекающиеся подмножества:

$$\begin{cases} X_{1j} = \{x_j \mid x_j < d_{1j}\}, \\ X_{2j} = \{x_j \mid d_{1j} \leq x_j < d_{2j}\}, \\ \dots \\ X_{k_j j} = \{x_j \mid x_j \geq d_{k_j-1,j}\}. \end{cases}$$

Каждая область сгущения значений $X_{\alpha j}$ имеет единственную вершину плотности распределения $\varphi(x_j)$ и представляется условными математическим ожиданием и средним квадратическим отклонением показателя x_j .

Разбиению одномерных областей допустимых значений всех показателей на непересекающиеся подмножества соответствует разбиение многомерной области допустимых значений X объектов моделируемой сущности (в нашем случае – нормального состояния сервера), т.е. различных комбинаций значений показателей, на непересекающиеся подмножества (кластеры), каждое из которых либо вообще не содержит ни одной комбинации значений показателей, либо содержит только такие комбинации значений показателей, у которых значения одноимённых качественных показателей равны, а значения каждого количественного показателя принадлежат одному и тому же подмножеству разбиения области допустимых значений этого показателя.

Таким образом:

$$X = \bigcup_{\substack{\alpha_1 \in \overline{1, k_1}, \\ \dots \\ \alpha_n \in \overline{1, k_n}}} (X_{\alpha_1 1} \times X_{\alpha_2 2} \times \dots \times X_{\alpha_n n}),$$

где

$$\begin{aligned} X_{\alpha_1 1} \times X_{\alpha_2 2} \times \dots \times X_{\alpha_n n} = \\ = \{(x_1, x_2, \dots, x_n) \mid (x_1 \in X_{\alpha_1 1}) \& (x_2 \in X_{\alpha_2 2}) \& \dots \& (x_n \in X_{\alpha_n n})\}, \end{aligned}$$

причём $(X_{\alpha_1 1} \times X_{\alpha_2 2} \times \dots \times X_{\alpha_n n}) \cap (X_{\beta_1 1} \times X_{\beta_2 2} \times \dots \times X_{\beta_n n}) \neq 0$ тогда и только тогда, когда $\alpha_i = \beta_i, i \in \overline{1, n}$.

Рассмотрим множество C , элементами которого являются подмножества рассмотренного выше разбиения области допустимых значений моделируемой сущности:

$$C = \{(X_{\alpha_1 1}, X_{\alpha_2 2}, \dots, X_{\alpha_n n}) \mid (X_{\alpha_1 1} \in C_1) \& (X_{\alpha_2 2} \in C_2) \& \dots \& (X_{\alpha_n n} \in C_n)\},$$

где

$$\begin{cases} C_1 = \{X_{11}, X_{21}, \dots, X_{k_1 1}\}, \\ C_2 = \{X_{12}, X_{22}, \dots, X_{k_2 2}\}, \\ \dots \\ C_n = \{X_{1n}, X_{2n}, \dots, X_{k_n n}\}. \end{cases}$$

Каждому элементу множества C однозначно соответствует свой вектор введённого линейного пространства. Из способа построения этого пространства следует, что в данном векторе в группе составляющих (битов), биективно соответствующей показателю x_j , существует единственная 1, остальные разряды этой группы равны 0. Векторам, у которых хотя бы в одной из рассматриваемых групп разрядов отсутствует 1 либо их число больше одной, не соответствуют элементы множества C . Векторам, у которых в каждой из рассматриваемых групп разрядов существует единственная 1, взаимно однозначно соответствуют элементы множества C .

В пространстве L определено расстояние Хэмминга между двумя произвольными векторами:

$$\begin{aligned} \rho_{\text{ham}}(\bar{y}_i, \bar{y}_j) &= \rho_{\text{ham}}((y_{1i}, y_{2i}, \dots, y_{k_i i}), (y_{1j}, y_{2j}, \dots, y_{k_j j})) = \\ &= |\{(y_{\alpha i}, y_{\alpha j}) \mid y_{\alpha i} + y_{\alpha j} = 1\}|. \end{aligned}$$

Расстояние между двумя элементами множества C определим как половину расстояния Хэмминга между соответствующими векторами пространства L :

$$\rho((X_{\alpha_1 1}, X_{\alpha_2 2}, \dots, X_{\alpha_n n}), (X_{\beta_1 1}, X_{\beta_2 2}, \dots, X_{\beta_n n})) = \frac{1}{2} \cdot \rho_{\text{ham}}(\bar{y}_i, \bar{y}_j). \quad (1)$$

Непосредственной подстановкой можно доказать, что половина расстояния Хэмминга удовлетворяет аксиомам расстояния. Из способа построения пространства L следует, что это расстояние численно равно количеству пар одноимённых составляющих $(X_{\alpha_j j}, X_{\beta_j j})$, у которых $X_{\alpha_j j} \neq X_{\beta_j j}$.

Расстояние между двумя комбинациями значений показателей сущности (т.е. между двумя объектами сущности) положим равным расстоянию между двумя элементами множества C (т.е. между двумя кластерами), которым эти комбинации значений принадлежат. Таким образом, расстояние между двумя комбинациями значений показателей равно числу пар одноимённых показателей, которые принадлежат различным подмножествам из разбиения области допустимых значений их комбинаций.

Расстояние между отдельной комбинацией значений показателей (объектом сущности) и произвольным элементом множества C (кластером) положим равным расстоянию между двумя соответствующими векторами пространства C .

Элементарным событиям (векторам введённого пространства L) припишем вероятности следующим образом: если вектор не соответствует ни одному элементу множества C (формальным критерием этого является отсутствие единиц или наличие более одной единицы в группе разрядов вектора, соответствующего хотя бы одному из показателей), ему приписывается вероятность, равная 0; если вектор соответствует элементу множества C (формальным критерием этого является наличие единственной единицы в каждой группе разрядов вектора, соответствующей одному из показателей), ему приписывается вероятность, равная отношению числа объектов сущности, соответствующих рассматриваемому элементу множества C , к общему числу m объектов сущности.

Во многих практических задачах кластеризации объекты сущности будут группироваться в небольшом числе элементов множества C . Большинство же других элементов множества C будут иметь нулевую или близкую к нулю вероятность.

Упорядочим элементарные события (двоичные вектора) в соответствии с убыванием их вероятностей. После этого введём новую дискретную случайную величину ξ , значения которой равны номерам элементарных событий в этом упорядоченном ряду, а вероятности – вероятностям соответствующих элементарных событий.

После вычисления математического ожидания μ_ξ и среднего квадратического отклонения σ_ξ этой случайной величины на основании неравенства Чебышёва

$$P(|\xi - \mu_\xi| \geq 3\sigma_\xi) \leq \frac{1}{9}$$

можно заключить, что подавляющее большинство объектов моделируемой сущности (не меньше 89 %) принадлежит тем элементам множества C , которым соответствуют значения случайной величины ξ , удовлетворяющие неравенству

$$1 \leq \xi \leq \lfloor \mu_\xi + 3\sigma_\xi \rfloor, \quad (2)$$

и доля выбросов не превышает 11 %.

Тогда за доверительную область объектов исследуемой сущности следует принять объединение доверительных областей только таких элементов множества C , которые соответствуют практически возможным значениям величины ξ . Детектирование аномального объекта, в свою очередь, сводится к проверке объекта на непринадлежность совокупной доверительной области сущности.

Доверительная область объектов отдельного элемента множества C представляет собой многомерный эллипсоид, описываемый уравнением:

$$\frac{(x_1 - \mu_{x_1})^2}{(3\sigma_{x_1})^2} + \frac{(x_2 - \mu_{x_2})^2}{(3\sigma_{x_2})^2} + \dots + \frac{(x_n - \mu_{x_n})^2}{(3\sigma_{x_n})^2} = 1, \quad (3)$$

положение и полуоси эллипса определяются средними значениями и средними квадратическими отклонениями показателей объектов, принадлежащих рассматриваемому кластеру (здесь законы распределения показателей полагаем неизвестными).

3. Реляционная модель состояния сервера

При эксплуатации любой компьютерной системы продуцируется большое число так называемых сигналов (значения сенсоров, команд, параметров логики управления и т.д.), они тесно взаимосвязаны, что определяется физикой и логикой производственных процессов. Вследствие этого воздействие на одни параметры процесса неизбежно влечёт за собой изменение других параметров. В совокупности показания «в моменте» всех источников сигналов системы (условных сенсоров) определяют её состояние – нормальное (штатное) или аномальное (потенциально опасное).

Формализовать характер связей между показаниями сенсоров в общем случае не представляется возможным, поэтому на практике по собранным показаниям сенсоров формируется реляционная модель нормального состояния системы (с неявным учетом корреляций между сигналами). Тогда под аномальными понимают нетипичные состояния системы (значительно удалённые от нормальных).

В данной работе для формирования обучающей выборки, описывающей состояние сервера, разработана следующая схема. Предварительно регистрируются характеристики запущенных процессов:

- pid – идентификатор процесса;
- name – название процесса;
- username – пользователь, от имени которого запущен процесс;
- ppid – идентификатор родителя процесса;
- parent_name – название родителя процесса;
- cpu_percent – процент потребления ресурсов процессора процессом;
- memory_percent – процент потребления ресурсов оперативной памяти процессом;
- num_threads – количество потоков процесса;
- terminal – идентификатор терминала, из которого был запущен процесс;
- nice – приоритет выполнения процесса;
- cmdline – команда, которой был запущен процесс;
- exe – путь к исполняемому файлу процесса;
- status – статус процесса;
- create_time – время запуска процесса;
- connections – количество открытых соединений процесса;
- open_files – количество открытых файлов процесса.

Данные о процессах собираются с некоторым настраиваемым шагом (в работе равным одной минуте) с помощью кроссплатформенной библиотеки psutil языка программирования Python.

Обучающий датасет производится агрегированием данных исходного датасета по следующему правилу: для каждой строки с временем t_c агрегируются все строки с временем t_r , которые отвечают условию

$$(t - \Delta) < t_r \leq t_c, \quad (4)$$

где Δ – некоторый небольшой заданный промежуток времени (в работе равный десяти минутам). Для того, чтобы процесс не учитывался несколько раз во время агрегации, все процессы, попадающие в заданный интервал перед основной агрегацией, группируются по идентификаторам процесса (pid) и по пользователям (username), от имени которых запущены процессы. При этом все количественные характеристики усредняются, а качественные соединяются в строку через запятую.

В результате методом скользящего окна была выработана следующая система суррогатных признаков для моделирования состояния сервера:

- cpu_percent_avg – средний процент потребления ресурсов процессора;

- `memory_percent_avg` – средний процент потребления ресурсов оперативной памяти;
- `num_threads_avg` – среднее количество потоков;
- `connections_avg` – среднее количество открытых соединений;
- `open_files_avg` – среднее количество открытых файлов;
- `cpu_percent_sum` – сумма процентов потребления ресурсов процессора;
- `memory_percent_sum` – сумма процентов потребления ресурсов оперативной памяти;
- `num_threads_sum` – сумма количества потоков;
- `connections_sum` – сумма количества открытых соединений;
- `open_files_sum` – сумма количества открытых файлов;
- `cpu_percent_max` – максимальный процент потребления ресурсов процессора;
- `memory_percent_max` – максимальный процент потребления ресурсов оперативной памяти;
- `num_threads_max` – максимальное количество потоков;
- `connections_max` – максимальное количество открытых соединений;
- `open_files_max` – максимальное количество открытых файлов;
- `idle_status_count` – количество процессов со статусом `idle`;
- `sleeping_status_count` – количество процессов со статусом `sleeping`;
- `running_status_count` – количество процессов со статусом `running`;
- `zombie_status_count` – количество процессов со статусом `zombie`;
- `disk_sleep_status_count` – количество процессов со статусом `disk_sleep`;
- `root_processes_count` – количество процессов, запущенных от имени суперпользователя (`root`);
- `system_processes_count` – количество системных процессов (имя пользователя начинается с `sys`);
- `time_of_day` – время (ночь, утро, день, вечер).

Признак автоматически исключается из датасета, если не является «говорящим», т.е. имеет одно значение (для качественного признака) или одну область сгущения значений (для количественного признака), – по нему объекты сущностно неразличимы. Для наглядности на рис. 4 приведены распределения значений признаков расчётного примера.

4. Программный модуль выявления аномальных состояний сервера

Формализованный выше математический аппарат для выявления выбросов и аномалий в многомерных данных лежит в основе разработанного в рамках работы программного модуля детектирования аномальных состояний сервера на базе операционной системы семейства Linux. Для демонстрации работы модуля был выбран сервер на базе операционной системы Ubuntu 20.04 со следующими характеристиками: процессор Intel(R) Xeon(R) Platinum 8171M CPU @ 2.60 GHz, ОЗУ 4 Гб.

На сервере был запущен Django-проект с использованием `docker-compose`, вместе с ним в Docker-контейнерах были развёрнуты база данных на основе СУБД PostgreSQL и web-сервер `nginx`. Также на сервере был запущен внешний `nginx`, который перенаправлял запросы Django-приложению и клиенту детектора. Для запуска сборщика данных, детектора аномалий и клиента в фоновом режиме использовался сервис `supervisor`.

Сбор первичных данных осуществлялся в течение четырёх дней. За весь период сервер использовался в обычном режиме в контролируемых условиях. Никаких сбоев и отказов не наблюдалось, поэтому собранные данные можно считать образцовыми. За всё время было собрано 121 МБ данных, или 889039 записей с информацией о запущенных процессах. Записанные в CSV-файл данные были выгружены с сервера и импортированы в локальную базу данных.

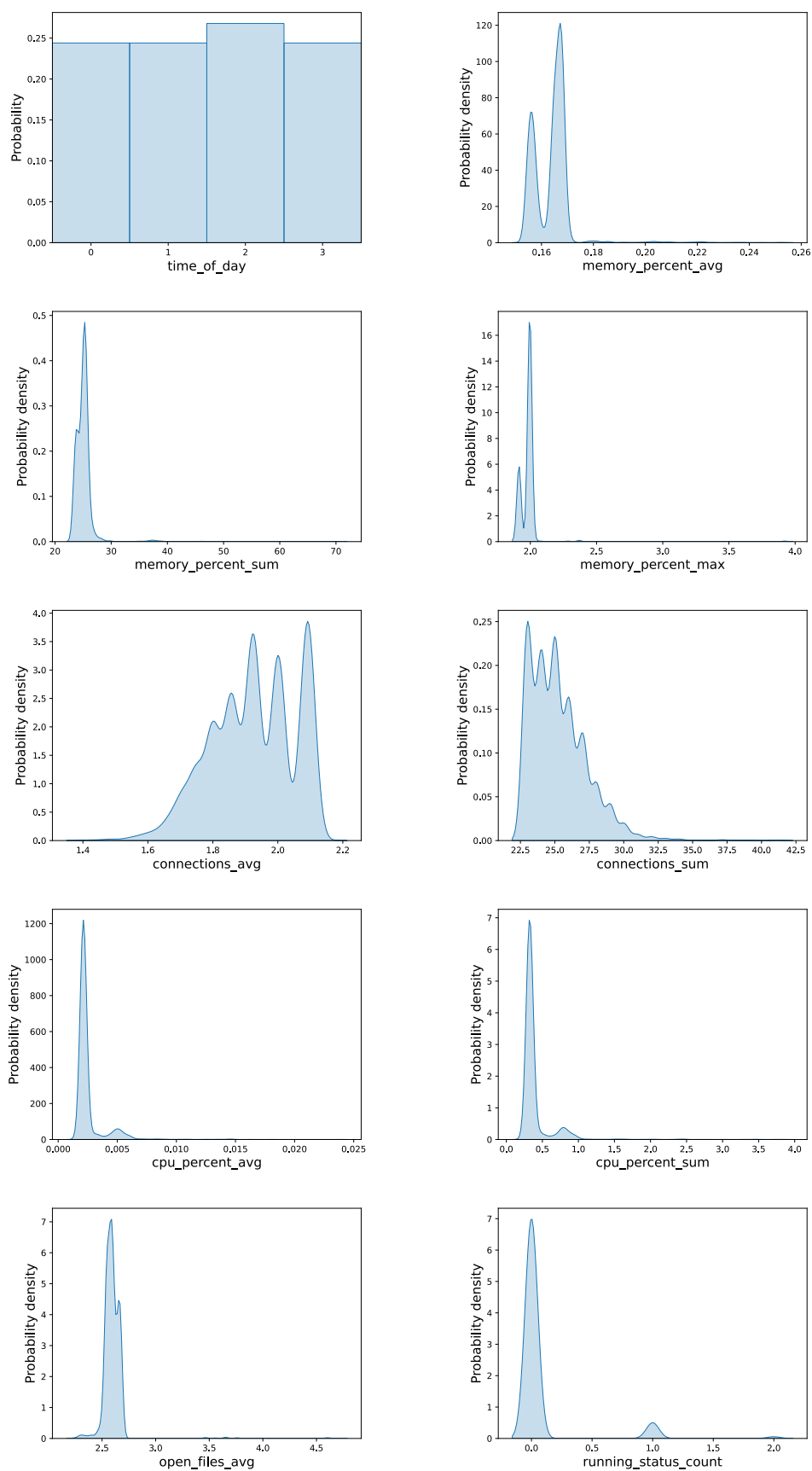


Рис. 4. Распределение значений признаков расчётной обучающей выборки

Затем они были поданы на вход детектора аномалий, причём без традиционной нормализации данных. После агрегации по правилу (4) был получен обучающий датасет из десяти признаков (их структура приведена на рис. 4). В процессе обучения на основании (2) детектором автоматически было выделено 313 кластеров нормальных состояний сервера (и отброшены выбросы – маломощные кластеры).

Размерность двоичных векторов, отождествляемых с кластерами, составила 32. Под детализацию области допустимых значений признака `time_of_day` программой было выделено 4 бита, признака `memory_percent_avg` – 2 бита, признака `memory_percent_sum` – 2 бита, признака `memory_percent_max` – 2 бита, признака `connections_avg` – 5 бит, признака `connections_sum` – 8 бит, признака `cpu_percent_avg` – 2 бита, признака `cpu_percent_sum` – 2 бита, признака `open_files_avg` – 3 бита и признака `running_status_count` – 2 бита.

Так, например, наибольшую частоту показал кластер, морфология которого представляется вектором 1000 01 01 01 00001 10000000 10 10 010 10. Здесь единица во второй группе битов трактуется как принадлежность значения признака `memory_percent_avg` интервалу, представляемому $\mu_{22} = 16.77e-2$ и $\sigma_{22} = 66.58e-5$.

Для проверки корректности работы детектора был проведён ряд экспериментов:

1. Обычная работа с сервером: посетили несколько раз развёрнутый на сервере сайт, выполнили несколько действий на сайте; зашли на сервер, посмотрели логи приложений, переключились между директориями.
2. Запуск процесса, который раньше никогда не запускался: запустили предварительно написанный на языке Python скрипт, который сгенерировал в большом объёме случайные данные, записали их во временную директорию `/tmp`, затем считали и удалили созданные файлы.
3. Подбор логина и пароля для ssh: скачали список из пятисот самых распространённых паролей, создали файл с распространёнными логинами пользователей и с помощью утилиты `hydra` запустили команду подбора пароля.
4. Реализация DoS-атаки на сервер: с помощью программы `Cyphon-DoS` послали большое количество запросов к серверу.

Детектор не выявил аномальные состояния при нормальной работе сервера и, наоборот, чётко детектировал аномальные состояния, которые проявились в процессе смоделированных компьютерных атак. Для повышения чувствительности алгоритма был установлен порог детектирования аномалии, равный на основании (1) одному отличающемуся признаку.

На более сложных данных детектор ожидаемо показывает более низкую эффективность. Это объясняется избыточностью доверительных областей кластеров (в пользу иллюстративности метода). Границы и ориентацию доверительных областей можно уточнить путем уменьшения числа стандартных отклонений в (3) или применения метода главных компонент к каждому кластеру.

5. Заключение

Одна из современных парадигм информационной безопасности заключается в том, что невозможно гарантированно предотвратить проникновение в систему, но важно как можно быстрее обнаружить подозрительное поведение системы и не позволить атаке развиваться до наступления недопустимого события. С этой целью разрабатывают специальные, требующие корректного использования, инструменты обнаружения аномального или подозрительного поведения системы и оповещения о нём.

В настоящей работе предложен вычислительно простой алгоритм выявления выбросов и аномалий на основе морфологического анализа внутренней структуры многомерных данных.

Важным преимуществом метода является возможность одновременной работы как с качественными, так и с количественными признаками. От аналогов его также отличает простота представления и интерпретации результатов. По сути, доверительная область изучаемых объектов аппроксимируется объединением доверительных областей качественно однородных объектов.

Разработанный на основе представленного метода программный модуль может быть использован в качестве вспомогательного в полноценных IDS-системах. Как правило, они чувствительны и к несанкционированным сбоям или отказам системы, которые также несут потенциальные угрозы информационной безопасности.

Литература

1. *Левцов В.* Анатомия таргетированной атаки [Электронный ресурс]. URL: <https://www.kaspersky.ru/blog/targeted-attack-anatomy/4388> (дата обращения: 28.06.2023).
2. *Лаврентьев А.* MLAD: обнаружение аномалий методами машинного обучения [Электронный ресурс]. URL: <https://ics-cert.kaspersky.ru/publications/reports/2018/01/16/mlad-machine-learning-for-anomaly-detection> (дата обращения: 28.06.2023).
3. Указ Президента Российской Федерации от 30.03.2022 № 166 "О мерах по обеспечению технологической независимости и безопасности критической информационной инфраструктуры Российской Федерации" [Электронный ресурс]. URL: <http://publication.pravo.gov.ru/Document/View/0001202203300001> (дата обращения: 28.06.2023).
4. ГОСТ Р ИСО 16269-4-2017 "Статистические методы. Статистическое представление данных. Часть 4. Выявление и обработка выбросов". М.: Стандартиформ, 2017. 53 с.
5. *Дьяконов А. Г., Головина А. М.* Выявление аномалий в работе механизмов методами машинного обучения // Аналитика и управление данными в областях с интенсивным использованием данных. 2017. С. 469–476.
6. *Han J., Kamber M., Pei J.* Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011. 740 p.
7. *Tan P.-N., Steinbach M., Karpatne A., Kumar V.* Introduction to Data Mining. Pearson, 2019. 839 p.
8. *Харченко Е. А.* Морфологический подход к принятию обоснованных решений по экспертным суждениям // Вестник ТвГУ. Серия: Прикладная математика. 2019. № 2. С. 42–56. <https://doi.org/10.26456/vtppmk531>.
9. *Харченко Е. А.* Алгоритм морфологического метода экспертных оценок для решения задачи прогнозирования // Компьютерные инструменты в образовании. 2023. № 2. С. 5–20. <https://doi.org/10.32603/2071-2340-2023-2-5-20>.

Петров Антон Денисович

магистрант кафедры информационной безопасности Московского политехнического университета (Московский Политех, 107023, Москва, ул. Большая Семёновская, д. 38); разработчик направления анализа защищённости ООО «Безопасная информационная зона» (ООО «БИ-Зон», 105066, Москва, ул. Ольховская, д. 4, корп. 2), e-mail: antonp2@yandex.ru, ORCID ID: 0009-0007-9546-8544.

Харченко Елена Алексеевна

старший преподаватель кафедры инфокогнитивных технологий Московского политехнического университета (Московский Политех, 107023, Москва, ул. Большая Семёновская, д. 38), e-mail: elenakhaa@yandex.ru, ORCID ID: 0000-0002-5082-4564.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: каждый автор внёс равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Morphological Method for Detecting Abnormal Server States

Anton D. Petrov^{1,2}, Elena A. Kharchenko¹

¹ Moscow Polytechnic University,

² «BiZone» Limited Liability Company

Abstract: The paper proposes a computationally simple algorithm for detecting outliers and anomalies based on morphological analysis of the internal structure of multidimensional data. An important advantage of the method is the possibility of simultaneous work with qualitative and quantitative signs. It is also distinguished from its analogues by the simplicity of presentation and interpretation of the results. The values' confidence range of the studied objects is approximated by combining the values' confidence ranges of qualitatively homogeneous objects (clusters). The belonging of objects to one cluster is determined by the causal relationships between the features characteristic of the subject area. The method is based on the construction of a finite probability space and each element of binary vector is uniquely assigned to the objects of the sample. Based on the Chebyshev inequality, low-power clusters are taken as emissions. Objects that do not belong to the aggregate confidence area are taken as anomalies. Comparison mechanisms based on the Hamming distance have developed: 1) cluster and cluster; 2) cluster and object; 3) object and object. To demonstrate the effectiveness of the method a software module for detecting abnormal server states based on the Linux operating system has been developed. It can also be used as an auxiliary in professional intrusion detection systems.

Keywords: multidimensional data, outliers, anomaly, clustering, segmentation, unsupervised learning, intrusion detection system.

For citation: Petrov A. D., Kharchenko E. A. Morphological method for detecting abnormal server states (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 3-15. <https://doi.org/10.55648/1998-6920-2024-18-1-3-15>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Petrov A. D., Kharchenko E. A., 2024

The article was submitted: 05.07.2023;
accepted for publication 12.08.2023.

References

1. Levtsov V. Anatomya targetirovannoj ataki [The anatomy of a targeted attack], available at: <https://www.kaspersky.ru/blog/targeted-attack-anatomy/4388> (accessed 28.06.2023).

2. Lavrentyev A. MLAD: obnaruzhenie anomalij metodami mashinnogo obucheniya [MLAD: Anomaly detection by machine learning methods], available at: <https://ics-cert.kaspersky.ru/publications/reports/2018/01/16/mlad-machine-learning-for-anomaly-detection> (accessed 28.06.2023).
3. Ukaz Prezidenta Rossijskoj Federacii ot 30.03.2022 № 166 "O merah po obespecheniyu tekhnologicheskoy nezavisimosti i bezopasnosti kriticheskoy informacionnoj infrastruktury Rossijskoj Federacii" [Decree of the President of the Russian Federation No. 166 dated 30.03.2022 "On Measures to ensure the Technological Independence and security of the Critical Information Infrastructure of the Russian Federation"], available at: <http://publication.pravo.gov.ru/Document/View/0001202203300001> (accessed 28.06.2023).
4. GOST R ISO 16269-4-2017 "Statisticheskie metody. Statisticheskoe predstavlenie dannyh. CHast' 4. Vyyavlenie i obrabotka vybrosovi" [ISO 16269-4-2017 "Statistical methods. Statistical data presentation. Part 4. Detection and treatment of outliers"]. Moscow, Standartinform, 2017. 53 p.
5. D'yakonov A. G., Golovina A. M. Vyyavlenie anomalij v rabote mekhanizmov metodami mashinnogo obucheniya [Anomaly detection in mechanisms using machine learning]. *Analitika i upravlenie dannymi v oblastyah s intensivnym ispol'zovaniem dannyh*, 2017, pp. 469-476.
6. Han J., Kamber M., Pei J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011. 740 p.
7. Tan P.-N., Steinbach M., Karpatne A., Kumar V. *Introduction to Data Mining*. Pearson, 2019. 839 p.
8. Kharchenko E. A. Morfologicheskij podhod k prinyatiyu obosnovannyh reshenij po ekspertnym suzhdeniyam [The morphological approach to making reasonable decisions based on expert judgements]. *Vestnik TvGU. Seriya: Prikladnaya Matematika*, 2019, no. 2, pp. 42-56. <https://doi.org/10.26456/vtpmk531>
9. Kharchenko E. A. Algoritm morfologicheskogo metoda ekspertnyh ocenok dlya resheniya zadachi prognozirovaniya [Algorithm of the morphological method of expert estimates for solving the forecasting problem]. *Computer tools in education*, 2023, no. 2, pp. 5-20. <https://doi.org/10.32603/2071-2340-2023-2-5-20>.

Anton D. Petrov

Master's Degree Student of the Department of Information Security, Moscow Polytechnic University (Moscow Poly, Russia, 107023, Moscow, B. Semenovskaya St. 38); Developer of the Security Analysis Direction, «BiZone» Limited Liability Company («BiZone» LLC, Russia, 105066, Moscow, Olkhovskaya St., Bld. 2, 4), e-mail: antonp2@yandex.ru, ORCID ID: 0009-0007-9546-8544.

Elena A. Kharchenko

Senior Lecturer of the Department of Infocognitive Technologies, Moscow Polytechnic University (Moscow Poly, Russia, 107023, Moscow, B. Semenovskaya St. 38), e-mail: elenakhaa@yandex.ru, ORCID ID: 0000-0002-5082-4564.

Моделирование гетероэпитаксиального роста Ge на структурированной подложке Si(100) методом Монте-Карло и распараллеливание вычислений *

С. А. Рудин¹, К. В. Павский^{1,2}, А. Л. Ревун^{1,2}, А. В. Двуреченский^{1,3}

¹ Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН)

² Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

³ Новосибирский государственный университет (НГУ)

Аннотация: Рассмотрены методы и алгоритмы моделирования гетероэпитаксиального роста Ge на Si(100) методом Монте-Карло. Проведен анализ вычислительной сложности составляющих алгоритмы блоков моделирования и предложено распараллеливание трудоемких циклов с использованием директив открытого стандарта OpenMP. Проведено моделирование роста Ge на структурированной подложке Si(100), содержащей упорядоченный массив ямок. Анализ результатов моделирования показал, что расположение nanoостровков Ge определяется конкуренцией между процессом зарождения трехмерных островков вследствие накопления упругой деформации в растущем слое и процессом диффузии Ge в ямки, которые служат для них стоком.

Ключевые слова: моделирование методом Монте-Карло, гетероэпитаксия, германий, кремний, структурированные подложки, пространственное упорядочение, OpenMP.

Для цитирования: Рудин С. А., Павский К. В., Ревун А. Л., Двуреченский А. В. Моделирование гетероэпитаксиального роста Ge на структурированной подложке Si(100) методом Монте-Карло и распараллеливание вычислений // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 16–28. <https://doi.org/10.55648/1998-6920-2024-18-1-16-28>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Рудин С. А., Павский К. В.,
Ревун А. Л., Двуреченский А. В., 2024

Статья поступила в редакцию 07.07.2023;
принята к публикации 18.08.2023.

1. Введение

Квантовые точки (КТ) GeSi привлекают значительное внимание из-за возможности их практического применения в электронных и электронно-оптических приборах нового поколения, создаваемых на базе существующей кремниевой технологии [1–3]. Наиболее распространенный способ синтеза подобных структур – эпитаксия, представляющая собой рост одного кристаллического материала на другом с сохранением кристаллической ориентации. Несоответствие кристаллических решеток осаждаемого материала и материала подложки приводит к росту напряженных слоев и накоплению деформации в них. При достижении критической толщины пленки формируются трехмерные nanoостровки. Такой рост называется ростом по механизму Странского–Крастанова [4–5]. Однако при осаждении Ge на гладкую подложку Si места зарождения nanoостровков Ge располагаются хаотично и не подда-

* Работа выполнена в рамках государственного задания ИФП СО РАН (ГЗ FWGW-2022-0011).

ются точному контролю, что сильно ограничивает область возможных применений. Упорядоченные ансамбли КТ обладают более однородным химическим составом и энергетическим спектром, а точный контроль расположения КТ в пространстве открывает возможность адресного обращения к каждой КТ. В последнее десятилетие было предложено использовать мелкие ямки, периодически расположенные на поверхности подложки [6–9], для создания мест преимущественного зарождения островков. Эти ямки служат ловушками для осажденных атомов Ge, что приводит к образованию трехмерных островков в определенных местах относительно ямок. В зависимости от условий роста и формы ямок можно получать различные конфигурации КТ: одна или несколько КТ внутри ямок [10–12], КТ на границе или вне ямок [10, 13–14].

Рост SiGe-пленок включает в себя достаточно сложные механизмы, не принимаемые во внимание в упрощенной модели роста по механизму Странского–Крастанова. Эти механизмы включают в себя комбинацию процессов осаждения, диффузии, перераспределения упругой энергии, морфологической нестабильности, поверхностной анизотропии, различных реконструкций на стенках ямок и траншей. Для изучения роста таких пленок была разработана атомистическая трехмерная модель Монте-Карло (МК) гетероэпитаксиального роста [15, 16]. С помощью данной модели были выявлены механизмы роста Ge на подложках Si(100) с созданными в условиях ионного облучения траншеями [17], массивами ямок, образующих фотонный кристалл [18], и зависимость положения КТ от формы дна ямок [19]. Одной из проблем использованной программной реализации модели является трудоемкость вычислений. Например, моделирование осаждения Ge на гладкую подложку Si(100) размерами 27х27х5.4 нм (200 тыс. атомов) при температуре 720 К со скоростью осаждения 0.0136 нм/с может занимать более двух недель. Поэтому актуальной задачей является поиск путей оптимизации вычислений, а также разработка параллельных алгоритмов для многоядерных вычислительных систем. В настоящей работе представлены первые результаты оптимизации алгоритмов расчета этой модели для систем с общей памятью.

2. Физическая модель МК-моделирования гетероэпитаксиального роста

В основе модели лежит трехмерная кристаллическая решетка типа алмаза с периодическими граничными условиями вдоль поверхности подложки. Каждый узел решетки может быть занят атомом или быть пустым. Каждый атом может смещаться относительно положения узла идеальной решетки. Эти смещения позволяют учитывать упругую деформацию в кристалле. Моделирование роста производится методом Монте-Карло как последовательность элементарных событий, выбираемых случайным образом в соответствии с их вероятностями. В модели возможны следующие элементарные события: осаждение (добавление нового атома) и диффузионный прыжок атома по поверхности. Вероятность осаждения атома вычисляется исходя из требуемой скорости роста. Вероятность диффузионного прыжка i -го атома зависит от энергии активации ε_i :

$$\varepsilon_i = n_{i1}E_1 + n_{i2}E_2,$$

где n_{i1}, n_{i2} – число первых и вторых соседей i -го атома; E_1, E_2 – численные константы. В данной модели к энергии активации добавляется энергия упругой деформации E_{el} , выражаемая в виде потенциала Китинга [20]:

$$E_{el} = A \sum_{i,j} \left(\left(\vec{r}_i - \vec{r}_j \right)^2 - d_{ij}^2 \right)^2 + B \sum_{i,j,k>j} \left(\left(\vec{r}_i - \vec{r}_j \right) \left(\vec{r}_i - \vec{r}_k \right) + \frac{1}{3} d_{ij} d_{ik} \right)^2,$$

где $r_i = r_i^{(0)} + u_i$ – радиус-вектор i -го атома (u_i – смещение атома относительно положения узла идеальной решетки); d_{ij}, d_{ik} – равновесные длины атомных связей $i-j$, $i-k$ (зависит от типа атомов); A и B – численные параметры.

Тогда вероятность прыжка атома из узла i в узел j :

$$p_{i \rightarrow j} = v_0 \exp \left(\frac{-\varepsilon_i + \min [\varepsilon_{el, i}(r_i)]}{k_B T} \right),$$

где v_0 – частота попыток (параметр); $\varepsilon_{el, i}(r_i)$ – часть упругой энергии кристалла, относящаяся к i -му атому; k_B – постоянная Больцмана; T – температура. Минимизация производится по положению i -го атома r_i , в то время как положения остальных атомов фиксированы.

Разрешены любые прыжки из заполненного узла в пустой в пределах второй координационной сферы, не приводящие к «отрыву» атома от кристалла. Критерием отсутствия отрыва считается наличие хотя бы двух вторых соседей у каждого атома.

После прыжка $i \rightarrow j$ (или осаждения нового атома в узел j) положение атома r_j выбирается случайно в соответствии с распределением Больцмана: плотность распределения имеет вид:

$$f(r_j) = F \exp \left(-\frac{\varepsilon_{el, j}(r_j)}{k_B T} \right),$$

где F – нормировочный множитель.

Периодический (например, после каждого элементарного события, параметр модели) повтор случайного выбора положения атома в соответствии с распределением Больцмана для N (параметр модели) атомов кристалла приводит к перераспределению упругой деформации в кристалле, что является аналогом термического отжига и служит для минимизации полной упругой энергии.

Таким образом, моделирование состоит из следующих шагов:

1. Создание начальной конфигурации атомов и вычисление вероятностей всех возможных элементарных событий (осаждения или прыжка атома).
 2. Случайный выбор одного из событий (исходя из найденных вероятностей) и его выполнение.
 3. Смещение N атомов кристалла из равновесных положений в соответствии с распределением Больцмана.
 4. Обновление вероятностей элементарных событий и возврат к п. 2.
- Более детальную информацию о физической модели можно найти в [18].

3. Программная реализация модели

Предложенные алгоритмы моделирования гетероэпитаксиального роста были реализованы в программе, написанной на языке C++. На рис. 1 изображена блок-схема моделирования гетероэпитаксиального роста методом Монте-Карло.

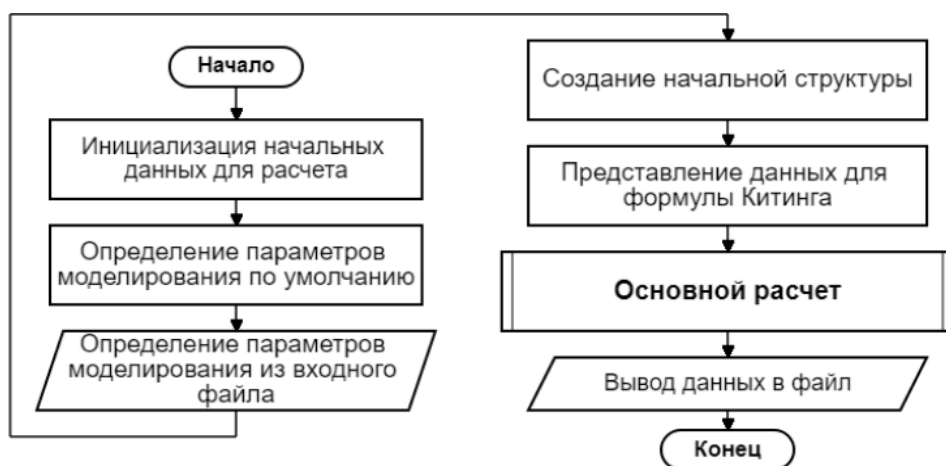


Рис. 1. Блок-схема моделирования гетероэпитаксии методом МК

На рис. 2 изображена блок-схема основного цикла и соотношение времени выполнения в процентах от общего времени счета двух основных блоков – «Шаг Монте-Карло» и «Термический отжиг» на структуре $5.4 \times 5.4 \times 5.4$ нм. Для других размеров структур соотношение между этими блоками практически не меняется – на каждый приходится около половины времени выполнения программы.

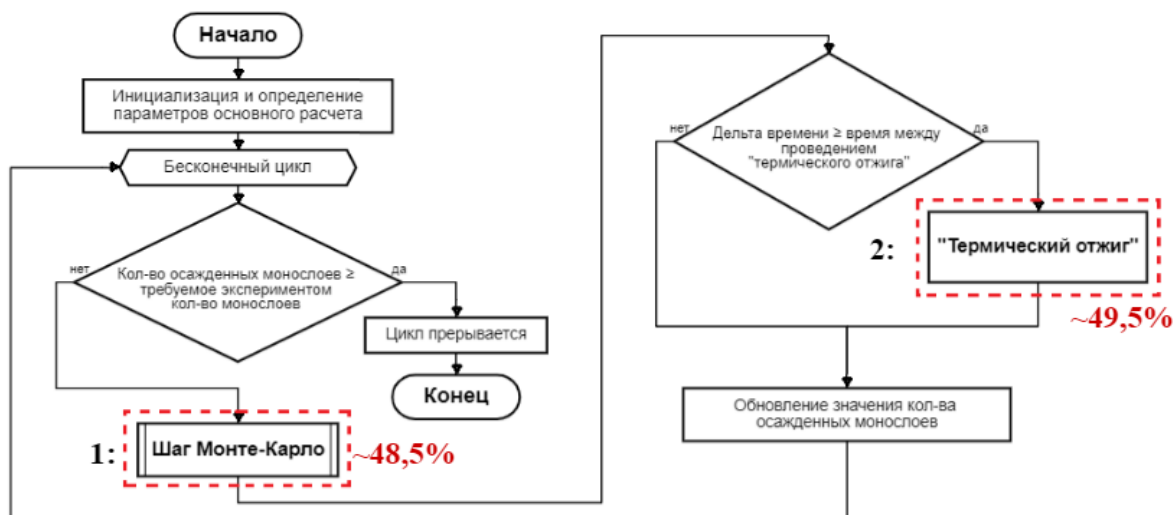


Рис. 2. Блок-схема «Основной расчет»

В блоке «Шаг Монте-Карло» происходит расчет вероятностей элементарных событий, случайный выбор и выполнение одного из них. В блоке «Термический отжиг» выполняется смещение N атомов кристалла из равновесных положений в соответствии с распределением Больцмана. На рис. 3 представлена блок-схема блока «Шаг Монте-Карло» с соответствующим соотношением в процентах от общего времени счета выполнения входящих в него блоков.

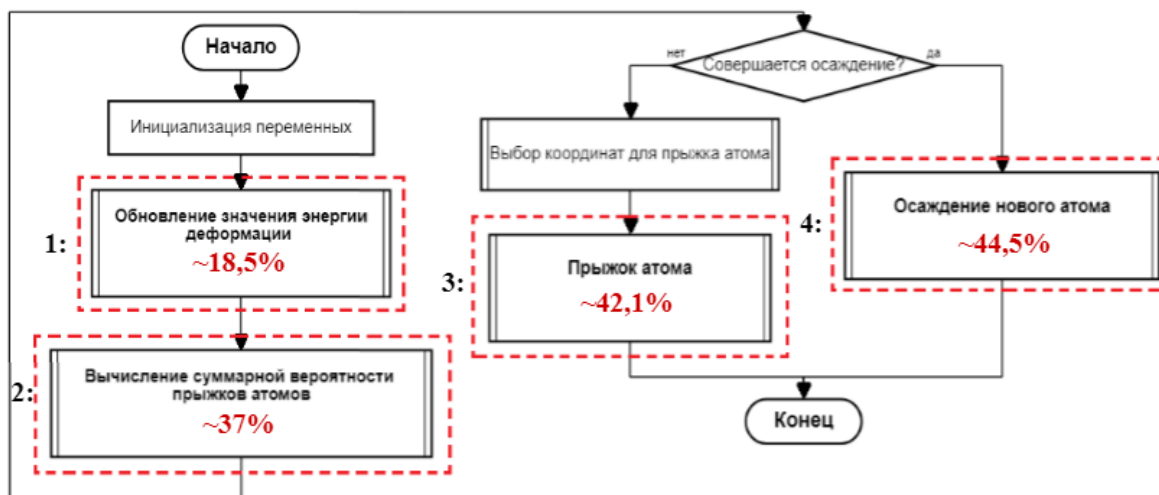


Рис. 3. Блок-схема «Шаг Монте-Карло»

Поэтому в данной работе основной акцент по распараллеливанию был сделан на блок «Термический отжиг» и блоки 1–4 «Шага Монте-Карло».

В каждом из выбранных блоков предложено распараллеливание трудоемких циклов с использованием директив открытого стандарта OpenMP [21] на системах с общей памятью.

3.1. Блок обновления значения энергии деформации

```

void calc_Edef(int x, int y, int z) {
/*инициализация переменных*/
/*проверка корректности входных параметров*/
/*определение соседей атома*/
/* Цикл (1) */
#pragma omp parallel for private(dir) reduction(+: Bx, By, Bz)
for (dir=0; dir<dir_number; dir++) {
    /*определение дополнительных параметров*/
    Bx+=Bxx*ax2+Bxy*ay2+Bxz*az2;
    By+=Byx*ax2+Byy*ay2+Byz*az2;
    Bz+=Bzx*ax2+Bzy*ay2+Bzz*az2;
}
/* Цикл (2) */
#pragma omp parallel for private(dir)
for (dir=0; dir<dir_number; dir++) {
/*инициализация и определение значения элементов массива*/
}
/*вычисление значения энергии деформации*/
/*обновление значения энергии деформации*/
}
  
```

В данном блоке были выделены два цикла, подходящих для распараллеливания.

В первом цикле (1) по соседям атома происходит редуцирование параметров B_x , B_y и B_z . Для его параллельного вычисления в цикле была применена директива «`pragma omp parallel for`». Условие «`private(dir)`» используется для ограничения целостности переменной итерирования. «`reduction(+: Bx, By, Bz)`» добавлено для корректного подсчета редуцированных параметров в теле параллельного блока.

В цикле (2) производится инициализация и определение значения элементов массива. Его распараллеливание аналогично циклу (1), кроме добавления параметра редукции переменной.

3.2. Блок вычисления суммарной вероятности прыжков атомов

```
double calc_P_jump_sum(void) {
/*инициализация переменных*/
#pragma omp parallel for private(x, y, z)reduction(+: real_sum)
for (z=2; z<Lz-2; z++) {
    for (y=((int)z%2; y<Ly; y+=2) {
for (x=((int)z%2)+2*(((int)z/2+(int)y/2)%2); x<Lx; x+=4) {
        real_sum += jump_probability(x, y, z);
    }
}
}
return real_sum;
}
```

Блок вычисления суммарной вероятности содержит один векторизуемый вложенный цикл. В нем проводится суммирование параметра `real_sum` по всем атомам решетки. Для параллельного исполнения цикла применяется прагма «`pragma omp parallel for`», а для ограничения одновременного доступа к переменным итерирования был добавлен параметр «`private(x, y, z)`». Корректный подсчет переменной `real_sum` обеспечивается с помощью условия «`reduction(+: real_sum)`».

3.3. Блок прыжка атома

```
int jump(int x,int y, int z, int dir, int* x2_, int* y2_, int*
z2_) {
/*инициализация переменных*/
/*определение координаты для совершения прыжка*/
/*поиск соседей атома*/
#pragma omp parallel for private(n)shared(bad_jump)
for (n=0; n<n_spisok; n++) {
    if (bad_jump == 1) continue;
    x2=spisok[n][0]; y2=spisok[n][1]; z2=spisok[n][2];
    set_config(x2, y2, z2);
    if (atoms(x2, y2, z2).type > 0 && !good_config(atoms(x2, y2,
z2).config)) {
        bad_jump = 1;
    }
}
return real_sum;
}
```

Данный блок содержит единственный цикл, подходящий для параллельного исполнения. В нем происходит проверка координат, найденных для прыжка. Для распараллеливания цикла была применена директива «`pragma omp parallel for`», «`private(n)`» – для ограничения целостности переменной итерирования. Для корректного одновременного выхода из цикла всех потоков (при обнаружении «плохого» прыжка) использовался параметр «`shared(bad_jump)`».

4. Результаты распараллеливания

Для оценки эффективности распараллеливания по блокам было выполнено моделирование структуры размером $13 \times 13 \times 4$ нм с осаждением 0.136 нм. На рис. 4–7 представлен график зависимости коэффициента ускорения K_n от количества рабочих потоков, исполняющихся на соответствующем числе вычислительных ядер системы:

$$K_n = \frac{T_1}{T_n},$$

где T_n – время исполнения блока на n ядрах.

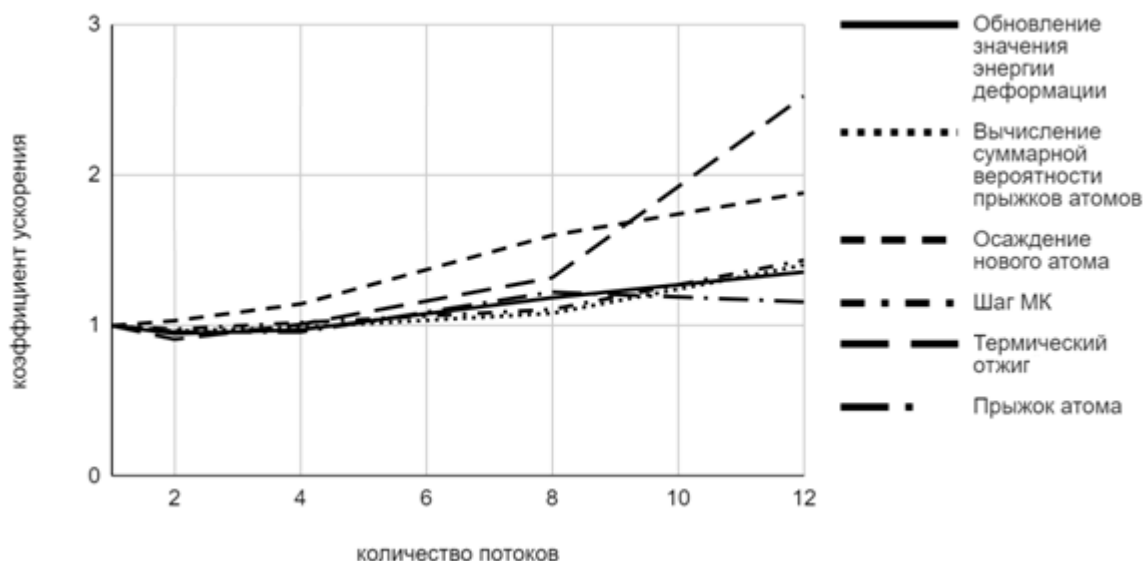


Рис. 4. График коэффициентов ускорения блоков

По графику видно, что наиболее эффективным оказалось распараллеливание блоков «Термический отжиг» и «Осаждение нового атома». Также можно заметить, что с ростом количества используемых потоков в подавляющем большинстве случаев коэффициент ускорения увеличивается.

Эффективность распараллеливания всей программы оценивалась для трех стадий осаждения Ge: 0.136 нм, 0.407 нм и 0.815 нм соответственно. Размеры подложки: $13 \times 13 \times 4$ нм. Время моделирования вычислялось по среднему времени за 10 запусков программы.

В первом случае было выполнено осаждение 0.136 нм. На рис. 5 изображен график зависимости коэффициента ускорения от количества используемых потоков. При использовании 12 потоков ускорение вычислений составило около 3 раз.

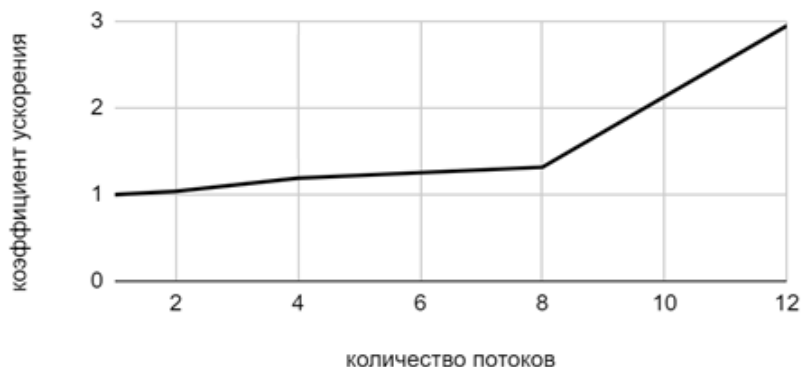


Рис. 5. График зависимости коэффициента ускорения от количества используемых потоков при осаждении 0.136 нм

Во втором случае было выполнено осаждение 0.407 нм (рис. 6). Максимальное полученное ускорение – 1.28 раза – было достигнуто при использовании 12 потоков.

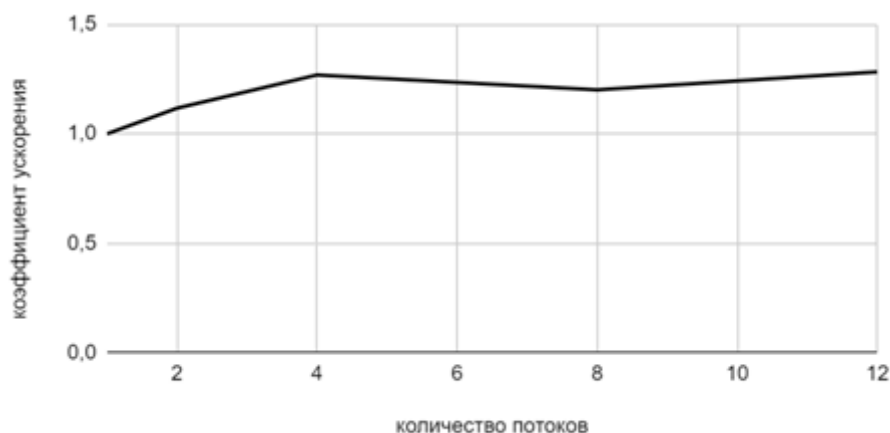


Рис. 6. График зависимости коэффициента ускорения от количества используемых потоков при осаждении 0.407 нм

В третьем случае было выполнено осаждение 0.815 нм (рис. 7). Максимальное полученное ускорение – 1.136 раза.

При увеличении количества осажденного материала увеличивается количество прыжков и, соответственно, количество событий «термического отжига». Это отражается на увеличении доли последовательных вычислений и, следовательно, приводит к падению коэффициента ускорения.

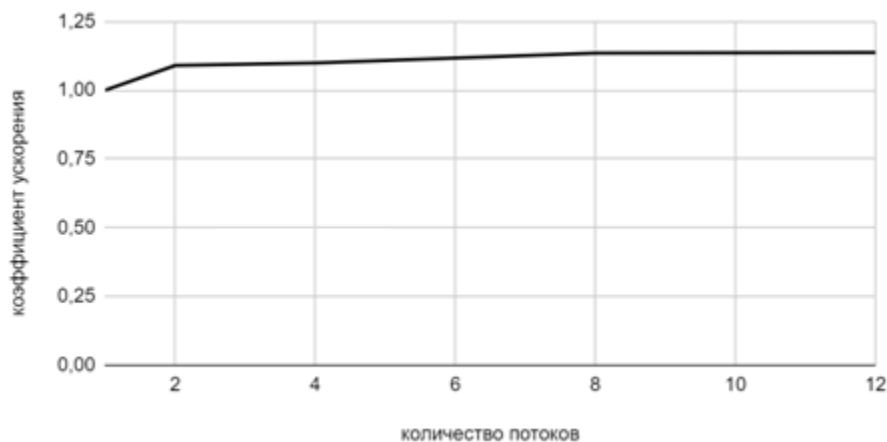


Рис. 7. График зависимости коэффициента ускорения от количества используемых потоков при осаждении 0.815 нм

5. Результаты моделирования

С целью изучения взаимосвязи эволюции Ge-слоя и поверхностной диффузии, а также выявления механизмов зарождения nanoостровков была проведена серия экспериментов по МК-моделированию осаждения Ge на подложку Si, содержащую упорядоченный массив ямок. В качестве модельной структуры использовались участки кристалла Si размерами от 10x10x13.5 нм до 27x27x13.5 нм. Поверхность подложки имела ориентацию (100), использовались периодические граничные условия. На поверхности подложки был создан массив ямок воронкообразной формы диаметром от 5.4 нм. Период между ямками менялся от 10 до 27 нм. Глубина ямок считалась бесконечной. И хотя линейные размеры модельных структур примерно на порядок меньше, чем в эксперименте, распределение механических

напряжений остается адекватным в силу самоподобного поведения упругих полей. Скорость осаждения Ge составляла 0.0136 нм/с, температура осаждения – 707 К, осаждено до 1.5 нм. Расчеты выполнялись на высокопроизводительном вычислительном комплексе Информационно-вычислительного центра Новосибирского государственного университета [22].

Моделирование методом Монте-Карло позволило установить, что места зарождения трехмерных наноструктур Ge определяются балансом между потоком адатомов Ge в ямки и их диффузией по поверхности. В случае близкого расположения ямок (10 – 13.5 нм) островки не зарождаются. При большем периоде (16.3 нм) потребовалось осадить ≈ 8 МС Ge, чтобы зародились трехмерные островки. С увеличением расстояния между ямками (20.6 и 23.9 нм) для зарождения наноструктур достаточно ≈ 5 МС Ge (рис. 8). В первом случае островки расположены на максимальном удалении от ямок, во втором – образуется больше островков на одну ямку и в третьем – образуется группа островков. Таким образом, ямки являются стоком для поступающего материала, и «сила» этого стока определяет места зарождения наноструктур.

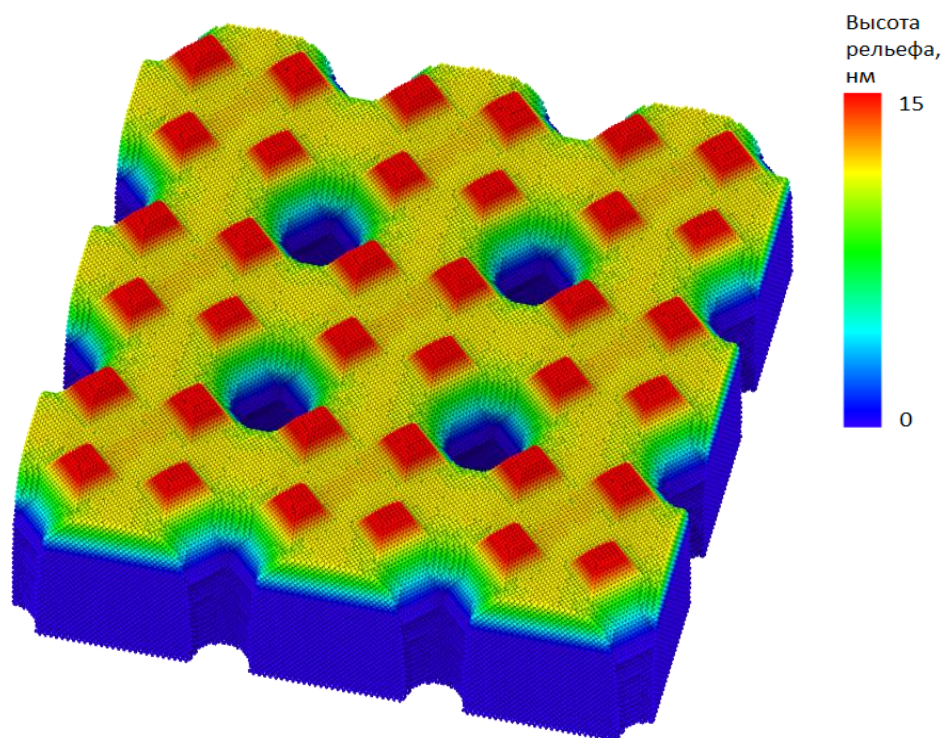


Рис. 8. Морфология поверхности после моделирования осаждения Ge на подложку Si(100), содержащую упорядоченный массив ямок. Размеры подложки 20.6x20.6x13.5 нм, диаметр ямок 5.4 нм, период между ямками 20.6 нм. Осажено 1.086 нм Ge со скоростью 0.0136 нм/с при температуре 707 К. Цветом показана высота рельефа.

6. Заключение

Были рассмотрены алгоритмы моделирования методом Монте-Карло гетероэпитаксиального роста. В схеме моделирования методом МК определены вычислительно сложные блоки «Шаг Монте-Карло» и «Термический отжиг». Результаты оптимизации и распараллеливания этих блоков позволили сократить время моделирования от 1.5 до 3 раз в зависимости от условий эксперимента. Проведенное моделирование роста Ge на подложке Si(100), содержащей упорядоченный массив ямок, показало, что полученные результаты согласуются с экспериментальными данными. Анализ результатов моделирования показал, что квантовые точки зарождаются в областях минимальной упругой энергии по внешнему периметру ямок. Эволюция поля упругой деформации обусловлена изменением огранки поверхности в

области ямок в процессе роста, что способствует накоплению Ge по периметру ямок. Этот процесс конкурирует с процессом миграции Ge вглубь ямки. Эффективность такого стока увеличивается с увеличением плотности ямок. Поэтому в случае близкорасположенных ямок зарождение островков не наблюдается.

Литература

1. *Shmidt O. G. and Eberl K.* Self-assembled Ge/Si dots for faster field-effect transistors // *IEEE Trans. El. Dev.* 2001. V. 48, № 6. P. 1175.
2. *Stangl J., Holy V., and Bauer G.* Structural properties of self-organized semiconductor nanostructures // *Rev. Mod. Phys.* 2004. V. 76. P. 725.
3. *Tsybeskov L. and Lockwood D. J.* Silicon-Germanium Nanostructures for Light Emitters and On-Chip Optical Interconnects // *Proc. IEEE.* 2009. V. 76, № 7. P. 1284.
4. *Eaglesham D. J. and Cerullo M.* Dislocation-free Stranski-Krastanow growth of Ge on Si(100) // *Phys. Rev. Lett.* 1990. V. 64. P. 1943.
5. *Mo Y.-W., Savage D. E., Swartzentruber B. S., and Lagally M. G.* Kinetic pathway in Stranski-Krastanow growth of Ge on Si(001) // *Phys. Rev. Lett.* 1990. V. 65. P. 1020.
6. *Zhong Z. and Bauer G.* Site-controlled and size-homogeneous Ge islands on prepatterned Si (001) substrates // *Appl. Phys. Lett.* 2004. V. 84. P. 1992.
7. *Zhong Z., Schwinger W., Schaffler F., Bauer G., Vastola G., Monalenti F., and Miglio L.* Delayed Plastic Relaxation on Patterned Si Substrates: Coherent SiGe Pyramids with Dominant {111} Facets // *Phys. Rev. Lett.* 2007. V. 98. P. 176102.
8. *Hackl F., Grydlik M., Brehm M., Groiss G., Schaffler F., Fromherz T., and Bauer G.* Microphotoluminescence and perfect ordering of SiGe islands on pit-patterned Si(001) substrates // *Nanotechnology.* 2011. V. 22. P. 165302.
9. *Pezzoli F., Stoffel M., Merdzhanova T., Rastelli A., and Schmidt O. G.* Alloying and Strain Relaxation in SiGe Islands Grown on Pit-Patterned Si(001) Substrates Probed by Nanotomography // *Nanoscale Res. Lett.* 2009. V. 4. P. 1073.
10. *Vastola G., Grydlik M., Brehm M., Fromherz T., Bauer G., Boioli F., Miglio L., and Montalenti F.* How pit facet inclination drives heteroepitaxial island positioning on patterned substrates // *Phys. Rev. B.* 2011. V. 84. P. 155415.
11. *Yang B., Liu F., and Lagally M.* Local Strain-Mediated Chemical Potential Control of Quantum Dot Self-Organization in Heteroepitaxy // *Phys. Rev. Lett.* 2004. V. 92. P. 025502.
12. *Ma Y., Huang S., Zeng C., Zhou T., Zhong Z., Zhou T., Fan Y., Yang X., Xia J. and Jiang Z.* Towards controllable growth of self-assembled SiGe single and double quantum dot nanostructures // *Nanoscale.* 2014. V. 6. P. 3941-3948.
13. *Zhong Z., Schmidt O., and Bauer G.* Increase of island density via formation of secondary ordered islands on pit-patterned Si (001) substrates // *Appl. Phys. Lett.* 2004. V. 87. P. 133111.
14. *Grydlik M., Brehm M., and Schaffler F.* Morphological evolution of Ge/Si(001) quantum dot rings formed at the rim of wet-etched pits // *Nanoscale Res. Lett.* 2012. V. 7. P. 601.
15. *Rudin S. A., Zinovyev V. A., Nenashev A. V., Polyakov A. Yu., Smagina Zh. V., and Dvurechenskii A. V.* Three-dimensional model of heteroepitaxial growth of germanium on silicon // *Optoelectronics Instr. and Data Proc.* 2013. V. 49. P. 461.
16. *Novikov P. L., Nenashev A. V., Rudin S. A., Polyakov A. S., and Dvurechenskii A. V.* Simulating the nucleation and growth of Ge quantum dots on Si using high-efficiency algorithms // *Nanotechnologies in Russia.* 2015. V. 10. P. 192.
17. *Smagina Zh. V., Zinovyev V. A., Rudin S. A., Novikov P. L., Rodyakina E. E., and Dvurechenskii A. V.* Nucleation sites of Ge nanoislands grown on pit-patterned Si substrate prepared by electron-beam lithography // *J. Appl. Phys.* 2018. V. 123. P. 165302.

18. Rudin S. A., Zinovyev V. A., Smagina Zh. V., Novikov P. L., Nenashev A. V., Pavsky K. V. Groups of Ge nanoislands grown outside pits on pit-patterned Si substrates // J. of Crystal Growth. 2022. V. 593. P. 126763.
19. Smagina Zh. V., Zinovyev V. A., Rudin S. A., Novikov P. L., Rodyakina E. E., and Dvurechenskii A. V. Nucleation sites of Ge nanoislands grown on pit-patterned Si substrate prepared by electron-beam lithography // J. Appl. Phys. 2018. V. 123. P. 165302.
20. Keating P. N. Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure // Phys. Rev. 1966. V. 145. P. 637.
21. Открытый стандарт для распараллеливания программ на языках Си, Си++ и Фортран [Электронный ресурс]. URL: <https://www.openmp.org/> (дата обращения: 22.06.2023).
22. Информационно-вычислительный центр Новосибирского государственного университета [Электронный ресурс]. URL: <http://nusc.ru> (дата обращения: 21.06.2023).

Рудин Сергей Алексеевич

младший научный сотрудник лаборатории неравновесных полупроводниковых систем, Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН), e-mail: rudin@isp.nsc.ru, ORCID ID: 0000-0002-9509-4008.

Павский Кирилл Валерьевич

д.т.н., доцент, заведующий лабораторией вычислительных систем, Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН); профессор кафедры вычислительных систем, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ), e-mail: pkv@isp.nsc.ru, ORCID ID: 0000-0002-2271-7358.

Ревун Артем Леонидович

инженер-программист лаборатории вычислительных систем, Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН); старший преподаватель кафедры вычислительных систем, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ), e-mail: pm@artemrevun.ru, ORCID ID: 0009-0006-5774-415X.

Двуреченский Анатолий Васильевич

чл.-корр РАН, профессор, д.ф.-м.н., заведующий лабораторией неравновесных полупроводниковых систем, Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН); профессор кафедры физики полупроводников, Новосибирский государственный университет (НГУ), e-mail: dvurech@isp.nsc.ru, ORCID ID: 0000-0001-5443-4769.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Monte Carlo Simulation of Heteroepitaxial Growth of Ge on Pit-patterned Si(100) Substrate and Parallelization of Calculations

Sergei A. Rudin¹, Kiril V. Pavsky^{1,2}, Artem L. Revun^{1,2}, Anatoly V. Dvurechenskii^{1,3}

¹ Rzhzanov Institute of Semiconductor Physics (ISP SB RAS)

² Siberian State University of Telecommunications and Information Science (SibSUTIS)

³ Novosibirsk State University (NSU)

Abstract: Methods and algorithms of Monte Carlo simulation of heteroepitaxial growth of Ge on Si are considered. The computational complexity analysis of simulation program blocks is carried out. Parallelization of time-consuming cycles using directives of the open standard OpenMP is proposed. The Monte Carlo simulation of heteroepitaxial growth of Ge on pit-patterned Si substrate was carried out. The position of Ge nanoislands is determined by the competition between the nucleation process of three-dimensional islands due to the accumulation of elastic strain in the growing layer and the process of diffusion of Ge atoms into the pits serving as a drain.

Keywords: Monte Carlo simulation, heteroepitaxy, germanium, silicon, pre-patterned substrate, spatial ordering, OpenMP.

For citation: Rudin S. A., Pavsky K. V., Revun A. L., Dvurechenskii A. V. Monte Carlo simulation of heteroepitaxial growth of Ge on pit-patterned Si(100) substrate and parallelization of calculations (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 16-28. <https://doi.org/10.55648/1998-6920-2024-18-1-16-28>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Rudin S. A., Pavsky K. V.,
Revun A. L., Dvurechenskii A. V., 2024

The article was submitted: 07.07.2023;
accepted for publication 18.08.2023.

References

1. Shmidt O. G. and Eberl K. Self-assembled Ge/Si dots for faster field-effect transistors. *IEEE Trans. El. Dev*, 2001, vol. 48, iss. 6, p. 1175.
2. Stangl J., Holy V., and Bauer G. Structural properties of self-organized semiconductor nanostructures. *Rev. Mod. Phys.*, 2004, vol. 76, p. 725.
3. Tsybeskov L. and Lockwood D. J. Silicon-Germanium Nanostructures for Light Emitters and On-Chip Optical Interconnects. *Proc. IEEE*, 2009, vol. 76, iss. 7, p. 1284.
4. Eaglesham D. J. and Cerullo M. Dislocation-free Stranski-Krastanow growth of Ge on Si(100). *Phys. Rev. Lett.*, 1990. vol. 64, p. 1943.
5. Mo Y.-W., Savage D. E., Swartzentruber B. S., and Lagally M. G. Kinetic pathway in Stranski-Krastanow growth of Ge on Si(001). *Phys. Rev. Lett.*, 1990, vol. 65, p. 1020.
6. Zhong Z. and Bauer G. Site-controlled and size-homogeneous Ge islands on prepatterned Si (001) substrates. *Appl. Phys. Lett.*, 2004, vol. 84, p. 1992.
7. Zhong Z., Schwinger W., Schaffler F., Bauer G., Vastola G., Monalenti F., and Migilo L. Delayed Plastic Relaxation on Patterned Si Substrates: Coherent SiGe Pyramids with Dominant {111} Facets. *Phys. Rev. Lett.*, 2007, vol. 98, p. 176102.
8. Hackl F., Grydlik M., Brehm M., Groiss G., Schaffler F., Fromherz T., and Bauer G. Microphotoluminescence and perfect ordering of SiGe islands on pit-patterned Si(001) substrates. *Nanotechnology*, 2011, vol. 22, p. 165302.
9. Pezzoli F., Stoffel M., Merdzhanova T., Rastelli A., and Schmidt O. G. Alloying and Strain Relaxation in SiGe Islands Grown on Pit-Patterned Si(001) Substrates Probed by Nanotomography. *Nanoscale Res. Lett.*, 2009, vol. 4, p. 1073.

10. Vastola G., Grydlik M., Brehm M., Fromherz T., Bauer G., Boioli F., Miglio L., and Montalenti F. How pit facet inclination drives heteroepitaxial island positioning on patterned substrates. *Phys. Rev. B.*, 2011, vol. 84, p. 155415.
11. Yang B., Liu F., and Lagally M. Local Strain-Mediated Chemical Potential Control of Quantum Dot Self-Organization in Heteroepitaxy. *Phys. Rev. Lett.*, 2004, vol. 92, p. 025502.
12. Ma Y., Huang S., Zeng C., Zhou T., Zhong Z., Zhou T., Fan Y., Yang X., Xia J. and Jiang Z. Towards controllable growth of self-assembled SiGe single and double quantum dot nanostructures. *Nanoscale.*, vol. 6, pp. 3941-3948.
13. Zhong Z., Schmidt O., and Bauer G. Increase of island density via formation of secondary ordered islands on pit-patterned Si (001) substrates. *Appl. Phys. Lett.*, 2004, vol. 87, p. 133111.
14. Grydlik M., Brehm M. and Schaffler F. Morphological evolution of Ge/Si(001) quantum dot rings formed at the rim of wet-etched pits. *Nanoscale Res. Lett.*, 2012, vol. 7, p. 601.
15. Rudin S. A., Zinovyev V. A. Nenashev A. V., Polyakov A. Yu., Smagina Zh. V., and Dvurechenskii A. V. Three-dimensional model of heteroepitaxial growth of germanium on silicon. *Optoelectronics Instr. and Data Proc.*, 2013, vol. 49, p. 461.
16. Novikov P. L., Nenashev A. V., Rudin S. A., Polyakov A. S., and Dvurechenskii A. V. Simulating the nucleation and growth of Ge quantum dots on Si using high-efficiency algorithms. *Nanotechnologies in Russia*, 2015, vol. 10, p. 192.
17. Smagina Zh. V., Zinovyev V. A., Rudin S. A., Novikov P. L., Rodyakina E. E., and Dvurechenskii A. V. Nucleation sites of Ge nanoislands grown on pit-patterned Si substrate prepared by electron-beam lithography. *J. Appl. Phys.*, 2018, vol. 123, p. 165302.
18. Rudin S. A., Zinovyev V. A., Smagina Zh. V., Novikov P. L., Nenashev A. V., Pavsky K. V. Groups of Ge nanoislands grown outside pits on pit-patterned Si substrates. *J. of Crystal Growth*, 2022, vol. 593, p. 126763.
19. Smagina Zh. V., Zinovyev V. A., Rudin S. A., Novikov P. L., Rodyakina E. E., and Dvurechenskii A. V. Nucleation sites of Ge nanoislands grown on pit-patterned Si substrate prepared by electron-beam lithography. *J. Appl. Phys.*, 2018, vol. 123, p. 165302.
20. Keating P. N. Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure. *Phys. Rev.*, 1966, vol. 145, p. 637.
21. *Otkrytiya standart dlya rasparallelivaniya programm na yazykakh Si, Si++ i Fortran* [Application programming interface (API) that supports multi-platform shared-memory multiprocessing programming in C, C++, and Fortran], available at: <https://www.openmp.org/> (accessed 22.06.2023).
22. *Informatsionno-vychislitel'nyi tsentr Novosibirskogo gosudarstvennogo universiteta* [Information and Computing Center of NSU], available at: <http://nusc.ru> (accessed 21.06.2023).

Sergei A. Rudin

Junior scientist of Laboratory of Nonequilibrium Semiconductor Systems, Rzhanov Institute of Semiconductor Physics Siberian Branch of Russian Academy of Sciences (ISP SB RAS), e-mail: rudin@isp.nsc.ru, ORCID ID: 0000-0002-9509-4008.

Kiril V. Pavsky

Dr. of Sci. (Engineering), Head of Computer Systems Laboratory, Rzhanov Institute of Semiconductor Physics Siberian Branch of Russian Academy of Sciences (ISP SB RAS); Professor of the Department of computer science, Siberian State University of Telecommunications and Information Science (SibSUTIS), e-mail: pkv@isp.nsc.ru, ORCID ID: 0000-0002-2271-7358.

Artem L. Revun

Software engineer of Computer Systems Laboratory, Rzhanov Institute of Semiconductor Physics Siberian Branch of Russian Academy of Sciences (ISP SB RAS); Senior teacher of the Department of computer science, Siberian State University of Telecommunications and Information Science (SibSUTIS), e-mail: pm@artemrevun.ru, ORCID ID: 0009-0006-5774-415X.

Anatoly V. Dvurechenskii

Professor, Doc. phys.-math. sci., Corresponding member of RAS, Head of Laboratory of Nonequilibrium Semiconductor Systems, Rzhanov Institute of Semiconductor Physics Siberian Branch of Russian Academy of Sciences (ISP SB RAS); Professor of the Department of Semiconductor Physics, Novosibirsk State (NSU), e-mail: dvurech@isp.nsc.ru, ORCID ID: 0000-0001-5443-4769.

Система анализа оптимизации доступа к общей памяти для реализации стандарта RMix*

К. Е. Крамаренко^{1,2}, А. Ю. Поляков³, А. В. Ефимов^{1,2}

¹ Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН)

² Сибирский гос. ун-в. телекоммуникаций и информатики (СибГУТИ)

³ NVIDIA

Аннотация: Рассмотрены алгоритмы синхронизации доступа к общей памяти с преобладанием операций чтения, которые характерны для реализаций RMix. Проведен анализ схемы синхронизации $N(\text{mutex} + \text{signal})$, результаты показали увеличение пропускной способности примерно на 20 % в режиме «только чтение» и дальнейшее направление проведения оптимизации.

Ключевые слова: распределенные вычислительные системы, интерфейс управления процессами, RMi, RMix, общая память, синхронизация доступа.

Для цитирования: Крамаренко К. Е., Поляков А. Ю., Ефимов А. В. Система анализа оптимизации доступа к общей памяти для реализации стандарта RMix // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 29–39. <https://doi.org/10.55648/1998-6920-2024-18-1-29-39>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Крамаренко К. Е., Поляков А. Ю.,
Ефимов А. В., 2024

Статья поступила в редакцию 23.06.2023;
принята к публикации 03.12.2023.

1. Введение

В архитектуре распределенной вычислительной системы (РВС) используется несколько элементарных машин (ЭМ), которые соединены сетью [1]. Элементарная машина (ЭМ) является основным компонентом РВС и может быть процессорным ядром или конфигурацией вычислительных узлов с различными типами процессоров. В передовых РВС количество ЭМ может достигать десятков миллионов.

Основная цель РВС – обработка вычислительных задач, которые могут состоять из одной или нескольких параллельных программ (ПП). Для решения этих задач требуются ресурсы РВС, которые описываются метаданными. Менеджер ресурсов РВС назначает ЭМ каждой ветви ПП. Запуск задачи содержит несколько этапов, включая развертывание среды исполнения (СИ), запуск прикладных процессов и определение компонентов среды исполнения [2–5]. Для управления выполнением используется интерфейс управления процессами (РМ), а самая актуальная и стандартизированная версия этого интерфейса называется RMix [6]. Информация о среде исполнения хранится в базе данных «ключ – значение» (KVDb), и для

* Работа выполнена в рамках государственного задания ИФП СО РАН (ГЗ FWGW-2022-0011).

доступа к ней используются примитивы чтения (Get), записи (Put) и синхронизации (Fence, Commit). Размещение KVDb в общей памяти вычислительного узла позволяет эффективно обращаться к ней без системных вызовов. В данной работе проведен анализ схемы синхронизации $N(\text{mutex}+\text{signal})$ и представлены описания существующих схем синхронизации доступа к общей памяти в реализациях стандарта PMIx. Работа является продолжением [21].

2. Синхронизация доступа к базе данных в стандарте PMIx

Компоненты стека системного программного обеспечения (ПО) в распределенной операционной системе, основанной на стандарте PMIx, включают следующие элементы: а) эталонная реализация PMIx (Open PMIx [7]); б) реализация среды исполнения (RTE); в) библиотека обмена сообщениями; и г) среда параллельного программирования (MPI [8], OpenSHMEM). На рис. 1 показано расположение этих компонентов на процессах распределенной операционной системы.

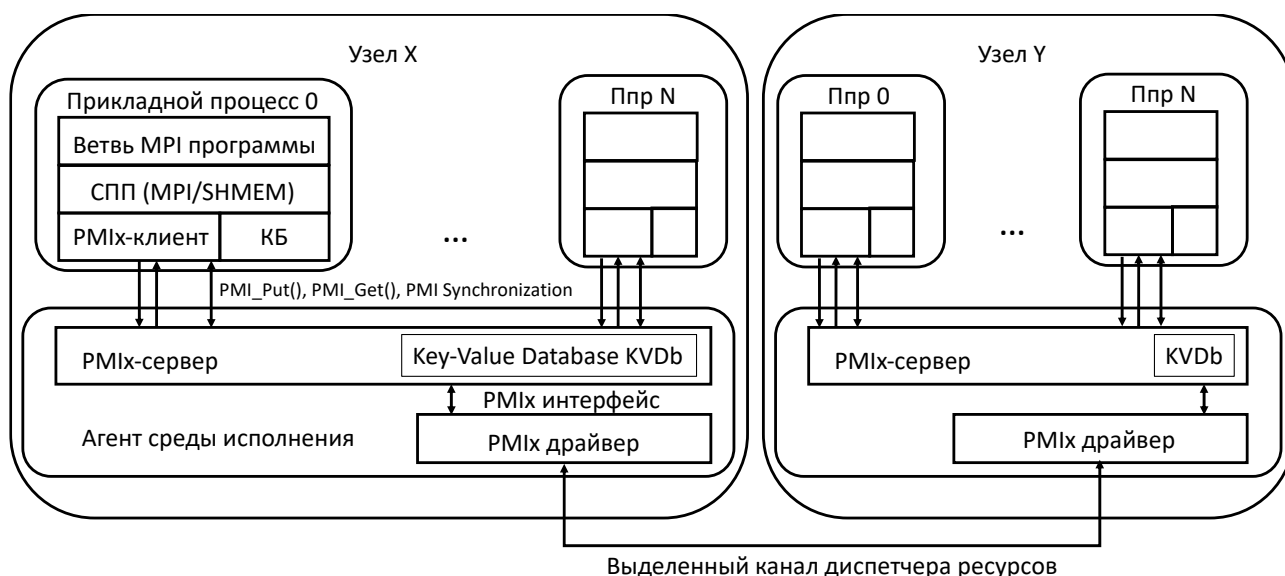


Рис. 1. Уровни системного программного обеспечения
(Ппр – прикладной процесс, СПП – среда параллельного программирования,
КБ – коммуникационная библиотека)

Прикладные процессы, составляющие задачу, взаимодействуют с системой СИ через интерфейс PMI [3, 4]. Это позволяет различным СИ совместимо работать с разными реализациями моделей и парадигм параллельного программирования. СИ может быть представлена как диспетчер ресурсов вычислительной системы [11], а также как специализированная библиотека [9, 10].

Агент СИ создается на каждом вычислительном узле, объединяющем несколько ЭМ общей оперативной памятью. Такой агент (рис. 1) включает в себя поток, выполняющий стандартные функции PMIx-сервера (реализован в библиотеке Open PMIx), и драйвер для взаимодействия с другими агентами, предоставляемый СИ. Прикладные процессы задачи включают реализацию клиентской части PMIx, предоставляемую Open PMIx.

Агент СИ обновляет базу данных PMI (KVDb), чтобы она всегда была актуальной. KVDb содержит как статическую, так и динамическую информацию о СИ. Статическая информация включает уникальный целочисленный идентификатор (ранг) процесса ПП, общее количество приложений и другие данные. Данная информация известна агенту СИ априори и не требует взаимодействия с агентами, расположенными на других узлах ВС.

Динамические параметры среды исполнения – это данные, которые специфичны для прикладных процессов. Эти данные могут включать информацию об адресах, получаемую от

коммуникационной подсистемы. Для работы динамического компонента необходима дополнительная синхронизация данных между агентами среды исполнения, которые могут располагаться на различных узлах вычислительной системы. Это может привести к накладным расходам на этапе инициализации вычислительной задачи.

PMIx отличается от предшественников PMI-1 и PMI-2 следующими важными особенностями [4]: а) значительное сокращение динамической информации о среде исполнения, за счет переноса большей ее часть в статический компонент, известный диспетчеру ресурсов; б) режим обмена данными по требованию.

Внутренний доступ к KVDb в рамках существующих реализаций PMI [12] является узким местом. Это прежде всего отражается на производительности операций Get и Put. Оптимизация доступа к KVDb путем его размещения в общей памяти является наиболее актуальным подходом, который дает следующие преимущества:

- Память KVDb масштабируется, что означает, что ее содержимое не дублируется на каждом процессе приложения.
- Общая память обеспечивает параллелизм, позволяя клиентам получать доступ к KVDb независимо от сервера.
- Доступ к операциям осуществляется с низкими накладными расходами, так как они выполняются без системных вызовов, которые требуют переключения контекста операционной системы.

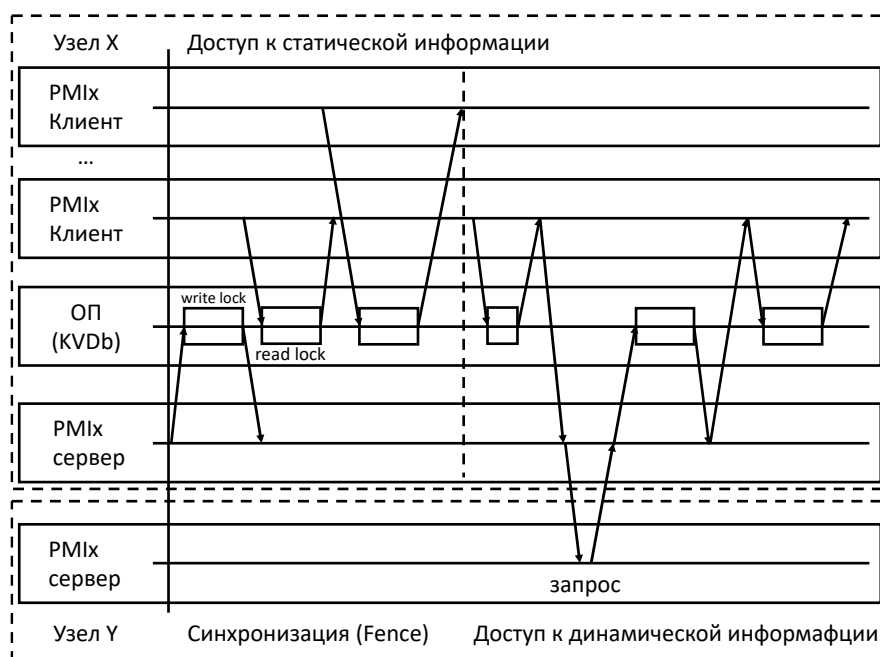


Рис. 2. Диаграмма доступа к KVDb

Для доступа к данным KVDb используются операции глобальной, локальной барьерной синхронизации процессов и операции чтения/записи. Во время выполнения синхронизации с помощью примитива "Fence" (см. рис. 2) поток "PMIx-сервер" блокирует общую память для записи (write lock) и помещает данные в KVDb. После завершения синхронизации приложения блокируют KVDb для чтения (read lock) и получают доступ к параметрам среды выполнения. Если запрашиваемый параметр отсутствует в локальной KVDb, блокировка на чтение снимается, и производится запрос к потоку СИ для получения данных, которые хранятся на других узлах ВС. Если агент СИ не выполняет операции записи в KVDb, приложения могут обращаться к ней одновременно. Для обеспечения такого доступа используются примитивы синхронизации [13–15].

3. Обзор существующих алгоритмов синхронизации доступа к оперативной памяти

Важной особенностью PMIx является параллельное и интенсивное чтение данных KVDb, выполняемое PMIx-клиентами, в то время как операция записи осуществляется относительно редко и только PMIx-сервером. Рассмотрим подробнее схемы синхронизации, обеспечивающие безопасный доступ к KVDb.

Первая схема (рис. 3) основана на примитиве «чтение – запись», предоставляемом стандартом POSIX [14]. Каждый клиент PMIx должен выполнить операцию захвата примитива синхронизации над одной и той же областью памяти для конкурентного доступа к KVDb при чтении. Это приводит к повышению нагрузки на подсистему обеспечения согласованности процессорных кэшей в узлах вычислительных систем, что приводит к увеличению накладных расходов [5].

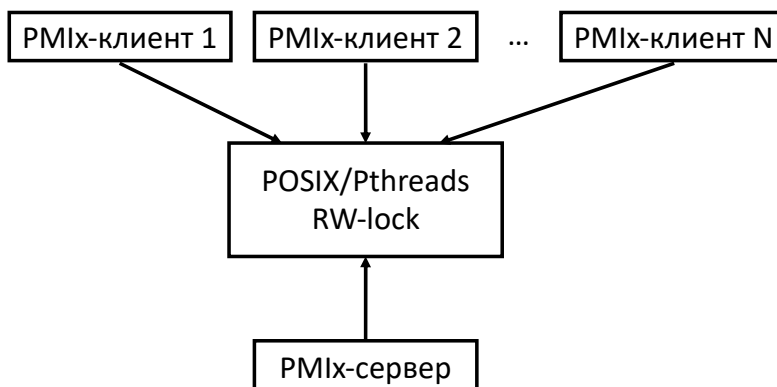


Рис. 3. Схема синхронизации на основе примитива «чтение – запись»

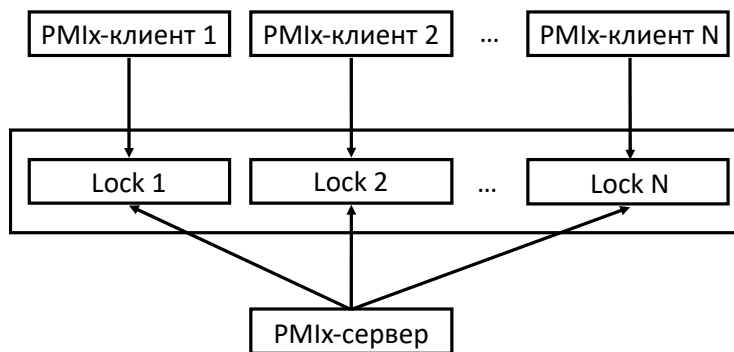


Рис. 4. Схема синхронизации 1N-mutex

В работе [15] предложена реализация схемы синхронизации 1N-mutex (рис. 4), которая не приводит к перегрузке подсистемы обеспечения когерентности процессорных кэшей. В схеме 1N-mutex для каждого клиента PMIx создается отдельный примитив синхронизации (mutex), при захвате которого гарантируется атомарный доступ к KVDb для чтения. Однако для записи требуется захват всех клиентских примитивов, что ограничивает масштабируемость операции в зависимости от количества клиентов PMIx.

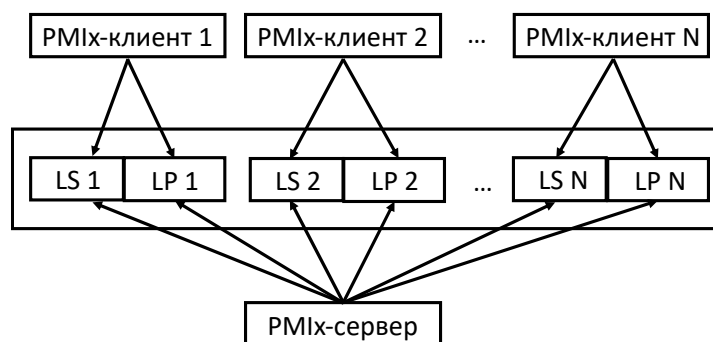


Рис. 5. Схема синхронизации 2N-mutex

В настоящее время в Open PMIx используется схема синхронизации 2N-mutex [5] (рис. 5). Каждому PMIx-клиенту соответствует пара синхронизационных примитивов (mutex). Первый примитив (LSx) служит сигнальным, а второй (LPx) – для доступа к оперативной памяти (ОП). При захвате примитива LPx PMIx-клиент освобождает примитив LSx. Для записи требуется захват всех сигнальных примитивов (LS1 – LSN), которые освобождаются PMIx-клиентами как можно быстрее. Затем сервер захватывает основные примитивы (LP1 – LPN), при этом благодаря блокировке сигнальных примитивов новые попытки доступа для чтения выполняться не будут.

4. Оптимизация алгоритмов синхронизации доступа к оперативной памяти

В данной работе предложена модификация схемы 2N-mutex, которая оптимизирует доступ к оперативной памяти в конкурентном режиме. Для снижения накладных расходов внесены изменения в способ сигнализации, который реализован в схеме 2N-mutex на основе примитива синхронизации LSx. В схеме синхронизации N(mutex+signal) (рис. 6) сигнальный примитив LSx (mutex) заменен переменной VSx, что позволяет сократить количество атомарных операций, необходимых для захвата примитива синхронизации при чтении, и повысить производительность.

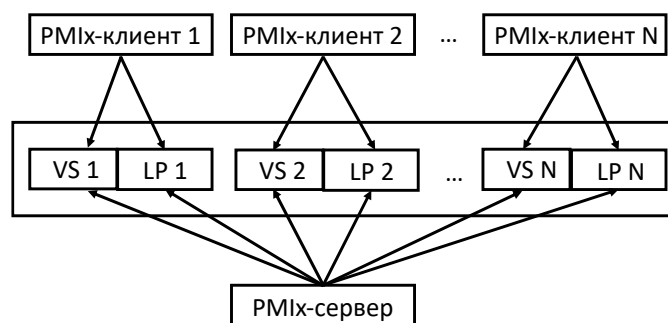


Рис. 6. Схема синхронизации N(mutex+signal)

На рис. 7 показан алгоритм блокировки примитивов на запись и чтение в схеме синхронизации N(mutex+signal). При доступе для записи (рис. 7a) все сигнальные переменные (VS1 – VSN) устанавливаются в значение «1». Затем PMIx-сервер последовательно ожидает освобождения и захватывает основные примитивы LPi. При захвате примитива синхронизации на чтение (рис. 7б) PMIx-клиент сначала проверяет состояние соответствующей сигнальной переменной (VSx). Если переменная установлена, то попытка захвата основного примитива синхронизации LPx не производится.

<pre> for i in 1 ... N: VS[i] = 1 for i in 1 ... N: lock(LP[i]) </pre> а	<pre> while (VS[x] = 1) : noop() ; lock(LP[x]) </pre> б
--	---

Рис. 7. Псевдокод захвата примитивов в схеме синхронизации $N(\text{mutex}+\text{signal})$ (PMIx-клиентом с номером x): (а) на запись; (б) на чтение

Стандарт POSIX Threads не предписывает соблюдать порядок доступа [16], поэтому использование сигнальных примитивов LSx в схеме $2N\text{-mutex}$ и сигнальных переменных VSx в схеме $N(\text{mutex}+\text{signal})$ и реализация примитива POSIX/mutex в библиотеке GNU C Library не обладает таким свойством.

5. Методика проведения испытаний

В данной работе разработан микробенчмарк [19], предназначенный для оценки эффективности схем синхронизации доступа к ОП. В рамках микробенчмарка реализованы функции проверки корректности схем синхронизации, а также нагрузочные тесты.

Для оценки эффективности схем синхронизации проводятся тесты в трех режимах: «только чтение», «только запись» и при конкурентном использовании. В последнем случае сервер выполняет заданное количество операций записи (в данной работе – 100 итераций), которые чередуются с ожиданием в 10 микросекунд. Во время работы сервера клиенты осуществляют непрерывный доступ для чтения, создавая помехи.

Эффективность схем синхронизации оценивается по следующим параметрам: пропускная способность в режиме «только чтение» (количество захваченных примитивов в секунду – кзп/с), пропускная способность в режиме «только запись» (кзп/с), пропускная способность в конкурентном режиме (кзп/с) и накладные расходы на конкурентный захват примитивов для записи (с).

Для проведения экспериментов были использованы две схемы синхронизации: $2N\text{-mutex}$, которая используется в промышленной реализации стандарта PMIx, и $N(\text{mutex}+\text{signal})$, предложенная в данной работе. В качестве вычислительной базы использовались NUMA-узлы с процессорами Intel (2 x Xeon E5-2680 v4) с 28 ядрами. Тесты запускались с использованием команды *mpirun* и параметра *--bind-to-core*, который гарантировал фиксированное назначение каждому потоку уникального вычислительного ядра.

Измерение эффективности схем синхронизации было проведено 10 раз для каждого значения N , после чего было вычислено среднее значение.

Результаты оценки производительности представлены на рис. 8.

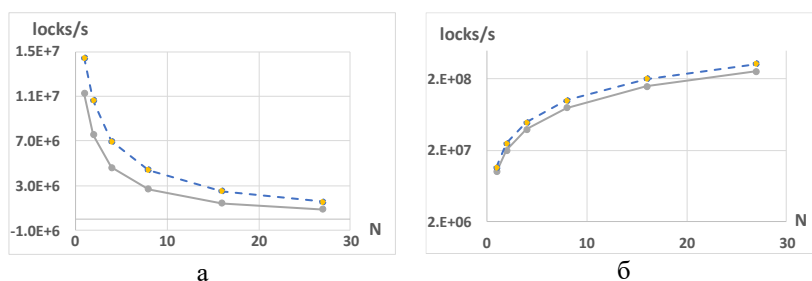


Рис. 8. Пропускная способность схем синхронизации:
а) в режиме «только чтение»; б) в режиме «только запись»;
—●— - $2N\text{-mutex}$; -♦- - $N(\text{mutex}+\text{signal})$

В предложенной схеме синхронизации удалось сократить количество атомарных операций, что привело к увеличению пропускной способности в режиме «только чтение» пример-

но на 20 %. В режиме «только запись» также наблюдается повышение пропускной способности на 3–5 %.

6. Реализация и анализ схемы синхронизации $N(\text{mutex}+\text{signal})$ в OpenPMIx

В работе была предложена схема синхронизации $N(\text{mutex}+\text{signal})$, которая была успешно реализована в Open-PMIx [20], соответствующей реализации стандарта PMI. Для анализа наиболее вычислительно сложных участков кода методов чтения и записи и последующей оптимизации был разработан программный инструмент. Оценка эффективности доступа к данным в KVDb производится следующим образом: функция чтения/записи данных в KVDb разбивается на несколько частей исходного кода в соответствии с задачами анализа. В начале и конце каждой части исходного кода вставляется функция, которая возвращает количество тактов, выполненных процессором с момента последней перезагрузки системы. Затем рассчитывается время выполнения каждой части и производится оценка вычислительной сложности.

Например, при анализе метода Get, функция Counter, возвращающая количество тактов, выполненных процессором, вставляется в соответствующие фрагменты исходного кода метода Get:

- При входе в метод Get (P1).
- Перед блокировкой примитива синхронизации (P2).
- После блокировки примитива синхронизации (P3).
- После нахождения ключа (P4).
- После распаковки ключа (P5).
- После разблокировки примитива синхронизации (P6).
- При выходе из метода Get (P7).

Чтобы получить результаты, нужно запустить тестовую программу, которая вызывает методы Get/Put из библиотеки OpenPMIx. После завершения программы значения количества тактов, затраченных на выполнение каждой части метода Get, записываются в лог-файл для последующего анализа.

7. Результаты

В данной работе был проведен анализ эффективности метода получения данных (Get) в KVDb для выявления наиболее вычислительно сложных участков кода.

В эксперименте использовались вычислительные узлы с процессорами Intel (2 x Xeon E5-2680 v4), имеющими 28 ядер. В качестве тестовой программы, вызывающей метод Get, была использована модифицированная утилита *pmix_intra_perf* (которая входит в состав исходного кода библиотеки OpenPMIx). Количество ключей, которые считывались из KVDb, составляло 100, при этом каждый ключ имел длину 50 символов. Количество клиентов N , считывающих ключи, варьировалось от 2 до 28.

Таблица 1. Результаты анализа метода Get

	P7-P1	P2-P1	P3-P2	P4-P3	P5-P4	P6-P5	P7-P6
$N = 2$							
Среднее	4182.085	931.845	284.815	1511.365	437.295	86.445	930.32
Медиана	3815.0	558.0	285.0	1569.0	330.0	75.0	961.0
$N = 8$							
Среднее	4231.64	759.485	175.11	1832.575	404.16	113.68	946.62
Медиана	3673.5	636.0	126.0	1605.0	345.0	78.0	810.0

$N = 16$							
Среднее	4493.59	803.79	271.90	1769.84	435.91	209.24	1002.89
Медиана	4143.0	612.0	123.0	1674.0	351.0	84.0	822.0
$N = 20$							
Среднее	4318.94	775.93	240.5	1735.19	441.96	143.48	981.85
Медиана	4068.0	600.0	123.0	1693.5	351.0	81.0	807.0

В табл. 1 представлены результаты анализа метода Get. В качестве значений указаны средние значения и медиана количества тактов процессора для 100 считанных ключей. Анализ полученных результатов показал, что около 50 % накладных расходов на выполнение метода Get приходится на инициализацию данных и их очистку. Для дальнейшей оптимизации необходимо сосредоточиться именно на этих участках кода.

8. Заключение

В работе проведен анализ схемы синхронизации $N(\text{mutex}+\text{signal})$. Полученные результаты показывают, что схема позволяет увеличить пропускную способность примерно на 20 % в режиме «только чтение» по сравнению со схемой $2N\text{-mutex}$. Анализ метода Get при реализации схемы синхронизации $N(\text{mutex}+\text{signal})$ в OpenPMIx показал, что около 50 % накладных расходов приходится на инициализацию данных и их очистку. Поэтому дальнейшая оптимизация должна быть направлена именно на эти участки кода.

Литература

1. Хорошевский В. Г. Архитектура вычислительных систем. М.: МГТУ им. Н. Э. Баумана, 2008. 520 с.
2. Yu W., Wu J., and Panda D. K. Fast and Scalable Startup of MPI Programs in InfiniBand Clusters // Proc. 11th International Conference on High Performance Computing (HiPC), Bangalore, India, December 19–22, 2004. P. 440–449.
3. Balaji P., Buntinas D., Goodell D., et al. PMI: A Scalable Parallel Processmanagement Interface for Extreme-scale Systems // Proc. 17th European MPI Users' Group Meeting Conference on Recent Advances in the Message Passing Interface (EuroMPI'10). Springer-Verlag, Berlin, Heidelberg. P. 31–41.
4. Castain R. H., Solt D., Hursey J., and Bouteiller A. PMIx: Process Management for Exascale Environments // Proc. 24th European MPI Users' Group Meeting (EuroMPI '17). ACM, New York, NY, USA, Article № 14, 10 p.
5. Polyakov A., Karasev B., Hursey J. et al. A Performance Analysis and Optimization of PMIx-based HPC Software Stacks // Proc. 26th European MPI Users' Group Meeting, September 2019, Article № 9. P. 1–10.
6. PMIx Consortium. 2017–2020. Process Management Interface for Exascale (PMIx) Standard (version 3.1). [Электронный ресурс]. URL: <https://github.com/pmix/pmix-standard/releases/download/v3.1/pmix-standard-3.1.pdf> (дата обращения: 10.04.2023).
7. Open PMIx master repository [Электронный ресурс]. URL: <https://github.com/openpmix/openpmix> (дата обращения: 10.04.2023).
8. Message Passing Interface Forum. 2015. MPI-3.1: Official document. [Электронный ресурс]. URL: <http://www.mpi-forum.org/docs> (дата обращения: 10.04.2023).

9. Argonne National Laboratory. 2014. Hydra Process Management Framework. [Электронный ресурс]. URL: https://wiki.mpich.org/mpich/index.php/Hydra_Process_Management_Framework (дата обращения: 10.04.2023).
10. PMIx Reference RunTime Environment (PRRTE). [Электронный ресурс]. URL: <https://github.com/openpmix/prrte> (дата обращения: 10.04.2023).
11. SchedMD, LLC. 2017. SLURM Workload Manager. [Электронный ресурс]. URL: <https://slurm.schedmd.com/> (дата обращения: 10.04.2023).
12. Chakraborty S., Subramoni H., Perkins J., Panda D. K. SHMEMPMI – Shared Memory Based PMI for Improved Performance and Scalability // Proc. 24th IEEE European MPI Users' Group Meeting (CCGrid). P. 60–69.
13. Calciu I., Dice D., Lev Y., Luchangco V. et. al. NUMA-aware reader-writer locks // Proc. 18th ACM SIGPLAN symposium on Principles and practice of parallel programming. ACM, New York, NY, USA, 2013. P. 157–166.
14. IEEE Standard for Information Technology–Portable Operating System Interface (POSIX(R)) Base Specifications, Issue 7. IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008) (Jan 2018). P. 1–3951.
15. Hsieh W. C., Weihl W. E. Scalable Reader-Writer Locks for Parallel Systems // Proc. Sixth IEEE International Parallel Processing Symposium, 1992. P. 656–659.
16. IEEE Standard for Information Technology–Portable Operating System Interface (POSIX(R)) Base Specifications, Issue 7. IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008) (Jan 2018). P. 1–3951. [Электронный ресурс]. URL: <https://doi.org/10.1109/IEEE-ESTD.2018.8277153> (дата обращения: 01.09.2023).
17. Mellor-Crummey J., Scott M. Algorithms for scalable synchronization on shared-memory multiprocessors // ACM Trans. Comput. Syst. 1991. № 9. P. 21–65.
18. Mellor-Crummey J. Algorithms for Scalable Lock Synchronization on Shared-memory Multiprocessors. Department of Computer Science Rice University. Lecture 18. 17 March 2009. P. 1–49. [Электронный ресурс]. URL: <https://cs.anu.edu.au/courses/comp8320/lectures/aux/comp422-Lecture18-HWLocks.pdf> (дата обращения: 10.04.2023).
19. Микробенчмарк. [Электронный ресурс]. URL: <https://bitly.su/Wtxe0Z> (дата обращения: 05.02.2023).
20. OpenPMIx. [Электронный ресурс]. URL: <https://shorturl.at/kwx14> (дата обращения: 16.04.2023).
21. Ефимов А. В., Поляков А. Ю., Крамаренко К. Е., Бочкарев Б. В. Оптимизация доступа к общей памяти для реализации стандарта PMIx // Вестник СибГУТИ. 2020. № 4. С. 28–38.

Крамаренко Константин Евгеньевич

инженер-программист ИФП СО РАН; ст. преподаватель СибГУТИ, тел. +7 383 2698 286, e-mail: kostya.kram@gmail.com, ORCID ID: 0000-0002-3844-5481.

Поляков Артём Юрьевич

старший архитектор программного обеспечения NVIDIA Corporation (Santa Clara, California, USA), e-mail: artemp@nvidia.com, ORCID ID: 0000-0001-6759-1448.

Ефимов Александр Владимирович

ведущий инженер-программист ИФП СО РАН (630090, Новосибирск, пр-кт. Ак. Лаврентьева, 13);

к.т.н., доцент СибГУТИ (630102, Новосибирск, ул. Кирова, 86), тел. +7 383 2698 293, e-mail: alexandr.v.efimov@gmail.com, ORCID ID: 0000-0002-8272-9924.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Shared Memory Access Optimization Analysis System for PMIx Standard Implementation

Konstantin E. Kramarenko, Artem Y. Polyakov, Alexander V. Efimov

¹ The Institute of Semiconductor Physics, Siberian Branch of the Russian Academy of Sciences (ISP SB RAS)

² Siberian State University of Telecommunications and Information Science (SibSUTIS)

³ NVIDIA

Abstract: The results in the field of system software stacks optimization for distributed computing systems based on the process management Interface (PMI) are obtained. Algorithms for synchronizing access to shared memory with a predominance of reading operations which are typical for PMIx implementation, are considered. The locking schemes $N(\text{mutex}+\text{signal})$ are proposed.

Keywords: distributed computing systems, resource management, process management interface, PMI, PMIx, shared memory.

For citation: Kramarenko K. E., Polyakov A. Y., Efimov A. V. Shared memory access optimization analysis system for PMIx standard implementation (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 29-39. <https://doi.org/10.55648/1998-6920-2024-18-1-29-39>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Kramarenko K. E., Polyakov A. Y.,
Efimov A. V., 2024

The article was submitted: 23.06.2023;
accepted for publication 03.12.2023.

References

1. Khoroshevskii V. G. *Arkhitektura vychislitel'nykh sistem* [Architectures of computer systems]. Moscow, MG TU im. Bauman, 2008. 520 p.
2. Yu W., Wu J., Panda D. K. Fast and Scalable Startup of MPI Programs in InfiniBand Clusters. *Proc. 11th International Conference on High Performance Computing (HiPC)*, Bangalore, India, 19-22 December, 2004, pp. 440-449.
3. Balaji P., Buntinas D., Goodell D., et al. PMI: A Scalable Parallel Processmanagement Interface for Extreme-scale Systems. *Proc. 17th European MPI Users' Group Meeting Conference on Recent Advances in the Message Passing Interface (EuroMPI'10)*, Springer-Verlag, Berlin, Heidelberg, pp. 31-41.
4. Castain R. H., Solt D., Hursey J., Bouteiller A. PMIx: Process Management for Exascale Environments. *Proc. 24th European MPI Users' Group Meeting (EuroMPI '17)*. ACM, New York, NY, USA, Article no. 14, 10 p.
5. Polyakov A., Karasev B., Hursey J. et al. A Performance Analysis and Optimization of PMIx-based HPC Software Stacks. *EuroMPI '19: Proceedings of the 26th European MPI Users' Group Meeting*, September, 2019, article no. 9, pp. 1-10.
6. *PMIx Consortium. 2017–2020. Process Management Interface for Exascale (PMIx) Standard (version 3.1)*, available at: <https://github.com/pmix/pmix-standard/releases/download/v3.1/pmix-standard-3.1.pdf> (accessed: 10.04.2023).
7. *Open PMIx master repository*, available at: <https://github.com/openpmix/openpmix> (accessed: 10.04.2023).

8. *Message Passing Interface Forum*. 2015. *MPI-3.1: Official document*, available at: <http://www.mpi-forum.org/docs> (accessed: 10.04.2023).
9. *Argonne National Laboratory*. 2014. *Hydra Process Management Framework*, available at: https://wiki.mpich.org/mpich/index.php/Hydra_Process_Management_Framework (accessed: 10.04.2023).
10. *PMIx Reference RunTime Environment (PRRTE)*, available at: <https://github.com/openpmix/prrte> (accessed: 10.04.2023).
11. *SchedMD, LLC*. 2017. *SLURM Workload Manager*, available at: <https://slurm.schedmd.com/> (accessed: 10.04.2023).
12. Chakraborty S., Subramoni H., Perkins J., Panda D. K. SHMEMPMI – Shared Memory Based PMI for Improved Performance and Scalability. *Proc. 24th IEEE European MPI Users' Group Meeting (CCGrid)*, pp. 60-69.
13. Calciu I., Dice D., Lev Y., Luchangco V. et. al. NUMA-aware reader-writer locks. *Proc. 18th ACM SIGPLAN symposium on Principles and practice of parallel programming*, ACM, New York, NY, USA, 2013, pp. 157-166.
14. IEEE Standard for Information Technology–Portable Operating System Interface (POSIX(R)) Base Specifications, iss. 7. *IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008)*, Jan, 2018, pp. 1-3951.
15. Hsieh W. C., Weihl W. E. Scalable Reader-Writer Locks for Parallel Systems. *Proc. Sixth IEEE International Parallel Processing Symposium*, 1992, pp. 656-659.
16. IEEE Standard for Information Technology–Portable Operating System Interface (POSIX(R)) Base Specifications, iss. 7. *IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008)*, Jan, 2018, pp. 1-3951, available at: <https://doi.org/10.1109/IEEESTD.2018.8277153> (accessed: 01.09.2023).
17. Mellor-Crummey J., Scott M. Algorithms for scalable synchronization on shared-memory multiprocessors. *ACM Trans. Comput. Syst*, 1991, no. 9, pp. 21-65.
18. Mellor-Crummey J. *Algorithms for Scalable Lock Synchronization on Shared-memory Multiprocessors*. Department of Computer Science Rice University, Lecture 18. 17 March, 2009, pp. 1-49. available at: <https://cs.anu.edu.au/courses/comp8320/lectures/aux/comp422-Lecture18-HWLocks.pdf> (accessed: 10.04.2023).
19. *Mikrobenchmark* [Microbenchmark], available at: <https://bitly.su/Wtxe0Z> (accessed: 05.02.2023).
20. *OpenPMIx*, available at: <https://shorturl.at/kwx14> (accessed: 16.04.2023).
21. Efimov A. V., Plyakov A. U., Kramarenko K. E., Bochkarev B. V. Optimizatsiya dostupa k obshchei pamyati dlya realizatsii standarta PMIx [Optimization of access to shared memory for the implementation of the PMIx standard]. *Vestnik SibGUTI*, 2020, no. 4, pp. 28-38.

Konstantin E. Kramarenko

Software engineer of ISP SB RAS;

Senior lecturer, SibSUTIS (Russia, 630102, Novosibirsk, Kirov St. 86), tel. + 7 383 2698 286,

e-mail: kostya.kram@gmail.com, ORCID ID: 0000-0002-3844-5481.

Artem Yu. Polyakov

Senior software architect of NVIDIA Corporation (Santa Clara, California, USA),

e-mail: artemp@nvidia.com, ORCID ID: 0000-0001-6759-1448.

Alexander V. Efimov

Senior software engineer ISP SB RAS (630090, Novosibirsk);

Ph.D., docent, SibSUTIS (Russia, 630102, Novosibirsk, Kirov St. 86), tel. +7 383 2698 293,

e-mail: alexandr.v.efimov@gmail.com, ORCID ID: 0000-0002-8272-9924.

Стохастическая модель функционирования вычислительных систем с накопителем при групповом обслуживании *

В. А. Павский¹, К. В. Павский^{2,3}

¹ Кемеровский государственный университет (КемГУ)

² Сибирский гос. ун-в. телекоммуникаций и информатики (СибГУТИ)

³ Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН)

Аннотация: Для оценки потенциальных возможностей вычислительных систем используются показатели осуществимости решения задач. Эти показатели характеризуют качество работы системы с учетом её надёжности и параметров поступающих задач. Одним из режимов работы вычислительной системы является режим обслуживания потоков задач. В силу особенностей этого режима при анализе функционирования вычислительные системы рассматривают как объект стохастический, который, в свою очередь, хорошо исследуется вероятностными методами. В работе предлагается математическая модель расчета показателей осуществимости решения задач на вычислительных системах с накопителем в режиме обслуживания потока задач. Построение модели выполнено в рамках теории массового обслуживания. Получены аналитические решения для оценки наполненности накопителя.

Ключевые слова: вычислительные системы, накопитель, показатели осуществимости решения задач, вероятность, теория массового обслуживания, групповое обслуживание.

Для цитирования: Павский В. А., Павский К. В. Стохастическая модель функционирования вычислительных систем с накопителем при групповом обслуживании // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 40–47. <https://doi.org/10.55648/1998-6920-2024-18-1-40-47>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

Павский В. А., Павский К. В., 2024

Статья поступила в редакцию 11.10.2023;
переработанный вариант – 29.11.2023;
принята к публикации 01.12.2023.

1. Введение

Масштабируемые вычислительные системы (ВС) относятся к наиболее развитым быстродействующим вычислительным средствам вычислительной техники [1]. Эффективность функционирования ВС при решении задач, безусловно, определяется производительностью всех компонентов, способами организации их совместного функционирования, а также надёжностью и живучестью систем [2, 3]. Поэтому разработка средств организации эффективного и надёжного функционирования вычислительной техники относится к важным проблемам [4–6].

* Работа выполнена в рамках государственного задания № 071-03-2023-001 от 19.01.2023.

Для оценки потенциальных возможностей вычислительных систем используются показатели осуществимости решения задач [1]. Эти показатели характеризуют качество работы системы с учетом надежности и параметров поступающих задач.

В зависимости от сложности задач и характера их поступления выделяют следующие режимы работы вычислительной системы [1]:

- решение сложной задачи;
- обработка наборов задач;
- обслуживание потоков задач.

В силу особенностей таких задач и процесса их решения на ВС функционирование систем рассматривается как объект стохастический, который, в свою очередь, хорошо исследуется вероятностными методами [5, 6]. В работе предлагается математическая модель для расчета показателей осуществимости решения задач в режиме обслуживания потока задач на вычислительных системах с накопителем. Построение модели выполнено в рамках теории массового обслуживания [7, 8].

На рис. 1 представлена модель системы с накопителем, на который поступает поток задач. Из задач формируются пакеты определенного размера с последующей их обработкой на системе, где α – параметр интенсивности поступающих задач, а β – параметр интенсивности решения задач, зависящий от производительности ВС [1, 9]. В работе представлены аналитические решения для оценки загруженности накопителя в стационарном режиме функционирования вычислительной системы.

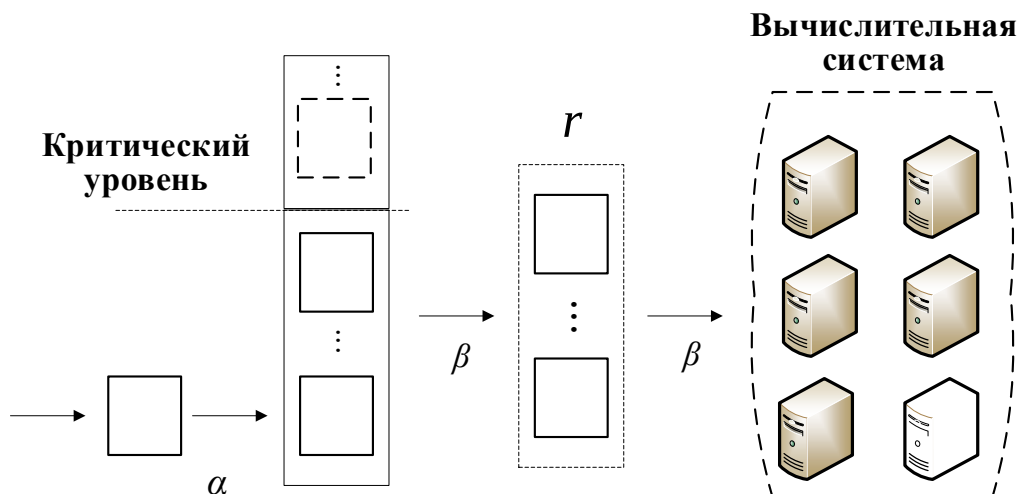


Рис. 1. Модель системы с накопителем

2. Математическая модель

Рассматривается система массового обслуживания (СМО), на которую поступает случайный поток требований с интенсивностью α . В каждый момент времени $t \in [0, \infty)$ СМО находится в одном из множества несовместных состояний C_k , где k – число требований, находящихся в системе, включая необслуженное (рис. 2). Группы из r требований обслуживаются с интенсивностью β , при этом $\alpha < r\beta$. Считаем, что все потоки событий – пуассоновские [1, 4].

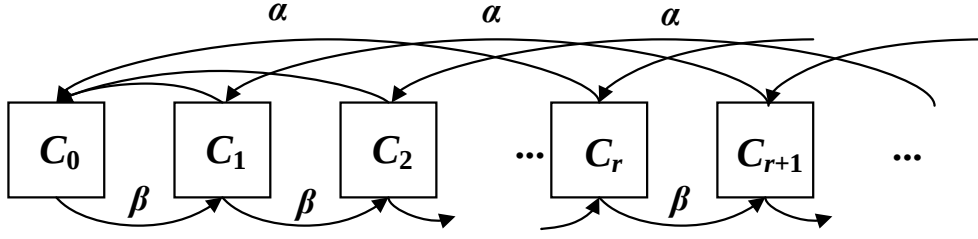


Рис. 2. Граф-схема состояний СМО

Пусть $P_k(t)$ – вероятность того, что в момент времени t , $t \in [0, \infty)$ СМО находится в состоянии C_k . В соответствии с граф-схемой составляется система дифференциальных уравнений и задаются начальные условия. Для переходного режима граф-схема (рис. 2) формализуется системой дифференциальных уравнений [7], где неизвестными функциями являются вероятности $P_k(t)$ того, что СМО находится в состоянии C_k в момент времени t :

$$\frac{dP_k(t)}{dt} = \begin{cases} -\alpha P_0(t) + \beta(P_1(t) + P_2(t) + \dots + P_r(t)), & \text{если } k = 0, \\ -(\alpha + \beta)P_k(t) + \beta P_{k+r}(t) + \alpha P_{k-1}(t), & \text{если } k > 0, \end{cases} \quad (1)$$

при начальных условиях $P_i(0) = 1$; $P_k(0) = 0$, $k \neq i$ и условии нормировки, которое является следствием формулировки модели:

$$\sum_{k=0}^{\infty} P_k(t) = 1.$$

Получены частные решения системы (1) при $r = 1$ и $r \rightarrow \infty$. В работе [1] приведено решение для $r = 1$. Решение системы (1) для $r \rightarrow \infty$ приводится далее. Вероятностные функции $P_0(t)$ находятся из линейного дифференциального уравнения первого порядка [10]:

$$\frac{dP_0(t)}{dt} = -\alpha P_0(t) + \beta \sum_{k=1}^{\infty} P_k(t).$$

Последующие функции $P_k(t)$ находятся рекуррентно из дифференциальных уравнений

$$\frac{dP_k(t)}{dt} = -(\alpha + \beta)P_k(t) + \alpha P_{k-1}(t) \text{ для } k > 1.$$

При $r \rightarrow \infty$ результатом решения системы (1) являются функции:

$$\begin{cases} P_k(t) = \frac{\beta \alpha^k}{(\alpha + \beta)^{k+1}} - \beta \alpha^k e^{-(\alpha + \beta)t} \sum_{j=0}^k \frac{t^{k-j}}{(k-j)!(\alpha + \beta)^{j+1}}, & k < i, \\ P_k(t) = \frac{\beta \alpha^k}{(\alpha + \beta)^{k+1}} + e^{-(\alpha + \beta)t} \left(\sum_{j=0}^k \frac{\alpha^{k-j}}{(k-j)!} t^{k-j} - \beta \alpha^k \sum_{j=0}^k \frac{t^{k-j}}{(k-j)!(\alpha + \beta)^{j+1}} \right), & k \geq i. \end{cases}$$

Полученные для вероятностей $P_k(t)$ формулы позволяют оценить скорость вхождения системы в стационарный режим работы.

Рассмотрим стационарный режим функционирования СМО. Имеем по определению $p_k = \lim_{t \rightarrow \infty} P_k(t)$, тогда из (1) получаем:

$$p_k = \begin{cases} \frac{\beta}{\alpha} (p_1 + p_2 + \dots + p_r), & \text{если } k = 0, \\ \frac{1}{(\alpha + \beta)} (\beta p_{k+r} + \alpha p_{k-1}), & \text{если } k > 0. \end{cases} \quad (2)$$

Для решения системы алгебраических уравнений (2) введем производящую функцию [7, 8, 11]:

$$F(z) = \sum_{k=0}^{\infty} p_k z^k, \text{ где } |z| \leq 1. \quad (3)$$

Умножая каждое уравнение k системы (2) на z^k , $k=0, 1, \dots$ и суммируя их, получаем выражение для производящей функции:

$$F(z) = \left(\beta \sum_{k=0}^r p_k z^k - (\alpha + \beta) p_0 z^r \right) / \left(\alpha (z - (\alpha + \beta) p_0) z^r + \beta \right) \quad (4)$$

После стандартных преобразований над уравнением (4) получаем решение [7]:

$$F(z) = \frac{1 - \frac{1}{z_0}}{1 - \frac{z}{z_0}}, \quad (5)$$

где $z_0 > 1$ иррациональный корень многочлена степени $r+1$,

$$\alpha (z - (\alpha + \beta) p_0) z^r + \beta = 0.$$

Учитывая, что $F(0) = p_0$, а также условие нормировки, для вероятностей p_k получаем представление:

$$p_k = p_0 (1 - p_0)^k, \quad k = 0, 1, 2, \dots \quad (6)$$

3. Оценка наполненности накопителя

Пусть $P_{\text{отк}}(s, t)$ – вероятность того, что накопитель будет переполнен сверх значения $(s-1)$, тогда

$$P_{\text{отк}}(s, t) = \sum_{k=s}^{\infty} P_k(t).$$

Для стационарного режима пусть $p_{\text{отк}}(s) = \lim_{t \rightarrow \infty} P_{\text{отк}}(s, t)$. Учитывая (6), получаем:

$$p_{\text{отк}}(s) = \sum_{k=s}^{\infty} p_k = (1 - p_0)^s, \quad s = 1, 2, \dots \quad (7)$$

Для (7) находим обратную функцию

$$s(p_{\text{отк}}) = \frac{\ln(p_{\text{отк}})}{\ln(1 - p_0)}. \quad (8)$$

Из (8) по заданной вероятности $p_{\text{отк}}$ получаем оценку для $s_{\text{кр}}$ критического уровня накопителя

$$s_{\text{кр}} \geq \left\lceil \frac{\ln(p_{\text{отк}})}{\ln(1 - p_0)} \right\rceil + 1, \quad (9)$$

где $[x]$ – целая часть числа x .

Решения (9) и (6) получены в общем виде, т.к. не определена вероятность p_0 в зависимости от размера группы. Ниже представлены частные решения для (9) и (6).

4. Частные решения для расчета наполненности накопителя

Приведём решения для расчета наполненности накопителя в зависимости от размера группы.

Обратимся к уравнению (2) при $k=0$ и формуле (6), тогда

$$p_0 = \frac{\beta}{\alpha} \left(p_0 (1 - p_0) + p_0 (1 - p_0)^2 + \dots + p_0 (1 - p_0)^r \right).$$

Далее получаем алгебраическое уравнение r -го порядка относительно неизвестного p_0 :

$$1 - \frac{\beta}{\alpha} \sum_{j=1}^r (1 - p_0)^j = 0. \quad (10)$$

Из уравнения (10) при значениях размера группы $r = 1$, $r = 2$ и $r \rightarrow \infty$ получаем:

при $r = 1$

$$p_0 = \frac{\beta - \alpha}{\beta};$$

при $r = 2$

$$p_0 = \frac{3}{2} - \frac{\sqrt{1 + 4 \frac{\alpha}{\beta}}}{2};$$

при $r \rightarrow \infty$

$$p_0 = \frac{\beta}{\alpha + \beta}.$$

Исходя из частных решений для (6), системы уравнений для вероятностей p_k состояний системы (2), получены решения для $s(p_{\text{отк}})$ при $r = 1$, $r = 2$ и $r \rightarrow \infty$ по заданной вероятности $p_{\text{отк}}$:

при $r = 1$

$$s(p_{\text{отк}}) = \frac{\ln(p_{\text{отк}})}{\ln(\alpha) - \ln(\beta)}.$$

при $r = 2$

$$s(p_{\text{отк}}) = \frac{\ln(p_{\text{отк}})}{\ln\left(\sqrt{1 + 4 \frac{\alpha}{\beta}} - 1\right) - \ln(2)}.$$

при $r \rightarrow \infty$

$$s(p_{\text{отк}}) = \frac{\ln(p_{\text{отк}})}{\ln \alpha - \ln(\alpha + \beta)}.$$

Замечание. Случай $r \rightarrow \infty$ предполагает, что ВС обладает достаточным ресурсом, чтобы при больших значениях r решать поставленные задачи за заданное время.

На рис. 3 представлены результаты расчета накопителя $s(p_{\text{отк}})$ при $r = 1$, $r = 2$ и $r \rightarrow \infty$. Пример показывает динамику уменьшения и предел размера накопителя в зависимости от увеличения размера группы.

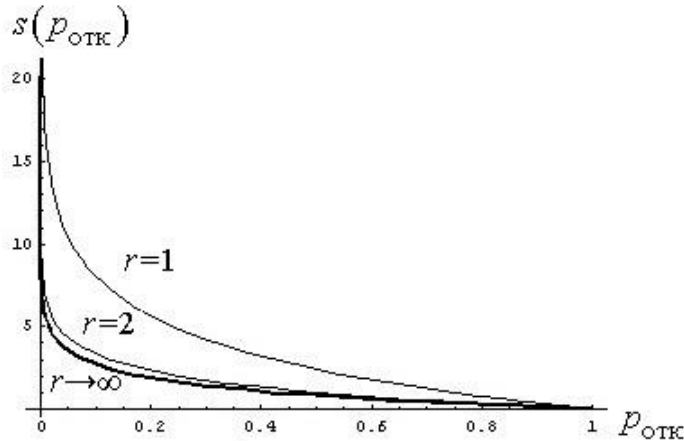


Рис. 3. Расчет размера накопителя $s(p_{\text{отк}})$ для группы $r = 1$, $r = 2$ и $r \rightarrow \infty$ при $\alpha = 18$ 1/ч, $\beta = 24$ 1/ч.

5. Заключение

Предложена математическая модель вычислительной системы с накопителем, на которую поступает случайный поток задач. Задачи потока обслуживаются группами согласно пуассоновскому распределению. Представлены частные решения, где неизвестными функциями являются вероятности состояний накопителя в переходном режиме функционирования системы и в предельном для размера группы случае. Предложены формулы оценки загрузки накопителя. Частные аналитические решения представлены для размера группы $r = 1$, $r = 2$ и $r \rightarrow \infty$. Полученные аналитические решения для стационарного режима функционирования вычислительной системы удобны для инженерных расчетов и при экспресс-анализе их функционирования.

Литература

1. *Хорошевский В. Г.* Архитектура вычислительных систем. М.: МГТУ им. Н. Э. Баумана, 2008. 520 с.
2. *Gupta S., Patel T., Engelmann C., and Tiwari D.* Failures in large scale systems: long-term measurement, analysis, and implications // Proc. International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, Colorado, November 12–17, 2017.
3. *Schroeder B., Gibson G.* A large-scale study of failures in high-performance computing systems // Proc. International Conference on Dependable Systems and Networks (DSN), Philadelphia, PA, USA, June 25–28, 2006.
4. *Хорошевский В. Г.* Модели анализа и организации функционирования больше-масштабных распределенных вычислительных систем // Электронное моделирование. 2003. Т. 25, № 6.
5. *Xie M., Dai Y. S., and Poh K. L.* Computing system reliability: models and analysis. New York: Kluwer academic publishers, 2004.
6. *Harchol-Balter M.* Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge University Press, 2013.
7. *Клейнрок Л.* Теория массового обслуживания. М.: Машиностроение, 1979. 432 с.
8. *Саати Т. Л.* Элементы теории массового обслуживания и ее приложения. М.: Книжный дом «ЛИБРОКОМ», 2010. 520 с.
9. *Павский К. В., Павский В. А.* Модель для расчета показателей осуществимости решения задач на распределенных вычислительных системах с накопителем // Вестник СибГУТИ. 2014. № 4. С. 4–8.
10. *Степанов В. В.* Курс дифференциальных уравнений. М.: Едиториал УРСС, 2006. 472 с.
11. *Феллер В.* Введение в теорию вероятностей и ее приложения: в 2-х т. Том 1. М.: «ЛИБРОКОМ», 2010. 528 с.

Павский Валерий Алексеевич

д.т.н., профессор, профессор кафедры общей математики и информатики, Кемеровский государственный университет (КемГУ), e-mail: pavva46@mail.ru, ORCID ID: 0000-0001-9460-7126.

Павский Кирилл Валерьевич

д.т.н., доцент, профессор кафедры вычислительных систем, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ); заведующий Лабораторией вычислительных систем, Институт физики полупроводников им. А. В. Ржанова СО РАН (ИФП СО РАН), e-mail: pkv@isp.nsc.ru, ORCID ID: 0000-0002-2271-7358.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Stochastic Model of Computer Systems Functioning with Accumulator Storage for Serving Groups

Valery A. Pavsky¹, Kiril V. Pavsky^{2,3}

¹ Kemerovo State University (KemSU)

² Siberian State University of Telecommunications and Information Science (SibSUTIS)

³ Rzhanov Institute of Semiconductor Physics (ISP SB RAS)

Abstract: To assess the potential capabilities of computer systems, feasibility indicators for solving problems are used. These indicators characterize the quality of the systems taking into account reliability and parameters of incoming tasks. One of the modes of computer systems functioning is the task flow servicing mode. Due to the peculiarities of this mode, when analyzing the functioning of computer systems, they are considered as a stochastic object, which in turn is well studied by stochastic methods. The paper proposes a mathematical model for calculating feasibility indicators for solving tasks in servicing mode of task flow on computer systems with accumulator storage. The model is constructed within the framework of queuing theory. Analytical solutions for assessing the fullness of accumulator storage are obtained.

Keywords: computer systems, feasibility indicators for solving tasks, accumulator storage, queuing theory, group servicing.

For citation: Pavsky V. A., Pavsky K. V. Stochastic model of computer systems functioning with accumulator storage for serving groups (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 40-47. <https://doi.org/10.55648/1998-6920-2024-18-1-40-47>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Pavsky V. A., Pavsky K. V., 2024

The article was submitted: 11.10.2023;
revised version: 29.11.2023;
accepted for publication 01.12.2023.

References

1. Horoshevskij V. G. *Arhitektura vychislitel'nyh sistem* [Architecture of computer systems]. Moscow, BMSTU, 2008. 520 p.
2. Gupta S., Patel T., Engelmann C., and Tiwari D. Failures in large scale systems: long-term measurement, analysis, and implications. *SC '17: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Article no. 44, Denver, Colorado, 12 -17 November, 2017.
3. Schroeder B., Gibson G. A large-scale study of failures in high-performance computing systems. *Proceedings of the International Conference on Dependable Systems and Networks (DSN2006)*, Philadelphia, PA, USA, 25-28 June, 2006, 10 p.
4. Horoshevskij V. G. Modeli analiza i organizacii funkcionirovaniya bol'shemasshtabnyh raspredelennyh vychislitel'nyh sistem [Models of analysis and organization of large-scale distributed computer systems]. *Jelektronnoe modelirovanie*, Kiev, 2003, vol 25, no. 6.
5. Xie M., Dai Y. S., and Poh K. L. *Computing system reliability: models and analysis*. New York, Kluwer academic publishers, 2004.

6. Harchol-Balter M. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.
7. Klejnrok L. *Teorija massovogo obsluzhivaniya* [Queueing theory]. Moscow, Mashinostroenie, 1979. 432 p.
8. Saati T. L. *Jelementy teorii massovogo obsluzhivaniya i ee prilozheniya* [Elements of queueing theory and its applications]. Moscow, Knizhnyj dom «LIBROKOM», 2010. 520 p.
9. Pavskij K. V., Pavskij V. A. Model' dlja rascheta pokazatelej osushhestvimosti resheniya zadach na raspredelennyh vychislitel'nyh sistemah s nakopitelem [Model for realizability calculation of tasks solution on distributed computer systems with buffer]. *Vestnik SibGUTI*, 2014, vol. 4, pp.4-8.
10. Stepanov V. V. *Kurs differencial'nyh uravnenij* [Course of differential equations]. Moscow, Editorial URSS, 2006. 472 p.
11. Feller V. *Vvedenie v teoriju verojatnostej i ee prilozheniya* [An introduction to probability theory and its applications. Vol. 1]., Moscow, « LIBROKOM», 2010. 528 p.

Valery A. Pavsky

Dr. of Sci. (Engineering), Professor, Professor of Department of the General mathematics and informatics, Kemerovo State University (KemSU), e-mail: pavva46@mail.ru, ORCID ID: 0000-0001-9460-7126.

Kiril V. Pavsky

Dr. of Sci. (Engineering), Assistant professor, Head of Computer Systems Laboratory, Rzhhanov Institute of Semiconductor Physics Siberian Branch of Russian Academy of Sciences (ISP SB RAS); Professor of the Department of computer science, Siberian State University of Telecommunications and Information Science (SibSUTIS), e-mail: pkv@isp.nsc.ru, ORCID ID: 0000-0002-2271-7358.

Волоконно-оптическая система морского базирования для сбора и передачи данных

В. М. Деревяшкин¹, А. О. Виркунин¹, Н. И. Горлов¹,
А. Е. Малащенко², А. И. Сидоренко²

¹ Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

² СКБ средств автоматизации морских исследований ДВО РАН (г. Южно-Сахалинск)

Аннотация: В статье представлены принципы построения волоконно-оптической системы морского базирования, предназначенной для двустороннего информационного обмена между подводными станциями и аппаратно-программными комплексами. Предложены методы контроля основных параметров подводных станций. Проанализированы основные принципы мониторинга физической среды передачи. При этом обоснована необходимость применения когерентного оптического рефлектометра, позволяющего улучшить чувствительность мониторинга до квантового предела за счет гетеродинного детектирования.

Ключевые слова: волоконно-оптическая система, донный модуль, аппаратно-программный модуль, коммутатор, модем, система управления, мониторинг, когерентный рефлектометр.

Для цитирования: Деревяшкин В. М., Виркунин А. О., Горлов Н. И., Малащенко А. Е., Сидоренко А. И. Волоконно-оптическая система морского базирования для сбора и передачи данных // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 48–58.

<https://doi.org/10.55648/1998-6920-2024-18-1-48-58>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Деревяшкин В. М., Виркунин А. О.,
Горлов Н. И., Малащенко А. Е.,
Сидоренко А. И., 2024

Статья поступила в редакцию 24.05.2023;
переработанный вариант – 04.07.2023;
принята к публикации 03.09.2023.

1. Введение

Развитие технологий оптической связи в настоящее время проявляется не только в традиционных телекоммуникационных системах, но и в различных технических приложениях. Ряд существенных преимуществ, таких как высокая скорость передачи, высокая надежность, обеспечение высокой помехоустойчивости, позволяют передавать информацию на большие расстояния, в том числе и в водной среде, обеспечивая при этом необходимые качественные показатели. Сферами применения таких систем являются оборона, экологический и климатический мониторинг водного пространства, освоение подводных месторождений полезных ископаемых, контроль гидротехнических сооружений и т.д.

Волоконно-оптическая система морского базирования (ВОСМБ) сбора и передачи данных предназначена для решения следующих задач:

1. Обеспечение двустороннего информационного обмена между донными модулями (ДМ), в которых находится сетевое оборудование, и аппаратно-программными модулями (АПМ) с оборудованием сбора данных. Оборудование для сбора данных может включать в себя различные датчики и станции, например, станции гидроакустического мониторинга,

датчики сейсмоактивности и т.д. Количество донных модулей, модулей сбора данных, а также расстояния между ними могут варьироваться.

2. Передача информации от аппаратно-программных модулей в береговой пункт приема и обработки данных.

3. Управление оборудованием донных модулей и аппаратно-программных модулей.

4. Обеспечение дистанционным электропитанием оборудования ДМ и АПМ.

5. Автоматизированный контроль технического состояния волоконно-оптического кабеля и выдача данных контроля в централизованную систему управления сетевыми элементами [1].

Для проверки основных технических решений была собрана гибридная волоконно-оптическая система, в которой ДМ последовательно соединены между собой оптическими кабелями связи. К донным модулям с помощью медного кабеля подключается аппаратура станций сбора данных. АПМ и ДМ могут соединяться между собой оптическими линиями связи.

2. Структурная схема гибридной волоконно-оптической системы

Для организации двустороннего обмена между ДМ и передачи данных в береговой пункт приема используются последовательно соединенные коммутаторы доступа Ethernet уровня L2 MES 2208P [2] производства компании ЭЛТЕКС. Передача информации по медному кабелю между донными модулями и аппаратно-программными модулями осуществляется с помощью высокоскоростных DSL-модемов MXL-2E [3] (ЭЛТЕКС).

Структурная схема ВОСМБ показана на рис. 1.

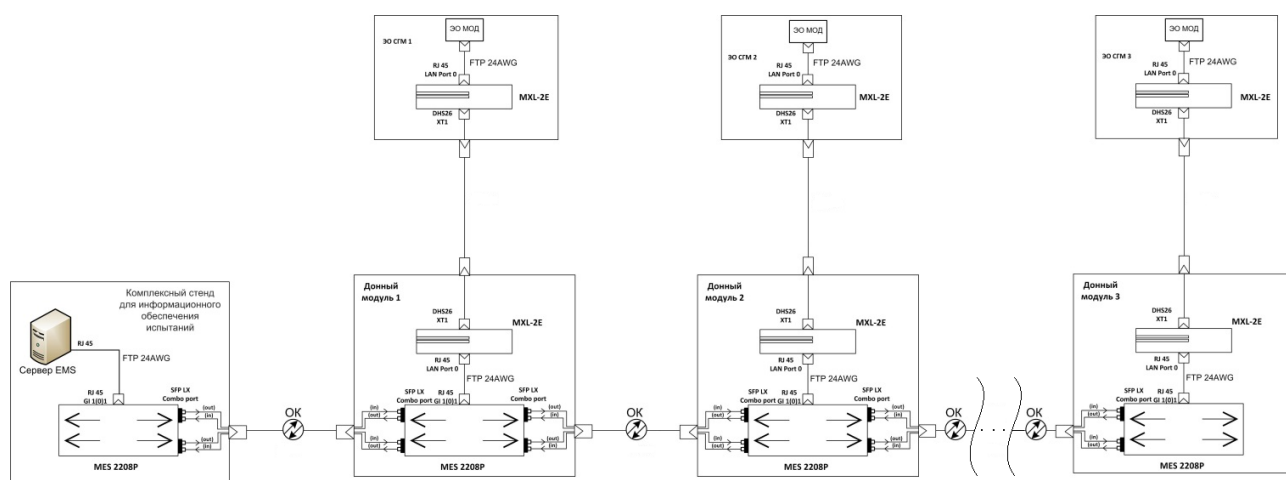


Рис. 1. Структурная схема ВОСМБ

Централизованное управление элементами сети осуществляется с помощью системы управления «Eltex.EMS» [4]. Она контролирует параметры такого оборудования, как коммутаторы MES и SHDSL-модемы, посредством активного мониторинга. Для обмена информацией с оборудованием на сети используется адаптированный SNMP-менеджер, реализующий наиболее частые и массовые операции по управлению параметрами оборудования.

Система «Eltex.EMS» устроена по клиент-серверной архитектуре. Единый сервер доступа предоставляет интерфейс, позволяющий производить независимое одновременное управление различными элементами сети.

Данная система управления имеет возможность установки пакетов под используемый вид оборудования, что актуально для реализуемого проекта, поскольку используются два типа оборудования из всей линейки, что позволяет оптимально нагружать серверные ресурсы,

видеть оперативные данные о состоянии сети и рационально использовать человеческие ресурсы. Таким образом, происходит максимально эффективная отдача от системы.

Специфика прокладки и особая агрессивность внешней среды накладывают жёсткие требования к волоконно-оптическим и медным кабелям:

- наличие бронепокровов, позволяющих выдерживать высокие разрывные и раздавливающие усилия, возникающие при подводной прокладке;
- наличие в кабеле гидрофобного (битумного) заполнителя, препятствующего прониканию в кабель воды;
- наличие в конструкции стальных и полипропиленовых жгутов, позволяющих справляться с постоянной эксплуатационной нагрузкой;
- минимально возможный вес кабеля, позволяющий работать с большими строительными длинами (до 80 км);
- при прокладке на больших глубинах кабель должен выдерживать высокое гидростатическое давление (свыше 40 МПа);
- токопроводящая жила (ТПЖ) кабеля, предназначенная для питания оборудования, должна обеспечивать минимальное электрическое сопротивление. При протяженной линии важным фактором также будет являться максимальное рабочее напряжение ТПЖ, которое позволит питать самые удаленные модули [5].

3. Испытания системы

С целью проверки работоспособности и надежности системы, а также основных параметров передачи проведены испытания системы на действующем макете (рис. 2) в береговой и морской фазах испытаний.

Основные параметры макета:

- количество ДМ – 4;
 - расстояние между донными модулями – 4 км;
 - линия связи между донными модулями – оптический кабель;
 - интерфейс связи между донными модулями – Ethernet;
 - линия связи между ДС и АПМ – медный кабель;
 - интерфейс связи между ДС и АПМ – SHDSL;
 - питание системы – дистанционное, постоянным напряжением до 700 В;
 - питание объекта – дистанционное, вторичное, постоянным напряжением 48 В;
 - максимальная допустимая потребляемая мощность питания узла – до 120 Вт.
- Все кабели подключаются к оборудованию с помощью специальных герморазъемов.



Рис. 2. Действующий макет ВОСМБ

Береговая фаза включала в себя следующие проверки и исследования:

1. Проверка аппаратной защиты портов (агрегирование портов) при эмуляции аварийных ситуаций.

Контроль за работоспособностью канала связи при эмуляции аварийных ситуаций осуществлялся с помощью специальной программы – генератора трафика (iperf). Физическое повреждение линии имитировалось с помощью перегиба патчкорда либо путем отключения порта приема/передачи.

Логические аварии имитировались путем внесения повышенного затухания в линию с помощью переменного аттенюатора.

2. Проверка системы предупреждения аварийных ситуаций при превышении допустимых значений уровня сигнала в линии.

Проверка предупреждения аварийных ситуаций заключалась в выставлении референсных значений сигнала, равных тем, которые будут сигнализировать о серьезной проблеме на линии. Задавались значения уровня сигнала с помощью специальной команды, в которой указывались следующие значения: низкий – критический, низкий – предупреждающий, высокий – предупреждающий, высокий – критический.

3. Исследование скорости передачи информации между модемами в зависимости от разрядности кода.

SHDSL-модем, используемый в данном макете, поддерживает формат кода ТС-РАМ. На данном оборудовании есть возможность менять разрядность кода от 4 до 128, соответственно, число поддерживаемых основных цифровых каналов меняется от 38 до 178. Разрядность кода выбиралась исходя из длины кабеля между модемами. Необходимо выставить достаточную разрядность для обеспечения необходимой скорости при соблюдении значения отношения сигнал/шум около 11 дБ. Для доступа к настройкам необходимо было войти в графический интерфейс модема, введя в адресной строке его IP-адрес.

4. Исследование зависимости отношения сигнал/шум от скорости передачи информации.

При проведении данных испытаний выбирались оптимальные значения отношения сигнал/шум. Необходимо было добиться значения около 11 дБ. Деградация сигнала начиналась

при значении SNR порядка 3–6 дБ. Значение SNR проверялось с помощью встроенных инструментов анализа в графическом интерфейсе модема.

5. Исследование времени задержки при эмуляции аварийных ситуаций в волоконно-оптическом кабеле.

Исследование проводилось путём имитации наихудших условий передачи сигнала при полностью загруженном канале. Для загрузки канала использовалась специальная программа – генератор трафика iperf.

6. Проверка энергетического запаса (пределов допустимых вносимых потерь) в волоконно-оптическом кабеле.

При проверке энергетического запаса определялись пороговые значения приема SFP-модуля с помощью переменного аттенюатора. При постепенном увеличении вносимого затухания до порогового значения передача сигнала прекращалась.

Морская фаза испытаний проводилась после успешных береговых испытаний. После погружения оборудования в водную среду контролировались следующие основные параметры системы:

- мощность приема/передачи сигнала между сетевыми элементами;
- температура;
- доступность сетевых элементов;
- корректная работа сетевых элементов;
- проверка ошибочно принятых пакетов;
- оценка SNR.

Данные параметры проверялись с помощью системы управления.

4. Мониторинг волоконно-оптической системы морского базирования

В процессе строительства и технической эксплуатации ВОСМБ необходимо проводить комплекс измерений с целью определения технического состояния и предупреждения повреждений волоконно-оптических кабелей (ВОК). В перечень обязательных к контролю параметров ВОК относятся: вносимые потери, коэффициенты затухания и отражения. Для проведения измерений предлагается применение универсальной измерительной системы для тестирования и мониторинга. Оборудование должно быть развернуто на каждой оконечной станции для обеспечения текущего мониторинга с целью технического обслуживания. Мониторинг состояния системы может быть получен путем периодического сбора данных о производительности с подводной станции. Кроме того, для удобства обслуживания оно должно поддерживать или предоставлять интерфейсы для поиска неисправностей. Для конфигурации соединения оборудования мониторинга (monitoring equipment) в подводной кабельной системе можно выбрать два различных способа:

- Терминальное передающее оборудование подключается между оборудованием мониторинга и кабельным оконечным оборудованием (см. рис. 3). Сигналы мониторинга и служебные сигналы объединяются в терминальном передающем оборудовании и передаются в подводную кабельную систему [6].

- Оборудование мониторинга подключается между терминальным передающим оборудованием и кабельным оконечным оборудованием (см. рис. 4), служебные сигналы и сигналы мониторинга объединяются в оборудовании мониторинга и передаются в подводную кабельную систему.

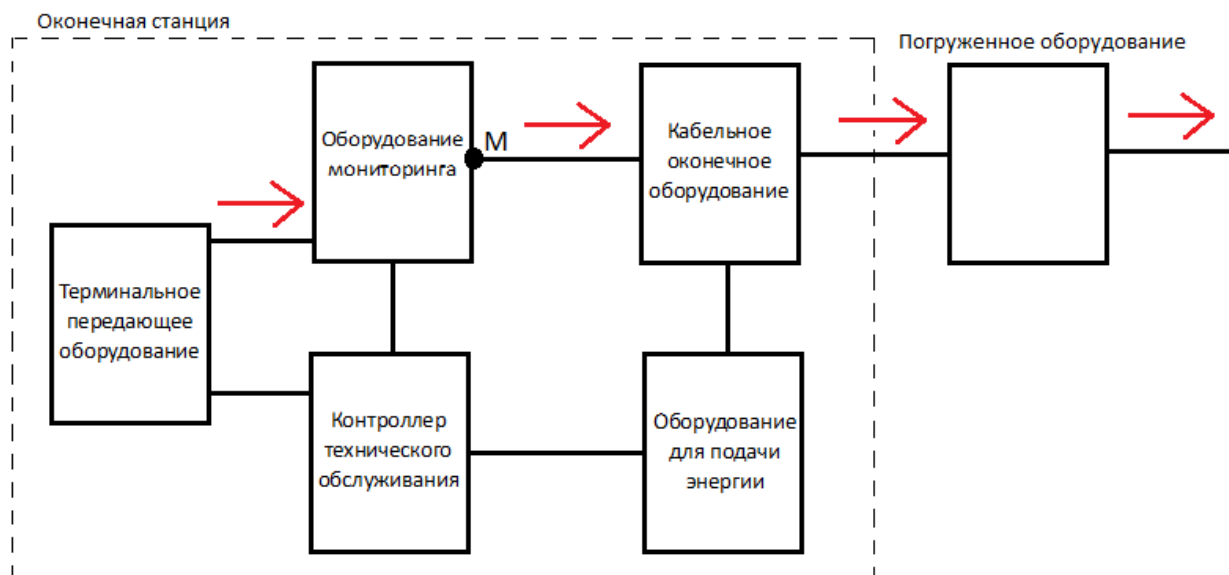


Рис. 3. Схема включения терминального передающего оборудования между оборудованием мониторинга и входом зондируемой кабельной системы

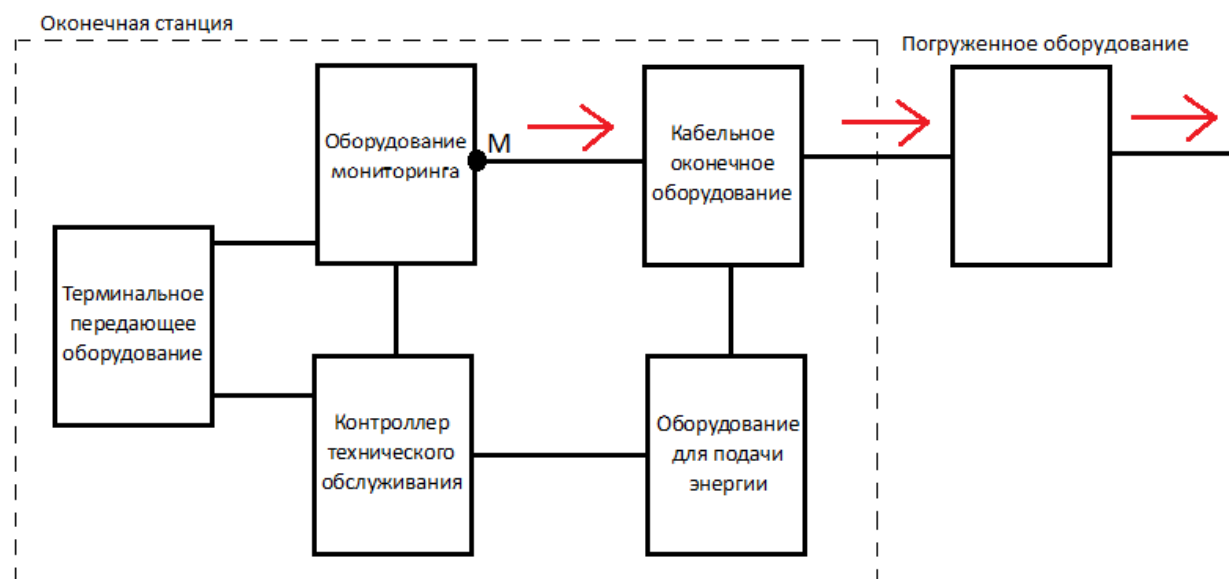


Рис. 4. Схема включения оборудования мониторинга между терминальным передающим оборудованием и входом зондируемой кабельной системы

Контрольные точки (рис. 3, 4) должны соответствовать параметрам, указанным для мониторинга. В точке М выходной интерфейс мониторинга должен соответствовать следующим параметрам:

- выходная мощность зондирующего света, длина волны и ширина импульса (для пассивного мониторинга);
- параметры модуляции (для активного контроля).

При пассивном мониторинге характеристики получают путем обнаружения оптических или электрических сигналов. Сигналы зонда посылаются на погруженную установку, а возвращенные сигналы анализируются для оценки состояния работоспособности системы. Например, рассеянный назад свет рефлектометра/COTDR, описанного в [7], обнаруживается и обрабатывается, обычно в виде кривых зависимостей интенсивности сигнала от расстояния

для анализа и диагностики состояния системы. Одно оборудование для пассивного мониторинга контролирует производительность подводной кабельной системы в одном направлении. Чтобы получить двунаправленную производительность, оборудование для пассивного мониторинга должно быть развернуто на каждой оконечной станции.

При пассивном мониторинге рекомендуется отслеживать изменения состояния в зондируемых оптических путях, такие как обрывы волокон, изменения усиления ретранслятора, а также затухание или отражения волокон. Поскольку пассивный мониторинг получает информацию о состоянии системы косвенным путем, при мониторинге состояния используется сравнение с базовыми показателями. Мониторинг состояния должен использовать те же параметры, что и базовые линии, чтобы обеспечить достоверность сравнения и найти изменения в работе, вызванные повреждениями или неисправностями, которые изменяют состояние системы. Базовые показатели должны быть собраны после надлежащего развертывания системы и должны обновляться после каждого ремонта или изменения конфигурации системы.

Поиск неисправностей часто выполняется вручную в нерабочем состоянии. Это может быть выполнено РМЕ на основе механизма рефлектометрии/COTDR или с помощью коммерческого оборудования рефлектометрии/COTDR. Применение когерентного оптического рефлектометра позволяет улучшить чувствительность мониторинга до квантового предела за счет гетеродинного детектирования. Высокоточное разрешение при поиске неисправностей всегда полезно для технического обслуживания. Рекомендуются удобные функции, например, автоматическое определение места повреждения, быстрое определение места повреждения с меньшей точностью, но за более короткое время. Подводное оборудование должно иметь обратные пути, как описано в [8], для поддержки рефлектометрии/COTDR для поиска неисправностей.

При активном мониторинге оборудование мониторинга запрашивает и собирает непосредственно статус производительности погруженного оборудования. Соответствующими параметрами производительности являются входная мощность, выходная мощность, ток насоса и т.д.

Абсолютная погрешность измерения координаты неоднородности в оптическом волокне определяется длительностью зондирующего импульса. Как и в некогерентной рефлектометрии, она может быть оценена по формуле:

$$\Delta L_{\text{изм.}} = c_0 t_{\text{и}} / 2n_g, \quad (1)$$

где: $\Delta L_{\text{изм.}}$ – абсолютная погрешность измерения координаты;

c_0 – скорость света в свободном пространстве;

$t_{\text{и}}$ – длительность зондирующего импульса;

n_g – групповой коэффициент преломления сердцевины.

Улучшение метрологических параметров оптического рефлектометра предлагается осуществить посредством когерентного приема обратно рассеянного сигнала. Оно позволяет существенно расширить динамический диапазон измерения. Кроме этого, становится возможным применение одного и того же гетеродина в оптическом передатчике и приемнике. Это в значительной степени снижает проблему стабилизации частоты оптического излучения. Техническая реализация метода когерентного приема представлена на рис. 5.

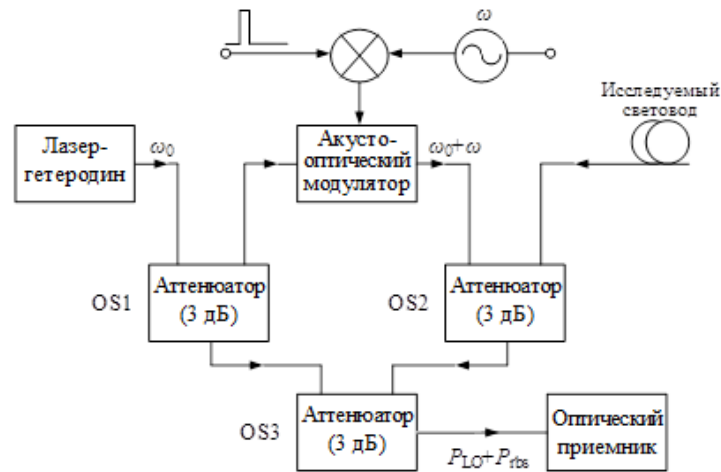


Рис. 5. Техническая реализация метода когерентного приема

При когерентном приеме доминирующими являются шумы оптического приемного устройства и гетеродина. Первые определяются темновым током диода i_d . Результирующий шум гетеродина других компонентов технической реализации для анализа отношения сигнал/шум целесообразно выразить через эквивалентную мощность P_{NEP} . При этих условиях соотношение сигнал/шум примет вид [9]:

$$SNR = \frac{2P_{LO}P_{rbs} \left[\frac{e\eta}{\eta} \right]}{2eB \left\{ i_d + \left[\frac{e\eta}{\eta w_0} \right] + P_{NEP} \left[\frac{e\eta}{\eta w_0} \right] \right\}}, \quad (2)$$

где e – заряд электрона, η – квантовая эффективность фотодетектора, P_{LO} – мощность сигнала гетеродина, P_{rbs} – мощность сигнала обратного рассеяния, B – полоса пропускания оптического приемника, P_{NEP} – эквивалентная мощность шума гетеродина.

Принимая во внимание, что мощность сигнала оптического гетеродина значительно превышает уровень собственного шума, последним можно пренебречь. При этих допущениях формула для отношения сигнал/шум примет следующий вид:

$$SNR = \frac{\eta P_{rbs}}{\eta w_0 B}. \quad (3)$$

Это отношение дает оценку квантовомеханического предела чувствительности когерентного оптического приемника.

Способность системы по пространственному разрешению зависит от длительности зондирующего импульса. Она, в свою очередь, определяется быстродействием акустооптического модулятора.

За счет гетеродинного детектирования чувствительность мониторинга увеличивается до квантового предела. Высокоточное разрешение при поиске неисправностей всегда полезно для технического обслуживания. Рекомендуются удобные функции, например, автоматическое определение места повреждения, быстрое определение места повреждения с меньшей погрешностью, но за более короткое время. Подводное оборудование должно иметь обратные пути, как описано в [3], для когерентного приема. При активном мониторинге оборудование мониторинга запрашивает и собирает непосредственно статус производительности по-

груженого оборудования. Соответствующими параметрами производительности являются входная мощность, выходная мощность и т.д.

5. Заключение

Предложенные принципы построения волоконно-оптических систем морского базирования для сбора и передачи данных могут быть востребованы в различных сферах жизнедеятельности человека, например, для гидроакустического мониторинга, датчиков сейсмоактивности, охраны периметра и т.д. Работоспособность системы проверена на действующем макете в ходе лабораторных испытаний на водном полигоне СКБ средств автоматизации морских исследований ДВО РАН (г. Южно-Сахалинск).

Выражение благодарности

Авторы выражают благодарность компании ООО «Предприятие «ЭЛТЕКС» за консультационную поддержку и содействие в изготовлении системы.

Литература

1. Бадеева Е. А., Целикин К. Д., Кабанов А. Г., Куклин Г. С. Волоконно-оптические системы военного назначения и их элементы // Модели, системы, сети в экономике, технике, природе и обществе. 2021. № 4. С. 106–120. DOI:10.21685/2227-8486-2021-4-10.
2. URL: <https://eltex-co.ru/catalog/management/eltex-ems/> (дата обращения: 09.04.2023).
3. URL: <https://eltex-co.ru/catalog/dsl/mx12e-2/> (дата обращения: 09.04.2023).
4. URL: https://pskovgeokabel.ru/articles/0_10/ (дата обращения: 10.04.2023).
5. Recommendation ITU-T G.979. Characteristics of monitoring systems for optical submarine cable systems. 2012.
6. IEC61746-1. Калибровка оптических рефлектометров с временной диаграммой направленности (OTDR). Часть 1: Рефлектометры для одномодовых волокон. 2009.
7. Recommendation ITU-T G.977. Characteristics of optically amplified optical fibre submarine cable systems. 2011.
8. Айбатов Д. Л., Морозов О. Г., Польский Ю. Е. Основы рефлектометрии: учебное пособие Казань: ЗАО «Новое знание», 2008. 116 с.

Деревяшкин Владимир Михайлович

к.т.н., доцент кафедры фотоники в телекоммуникациях, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), e-mail: dvm@sibgut.ru, ORCID ID: 0009-0006-3069-0905.

Виркунин Андрей Олегович

преподаватель кафедры фотоники в телекоммуникациях, Сибирский государственный университет телекоммуникаций и информатики, e-mail: stand.nsk@gmail.com, ORCID ID: 0009-0006-1695-664X.

Горлов Николай Ильич

д.т.н., профессор, профессор кафедры фотоники в телекоммуникациях, Сибирский государственный университет телекоммуникаций и информатики, e-mail: gorlovnik@yandex.ru, ORCID ID: 0000-0003-3614-8325.

Малашенко Анатолий Емельянович

к.т.н., ведущий научный сотрудник Федерального государственного бюджетного учреждения науки «Специальное конструкторское бюро средств автоматизации морских исследований Дальневосточного отделения Российской академии наук» (СКБ САМИ ДВО РАН, 693023, Южно-Сахалинск, ул. Горького, 25), e-mail: a.malashenko@skbsami.ru, ORCID ID: 0009-0000-3768-9196.

Сидоренко Артём Игоревич

младший научный сотрудник Федерального государственного бюджетного учреждения науки «Специальное конструкторское бюро средств автоматизации морских исследований Дальневосточного отделения Российской академии наук» (СКБ САМИ ДВО РАН, 693023, Южно-Сахалинск, ул. Горького, 25), e-mail: sidorenko.artem@inbox.ru, ORCID ID: 0000-0002-2094-7854.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Marine-based Fiber-optic Systems for Data Collection and Transmission

Vladimir M. Derevyashkin¹, Andrey O. Virkunin¹, Nikolay I. Gorlov¹,
Anatoliy E. Malashenko², Artem I. Sidorenko²

¹ Siberian State University of Telecommunications and Information Science (SibSUTIS)

² Special Design Bureau for Marine Research Automation Tools (SKB SAMI FEB RAS)

Abstract: The article presents the principles of sea-based fiber-optic system construction for data collection and transmission intended for two-way information exchange between underwater stations and a hardware and software complex with a controlled flow of information, control of equipment of bottom modules and an apparatus-software module. Methods of the main parameters control of underwater stations are proposed. The basic principles of monitoring the physical transmission medium are analyzed. At the same time, the necessity of using a coherent optical reflectometer is justified that makes it possible to improve the sensitivity of monitoring the quantum limit due to deep detection.

Keywords: fiber-optic system, bottom module, monitoring, coherent reflectometer.

For citation: Derevyashkin V. M., Virkunin A. O., Gorlov N. I., Malashenko A. E., Sidorenko A. I. Marine-based fiber-optic systems for data collection and transmission (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 48-58. <https://doi.org/10.55648/1998-6920-2024-18-1-48-58>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Derevyashkin V. M., Virkunin A. O.,
Gorlov N. I., Malashenko A. E.,
Sidorenko A. I., 2024

The article was submitted: 24.05.2023;
revised version: 04.07.2023;
accepted for publication 03.09.2023.

References

1. Badeeva E. A., Tselikin K. D., Kabanov A. G., Kuklin G. S. Volokonno-opticheskie sistemy voennogo naznacheniya i ikh elementy [Military fiber optic systems and their elements]. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve*, 2021, no. 4. pp. 106-120. doi:10.21685/2227-8486-2021-4-10.
2. Available at: <https://eltex-co.ru/catalog/management/eltex-ems/> (accessed: 04/09/2023).
3. Available at: <https://eltex-co.ru/catalog/dsl/mxl2e-2/> (accessed: 04/09/2023).
4. Available at: https://pskovgeokabel.ru/articles/0_10/ (accessed: 04/10/2023).
5. Recommendation ITU-T G.979, "Characteristics of monitoring systems for optical submarine cable systems", 2012.
6. IEC61746-1 (2009), *Kalibrovka opticheskikh reflektometrov s vremennoi diagrammoi napravlenosti (OTDR). Chast' 1: Reflektometry dlya odnomodovykh volokon* [Calibration of optical time-diffraction reflectometers (OTDR) - Part 1: OTDRs for single-mode fibers].
7. Recommendation ITU-T G.977 (2011), *Characteristics of optically amplified optical fiber sub-marine cable systems*.
8. Aibatov D. L., Morozov O. G., Polsky Yu. E. *Osnovy reflektometrii* [Fundamentals of reflectometry]. Kazan', ZAO «Novoe znanie», 2008. 116 p.

Vladimir M. Derevyashkin

Cand. of Sci. (Engineering), Associate Professor of the Department of Photonics in Telecommunications, Siberian State University of Telecommunications and Informatics (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), e-mail: dvm@sibgut.ru, ORCID ID: 0009-0006-3069-0905.

Andrey O. Virkunin

Lecturer, Department of Photonics in Telecommunications, Siberian State University of Telecommunications and Informatics (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), e-mail: stand.nsk@gmail.com, ORCID ID: 0009-0006-1695-664X.

Nikolay I. Gorlov

Dr. of Sci. (Engineering), Professor; Professor of the Department of Photonics in Telecommunications, Siberian State University of Telecommunications and Informatics (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), e-mail: gorlovnik@yandex.ru, ORCID ID: 0000-0003-3614-8325.

Anatoliy E. Malashenko

Cand. of Sci. (Engineering), Leading Researcher of the Federal State Budgetary Institution of Science "Special Design Bureau for Automation of Marine Research of the Far Eastern Branch of the Russian Academy of Sciences" (SKB SAMI FEB RAS, Russia, 693023, Yuzhno-Sakhalinsk, Gorkogo St. 25), e-mail: a.malashenko@skbsami.ru, ORCID ID: 0009-0000-3768-9196.

Artem I. Sidorenko

Junior researcher, Federal State Budgetary Institution of Science "Special Design Bureau for Marine Research Automation Tools of the Far Eastern Branch of the Russian Academy of Sciences" (SKB SAMI FEB RAS, Russia, 693023, Yuzhno-Sakhalinsk, Gorky St. 25), e-mail: sidorenko.artem@inbox.ru, ORCID ID: 0000-0002-2094-7854.

Использование формантных характеристик гласных и согласных звуков русского языка для аутентификации пользователей

Е. П. Белова

Уфимский университет науки и технологий (УУНиТ)

Аннотация: В статье предлагается метод аутентификации пользователей на основе формантных характеристик. Разработан биометрический образ, состоящий из формантных характеристик 46 звуков русского языка (6 гласных, 4 дифтонгов и 36 согласных). Предложена архитектура системы аутентификации по формантным характеристикам и частоте лидирующей форманты гласных и согласных звуков. Разработан модуль выделения формантных характеристик гласных и согласных звуков. Представлены структурная схема модуля выделения формантных характеристик гласных и согласных звуков и блок-схемы его подмодулей. Алгоритм «случайный лес», метод k-ближайших соседей, логистическая регрессия, многослойный перцептрон и метод опорных векторов объединены в группу алгоритмов. Приведена структурная схема обучения и тестирования моделей на основе группы алгоритмов. На основе группы алгоритмов обучены модели. Получены результаты эксперимента и проведён их анализ.

Ключевые слова: биометрическая аутентификация по голосу, биометрическая система аутентификации, биометрический образ пользователя, формантные характеристики, гласные звуки, дифтонги, согласные звуки.

Для цитирования: Белова Е. П. Использование формантных характеристик гласных и согласных звуков русского языка для аутентификации пользователей // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 59–69. <https://doi.org/10.55648/1998-6920-2024-18-1-59-69>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Белова Е. П., 2024

Статья поступила в редакцию 07.08.2023;
переработанный вариант – 29.09.2023;
принята к публикации 01.10.2023.

1. Введение

В последнее время популярность набирают биометрические методы аутентификации. В отличие от парольных методов аутентификации и токена, они не требуют запоминания или хранения. В основе каждого биометрического метода аутентификации находится биометрический образ личности, основанный на одном или нескольких биометрических характеристиках. Каждый биометрический образ уникален, что обеспечивает высокую степень безопасности информационной системы. В большинстве случаев исключается использование биометрических характеристик пользователя другими лицами.

Биометрическая аутентификация по голосу не требует дополнительного оборудования и достаточно проста для применения пользователями. Одним из перспективных методов аутентификации по голосу является аутентификация по формантным характеристикам звуков речи человека.

Эффективный алгоритм выделения формантных характеристик представлен в работе [1]. Использование формантного анализа для задач биометрической аутентификации исследуется в статьях [2, 3]. Применению формантных характеристик звуков речи посвящены [4–6].

Исследованию формантных характеристик гласных звуков посвящены работы [7–8].

Целью данной работы является исследование формантных характеристик гласных и согласных звуков для повышения эффективности работы биометрической системы аутентификации.

2. Биометрический образ личности, основанный на формантных характеристиках гласных и согласных звуков

Целью данной работы является исследование формантных характеристик гласных и согласных звуков для повышения эффективности работы биометрической системы аутентификации.

Исходя из общей теории фонетики [9–11] и акустической теории речевого производства [12], под формантой понимается частотный диапазон звуков, содержащий усиленный звуковой компонент.

Форманта обычно образуется из-за резонанса в дыхательных путях при произнесении звука.

Под формантными характеристиками подразумеваются параметры формант, определяющие их частоту, амплитуду и ширину.

Биометрический образ личности в данном исследовании составляют следующие формантные характеристики:

1. Частоты первых 7 формант – $F_1, F_2, F_3, F_4, F_5, F_6, F_7$.

2. Начало и конец частотного диапазона каждой из 7 формант – $F_{1s}, F_{1f}, F_{2s}, F_{2f}, F_{3s}, F_{3f}, F_{4s}, F_{4f}, F_{5s}, F_{5f}, F_{6s}, F_{6f}, F_{7s}$ и F_{7f} .

3. Частота лидирующей форманты – F_l .

Под частотой лидирующей форманты подразумевается частота, которая соответствует наибольшему значению амплитуды выброса в любой из фиксируемых на спектрограмме формант [8].

В данном исследовании используются формантные характеристики 46 звуков (6 гласных, 4 дифтонгов и 36 согласных звуков), то есть все звуки русского языка.

Стоит отметить, что согласные звуки имеют менее выраженные форманты. Поэтому они рассматриваются главным образом посредством их артикуляции.

На рис. 1 приведена спектрограмма звука «Э».

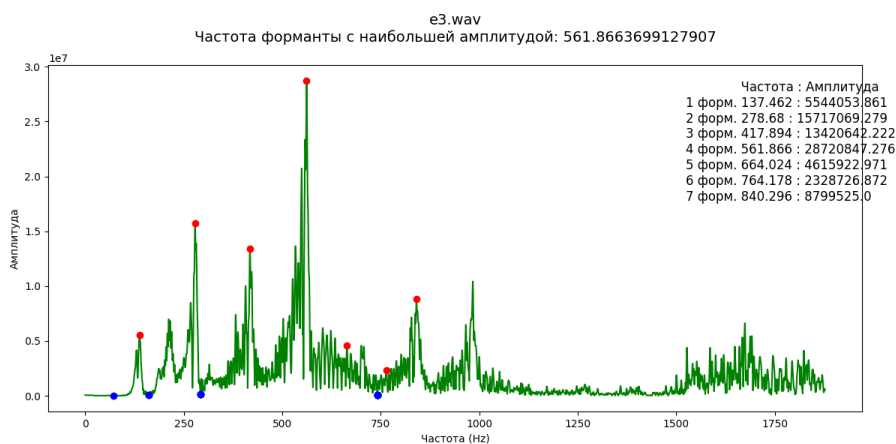


Рис. 1. Спектрограмма звука «Э»

На рис. 2 представлена спектрограмма звука «Т».

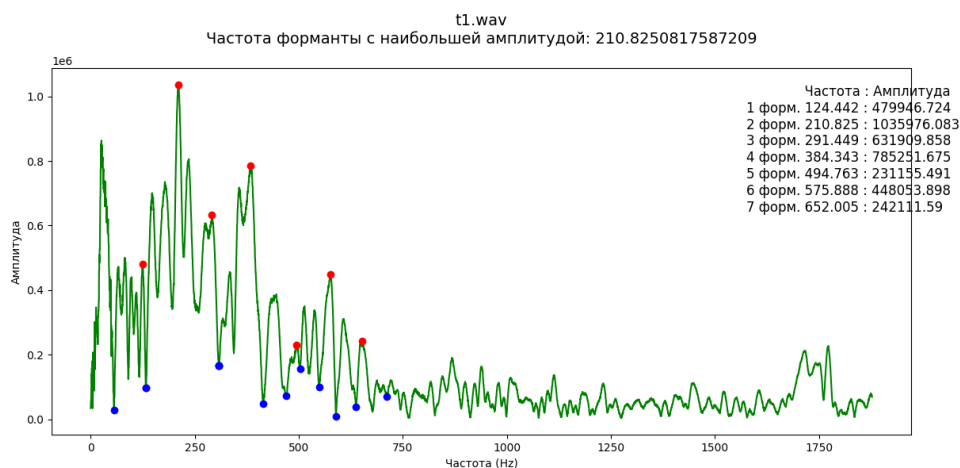


Рис. 2. Спектрограмма звука «Т»

Как видно из рисунков, спектрограмма звука «Т» имеет ярко выраженные форманты.

3. Разработка модуля выделения формантных характеристик гласных и согласных звуков

На рис. 3 представлена архитектура системы аутентификации по формантным характеристикам и частоте лидирующей форманты гласных и согласных звуков.

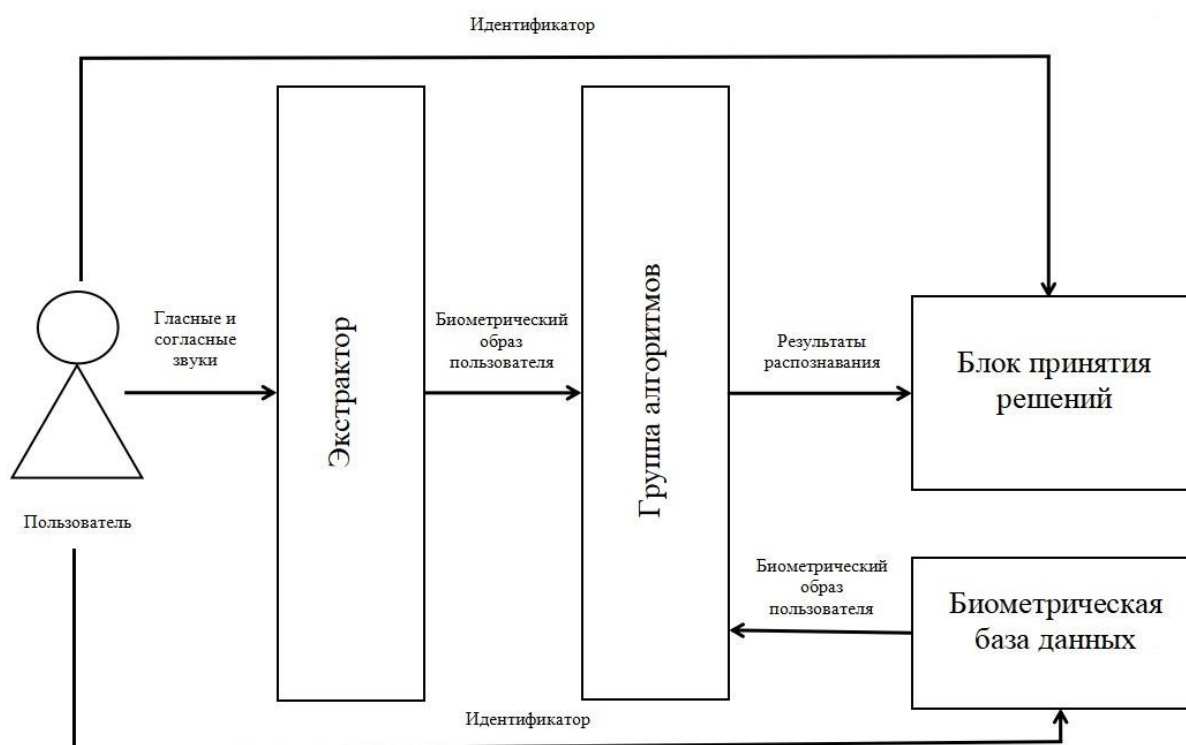


Рис. 3. Архитектура системы аутентификации по формантным характеристикам и частоте лидирующей форманты гласных и согласных звуков

Группа алгоритмов получает биометрический образ пользователя из экстрактора, также называемого модулем выделения формантных характеристик гласных и согласных звуков.

Структурная схема модуля выделения формантных характеристик гласных и согласных звуков приведена на рис. 4.



Рис. 4. Структурная схема модуля выделения формантных характеристик гласных и согласных звуков

Модуль выделения формантных характеристик гласных и согласных звуков состоит из следующих подмодулей:

- main – задаёт основное меню;
- recognition – строит спектрограммы, ищет форманты и определяет их характеристики;
- detect_peaks – непосредственно занимается поиском пиков.

Также модуль содержит папку Results, в которую записываются результаты.

Блок-схема подмодуля main представлена на рис. 5.

Подмодуль main создаёт окно программы, папку для записи результатов и кнопки меню. После того как происходит запись звука, подмодуль main запускает в работу подмодуль recognition.

Блок-схема подмодуля recognition представлена на рис. 6.

Подмодуль recognition запускает алгоритм поиска пиков через подмодуль detect_peaks и согласно полученным результатам строит спектрограммы.

На рис. 7 представлена блок-схема подмодуля detect_peaks.

Подмодуль detect_peaks осуществляет поиск пиков при помощи быстрого преобразования Фурье.

Выгрузка формантных характеристик гласных и согласных звуков в таблицу формата csv идёт последовательно. Порядок следования гласных и согласных звуков задаётся в подмодуле main.

Для обработки в приложении названия гласных и согласных звуков должны быть написаны латиницей. Латинские названия звуков представлены в столбце «Название папки».

В результате работы модуля выделения формантных характеристик гласных и согласных звуков получена выборка.

4. Проектирование группы алгоритмов

Группа алгоритмов состоит из:

- 1) алгоритма «случайный лес»;
- 2) метода k-ближайших соседей;
- 3) логистической регрессии;
- 4) многослойного перцептрона;
- 5) метода опорных векторов.



Рис. 5. Блок-схема подмодуля main

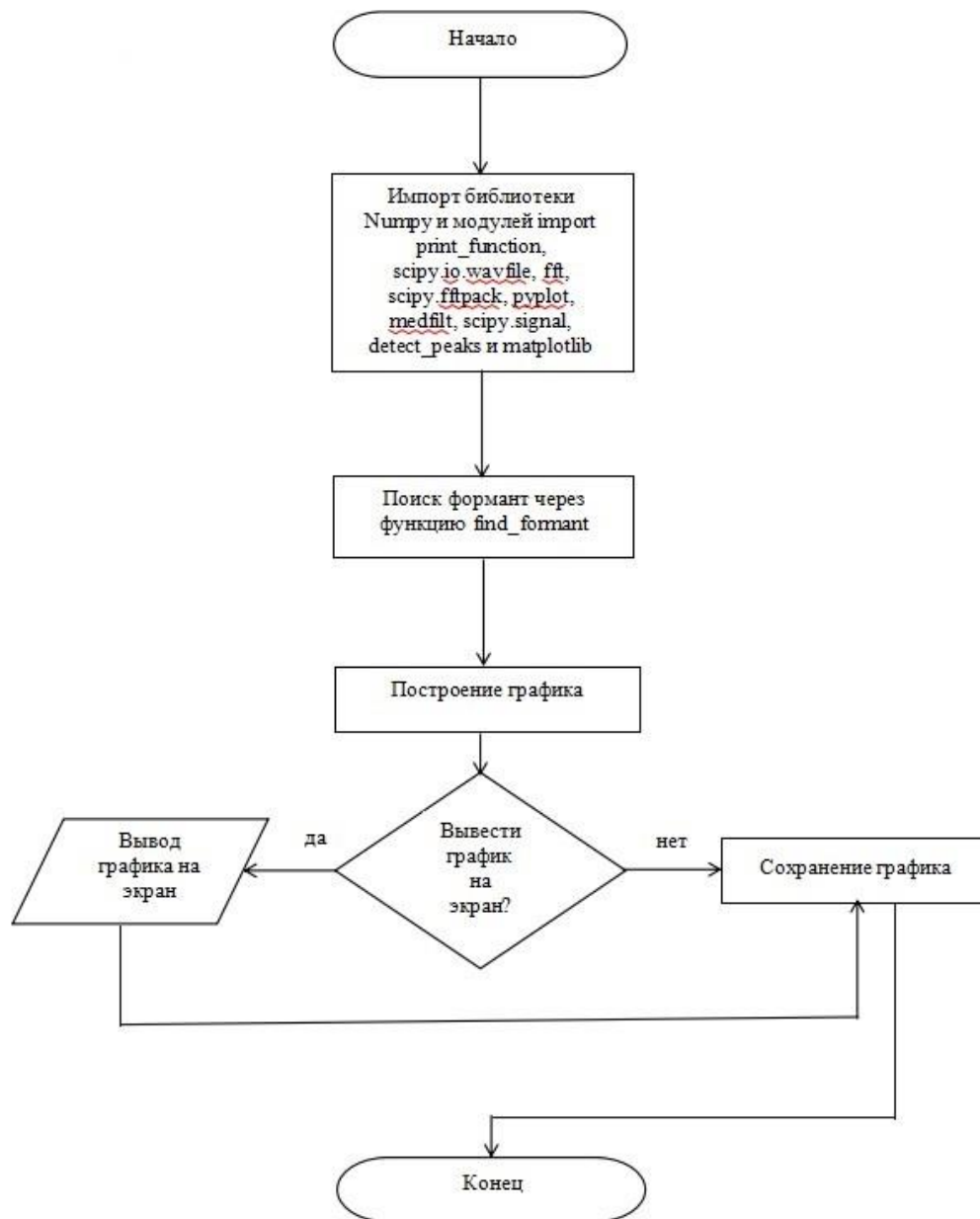


Рис. 6. Блок-схема подмодуля recognition



Рис. 7. Блок-схема подмодуля detect_peaks

На рис. 8 представлена структурная схема обучения и тестирования моделей на основе группы, состоящей из метода опорных векторов (SVM), логистической регрессии (Log Reg), метода k-ближайших соседей (KNN), алгоритма «случайный лес» (RF) и многослойного перцептрона (MLP).

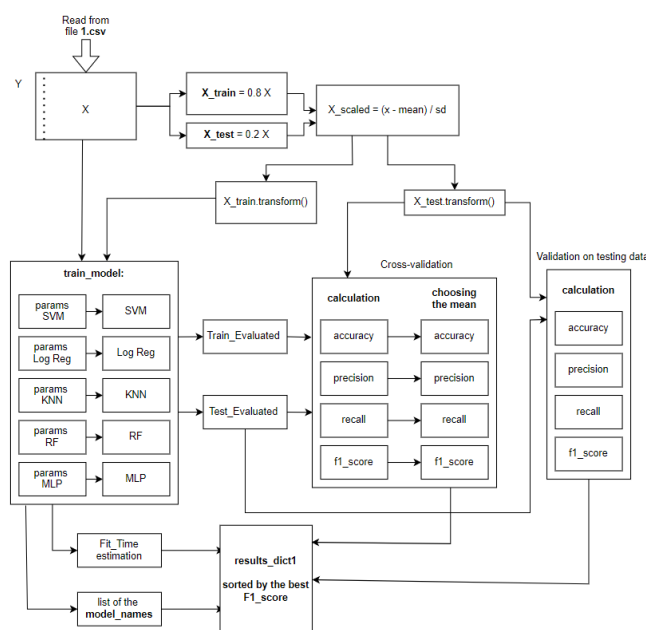


Рис. 8. Структурная схема обучения и тестирования моделей на основе группы алгоритмов

Проведение процедуры аутентификации при помощи группы алгоритмов состоит из следующих этапов:

1. На вход группы алгоритмов подаётся выборка, состоящая из формантных характеристик гласных (в том числе и дифтонгов) и согласных звуков.
2. Осуществляется разделение выборки на обучающую и тестовую в соотношении 80 % к 20 %.
3. Представленные данные проходят через процедуру усреднения.
4. Обучающая выборка служит для обучения моделей на основе метода опорных векторов (SVM), логистической регрессии (Log Reg), метода k-ближайших соседей (KNN), алгоритма «случайный лес» (RF) и многослойного перцептрона (MLP), для этого она поступает на входы соответствующих формируемых моделей.
5. Для каждого выбранного алгоритма во время обучения происходит расчёт метрик и выбор лучшей из них.
6. Формируется отчёт по обучению моделей на основе группы алгоритмов.
7. На вход обученных моделей поступает тестовая выборка.
8. Данные тестовой выборки при помощи метрик сравниваются с результатами обучения.
9. На основе полученных результатов формируется отчёт по тестированию моделей.

5. Анализ полученных результатов

Для анализа эффективности работы группы алгоритмов выбраны следующие параметры:

1. Среднее время подсчёта – это время, за которое свёрточная искусственная нейронная сеть решает 1 пример. Измеряется в секундах.
2. Среднее значение точности – это значение точности работы искусственной нейронной сети на обучающих и тестовых примерах.
3. Среднее значение точности на тестовой выборке.

В табл. 1 представлены результаты использования биометрического образа пользователя, основанного на формантных характеристиках гласных и согласных звуков.

Таблица 1. Результаты использования биометрического образа пользователя, основанного на формантных характеристиках гласных и согласных звуков

Model Name	Среднее время подсчёта	Среднее значение точности	Среднее значение точности на тестовой выборке
Алгоритм «случайный лес»	0.856373	0.871467	0.891304
Метод k-ближайших соседей	0.003199	0.704348	0.729348
Логическая регрессия	0.138321	0.533424	0.513043
Многослойный перцептрон	34.335736	0.797554	0.816304
Метод опорных векторов	0.149714	0.508152	0.548913

Большее количество времени занял подсчёт результатов у многослойного перцептрона – 34.335736 с. на решение одного примера.

Максимальное среднее значение точности (0.871467) и максимальное среднее значение точности на тестовой выборке (0.891304) получены с помощью алгоритма «случайный лес».

6. Заключение

Данное исследование посвящено изучению использования формантных характеристик гласных и согласных звуков русского языка для процедуры аутентификации по голосу. Всего использовано 46 звуков: 10 гласных (в том числе 4 дифтонга) и 36 согласных.

Для выделения формантных характеристик использовался модуль выделения формантных характеристик гласных и согласных звуков.

Процедура биометрической аутентификации проходила при помощи группы алгоритмов: алгоритма «случайный лес», метода k-ближайших соседей, логистической регрессии, многослойного перцептрона и метода опорных векторов. Каждый из этих алгоритмов имеет свои достоинства для решения задач по распознаванию пользователя.

Результаты, полученные при тестирования разработанных по данным алгоритмам моделей, находятся на оптимальном уровне.

Для повышения эффективности распознавания в следующих работах группа алгоритмов будет объединена в ансамбль. То есть все модели, обученные по алгоритмам, будут взаимосвязаны между собой и будут обучаться с учётом результатов друг друга. Планируется проведение работ по уменьшению ошибки 1-го рода, также будет проведён анализ и вычисление ошибки 2-го рода.

Литература

1. Оганесян А. Г. Эффективный алгоритм выделения формант из спектра речевого сигнала // Вычислительная техника и информатика. 2006. Т. LIX, № 1. С. 177–183.
2. Сысоев О. В., Глаздовский В. В., Курпатов В. Л. Исследование возможности применения формантных характеристик в задачах биометрической аутентификации // Сибирский информационно-коммуникационный журнал. 2017. № 1 (63). С. 93–97.
3. Шохин А. А., Седельников С. Ю. Методика формантного анализа для биометрической аутентификации пользователей // Информатика и вычислительная техника. 2018. № 4. С. 36–41.
4. Peterson G. E. and Barney H. L. Control methods used in a study of the vowels // The Journal of the Acoustical Society of America. 1952. V. 24 (2). P. 175–184.
5. Бабосов Е. Г., Авдеев А. В., Зуева Е. В. Исследование возможности использования формантных характеристик звуков речи для аутентификации пользователей // Медико-биологические и технические системы. 2015. № 3 (79). С. 172–178.
6. Никитин А. И., Поливодский А. А. Исследование возможности использования формантных характеристик гласных в процессе аутентификации // Новые информационные технологии. 2016. № 1. С. 37–43.
7. Машкина И. В., Белова Е. П. Разработка нейросетевой базы данных биометрических образов для системы аутентификации по голосу // Проблемы информационной безопасности. Компьютерные системы. 2019. № 2. С. 86–93.
8. Belova Ye. P., Mashkina I. V. Architecture and Work Algorithm of Biometric Authentication System by Voice // Automatic Control and Computer Sciences. 2021. V. 55, №. 8. P. 924–931. DOI: 10.3103/S014641162108006X.
9. Ladefoged P. and Johnson K. A course in phonetics. 6th ed. Wadsworth Cengage Learning, 2011.
10. Stevens K. N. Acoustic phonetics (Current studies in linguistics series). MIT Press, 1998.
11. Ohala J. J. Phonetic explanations for the development of tones // Language. 1972. V. 48 (2). P. 245–261.
12. Fant G. Acoustic theory of speech production. Mouton, 1960.

Белова Елена Петровна

старший преподаватель кафедры управления информационной безопасностью, Уфимский университет науки и технологий (УУНиТ), e-mail: yelenabelova92@gmail.com, ORCID ID: 0000-0001-5488-943X.

Автор прочитала и одобрила окончательный вариант рукописи.

Автор заявляет об отсутствии конфликта интересов.

Using Formant Characteristics of Russian Vowels and Consonants for User Authentication

Yelena P. Belova

Ufa University of Science and Technology (UUST)

Abstract: The article proposes a method for user authentication based on formant characteristics. A biometric image has been developed consisting of the formant characteristics of 46 sounds of the Russian language (6 vowels, 4 diphthongs and 36 consonants). An authentication system architecture based on formant characteristics and frequency of the leading formant of vowels and consonants is proposed. A Module for extracting the formant characteristics of vowels and consonants has been developed. A block diagram of the Module for extracting the formant characteristics of vowels and consonants and block diagrams of its submodules are presented. The Random Forest algorithm, k-nearest neighbors, logistic regression, multilayer perceptron and Support Vector Machine are combined into a cluster of algorithms. A block diagram of training and testing models based on a cluster of algorithms is given. Models were trained on the basis of algorithms clusters. The results of the experiment were obtained and their analysis was carried out.

Keywords: biometric voice authentication, biometric authentication system, biometric user image, formant characteristics, vowels, diphthongs, consonants.

For citation: Belova Ye. P. Using formant characteristics of Russian vowels and consonants for user authentication (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 59-69. <https://doi.org/10.55648/1998-6920-2024-18-1-59-69>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Belova Ye. P., 2024

The article was submitted: 07.08.2023;
revised version: 29.09.2023;
accepted for publication 01.10.2023.

References

1. Oganessian A. G. Effektivnyi algoritm vydeleniya formant iz spektra rechevogo signala [Efficient algorithm for extracting formants from the spectrum of a speech signal]. *Vychislitel'naya tekhnika i informatika*, Moscow, 2006, vol. LIX, no.1, pp. 177-183.
2. Sysoev O. V., Glauzdovskii V. V., Kurpatov V. L. Issledovanie vozmozhnosti primeneniya formantnykh kharakteristik v zadachakh biometricheskoi autentifikatsii [Study of the possibility of using formant characteristics in biometric authentication tasks] *Sibirskii informatsionno-kommunikatsionnyi zhurnal*, 2017, no. 1 (63), pp. 93-97.

3. Shokhin A. A., Sedel'nikov S. Yu. Metodika formantnogo analiza dlya biometricheskoi autentifikatsii pol'zovatelei [Methods of formant analysis for biometric user authentication] *Informatika i vychislitel'naya tekhnika*, 2018, no. 4, pp. 36-41.
4. Peterson G. E., Barney H. L. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 1952, no. 24(2), pp.175-184.
5. Babosov E. G., Avdeev A. V., Zueva E. V. Issledovanie vozmozhnosti ispol'zovaniya formantnykh kharakteristik zvukov rechi dlya autentifikatsii pol'zovatelei [Study of the possibility of using the formant characteristics of speech sounds for user authentication] *Mediko-biologicheskie i tekhnicheskie sistemy*, 2015, no. 3 (79), pp. 172-178.
6. Nikitin A. I., Polivodskii A. A. Issledovanie vozmozhnosti ispol'zovaniya formantnykh kharakteristik glasnykh v protsesse autentifikatsii [Study of the possibility of using the formant characteristics of vowels in the process of authentication]. *Novye informatsionnye tekhnologii*, 2016, no. 1, pp. 37-43.
7. Mashkina I. V., Belova E. P. Razrabotka neirosetevoi bazy dannykh biometricheskikh ob-razov dlya sistemy autentifikatsii po golosu [Development of a neural network database of biometric images for a voice authentication system]. *Problemy informatsionnoi bezopasno-sti. Komp'yuternye sistemy*, St. Petersburg, 2019, no. 2, pp. 86-93.
8. Belova Ye. P., Mashkina I. V. Architecture and Work Algorithm of Biometric Authentication System by Voice. *Automatic Control and Computer Sciences*, 2021, vol. 55, no. 8, pp. 924-931, DOI: 10.3103/S014641162108006X.
9. Ladefoged P., Johnson K. *A course in phonetics* (6th ed.), Wadsworth Cengage Learning, 2011.
10. Stevens K. N. *Acoustic phonetics* (Current studies in linguistics series), MIT Press, 1998.
11. Ohala J. J. Phonetic explanations for the development of tones. *Language*, 1972, no. 48(2), pp. 245-261.
12. Fant, G. *Acoustic theory of speech production*, Mouton, 1960.

Yelena P. Belova

Senior lecturer of the Department of Information Security Management, Ufa University of Science and Technology (UUST), e-mail: yelenabelova92@gmail.com, ORCID ID: 0000-0001-5488-943X.

Лингвистическое обеспечение информационных систем и процессов

И. В. Перцев, Е. И. Ситняковская

Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: Статья рассматривает процесс проектирования словарей данных, которые представляют собой наборы лексических единиц с определенной информацией о них. Описываются методы создания словарей индексирования и поиска информации, которые позволяют эффективно организовывать поиск по текстовым данным с использованием специализированных индексов.

Ключевые слова: лингвистическое обеспечение, информационные системы, словари, индексирование, поиск информации, тезаурусы, лексические комплексы, семантический анализ, синтаксический анализ, базы данных.

Для цитирования: Перцев И. В., Ситняковская Е. И. Лингвистическое обеспечение информационных систем и процессов // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 70–77. <https://doi.org/10.55648/1998-6920-2024-18-1-70-77>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Перцев И. В., Ситняковская Е. И.

Статья поступила в редакцию 15.12.2023;
принята к публикации 10.01.2024.

1. Введение

Лингвистическое обеспечение информационных систем играет важную роль в разработке и эффективном функционировании таких систем. Оно позволяет улучшить качество и эффективность работы с текстами, обеспечивая точный и быстрый поиск информации, а также удобное взаимодействие пользователей с системой.

Понятие «лингвистическое обеспечение» включает в себя различные аспекты и задачи, связанные с использованием языка и текстов в информационных системах. Оно заключается в применении лингвистических методов анализа, классификации и интерпретации текстов, а также в разработке соответствующих инструментальных средств, таких как словари данных, словари индексирования и поиска информации, тезаурусы и другие лексические комплексы [1].

Проектирование словарей индексирования и поиска данных является ключевым инструментом в обеспечении быстрого и точного поиска информации. Будут рассмотрены инвертированные списки терминов, деревья поиска и семантический анализ для создания эффективных словарей и индексов для поиска информации.

Тезаурусы и другие лексические комплексы используются для структурирования и классификации информации. Также будет исследовано применение методов семантического, синтаксического и прагматического анализа текстовой информации для представления данных в базах данных и организации интерфейсов информационных систем с пользователями [2].

Так как в современном информационном обществе огромное количество данных представлено в текстовой форме, эффективное управление и поиск этой информации являются важной задачей. Лингвистическое обеспечение позволяет осуществлять адекватный анализ текстов, обнаруживать и извлекать смысловую информацию, а также предоставлять удобный доступ к данным. Кроме того, лингвистическое обеспечение способствует улучшению качества информационных систем и процессов. Оно позволяет автоматизировать задачи, связанные с обработкой текстов, автоматически разбирать и классифицировать информацию, а также создавать удобные интерфейсы для взаимодействия пользователей с системой. Всё это способствует повышению эффективности работы с текстовой информацией и улучшению результатов в различных областях, таких как поиск информации, анализ текста, машинный перевод, автоматическая обработка естественного языка и другие.

Таким образом, лингвистическое обеспечение информационных систем играет важную роль в поиске, обработке и управлении текстовой информацией. Его использование способствует улучшению качества работы систем и повышению эффективности процессов.

2. Методы проектирования словарей

Методы проектирования словарей данных играют важную роль в обеспечении эффективной организации и управления данными в базах данных.

Первоначальный шаг в проектировании словарей данных – определение требований и понимание предметной области, с которой будет работать словарь данных. Это позволит определить основные термины и связи между ними [3]. Словарь данных может быть организован в виде иерархической структуры, где различные термины и их связи могут быть представлены в виде древовидной структуры. Также можно использовать другие методы организации, такие как сети или списки терминов с определениями и атрибутами. Для каждого термина в словаре данных необходимо предоставить четкое и однозначное определение. Определение должно быть достаточно подробным, чтобы пользователь мог понять значение и использование термина, а также связи с другими терминами. Важно вести документацию словаря данных, чтобы обеспечить ее актуальность и доступность для пользователей. Также можно предусмотреть механизмы обновления и внесения изменений в словарь данных при необходимости. Существуют различные средства, которые могут быть использованы для проектирования и поддержки словарей данных. Например, специализированные программы или инструменты баз данных могут предоставлять функциональность для определения и организации терминов, а также для отслеживания изменений.

Проектирование словарей включает определение требований, создание структуры и организацию терминов, описание терминов, документирование и поддержку словаря данных, а также использование средств автоматизации. Определение требований и анализ предметной области позволяют определить основные термины и связи между ними. Создание структуры и организация терминов способствуют логичной и понятной систематизации данных. Описание терминов обеспечивает единообразное понимание и использование терминологии в системе. Документирование и поддержка словаря данных помогают поддерживать актуальность и качество словаря. Использование средств автоматизации, таких как специальные программы или инструменты баз данных, повышает эффективность процесса проектирования словарей данных.

В целом методы проектирования словарей данных являются неотъемлемой составляющей обеспечения эффективной организации и управления данными в информационных системах. Качественно проведенное проектирование словарей данных позволяет улучшить эффективность работы с данными в информационных системах.

Примеры методов и средств проектирования словарей данных:

1. Создание табличных словарей. В этом методе каждый термин и его определение хранятся в отдельной записи таблицы с соответствующими полями. Такой подход удобен для поиска и управления данными.

2. Использование онтологий и тезаурусов. Онтологии и тезаурусы позволяют описать термины и их отношения между собой. Это помогает более полно и точно описать данные и взаимосвязи между ними [4].

3. Иерархические словари. Этот метод предполагает организацию терминов в иерархическую структуру, где каждый термин имеет родительский и дочерний элементы. Такая организация позволяет более удобно классифицировать данные и проводить поиск по иерархии.

4. Семантические словари. Семантический словарь предоставляет информацию о значениях и смысле терминов. Он описывает лингвистические и семантические свойства слова, его синтаксические части, семантические роли и другие связи. Семантические словари помогают улучшить поиск и анализ данных.

5. Использование графических средств. Графические средства, такие как схемы и диаграммы, могут использоваться для визуализации словарей данных. Это позволяет лучше представить связи и ассоциации между терминами и сущностями, а также облегчить понимание данных.

Хороший словарь данных должен иметь четкие и универсальные определения терминов, чтобы избежать путаницы и неоднозначности. Словари данных должны учитывать особенности предметной области и контекст, в котором будут использоваться данные. Должны быть определены специфические термины и понятия, которые относятся к данной области. Словари данных должны быть гибкими и способными к дальнейшему расширению, чтобы удовлетворить потребности в новых терминах и понятиях, должны быть согласованы и сопоставимы с другими используемыми словарями и схемами данных, чтобы обеспечить единообразное понимание и использование данных. Словари данных должны быть легкодоступными и удобными в использовании, чтобы пользователи могли быстро находить нужную информацию и применять ее при работе с данными.

Таким образом, методы и средства проектирования словарей данных предоставляют возможности для эффективного хранения, организации и поиска информации. Соблюдение принципов проектирования словарей данных позволяет создать удобные и надежные инструменты для работы с текстовыми данными.

3. Этапы работы системы лингвистического обеспечения

Система работы лингвистического обеспечения информационных систем и процессов включает несколько этапов:

1. Определение требований и анализ предметной области: на этом этапе осуществляется изучение конкретной предметной области и выявление основных терминов и связей между ними.

2. Проектирование словарей данных: создаются словари, которые содержат определения терминов, связи между ними и другую сопутствующую информацию.

3. Создание словарей индексирования и поиска информации: разрабатываются специализированные словари и индексы, которые позволяют эффективно организовывать поиск и индексацию текстовой информации.

4. Применение методов семантического, синтаксического и прагматического анализа: эти методы используются для обработки текстовой информации и ее представления в базах данных. Семантический анализ позволяет выявить смысловую информацию, синтаксический анализ занимается разбором грамматической структуры предложений, а прагматический анализ учитывает контекст и цели коммуникации.

5. Организация интерфейсов информационных систем с пользователями: разрабатываются удобные и понятные интерфейсы, обеспечивающие эффективное взаимодействие пользователя с информационной системой.

Таким образом, система работы лингвистического обеспечения информационных систем и процессов включает этапы определения требований и анализа предметной области, проектирования словарей данных, создания словарей индексирования и поиска, применения методов анализа текстовой информации, а также организации интерфейсов с пользователями [5]. Каждый этап имеет свою значимость и важность для эффективной работы информационной системы.

Словари индексирования и поиска информации являются важной составляющей информационных систем, так как они позволяют обеспечить быстрый и точный поиск данных. Они упрощают обработку и анализ больших объемов информации, позволяют организовать эффективное хранение данных и предоставляют возможность пользователям быстро находить нужную информацию. Словарь индексирования – это специальная структура данных, используемая для организации и упорядочивания информации для эффективного поиска. Он содержит отображение между ключевыми словами или терминами и связанными с ними документами или записями.

4. Средства индексирования и поиска информации

Приведём некоторые примеры методов и средств проектирования словарей индексирования и поиска информации.

Метод инвертированного списка терминов. Это один из наиболее популярных методов, используемых для построения словарей индексирования и поиска. Здесь каждый термин или ключевое слово связывается с соответствующими документами или записями. Это позволяет быстро находить необходимые данные по ключевым словам.

Деревья поиска. Это структуры данных, используемые для организации и упорядочивания данных. Они позволяют эффективно искать данные в больших объемах информации. Примерами деревьев поиска являются бинарные деревья поиска, или В-деревья.

Методы семантического анализа. Семантический анализ позволяет учитывать смысловую информацию при индексировании и поиске данных. Он помогает улучшить качество поиска и точность результатов поиска. Примеры методов семантического анализа включают индексирование по семантическим полям или использование онтологий.

Тезаурус является специальным словарем, который содержит систематизированное описание и организацию концептуальных связей между терминами или понятиями в определенной предметной области. Лексический комплекс, с другой стороны, представляет собой набор лексических данных, включающих разнообразные смыслы, значения, синтаксические свойства и другие характеристики слов или фраз.

Тезаурусы и лексические комплексы играют важную роль в информационных системах, они обеспечивают точность и полноту поиска информации. Введение тезаурусов и лексических комплексов позволяет учесть синонимы, близкие по значению термины и альтернативные формулировки, что помогает расширить охват поиска и улучшить точность результатов, обеспечить поддержку классификации и организации данных.

Тезаурусы и лексические комплексы могут использоваться для структурирования, классификации и организации информации в информационных системах, создавая иерархические или сетевые связи между связанными терминами. Тезаурусы и лексические комплексы помогают понимать и интерпретировать запросы пользователей, что позволяет более точно и полно находить необходимую информацию. К примеру, семантический анализ позволяет идентифицировать отношения между терминами и определить их семантические характеристики. Это могут быть связи типа «является», «часть – целое», «синонимы» и другие.

Методы семантического анализа могут включать использование лингвистических моделей, машинного обучения и статистических методов. Одним из методов проектирования тезаурусов является использование иерархических отношений, где термины организованы в структуру с родительскими и дочерними элементами. Это позволяет классифицировать термины по определенным критериям и создавать иерархическую структуру понятий. Также можно рассмотреть онтологическое моделирование. Онтологии представляют собой формальные описания понятий и связей в предметной области. Создание онтологий позволяет формализовать семантические отношения между терминами и обеспечить единообразное понимание. Применение онтологического моделирования включает использование RDF (расширяемый язык разметки ресурсов), OWL (язык описания онтологий) и других стандартов [6]. Семантические сети и графы могут быть применены для визуализации и организации тезаурусов и лексических комплексов. Они позволяют визуально представить связи между терминами и концептуальные отношения, упрощая восприятие и понимание данных.

Семантический анализ – это процесс извлечения и интерпретации смысловой информации из текстов. В контексте обработки текста семантический анализ означает понимание значения слов, синтаксиса и семантических отношений между ними.

Семантический анализ может быть использован для представления текстовой информации в базах данных. Он позволяет обработать и структурировать тексты таким образом, чтобы они стали доступными для поиска и извлечения нужной информации из базы данных. Семантический анализ обеспечивает более точное сопоставление запросов пользователей с содержанием текстовых данных в базе. Например, ключевые слова и извлечение сущностей. Методы извлечения ключевых слов и сущностей позволяют выделить наиболее значимые слова или фразы из текста, которые могут служить ключевыми понятиями для индексации и поиска данных [7].

5. Организация интерфейсов информационных систем

Организация интерфейсов информационных систем играет важную роль в обеспечении удобного и эффективного взаимодействия пользователя с системой [8]. Интерфейсы являются способом представления функциональности и данных системы, а также механизмом взаимодействия и обмена информацией с пользователем. Хорошо организованные интерфейсы делают работу с системой более удобной и понятной для пользователей, что приводит к повышению эффективности и производительности. Четко организованный и интуитивно понятный интерфейс снижает возможность ошибок пользователей и помогает предотвратить их возникновение, что повышает надежность всей системы. Хорошо спроектированный интерфейс позволяет пользователям быстро освоить функциональность системы и легко адаптироваться к ней, уменьшая время и усилия, необходимые для обучения. Грамотно организованный интерфейс учитывает потребности, навыки и предпочтения пользователей, что повышает удовлетворенность пользователей системой [9].

Предварительное исследование пользовательского опыта, проведение опросов и наблюдений помогает понять потребности и предпочтения пользователей, что является основой для организации интерфейса. Интерфейсы должны быть интуитивно понятными и легкими для использования, чтобы пользователи могли без лишних усилий выполнять необходимые задачи. В качестве методов можно использовать принципы дизайна, такие как простота, согласованность, видимость и гибкость. Существуют стандарты и руководства по проектированию интерфейсов, которые содержат рекомендации и принципы для создания удобных и эффективных интерфейсов, такие как стандарты ISO (например, ISO 9241-110:2018 "Human-centred design for interactive systems") и руководства по дизайну пользовательских интерфейсов [10].

Проведение тестирования интерфейса с участием пользователей позволяет выявить проблемы и улучшить его дальнейшими итерациями. Важно получать обратную связь пользователей и активно учитывать ее при переработке интерфейса.

6. Заключение

Дальнейшие исследования могут быть направлены на разработку новых методов и средств проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и лексических комплексов, чтобы обеспечить еще более эффективную и удобную работу с текстовой информацией. Важным направлением исследований является разработка улучшенных методов семантического анализа текстовой информации, которые позволят более точно и полно понимать смысл и контекст текстов и улучшить результаты поиска и анализа.

В совокупности лингвистическое обеспечение информационных систем играет важную роль в обработке и управлении текстовой информацией. Оно основано на методах и средствах проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и лексических комплексов, а также на методах семантического, синтаксического и прагматического анализа текстовой информации. Вместе они помогают повысить качество работы информационных систем и обеспечить удобство взаимодействия пользователей с системой. Важно также провести дальнейшие исследования в области эффективного использования лингвистического обеспечения в конкретных областях, таких как медицина, право, наука и т.д., чтобы разработать специализированные методы и инструменты для этих областей.

Литература

1. Aitchison J., Gilchrist A., Bawden D. *Thesaurus Construction and Use: A Practical Manual*. Aslib-IMI, 2000. 240 p.
2. Beale S. An open-source terminology and knowledge representation system: Ohsumed thesaurus design and implementation // *Computers and Biomedical Research*. 1996. V. 29, Is. 3. P. 193–223.
3. Hodge G. Systems of thesaurus compilation: a pilot study of thesaurus design and construction // *Journal of Documentation*. 2000. V. 56 (2). P. 145–168.
4. Shum S. B., Motta E., Domingue J. Ontology-based knowledge representation for aerospace design // *Journal of Knowledge Engineering and Knowledge Management*. 2001. V. 19 (5). P. 514–536.
5. Jurafsky D., Martin J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition)*. Pearson, 2023. 577 p.
6. ISO 25964-1:2011 *Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*.
7. Studer R., Benjamins V. R. Knowledge engineering: Principles and methods // *Data & Knowledge Engineering*. 1998. V. 25 (1–2). P. 161–197.
8. Shneiderman B., Plaisant C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 5th Ed. Pearson, 2020. 677 p.
9. Dix A., Finlay J., Abowd G., Beale R. *Human-Computer Interaction*. 3rd Ed. Prentice Hall, 2004. 52 p.
10. ISO 9241-110:2018 *Ergonomics of human-system interaction. Part 110: Interaction principles*.

Перцев Игорь Владимирович

к.т.н., доцент, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 272, e-mail: pertsev@sibguti.ru, ORCID ID: 0009-0005-2069-3937.

Ситняковская Елена Игоревна

к.т.н., доцент, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 272, e-mail: sitlen@sibguti.ru, ORCID ID: 0009-0000-0964-787X.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: *Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.*

Linguistic Support of Information Systems and Processes

Igor V. Pertsev, Elena I. Sitnyakovskaya

Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: The article examines the process of dictionary data design, which are sets of lexical units with specific information about them. The methods for creating indexing and search dictionaries are described, which enable efficient organization of search in textual data using specialized indexes.

Keywords: Linguistic support of information systems, design of data dictionaries, indexing and search dictionaries, thesauri and lexical complexes, semantic analysis of textual information, syntactic analysis of textual information, representation of textual information in databases.

For citation: Pertsev I. V., Sitnyakovskaya E. I. Linguistic support of information systems and processes (in Russian). Vestnik SibGUTI, 2024, vol. 18, no. 1, pp. 70-77. <https://doi.org/10.55648/1998-6920-2024-18-1-70-77>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Pertsev I. V. Sitnyakovskaya E. I., 2024

The article was submitted: 15.12.2023;
accepted for publication 10.01.2024.

References

1. Aitchison J., Gilchrist A., Bawden D. *Thesaurus Construction and Use: A Practical Manual*. Aslib-IMI, 2000. 240 p.
2. Shneiderman B., Plaisant C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition)*. Pearson, 2020. 677 p.
3. Beale S. An open-source terminology and knowledge representation system: Ohsumed thesaurus design and implementation. *Computers and Biomedical Research*, 1996, vol. 29, iss. 3, pp. 193-223.
4. Dix A., Finlay J., Abowd G., Beale R. *Human-Computer Interaction (3rd Edition)*. Prentice Hall, 2004. 52 p.
5. Hodge G. Systems of thesaurus compilation: a pilot study of thesaurus design and construction. *Journal of Documentation*, 2000, vol. 56(2), pp. 145-168.
6. ISO 25964-1:2011 Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval.
7. ISO 9241-110:2018 Ergonomics of human-system interaction. Part 110: Interaction principles.
8. Shum S. B., Motta E., Domingue J. Ontology-based knowledge representation for aerospace design. *Journal of Knowledge Engineering and Knowledge Management*, 2001, vol. 19(5), pp. 514-536.

9. Jurafsky D., Martin J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition)*. Pearson, 2023. 577 p.
10. Studer R., Benjamins V. R. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 1998. vol. 25(1-2), pp. 161-197.

Igor V. Pertsev

Cand. of Sci. (Engineering), Associate Professor; Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), phone: +7 383 2698 272, e-mail: pertsev@sibguti.ru, ORCID ID: 0009-0005-2069-3937.

Elena I. Sitnyakovskaya

Cand. of Sci. (Engineering), Associate Professor; Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), phone: +7 383 2698 272, e-mail: sitlen@sibguti.ru, ORCID ID: 0009-0000-0964-787X.

Вопросы нелинейной акустики в проектировании громкоговорителей

М. С. Шушнов, Т. В. Шушнова, А. М. Оноприенко, Н. О. Абросимова

Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: В статье рассмотрены вопросы нелинейной акустики. Составлена математическая модель нелинейных искажений акустического тракта применительно к работе громкоговорителя в поршневом режиме. Приведены и проанализированы результаты экспериментальных измерений. Даны рекомендации по снижению искажений в конструкциях высококачественных акустических систем.

Ключевые слова: звук, искажения, нелинейность, громкоговоритель, динамическая головка, эффект Доплера, мониторинг акустика, психоакустика.

Для цитирования: Шушнов М. С., Шушнова Т. В., Оноприенко А. М., Абросимова Н. О. Вопросы нелинейной акустики в проектировании громкоговорителей // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 78–90. <https://doi.org/10.55648/1998-6920-2024-18-1-78-90>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Шушнов М. С., Шушнова Т. В.,
Оноприенко А. М., Абросимова Н. О., 2024

Статья поступила в редакцию 10.09.2023;
принята к публикации 01.10.2023.

1. Введение

Для звукооператора, саундпродюсера при выполнении трудовых функций необходимо слышать неискаженную звуковую картину, с хорошо заметными недостатками фонограммы, а для систем массмедиа, к которым относятся телерадиовещание, концертная деятельность в учреждениях культуры и др., важнее передать звуковую информацию прежде всего в красивой, приятной на слух форме, чем достоверно. Несмотря на кажущееся различие в задачах звуковых решений, важно соотносить субъективные критерии оценки качества звуковоспроизведения, основанные на психоакустике, с электроакустическими свойствами громкоговорителей.

В подавляющем большинстве работ по акустике, в частности по громкоговорителям, считается, что акустическое давление и скорости частиц достаточно малы, чтобы можно было считать процессы излучения и распространения звуковых волн линейными. Если система или процесс линейны, то существует несколько правил, определяющих, что происходит с сигналами, когда они проходят через систему или процесс. Эти правила включают в себя принцип суперпозиции, который гласит, что реакция на сигнал $A+B$ равна реакции на сигнал A + реакция на сигнал B . Большинство инструментов и методов анализа, таких как анализ Фурье и функций частотной и фазовой характеристик, полностью полагаются на принцип суперпозиции и, следовательно, на линейность системы.

Когда система или процесс нелинейны, принцип суперпозиции больше не применим, и нельзя использовать обычные методы анализа. Рассмотрим причины и условия, при которых акустическое излучение и распространение могут стать нелинейными, а также некоторые

примеры степени нелинейного акустического поведения, встречающегося в громкоговорителях.

2. Нелинейная акустика

2.1. Нелинейное распространение звуковых волн

Скорость звука в воздухе зависит от термодинамических свойств воздуха. Акустическая волна состоит из чередующихся положительных и отрицательных давлений выше и ниже статического давления, и, поскольку это изоэнтропический процесс, соотношение между мгновенным давлением и плотностью становится все более нелинейным по мере увеличения уровня звукового давления (SPL). В линейной акустической теории связь между давлением и плотностью считается линейной, что справедливо, если изменения давления малы по сравнению со статическим давлением. Линейная зависимость между давлением и плотностью означает, что ни температура, ни скорость звука не меняются.

Звуковые колебания в воздухе приводят к его сжатию и разрежению. В областях сжатия давление воздуха возрастает, а в областях разрежения – понижается. Когда изменения давления значительны по сравнению со статическим давлением, нельзя игнорировать изменения мгновенной температуры и, следовательно, скорости звука. Кроме этого, когда акустическая волна существует в воздушном потоке, скорость распространения волны увеличивается в направлении потока и уменьшается в направлении против потока; создается эффект «конвекции» акустической волны вместе с потоком. Хотя постоянный поток воздуха обычно не встречается там, где работают громкоговорители, скорость частиц, связанную с распространением акустических волн, можно рассматривать как переменный нестационарный поток. Опять же, если скорости частиц малы по сравнению со скоростью звука, эффектом можно пренебречь, но в ситуациях, когда скорости частиц значительны по сравнению со скоростью звука, зависимость скорости распространения от скорости частицы должна быть принята во внимание. Результатом этого является то, что скорость распространения акустической волны увеличивается при увеличении давления и скорости частиц и уменьшается при уменьшении давления и скорости частиц.

Для плоской бегущей волны положительные давления сопровождаются положительными скоростями частиц, поэтому скорость распространения в положительном полупериоде акустической волны больше, чем в отрицательном. Затем положительный полупериод распространяется быстрее, чем отрицательный, и форма сигнала искажается по мере его распространения [1, 2].

Нелинейная трансформация акустической волны приводит к искажению формы волны, а первоначально синусоидальный профиль на некотором расстоянии становится близким к пилообразному. Искажение формы акустической волны приводит к появлению в ее спектре новых гармонических составляющих, то есть имеют место нелинейные искажения. Если в пространстве перекрываются две интенсивные волны, происходит их нелинейное взаимодействие, в частности, низкочастотная акустическая волна может модулировать высокочастотную.

2.2. Поршневой режим низкочастотных громкоговорителей

При уровнях звука, обычно возникающих при работе громкоговорителей, влияние давления и скорости частиц на мгновенную скорость звука настолько мало, что им можно пренебречь, и результирующее линейное приближение является достаточно точным. Однако бывают ситуации, когда это не так. Двумя распространенными примерами являются высокое звуковое давление в горловине рупорных громкоговорителей и высокие скорости диффузора низкочастотных динамиков с длинным ходом.

Использование небольших низкочастотных динамиков с длинным ходом диффузора в компактных мощных громкоговорителях также может привести к нелинейным искажениям. Выходная мощность диффузора громкоговорителя пропорциональна квадрату объемной скорости диффузора, поэтому для данной выходной звуковой мощности требуемая скорость диффузора пропорциональна обратной площади диффузора. Рассмотрим два громкоговорителя: один с диаметром диффузора 160 мм, другой – 25 мм. Чтобы излучать такое же количество акустической мощности на низких частотах, динамику меньшего размера требуется скорость, в 40 раз превышающая скорость большого динамика, поскольку он занимает 1/40 площади. Среднеквадратичная скорость большого динамика при излучении уровня звукового давления 104 дБ на расстоянии 1 м при частоте 100 Гц составляет примерно 1 м/с, в то время как этот же уровень звукового давления от меньшего динамика требует около 37 м/с. Величину 1 м/с можно считать незначительной по сравнению со скоростью звука (средней величиной прямо считать 340 м/с), а 37 м/с представляет собой вариацию скорости звука уже около 11 % [3], что должно быть принято во внимание.

2.3. Модель доплеровских искажений в электроакустической системе

Вторичным эффектом, являющимся прямым следствием значительных по сравнению со скоростью звука вариаций скоростей частиц, является так называемое доплеровское искажение. Если одновременно с излучением сигнала частотой 60 Гц небольшой динамик также излучал сигнал с частотой 1 кГц, циклическое приближение и опускание диффузора из-за низкочастотного сигнала при линейном ходе диффузора X_{LIN} , равном ± 6 мм, будет частотно модулировать (ЧМ) излучение более высокочастотного сигнала примерно на 35 Гц. С ростом излучаемой вторым громкоговорителем частоты этот эффект станет еще заметнее, и на частоте 3400 кГц составит уже около 410 Гц.

Для подтверждения теоретических утверждений приведем результаты моделирования смещения диффузора (CD – Cone Displacement) для трех динамических головок (ДГ) разных моделей в акустическом оформлении типа «закрытый ящик» (ЗЯ) (рис. 1) и фазоинверторного типа (ФИ) (рис. 2).

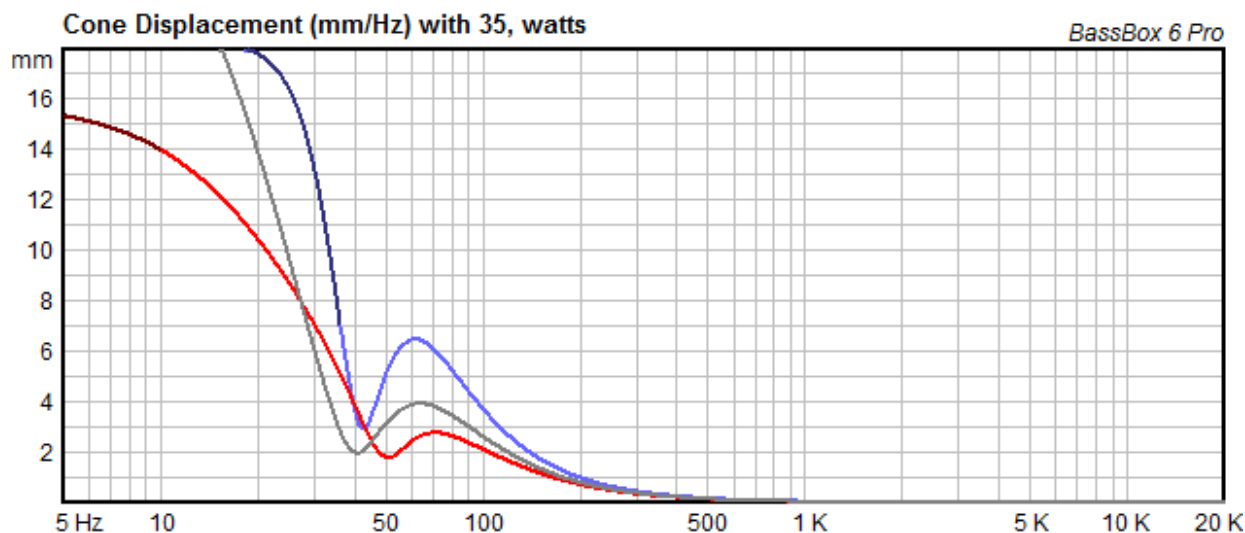


Рис. 1. Смещение диффузора ДГ для акустического оформления ЗЯ

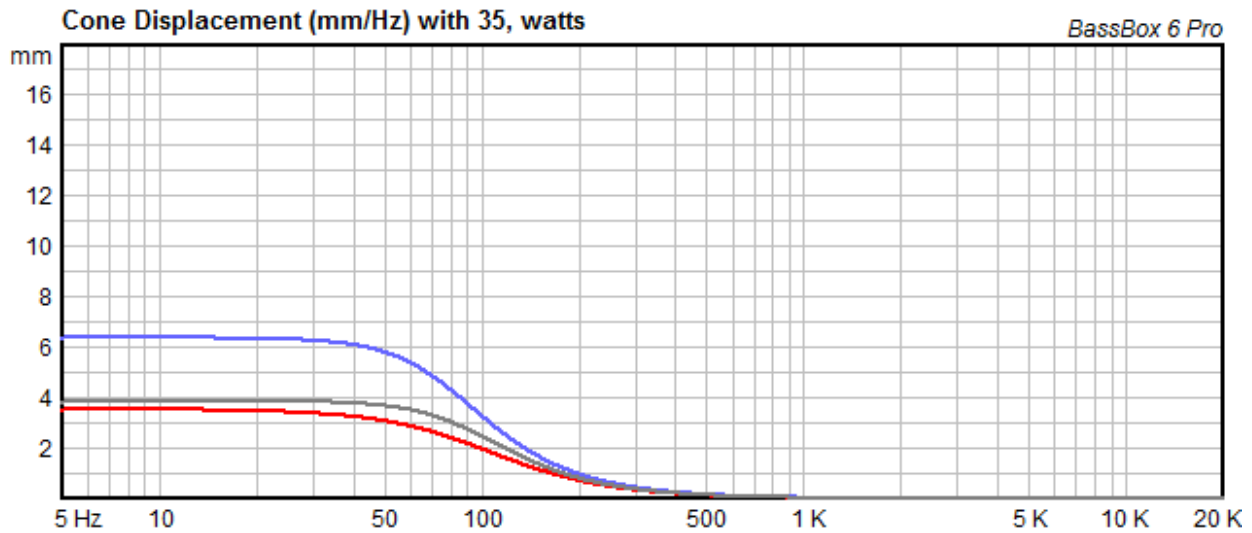


Рис. 2. Смещение диффузора ДГ для акустического оформления ФИ

Математически во временной области упрощенно можно описать процесс доплеровских искажений в следующем виде:

$$y(t) \sim x(t) + \sum_{i=0}^{\infty} K_D(f, X)(x(t))^i, \quad (1)$$

где $y(t)$ – акустическая волна, создаваемая громкоговорителем; $x(t)$ – смещение диффузора громкоговорителя; K_D – двумерная функция связи влияния доплеровского эффекта на акустическую волну, определяемая частотой колебаний волны f и амплитудой волны X .

Следует заметить, что выражение (1) носит теоретический характер и учитывает влияние нелинейностей всех порядков, при этом нелинейностями высоких порядков можно пренебречь ввиду незначительности их влияния.

Результат действия доплеровских искажений при работе громкоговорителя с двухтоновым сигналом с амплитудами U_{f1} и U_{f2} в пересчете на электрический тракт и при условии, что частоты сигналов соотносятся как $f_2 > f_1$, можно записать по аналогии с [4] в виде:

$$\begin{aligned} u(t) &= u_{f2}(t) \cos\left(2\pi \int_0^t f(\tau) d\tau\right) = u_{f2}(t) \cos\left(2\pi \int_0^t [f_2 + f_{\Delta} u_{f1}(\tau)] d\tau\right) = \\ &= u_{f2}(t) \cos\left(2\pi f_2 t + 2\pi f_{\Delta} \int_0^t \Delta u_{f1}(\tau) d\tau\right). \end{aligned} \quad (2)$$

В (2) $f_{\Delta} = K_D U_{f1}$, $f(\tau)$ – мгновенная частота, а f_{Δ} – наибольшее возможное смещение частоты f_2 в предположении, что u_{f1} принимает значения в интервале $[-1, +1]$ при симметричном линейном ходе диффузора ДГ X_{LIN} , что справедливо для большинства ДГ.

При приближении к границам хода диффузора ДГ X_{MAX} интервал u_{f1} может оказаться несимметричным относительно нуля, однако в этом режиме ДГ не может эксплуатироваться из-за риска разрушения подвижной системы ДГ, поэтому (2) справедливо для случаев безопасной эксплуатации ДГ.

Для нахождения спектра акустического сигнала с доплеровскими искажениями запишем его математическую модель (2) в более простом виде [5]:

$$u(t) = u_{f_2}(t) \cos(2\pi f_2 t - M_D \cos(2\pi f_1 t) + \varphi), \quad (3)$$

где $M_D = \left\lceil \frac{\Delta f_2}{f_1} \right\rceil$ – индекс доплеровской модуляции, а $\Delta f_2 \approx f_\Delta$ – максимальное изменение частоты f_2 . Для рассмотренного выше примера для соотношения частот $f_2 / f_1 > 10$ имеет место $M_D > 1$, то есть широкополосная частотная модуляция.

Для подтверждения обоснованности утверждения в табл. 1 приведены результаты расчета доплеровского смещения и индекса частотной модуляции.

Таблица 1. Результаты расчета доплеровского смещения частоты f_2 второго сигнала при воздействии первого сигнала с частотой f_1 , равной 60 Гц, при линейном ходе X_{LIN} , равном ± 6 мм

f_2 , Гц	λ_2 , м	λ_2^+ , м	λ_2^- , м	f_2^+ , Гц	f_2^- , Гц	f_Δ , Гц	M_D
100	3.400	3.406	3.394	99.82	100.18	0.35	0.00
250	1.360	1.366	1.354	248.90	251.11	2.21	0.02
400	0.850	0.856	0.844	397.20	402.84	5.65	0.05
600	0.567	0.573	0.561	593.71	606.42	12.71	0.11
1000	0.340	0.346	0.334	982.66	1017.96	35.31	0.29
2000	0.170	0.176	0.164	1931.82	2073.17	141.35	1.18
3400	0.100	0.106	0.094	3207.55	3617.02	409.47	3.41
5000	0.068	0.074	0.062	4594.59	5483.87	889.28	7.41
6400	0.053	0.059	0.047	5750.53	7214.85	1464.33	12.20
8000	0.043	0.049	0.037	7010.31	9315.07	2304.76	19.21

Хотя большая часть передаваемой энергии приходится на отрезок частот $[f_2 - f_\Delta, f_2 + f_\Delta]$, используя анализ Фурье, можно показать, что спектр ЧМ-сигнала в действительности неограничен, но амплитуда его компонент с большим отклонением частоты уменьшается, и этими компонентами обычно пренебрегают [3].

В общем виде частотный спектр (3) будет определяться функциями Бесселя, а его ширина будет тем больше, чем выше M_D . Запишем спектр (3) через гармонические составляющие:

$$u(t) = u_{f_2}(t) J_0(M_D) \cos(2\pi f_2 t + \varphi_0) + \sum_{i=1}^{\infty} u_{f_2}(t) J_i(M_D) \cos(2\pi(f_2 - if_1)t + \varphi_i), \quad (4)$$

где $J(M_D)$ – коэффициенты пропорциональности, зависящие от индекса доплеровской модуляции, определяемые по функциям Бесселя.

2.4. Имитационное моделирование доплеровских искажений

Выполним моделирование доплеровских искажений и оценим практические ширины спектров продуктов модуляции. Используем методы натурного и имитационного компьютерного моделирования по (4) для двухчастотного воздействия.

Для имитационного компьютерного моделирования использовалась программная среда Micro-CAP 12. Результаты моделирования спектра искажений в окрестности f_2 показаны на рис. 3–5.

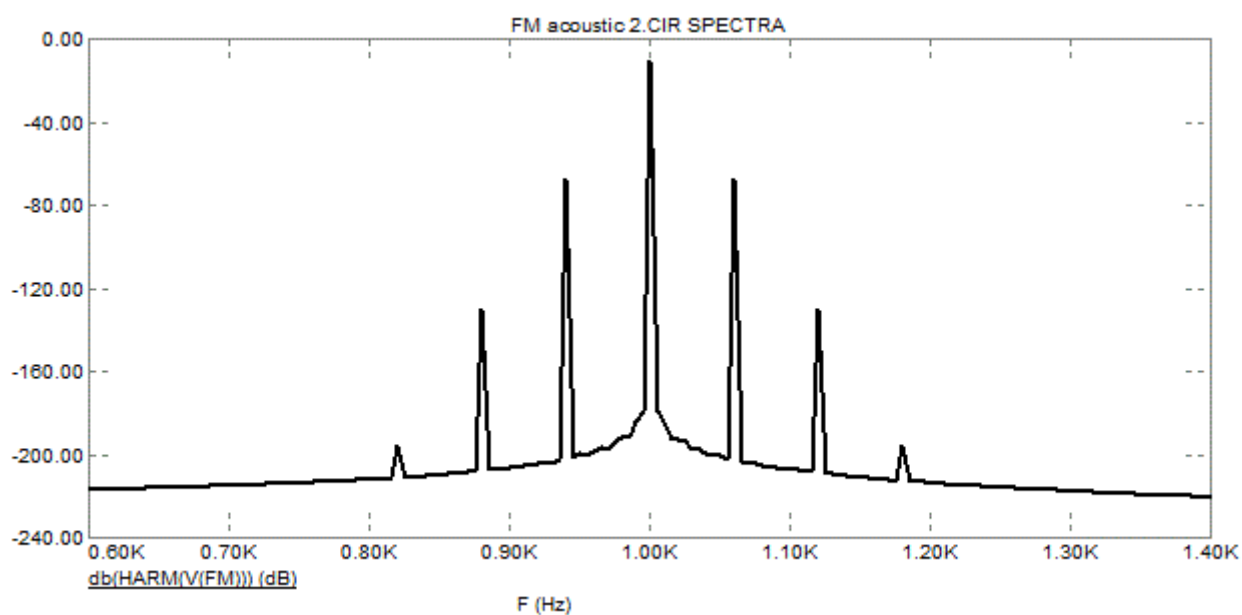


Рис. 3. Спектр акустического сигнала при воздействии частот 60 Гц и 1000 Гц при $M_D = 0.29$

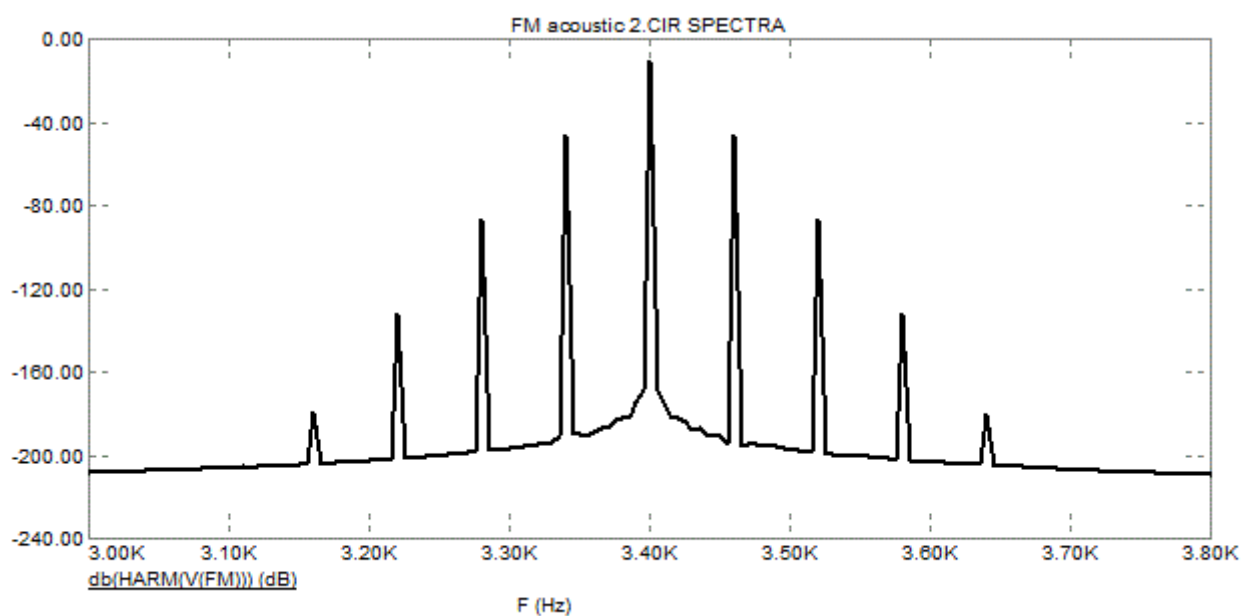


Рис. 4. Спектр акустического сигнала при воздействии частот 60 Гц и 3400 Гц при $M_D = 3.41$

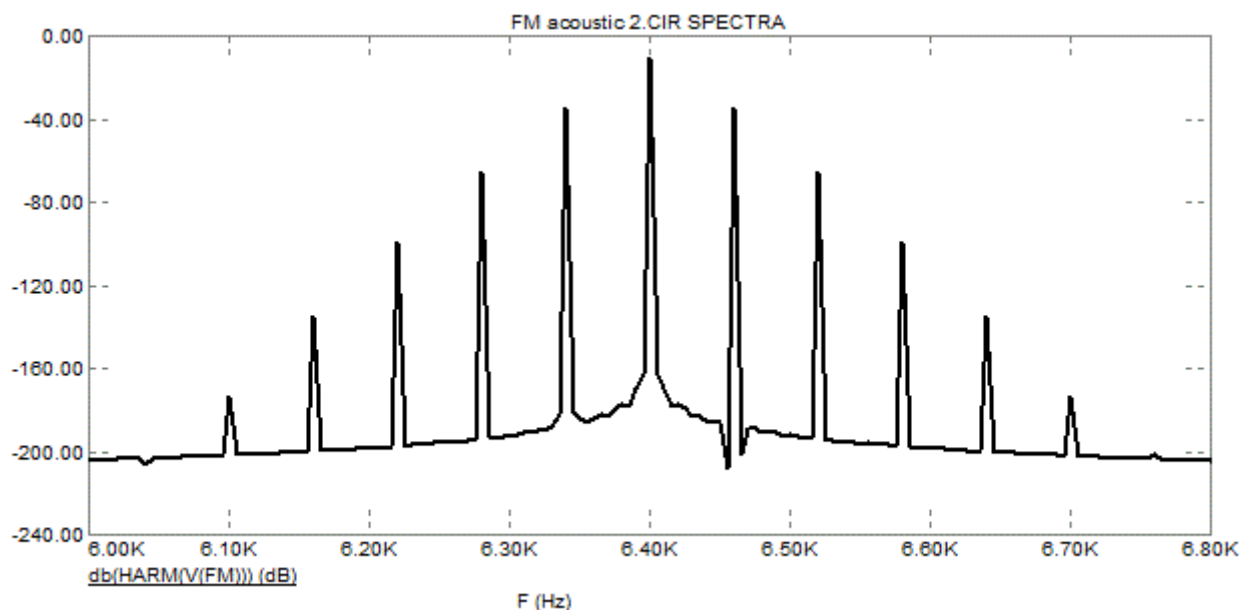


Рис. 5. Спектр акустического сигнала при воздействии частот 60 Гц и 6400 Гц при $M_D = 12.2$

Из рис. 3–5 видно расширение спектра из-за роста M_D вследствие повышения f_2 . Результаты моделирования на рис. 3–5 показывают хорошее соответствие с табл. 1.

2.5. Натурное моделирование доплеровских искажений

Выполним натурное моделирование доплеровских искажений и оценим практические ширины спектров продуктов модуляции. Используем метод натурального моделирования по (4) для двухчастотного воздействия.

Натурное моделирование осуществлялось по схеме на рис. 6. Использовался набор тестовых сигналов чистого тона с частотой f_1 , равной 16 Гц и 60 Гц, и частотой f_2 , равной 1000 Гц, 3400 Гц и 6400 Гц. Сигналы синтезированы и воспроизведены программным обеспечением (ПОС) Audacity 2.0.6 через встроенный плагин «Создание – Волна». Для ввода-вывода аудиосигнала использовался цифровой звуковой интерфейс (ЦЗИ) E-MU 1920m в режиме разрядности 24 бита и частотой дискретизации 44.1 кГц. Для усиления аудиосигнала по мощности применялся усилитель мощности звуковой частоты (УМЗЧ) с выходной мощностью 75 Вт на нагрузку 4 Ом. Снятие акустического сигнала осуществлялось при помощи измерительного микрофона-пушки (ИМ) Behringer ECM8000. Все соединительные линии были симметричными (балансное подключение). Для спектрального анализа записанного аудиосигнала использовалось бесплатное программное обеспечение (ПОА) RMAA v.5.5. В ходе натурального моделирования были получены результаты – спектры в окрестности частоты, приведенные на рис. 7–11.

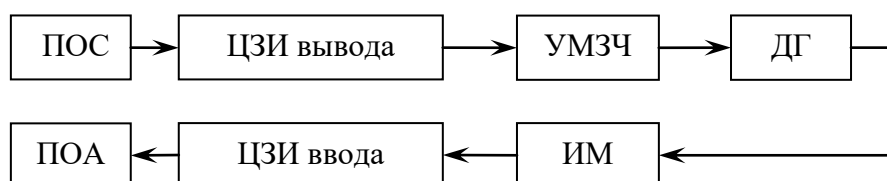


Рис. 6. Схема установки натурального моделирования

На рис. 7 показан спектр акустического сигнала при $f_1 = 60$ Гц и $f_2 = 1000$ Гц с соотношением амплитуд 2:1. Из рис. 7 видно, что спектр ЧМ узкий и подобен спектру продуктов интермодуляционных искажений низкого, преимущественно до 4-го, порядка.

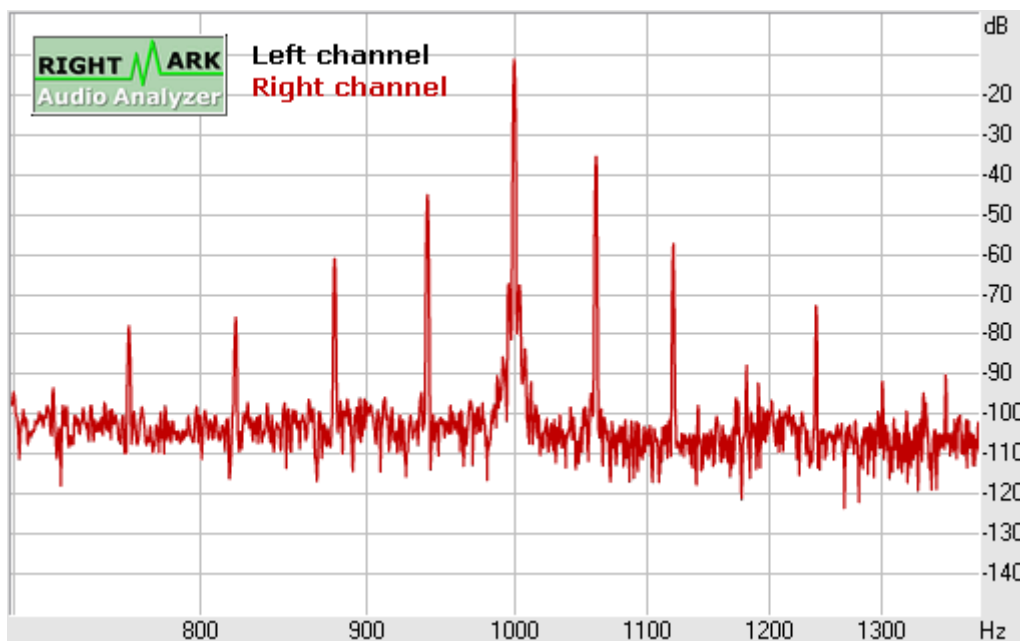


Рис. 7. Спектр акустического сигнала при воздействии частот 60 Гц и 1000 Гц с соотношением амплитуд 2:1

На рис. 8 показан спектр акустического сигнала при $f_1 = 60$ Гц и $f_2 = 3400$ Гц с соотношением амплитуд 2:1. Из рис. 8 видно, что спектр ЧМ стал шире с ростом частоты второго сигнала, присутствуют продукты интермодуляционных искажений низкого и высокого, преимущественно до 6-го, порядков.

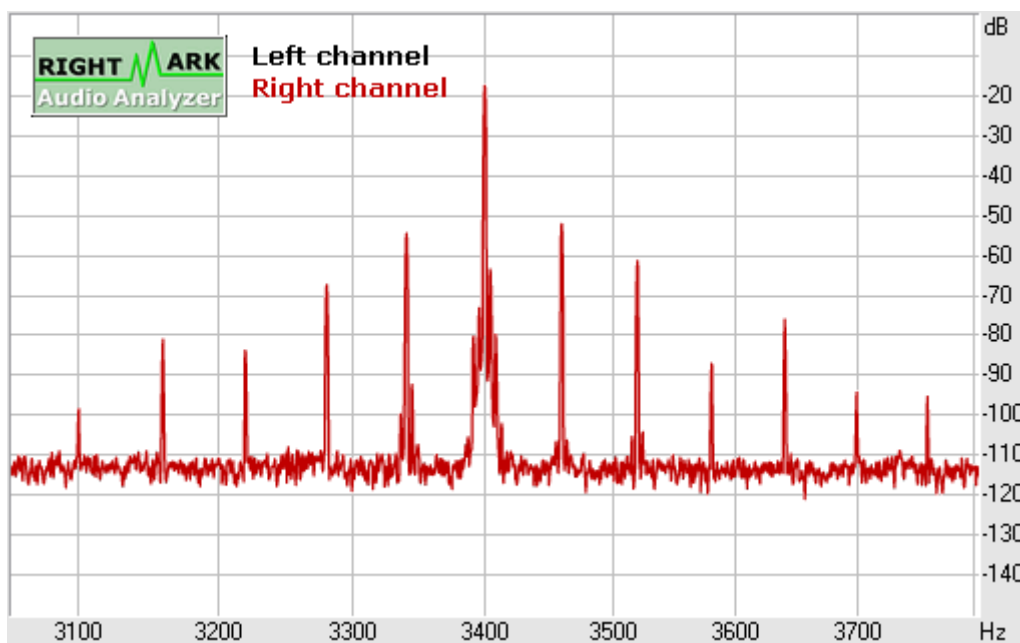


Рис. 8. Спектр акустического сигнала при воздействии частот 60 Гц и 3400 Гц с соотношением амплитуд 2:1

На рис. 9 показан спектр при $f_1 = 60$ Гц и $f_2 = 6400$ Гц с соотношением амплитуд 2:1. Из рис. 9 видно, что спектр ЧМ стал шире с ростом частоты второго сигнала, присутствуют продукты интермодуляционных искажений низкого и высокого порядков с высоким уровнем.

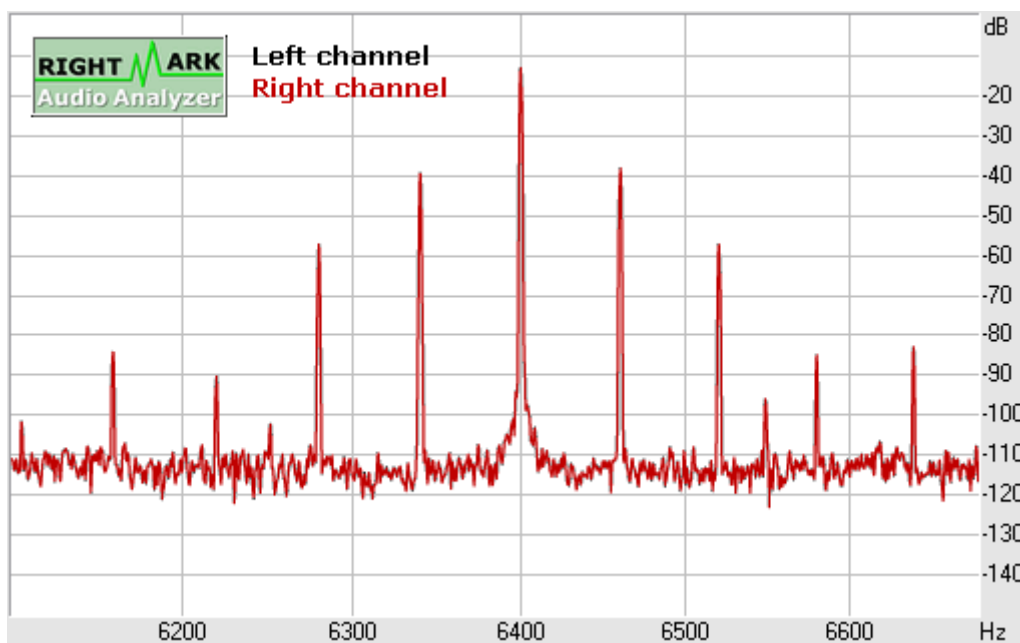


Рис. 9. Спектр акустического сигнала при воздействии частот 60 Гц и 6400 Гц с соотношением амплитуд 2:1

На рис. 10 показан спектр при $f_1 = 16$ Гц и $f_2 = 6400$ Гц с соотношением амплитуд 2:1. Из рис. 10 видно, что спектр ЧМ стал шире с понижением частоты первого сигнала, присутствуют продукты интермодуляционных искажений низкого и высокого, преимущественно до 8-го, порядков с очень высоким уровнем. Следует заметить, что частота 16 Гц не ощущается слушателем, но, несмотря на это, доплеровские искажения из-за ее воздействия имеются, причём значительные по величине.

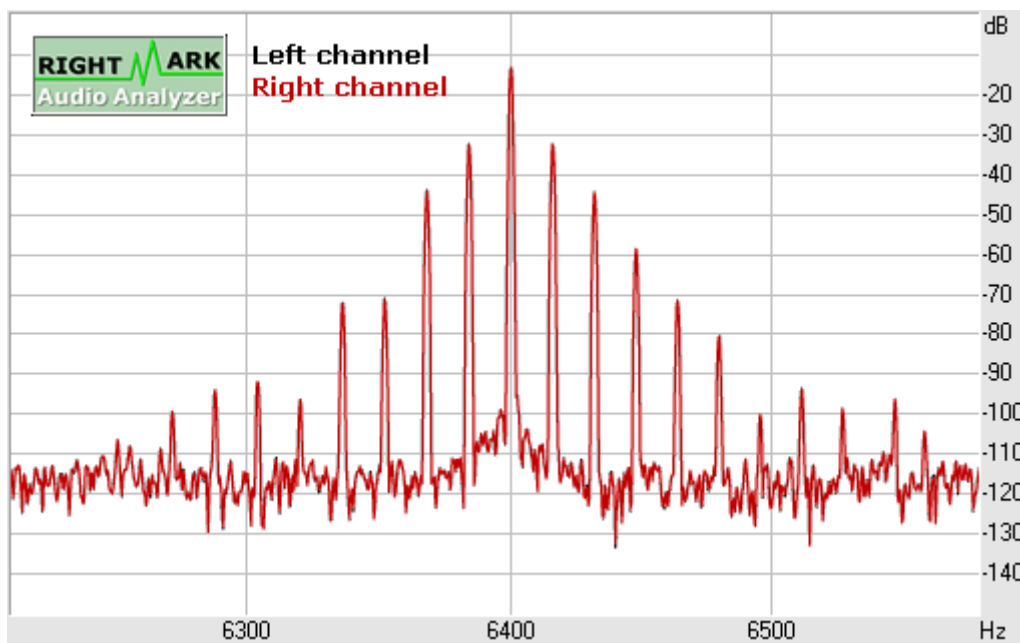


Рис. 10. Спектр акустического сигнала при воздействии частот 16 Гц и 6400 Гц с соотношением амплитуд 2:1

Повышение уровня искажений на рис. 7–10 связано с ростом индекса частотной модуляции M_D .

Аналогичные исследования были проведены с ДГ другой модели, допускающей значение $X_{LIN} = \pm 6$ мм. Результаты натурного моделирования показали тождественные результаты. Так, можно наблюдать практически полное соответствие спектра на рис. 11 спектру на рис. 8.

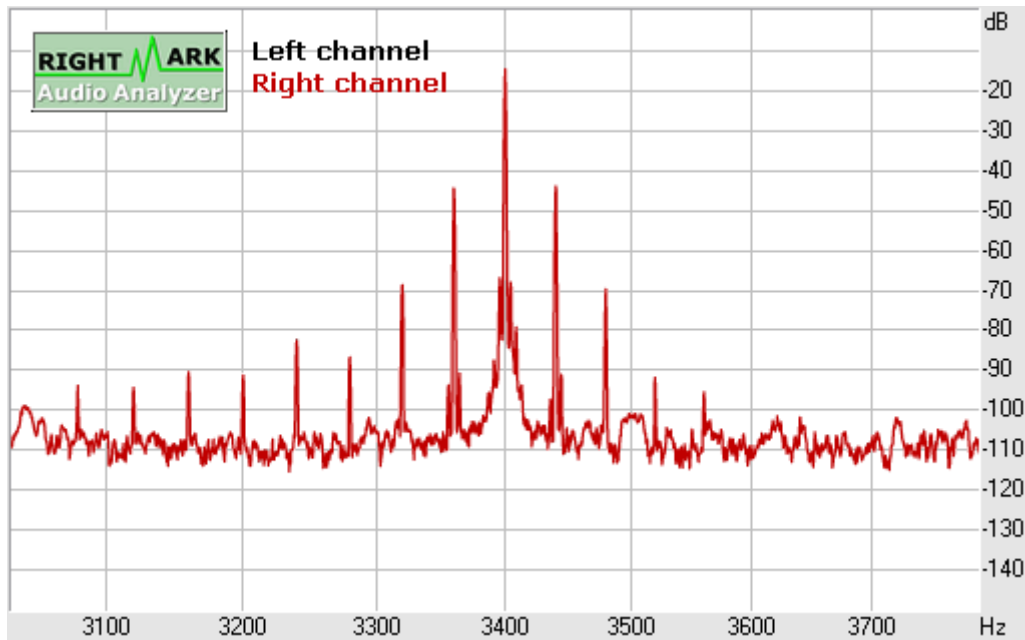


Рис. 11. Спектр акустического сигнала при воздействии частот 60 Гц и 3400 Гц с соотношением амплитуд 2:1 для ДГ другой модели

На рис. 11 показан спектр при $f_1 = 60$ Гц и $f_2 = 3400$ Гц с соотношением амплитуд 2:1. Из рис. 11 видно, что спектр ЧМ широкий, присутствуют продукты интермодуляционных искажений низкого и высокого, преимущественно до 8-го, порядков.

3. Анализ доплеровских искажений

Результаты расчетов и моделирования в достаточной степени соответствуют друг другу. Расхождения связаны со сложностью точной установки X_{LIN} в ходе натурного моделирования. Поскольку увеличение отношения составляющих частот приводит к росту ширины спектра интермодуляционных искажений акустического сигнала, следовательно, ширина спектра зависит от M_D . Так, для $M_D < 1$ ширина спектра сигнала соответствует ширине спектра амплитудно-модулированного сигнала, и в этом случае частотная модуляция является узкополосной. При увеличении M_D ширина спектра интермодуляционных искажений из-за доплеровского эффекта увеличивается, а ЧМ в этом случае становится широкополосной.

Как было показано ранее, индекс доплеровской модуляции при частоте 60 Гц для частот 1000 Гц и 3400 Гц может достигать 0.62 и 8.5 единиц соответственно. Поскольку как расширение спектра интермодуляционных искажений, так и увеличение амплитуд интермодуляционных частот определяет качество звучания громкоговорителя, то следует избегать $M_D > 1$ путем правильного подбора частот раздела в многополосном громкоговорителе.

Эффект доплеровского модуляционного воздействия U_{f2} на U_{f1} не проявляется из-за снижения подвижности диффузора динамика с ростом частоты, что видно из рис. 1, 2. Таким

образом, при необходимости получения высоких SPL доплеровский эффект динамиков с длинным ходом приводит к нелинейным искажениям акустической волны – доплеровскому эффекту. Следует учитывать это явление в проектировании громкоговорителей и акустических систем для высоких уровней SPL и при высоких требованиях к качеству звука.

Например, при работе в студиях музыкальной направленности, концертных залах при исполнении классических произведений требуемый проектный уровень SPL необходимо обеспечивать громкоговорителями с диффузорами большей площади (снижая подвижность диффузора и уменьшая тем самым K_D) или системами громкоговорителей (портальными акустическими системами), увеличивая площадь излучения акустической волны.

Низкочастотные ДГ с длинным ходом, которые должны использоваться при высоких уровнях звукового давления, не должны работать в среднечастотном и высокочастотном диапазонах, то есть частота раздела в многополосном громкоговорителе должна быть относительно низкой, если нужно избежать доплеровских искажений.

При невозможности увеличения площади диффузора динамической головки (ограниченные по массе размеры громкоговорителя или по габаритам места его установки) или при невозможности использования портальных систем (вес, занимаемая площадь) следует строить многополосные громкоговорители – с отработкой частотной области отдельной динамической головкой, отнесенной в пространстве. Частоту среза разделительного фильтра для НЧ-звена необходимо выбирать с учетом возможного появления ЧМ, избегая высоких значений M_D .

4. Заключение

При проектировании высококачественных громкоговорителей, предназначенных для студий звукозаписи, необходимо ограничивать диапазон воспроизводимых частот в области НЧ путем введения в схему УМЗЧ фильтра ВЧ с частотой среза ниже резонансной частоты громкоговорителя. Это следует выполнить, так как эффективность излучения звуковой волны ДГ на частотах ниже резонансной резко снижается, что видно по спаду АЧХ ДГ, но подвижность (смещение) диффузора остается постоянной, и его колебания будут вызывать доплеровский эффект (рис. 10).

Расширение рабочего диапазона частот ДГ путем введения в УМЗЧ цепей коррекции с подъемом в области НЧ (без ограничения нижней рабочей частоты ФВЧ) приводит к росту индекса ЧМ, так как этот метод вызывает увеличение подвижности (смещения) диффузора ДГ в области НЧ, в том числе ниже резонансной частоты ДГ, что приводит к ухудшению качества звуковоспроизведения.

Следует заметить, что в данном исследовании не были затронуты вопросы интермодуляционных и нелинейных искажений в электрическом тракте (УМЗЧ и ЗИ). Это не означает, что ими можно полностью пренебречь, но для современных усилительных устройств и устройств цифро-аналогового и аналого-цифрового преобразования величины интермодуляционных и нелинейных искажений существенно ниже и проявляются иначе, чем вызванные доплеровским эффектом.

Литература

1. Рыскин Н. М., Трубецков Д. И. Нелинейные волны: учеб. пособие для вузов по физ. спец. М.: Наука. Физматлит, 2000. 272 с.
2. Степанов Н. С. Волны / Большая российская энциклопедия. Т. 5. М., 2006. С. 645–649.
3. Newell P. R. et al. Loudspeakers: For Music Recording and Reproduction. Focal, 2007.
4. Faruque S. Radio Frequency Modulation Made Easy: Springer Cham, 2017. P. 33–37.
5. Thomas T. G., Sekhar S. C. Communication Theory. Tata-McGraw Hill 2005. P. 136.

Шушнов Максим Сергеевич

к.т.н., доцент, заведующий кафедрой цифрового телерадиовещания и систем радиосвязи, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, 86), e-mail: efemerian@gmail.com, ORCID ID: 0000-0002-1713-5177.

Шушнова Татьяна Владимировна

старший преподаватель кафедры цифрового телерадиовещания и систем радиосвязи, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, 86), e-mail: t.shushnova@gmail.com, ORCID ID: 0009-0006-9977-1818.

Онопrienко Анна Михайловна

старший преподаватель кафедры цифрового телерадиовещания и систем радиосвязи, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, 86), e-mail: anna_barabanova@mail.ru, ORCID ID: 0009-0005-9972-423X.

Абросимова Надежда Олеговна

аспирант 2-го года обучения, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, 86), e-mail: skachok94@mail.ru, ORCID ID: 0009-0000-3057-8215.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Non-linear Acoustic Problems in Loudspeaker Design

Maxim S. Shushnov, Tatiana V. Shushnova, Anna M. Onoprienko, Nadezhda O. Abrosimova

Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: The article deals with the issues of non-linear acoustics. A mathematical model of non-linear distortions of the acoustic path has been compiled in relation to the operation of a loudspeaker in a piston mode. The results of experimental measurements are presented and analyzed. Recommendations are given to reduce distortion in the construction of high-quality acoustic systems.

Keywords: sound, non-linear distortion, loudspeaker, woofer, audio system, monitor, optimization.

For citation: Shushnov M. S., Shushnova T. V., Onoprienko A. M., Abrosimova N. O. Voprosy nelineynoy akustiki v proyektirovanii gromkogovoriteley (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 78-90. <https://doi.org/10.55648/1998-6920-2024-18-1-78-90>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Shushnov M. S., Shushnova T. V.,
Onoprienko A. M., Abrosimova N. O., 2024

The article was submitted: 10.09.2023;
accepted for publication 01.10.2023.

References

1. Ryskin N. M., Trubetskov D. I. *Nelineynyye volny* [Nonlinear waves]. Moscow, Nauka, Fizmatlit, 2000. 272 p.
2. Stepanov N. S. *Volny* [Waves]. *Bol'shaya rossiiskaya entsiklopediya*, Moscow, 2006, vol. 5, pp. 645-649.
3. Newell P. R. et al. *Loudspeakers: For Music Recording and Reproduction*. Focal, 2007.
4. Faruque Saleh. *Radio Frequency Modulation Made Easy*. Springer Cham, 2017, pp. 33-37.
5. Thomas T. G., Sekhar S. C. *Communication Theory*. Tata-McGraw Hill, 2005, 136 p.

Maxim S. Shushnov

Cand. of Sci. (Engineering), Associate Professor, Head of the Department of Digital Television, Radio Broadcasting and Radio Communication Systems, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), e-mail: efemerian@gmail.com, ORCID ID: 0000-0002-1713-5177.

Tatiana V. Shushnova

Senior lecturer of the Department of Digital Television, Radio Broadcasting and Radio Communication Systems, Siberian State University of Telecommunications and Information Science, e-mail: t.shushnova@gmail.com, ORCID ID: 0009-0006-9977-1818.

Anna M. Onoprienko

Senior lecturer of the Department of Digital Television, Radio Broadcasting and Radio Communication Systems, Siberian State University of Telecommunications and Information Science, e-mail: anna_barabanova@mail.ru, ORCID ID: 0009-0006-9977-1818.

Nadezhda O. Abrosimova

Postgraduate student of the 2nd year of study at Siberian State University of Telecommunications and Information Science, e-mail: skachok94@mail.ru, ORCID ID: 0009-0000-3057-8215.

Оценка влияния процедуры пространственного сглаживания корреляционной матрицы на эффективность метода 2D-MUSIC *

Г. А. Фокин, И. В. Гришин

Санкт-Петербургский ун-в. телекоммуникаций им. проф. М. А. Бонч-Бруевича (СПбГУТ)

Аннотация: Возникновение корреляционных связей между радиосигналами, поступающими на антенную решётку базовой станции, снижает эффективность подпространственных методов позиционирования пользовательских устройств в беспроводных сетях связи. С целью декорреляции источников излучения может быть применён алгоритм пространственного сглаживания, заключающийся в разбиении антенной решётки базовой станции на подрешётки и последующем усреднении полученных для данных подрешёток матриц корреляции сигнала. В данной работе исследуется влияние различных способов разбиения плоской эквидистантной антенной решетки на подрешётки на рабочие характеристики метода множественной классификации сигналов 2D-MUSIC в сценарии сверхплотных сетей, работающих в диапазоне FR2.

Ключевые слова: алгоритм множественной классификации сигналов, азимут, угол места, пространственный спектр сигнала, пространственное сглаживание.

Для цитирования: Фокин Г. А., Гришин И. В. Оценка влияния процедуры пространственного сглаживания корреляционной матрицы на эффективность метода 2D-MUSIC // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 91–108. <https://doi.org/10.55648/1998-6920-2024-18-1-91-108>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Фокин Г. А., Гришин И. В., 2024

Статья поступила в редакцию 26.07.2023;
переработанный вариант – 01.10.2023;
принята к публикации 01.10.2023.

1. Введение

Внедрение сетей пятого поколения предъявляет ряд жестких требований к параметрам сетей, таким как высокая пропускная способность сети (до 20 Гбит/с), малое время задержки при передаче данных (возможность минимизации до 1 мс), высокая энергоэффективность и другие. Данный ряд требований не может быть достигнут в случае развертывания сетей связи, основанных на использовании макросот. Поэтому на уровне сети операторами связи осуществляется уплотнение сетей путём перехода на малые соты, радиус действия которых может быть ограничен несколькими десятками метров [1]. Такое решение обеспечивает плотное покрытие сетью, высокую пропускную способность и высокое качество обслуживания.

* Исследование выполнено при финансовой поддержке Российского научного фонда (грант № 22-29-00528), <https://rscf.ru/project/22-29-00528/>.

Увеличение скорости передачи в сетях связи пятого поколения также будет достигаться за счёт использования широких полос частот выше 6 ГГц, определенных в Регламенте радиосвязи МСЭ для IMT (диапазон FR2 24.25–52.60 ГГц).

Переход в диапазон частот 24.25–52.60 ГГц открывает возможности по применению антенных решёток с большим количеством элементов при малых размерах таких систем [2–4]. Увеличение числа элементов в антенных решётках позволяет формировать узкие диаграммы направленности, а возможность независимого управления фазой излучения антенных элементов позволяет оперативно изменять диаграмму направленности без внесения изменений в конструкцию антенн или их перемещения [5–7]. Процедуры управления лучом, направленным на пользовательское устройство, определяются стандартами NR и позволяют обеспечить прозрачную для пользовательского устройства мобильность для случая бесшовной передачи вызова между лучами разных базовых станций [8–10]. Для предотвращения потери пользовательским устройством (UE) узкого луча, формируемого антенной решёткой базовой станции, в процессе перемещения UE в пределах микросоты базовая станция (gNB) должна располагать актуальными данными о количестве и местоположении всех UE в пределах соты [11]. Таким образом, задача позиционирования пользовательских устройств в сверхплотных сетях связи является актуальной и востребованной. Для определения направления на источник сигнала применяются сверхразрешающие методы, такие как MVDR, ESPRIT, MUSIC (root-MUSIC) и другие.

В данной работе рассматривается двумерный метод сверхразрешения 2D-MUSIC. Целью работы является анализ влияния процедуры пространственного сглаживания на рабочие характеристики данного метода в условиях наличия корреляционных связей между поступающими на антенную решётку базовой станции сигналами.

Работа состоит из трёх частей: в первой части формируется математическая модель передачи данных с пользовательских устройств на базовую станцию в сверхплотных сетях, работающих в диапазоне частот 24.25–52.60 ГГц. Приводится описание метода множественной классификации сигналов для плоской эквидистантной антенной решётки, формулируется суть проблемы наличия корреляционных связей между сигналами.

Во второй части приводится описание алгоритма пространственного сглаживания, направленного на декорреляцию принимаемых сигналов.

В третьей части работы приводятся результаты вычислительных экспериментов с последующей их интерпретацией.

2. Математическая модель принимаемого сигнала

Пусть количество пользовательских устройств, находящихся в зоне действия базовой станции, равно K . Приёмопередающие антенны пользовательских устройств, а также базовой станции представлены плоскими эквидистантными антенными решётками (АР).

Для каждой антенны плоскость раскрыва совпадает с плоскостью xOy декартовой системы координат xuz , ось координат Oz является внешней нормалью к раскрыву.

Количество элементов для антенн пользовательских устройств и базовой станции определяется как $M = M_x M_y$ для UE и $N = N_x N_y$ для gNB соответственно. В целях упрощения последующих записей АР всех UE приняты идентичными. В данном случае M_x , M_y и N_x , N_y определяют количество элементов в горизонтальном (вдоль оси Ox) и вертикальном (вдоль оси Oy) рядах АР UE и gNB. Расстояние между антенными элементами передающих и приёмной антенн выбрано равным половине длины волны: $d = 0.5\lambda = 0.5cf_0^{-1}$, где величина $c \approx 3 \cdot 10^8$ м/с соответствует скорости распространения электромагнитной волны; f_0 – частота несущего колебания, Гц.

На рис. 1 представлена антенная решётка gNB. В данном случае принятая нумерация элементов решётки начинается с левого верхнего элемента и заканчивается на правом нижнем элементе решётки, принцип нумерации элементов сверху вниз и затем слева направо, что может быть описано выражением: $n = (n_y - 1)N_x + n_x$, $n = \overline{1, N}$, где n_x – номер горизонтального ряда, n_y – номер вертикального ряда, на пересечении которых расположен рассматриваемый элемент.

За начало системы декартовых координат принята точка центра масс антенной решетки базовой станции, нормаль апертуры AP gNB направлена вдоль оси Oz. Количество элементов в рядах может быть чётным или нечётным. Координаты крайних элементов ряда принимают значения $\pm 0.25\lambda(N_x - 1)$ и $\pm 0.25\lambda(N_y - 1)$. Система нумерации и расчётов координат элементов для антенных решёток пользовательских устройств построена по аналогичному принципу.

На основе сведений, приведённых выше, комплексные фазирующие векторы для AP UE и gNB в случае передачи сигнала от k -го источника сигнала к базовой станции могут быть описаны выражениями:

$$\mathbf{a}_{UE,k}(\theta_{UE,k}, \varphi_{UE,k}) = M^{-1/2} \exp(-j\mathbf{x}_{UE,k}^T \mathbf{b}_{UE,k}),$$

$$\mathbf{a}_{gNB,k}(\theta_{gNB,k}, \varphi_{gNB,k}) = N^{-1/2} \exp(-j\mathbf{x}_{gNB,k}^T \mathbf{b}_{gNB,k}),$$

где $\varphi_{UE,k}$, $\varphi_{gNB,k}$ – угол ориентации k -го луча в горизонтальной плоскости относительно нормали апертуры плоской AP пользовательского устройства или базовой станции (азимут); $\theta_{UE,k}$, $\theta_{gNB,k}$ – угол ориентации k -го луча в вертикальной плоскости относительно нормали апертуры плоской AP пользовательского устройства или базовой станции (угол места); $\mathbf{x}_{UE} = \|\mathbf{x}_{UE,1} \dots \mathbf{x}_{UE,M}\|$, $\mathbf{x}_{gNB} = \|\mathbf{x}_{gNB,1} \dots \mathbf{x}_{gNB,N}\|$ – матрица, состоящая из векторов декартовых координат элементов AP UE или gNB ($\mathbf{x}_{UE,m} = [x_{UE,m}, y_{UE,m}]^T$, $\mathbf{x}_{gNB,n} = [x_{gNB,n}, y_{gNB,n}]^T$, $m \in 1 \dots M$, $n \in 1 \dots N$); $\|\cdot\|$ – условное обозначение матриц, составленных не из числовых значений; $k = \overline{1, K}$ – порядковый номер UE, находящегося в радиусе действия gNB; $\mathbf{b}_{index} = \frac{2\pi}{\lambda} [\sin(\theta_{index}) \cos(\varphi_{index}), \sin(\theta_{index}) \sin(\varphi_{index})]^T$ – векторное волновое число, где $index$ соответствует UE или gNB.

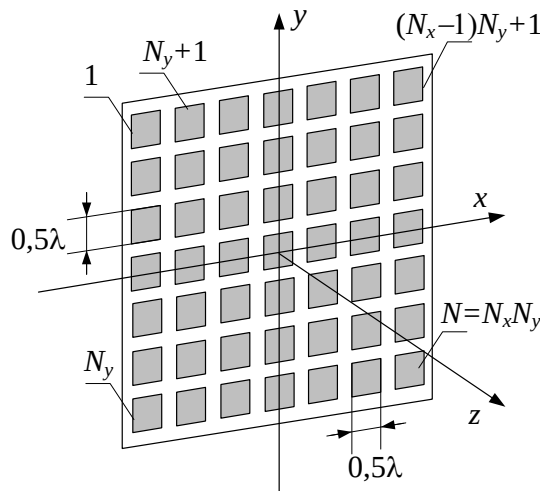


Рис. 1. Антенная решетка базовой станции

При передаче сигналов в диапазоне FR2 (24.25 ГГц – 52.60 ГГц) справедливо полагать, что сигналы являются узкополосными, и источники излучения, представленные UE, находятся в дальней зоне, т.е. $r_k \gg 2D^2\lambda^{-1}$, где D – размер апертуры.

При передаче i -го символа от k -го UE к gNB единичной амплитуды вектор комплексных амплитуд сигнала, принимаемого элементами AP gNB, имеет вид:

$$\begin{aligned} \dot{\mathbf{y}}_k^i &= \sum_{m=1}^{L_k} \dot{\mathbf{H}}_{k,m} \dot{\mathbf{s}}_k^i + \dot{\mathbf{w}}_k^i = \sqrt{NM} \sum_{l=1}^{L_k} \dot{g}_{k,l} \dot{\mathbf{a}}_{gNB,k,l} \dot{\mathbf{a}}_{UE,k,l}^H \dot{\mathbf{s}}_k^i + \dot{\mathbf{w}}_k^i = \\ &= \sqrt{NM} \dot{\mathbf{A}}_{gNB,k} \dot{\mathbf{G}}_k \dot{\mathbf{A}}_{UE,k}^H \dot{\mathbf{s}}_k^i + \dot{\mathbf{w}}_k^i = \dot{\mathbf{H}}_k \dot{\mathbf{s}}_k^i + \dot{\mathbf{w}}_k^i \end{aligned} \quad (1)$$

где $\dot{g}_{k,l} = \alpha_{k,l} \exp(j2\pi f_0 \tau_{k,l})$ – комплексный коэффициент отклика радиоканала для l -го пути распространения сигнала от UE_k ; $\dot{\mathbf{G}}_k = \sqrt{NM} \text{diag}\{\dot{g}_{k,l}\}$; $\text{diag}\{\}$ – диагональная матрица; $\dot{\mathbf{A}}_{UE,k}^H = \|\dot{\mathbf{a}}_{UE,k,1}^H, \dot{\mathbf{a}}_{UE,k,2}^H \dots \dot{\mathbf{a}}_{UE,k,L}^H\|^T$; $\dot{\mathbf{A}}_{gNB,k} = \|\dot{\mathbf{a}}_{gNB,k,1}, \dot{\mathbf{a}}_{gNB,k,2} \dots \dot{\mathbf{a}}_{gNB,k,L}\|^T$ – матрицы, состоящие из векторов АФР для AP UE и gNB; $\dot{\mathbf{w}}_k^i = [\dot{w}_{k,1}^i, \dot{w}_{k,2}^i \dots \dot{w}_{k,N}^i]^T$ – комплексный вектор собственных шумов приёмных устройств, рассматриваемый как аддитивный белый гауссовский шум (АБГШ).

Выражение (1) описывает принимаемый сигнал от UE_k к gNB как сумму компонент, поступающих с разных направлений и обладающих разными амплитудами и фазами. Отражение радиосигналов от многочисленных объектов приводит к появлению многолучевых не прямых трасс распространения сигнала наряду с трассой, находящейся в пределах зоны прямой видимости [12]. Матрица отклика радиоканала $\dot{\mathbf{H}}_{k,1}$ характеризует трассу прямой видимости от UE_k к gNB, матрицы $\dot{\mathbf{H}}_{k,m_k}$, $m_k = \overline{2, L_k}$ характеризуют многолучевые не прямые трассы.

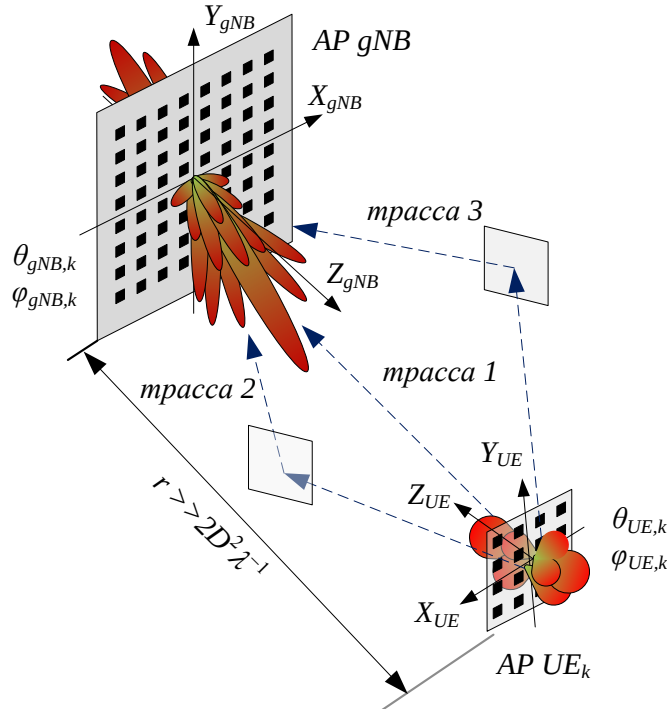


Рис. 2. Сценарий многолучевого распространения сигналов

Сигнал, поступающий на AP gNB от K пользовательских устройств в отсчётный момент времени i , описывается выражением:

$$\dot{\mathbf{y}}^i = \sum_{k=1}^K \dot{\mathbf{y}}_k^i = \|\dot{\mathbf{H}}_1, \dot{\mathbf{H}}_2 \dots \dot{\mathbf{H}}_K\| \|\dot{\mathbf{s}}_1^{iT}, \dot{\mathbf{s}}_2^{iT} \dots \dot{\mathbf{s}}_K^{iT}\|^T + \dot{\mathbf{w}}^i = \dot{\mathbf{H}}_\Sigma \dot{\mathbf{s}}^i + \dot{\mathbf{w}}^i. \quad (2)$$

С другой стороны, выражение (2) также может быть представлено в виде:

$$\dot{\mathbf{y}}^i = \dot{\mathbf{H}}_\Sigma \dot{\mathbf{s}}_\Sigma^i + \dot{\mathbf{w}}^i, \quad (3)$$

где $\dot{\mathbf{H}}_\Sigma \in \mathbb{C}^{N \times M}$ – некоторая результирующая матрица отклика канала; $\dot{\mathbf{s}}_\Sigma^i = \sum_{k=1}^K \dot{\mathbf{s}}_k^i$, $\dot{\mathbf{s}}_\Sigma^i \in \mathbb{C}^M$ – комплексный вектор суммы сигналов, передаваемых от K источников в момент времени i .

Параметры канала передачи $\dot{\mathbf{H}}_\Sigma$ на интервале времени $\overline{0\Delta t, I\Delta t}$ могут считаться постоянными, тогда для временного интервала $\overline{0\Delta t, I\Delta t}$ сигнал, поступающий на AP gNB, будет описываться выражением:

$$\dot{\mathbf{Y}} = \|\dot{\mathbf{y}}^0 \dots \dot{\mathbf{y}}^i \dots \dot{\mathbf{y}}^I\| = \dot{\mathbf{H}}_\Sigma \dot{\mathbf{S}} + \dot{\mathbf{W}}, \quad (4)$$

где $\dot{\mathbf{S}} = \|\dot{\mathbf{s}}^0 \dots \dot{\mathbf{s}}^i \dots \dot{\mathbf{s}}^I\|$ – матрица векторов комплексных амплитуд принимаемого сигнала AP gNB на всём интервале наблюдения; $\dot{\mathbf{W}} = \|\dot{\mathbf{w}}^0 \dots \dot{\mathbf{w}}^i \dots \dot{\mathbf{w}}^I\|$, $\dot{\mathbf{W}} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$; $\mathbf{I}_N$ – единичная матрица размерности N ; σ^2 – дисперсия шумовой составляющей.

Задачей базовой станции является обработка принимаемого сигнала $\dot{\mathbf{Y}}$, представляющего аддитивную смесь искаженных сигналов от K пользовательских устройств, результатам которой должна быть информация об оценках количества источников сигналов $\hat{K}$ и угловых координатах (азимуте и угле места $\hat{\theta}_{UE,k}, \hat{\phi}_{UE,k}$, $k = \overline{1, \hat{K}}$) данных источников, под которыми понимаются UE, для последующего формирования диаграммы направленности.

Одним из наиболее широко применяемых методов обнаружения координат UE в сетях мобильной связи является метод MUSIC, принцип работы которого базируется на ортогональности подпространств сигналов и шумов.

Корреляционная матрица принимаемого сигнала (4) $\dot{\mathbf{Q}}_y$ имеет вид:

$$\begin{aligned} \dot{\mathbf{Q}}_y &= E[\dot{\mathbf{Y}} \dot{\mathbf{Y}}^H] = \dot{\mathbf{H}}_\Sigma E[\dot{\mathbf{S}} \dot{\mathbf{S}}^H] \dot{\mathbf{H}}_\Sigma^H + E[\dot{\mathbf{W}} \dot{\mathbf{W}}^H] = \dot{\mathbf{H}}_\Sigma \dot{\mathbf{Q}}_s \dot{\mathbf{H}}_\Sigma^H + \sigma^2 \mathbf{I}_N = \\ &= \dot{\mathbf{U}} \Lambda \dot{\mathbf{U}}^H = \dot{\mathbf{U}}_s \Lambda_s \dot{\mathbf{U}}_s^H + \sigma^2 \dot{\mathbf{U}}_w \dot{\mathbf{U}}_w^H = \dot{\mathbf{U}}_s \tilde{\Lambda}_s \dot{\mathbf{U}}_s^H + \sigma^2 \mathbf{I}_N \end{aligned} \quad (5)$$

где $\dot{\mathbf{Q}}_s = E[\dot{\mathbf{S}} \dot{\mathbf{S}}^H]$ – корреляционная матрица передаваемых сигналов; $\dot{\mathbf{U}} \Lambda \dot{\mathbf{U}}^H$ – спектральное разложение корреляционной матрицы $\dot{\mathbf{Q}}_y$; $\dot{\mathbf{U}} = \|\dot{\mathbf{u}}_1, \dot{\mathbf{u}}_2, \dots, \dot{\mathbf{u}}_N\|$ – матрица собственных векторов $\dot{\mathbf{Q}}_y$; $\dot{\mathbf{U}}_s = \|\dot{\mathbf{u}}_1, \dot{\mathbf{u}}_2, \dots, \dot{\mathbf{u}}_{\hat{K}}\|$, $\dot{\mathbf{U}}_w = \|\dot{\mathbf{u}}_{\hat{K}+1}, \dot{\mathbf{u}}_2, \dots, \dot{\mathbf{u}}_N\|$ являются матрицами собственных векторов сигнального и шумового подпространств; $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_N\}$ – диагональная матрица собственных значений матрицы $\dot{\mathbf{Q}}_y$; $\Lambda_s = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_{\hat{K}}\}$ – диагональная матрица собственных значений $\dot{\mathbf{Q}}_y$, принадлежащая подпространству сигналов; $\lambda_{s,k} + \sigma^2$, $k = \overline{1, \hat{K}}$.

Определение размерности подпространства сигналов $\hat{K}$ может быть осуществлено одним из известных методов [13, 14]. В частности, довольно широкое применение нашёл критерий Акаике:

$$AIC(\hat{K}) = -I \cdot (N - \hat{K}) \ln \left[\frac{\prod_{k=\hat{K}+1}^N \lambda_k^{(N-\hat{K})^{-1}}}{(N - \hat{K})^{-1} \sum_{k=\hat{K}+1}^{N_R} \lambda_k} \right] + \hat{K} \cdot (2N - \hat{K}). \quad (6)$$

В силу ортогональности векторов отклика AP gNB и собственных векторов шумового подпространства координаты UE могут быть найдены через функцию пространственного спектра вида:

$$f(\hat{\theta}_{UE}, \hat{\phi}_{UE}) = \frac{\dot{\mathbf{a}}_{UE}^H(\hat{\theta}_{UE}, \hat{\phi}_{UE}) \dot{\mathbf{a}}_{UE}(\hat{\theta}_{UE}, \hat{\phi}_{UE})}{\dot{\mathbf{a}}_{UE}^H(\hat{\theta}_{UE}, \hat{\phi}_{UE}) \dot{\mathbf{U}}_w \dot{\mathbf{U}}_w^H \dot{\mathbf{a}}_{UE}(\hat{\theta}_{UE}, \hat{\phi}_{UE})}. \quad (7)$$

В силу ортогональности векторов амплитудно-фазового распределения и сигнального подпространства для оценок угловые координаты $\hat{\theta}_{UE,k}$, $\hat{\phi}_{UE,k}$, близкие по значению к координатам искомым UE, будут формировать локальные максимумы функции пространственного спектра $f(\hat{\theta}_{UE}, \hat{\phi}_{UE})$, что позволяет определить направление на источник сигнала.

Метод 2D-MUSIC применим для антенных решеток с различной геометрией и показывает высокие рабочие характеристики в случае некогерентных и некоррелированных сигналов. При наличии корреляции между поступающими на AP gNB сигналами некоторые собственные значения корреляционной матрицы (5), относящиеся к подпространству сигналов, становятся соизмеримыми со значениями шумового подпространства $\lambda_{\hat{K}+1}, \lambda_N$. При полной когерентности принимаемых сигналов оценка размерности сигнального подпространства оказывается меньше количества источников излучения $\hat{K} < K$, и разрешение таких UE становится невозможным.

Для компенсации данного негативного эффекта применяется алгоритм пространственного сглаживания, описание которого будут представлено в следующем разделе.

3. Алгоритм пространственного сглаживания

Суть метода пространственного сглаживания заключается в разбиении антенной решётки базовой станции gNB на $B = B_x B_y$ перекрывающихся подрешёток (рис. 3), для каждой из которых производится расчёт корреляционных матриц и последующее их усреднение [13].

Процедура разбиения плоской антенной решётки на B подрешёток может быть представлена с помощью матриц селекции для вертикального и горизонтального рядов (рис. 3):

$$\mathbf{J}_{b_x} = [\mathbf{0}_{N_{sub,x} \times (b_x - 1)}, \mathbf{I}_{N_{sub,x}}, \mathbf{0}_{N_{sub,x} \times (N_x - (N_{sub,x} + b_x - 1))}], \quad b_x = \overline{1, B_x}$$

$$\mathbf{J}_{b_y} = [\mathbf{0}_{N_{sub,y} \times (b_y - 1)}, \mathbf{I}_{N_{sub,y}}, \mathbf{0}_{N_{sub,y} \times (N_y - (N_{sub,y} + b_y - 1))}], \quad b_y = \overline{1, B_y},$$

где $\mathbf{0}_{a \times b}$ – нулевая матрица размерности $a \times b$. Тогда матрица выбора подрешётки принимает вид: $\mathbf{J}_b = \mathbf{J}_{b_x} \otimes \mathbf{J}_{b_y}$, где $b = (b_x - 1) B_y + b_y$; $\otimes$ – произведение Кронекера; $N_{sub,x}$, $N_{sub,y}$ – количество антенных элементов в подрешётке вдоль осей Ox , Oy .

Значения количества подрешёток B_x , B_y , размещаемых вдоль осей Ox , Oy , рассчитываются согласно формулам:

$$B_x = \left\lfloor \frac{N_x - \Delta_x}{N_{sub,x} - \Delta_x} \right\rfloor, \quad B_y = \left\lfloor \frac{N_y - \Delta_y}{N_{sub,y} - \Delta_y} \right\rfloor,$$

где Δ_x , Δ_y – количество антенных элементов, перекрываемых соседними подрешётками вдоль осей Ox , Oy ; $\lfloor \cdot \rfloor$ соответствует операции округления до нижележащего целого.

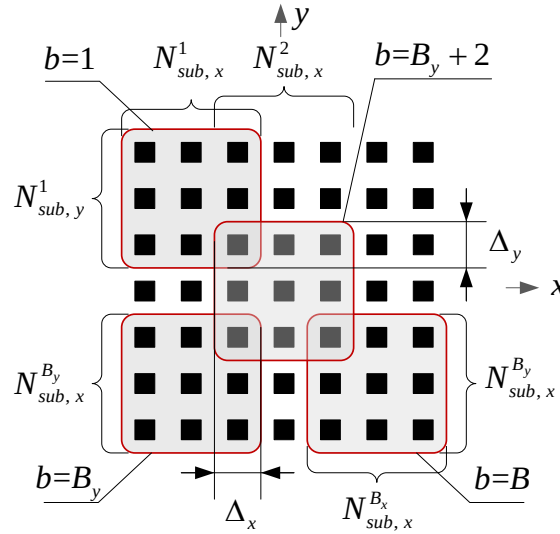


Рис. 3. Разбиение антенной решётки на подрешётки

Сигнал, поступающий на подрешётку b , может быть описан выражением:

$$\dot{\mathbf{Y}}_b = \mathbf{J}_b \dot{\mathbf{Y}} = \mathbf{J}_b \dot{\mathbf{H}}_\Sigma \dot{\mathbf{S}} + \mathbf{J}_b \dot{\mathbf{W}} = \dot{\mathbf{H}}_{sub,b} \dot{\mathbf{S}} + \mathbf{J}_b \dot{\mathbf{W}} = \dot{\mathbf{H}}_{sub,1} \dot{\Phi}_b \dot{\mathbf{S}} + \mathbf{J}_b \dot{\mathbf{W}}, \quad (8)$$

где $\dot{\mathbf{H}}_{sub,b}$ – матрица отклика канала для подрешётки b ; $\dot{\Phi}_b = \text{diag}\{1, \psi_b \dots \psi_b\}$ – диагональная унитарная матрица, где $\psi_b = \exp(-j\Delta_b^T \mathbf{p}_{gNB})$ – оператор, описывающий фазовый сдвиг между сигналами, поступающими на подрешётки 1 и b ; вектор разницы в координатах между одинаково расположенными элементами b -й и 1-й подрешёток имеет следующий вид:

$$\Delta_b^T = \left[\frac{\lambda}{2} \text{mod}(b-1, N_{sub,x})(N_{sub,x} - \Delta_x), \frac{\lambda}{4}(N_{sub,y} - \Delta_y) \lfloor (b-1) / N_{sub,y} \rfloor \right].$$

Корреляционная матрица принимаемого сигнала для подрешётки b будет описываться выражением:

$$\begin{aligned} \dot{\mathbf{Q}}_b &= \mathbf{J}_b \dot{\mathbf{Q}}_y \mathbf{J}_b^T = \dot{\mathbf{H}}_{sub,1} \dot{\Phi}_b E \left[\dot{\mathbf{S}} \dot{\mathbf{S}}^H \right] \dot{\Phi}_b^H \dot{\mathbf{H}}_{sub,1}^H + \mathbf{J}_b E \left[\dot{\mathbf{W}} \dot{\mathbf{W}}^H \right] \mathbf{J}_b^T = \\ &= \dot{\mathbf{H}}_{sub,1} \dot{\Phi}_b \dot{\mathbf{Q}}_s \dot{\Phi}_b^H \dot{\mathbf{H}}_{sub,1}^H + \sigma^2 \mathbf{I}_{N_{sub}} \end{aligned} \quad (9)$$

Модифицированная корреляционная матрица получается методом усреднения значений корреляционных матриц поступающих на подрешетки сигналов:

$$\bar{\mathbf{Q}}_{ss} = \frac{1}{B} \sum_{b=1}^B \dot{\mathbf{Q}}_b = \dot{\mathbf{H}}_1 \bar{\mathbf{Q}}_s \dot{\mathbf{H}}_1^H + \sigma^2 \mathbf{I}_{N_{sub}}, \quad (10)$$

где матрица $\bar{\mathbf{Q}}_s$ имеет вид:

$$\bar{\mathbf{Q}}_s = B^{-1} \left[\mathbf{I}, \dot{\Phi}_2 \dots \dot{\Phi}_B \right] \text{diag} \{ \dot{\mathbf{Q}}_s, \dot{\mathbf{Q}}_s \dots \dot{\mathbf{Q}}_s \} \left[\mathbf{I}, \dot{\Phi}_2^H \dots \dot{\Phi}_B^H \right]^T. \quad (11)$$

4. Численные результаты эксперимента и их интерпретация

В рамках данной работы была разработана программа имитационного моделирования, с помощью которой был проведён ряд вычислительных экспериментов. Целью проведения данных экспериментов являлся анализ влияния процедуры пространственного сглаживания на рабочие характеристики метода 2D-MUSIC при различных способах разбиения AP gNB на подрешётки в условиях наличия корреляции между отдельными источниками сигналов.

Эксперименты проводились для плоских эквидистантных АР (рис. 1) с расстоянием между элементами, равным половине длины волны несущего колебания: 0.5λ . Начала систем координат совпадают с центром масс АР gNB. Оценка корреляционной матрицы осуществлялась по 150 выборкам входного процесса. Исходные данные представлены ниже:

- частота несущего колебания $f_0 = 25$ ГГц (FR2);
- количество пользовательских устройств $K = 14$;
- высота подвеса антенны $h_{\text{gNB}} = 20$ м;
- высота положения пользовательского устройства $1.0 \leq h_{\text{UE}} \leq 2.0$ м;
- количество элементов АР gNB $N = N_x N_y = 16 \times 16 = 256$;
- количество элементов АР UE $M = M_x M_y = 4 \times 4 = 16$;
- расстояние между элементами в подрешётке АР gNB $d_{\text{sub}} = 0.5\lambda = 0.012$ м;
- метод определения количества UE – критерий Акаике (AIC).

Радиус соты при моделировании определялся согласно методике расчёта Coverage Dimensioning (рис. 4), для которой модель потерь на трассе рассчитывалась согласно TR38.901 [15]. Для диапазона FR2 использовались следующие значения аргументов для расчёта бюджета радиолинии:

- запас на медленные замирания – 8 дБ;
- типовые потери на проникновение через растительность – 17 дБ;
- типовые потери на проникновение через тело человека – 15 дБ;
- запас на интерференцию DL/UL – 1/0.5 дБ;
- значение сигнал/(шум + интерференция) SINR = 15 дБ;
- запас на осадки дождь/снег – 3 дБ.

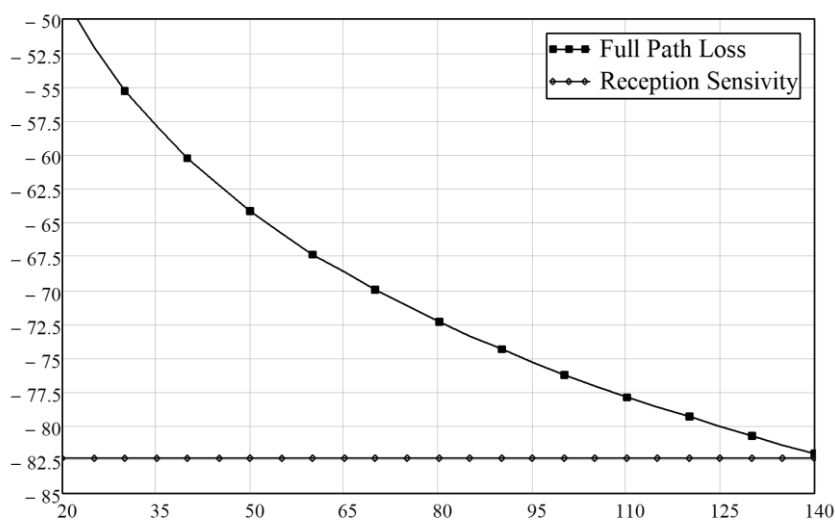


Рис. 4. Расчёт бюджета радиолинии

Как видно из рис. 4, радиус соты в данном случае может достигать $R_{\text{cell}} \approx 140$ м. В случае распространения сигнала в открытом пространстве расстояние между передающей и приёмной станциями, при котором обеспечивается приём сигналов, может превышать 800 метров на значительно большие расстояния (в данном случае – до 650 м с учётом двух переотражений от поверхностей, что характерно для городов). В рамках вычислительного эксперимента радиус соты составил 120 м, что обеспечивает уровень сигнала выше порога чувствительности приёмника.

Значения координат пользовательских устройств относительно gNB приведены в табл. 1.

Таблица 1. Координаты пользовательских устройств

UE	X, м	Y, м	Z, м	h_{UE} , м	UE-gNB, м	θ°	φ°
1	-50.7	48.8	-18.5	1.5	72.8	-14.7	46.1
2	-85.7	69.2	-18.9	1.1	111.8	-9.7	51.1
3	-35.4	44.8	-18.7	1.3	60.1	-18.1	38.3
4	-47.0	20.0	-18.5	1.5	54.4	-19.9	67.0
5	-10.9	103.0	-18.7	1.3	105.3	-10.2	6.0
6	-22.1	20.2	-18.3	1.7	35.1	-31.4	47.6
7	63.5	50.6	-18.4	1.6	83.2	-12.8	-51.5
8	-54.0	44.4	-18.2	1.8	72.2	-14.6	50.6
9	-61.6	36.6	-18.4	1.6	74.0	-14.4	59.3
10	-23.3	58.0	-17.9	2.1	65.0	-16.0	21.9
11	-17.6	47.8	-18.1	1.9	53.9	-19.6	19.9
12	73.9	80.9	-18.9	1.1	111.2	-9.8	-42.4
13	-50.8	44.2	-17.8	2.2	69.6	-14.8	49.0
14	-12.7	17.3	-18.6	1.4	28.4	-40.9	36.4

В табл. 1 жирным шрифтом обозначены UE с близкими значениями угловых координат: (1, 8, 13), (10, 11). На рис. 5 представлены диаграммы направленности АР 4×4 пользовательских устройств. Вычислительный эксперимент проводился для сценария, представленного на рис. 6, на котором изображена трасса не прямой видимости для UE₂. Для данного сценария длина трассы составила 534.8 м, разность хода лучей составляет 423 м, что соответствует временной задержке в 1.41 мкс и превышает длину циклического префикса в символах OFDM, которая составляет 0.6 мкс при использовании нумерологии $\mu = 3$.

В ходе эксперимента предполагалось, что данные от UE₂ и UE₆ передаются в ресурсных блоках с одинаковыми номерами соседних символов OFDM. Уровень сигнала от UE₆ на приёмной стороне устанавливался приблизительно равным уровню компоненты многолучевого распространения UE₂, что является возможным за счёт внесения потерь на прохождение радиосигнала через растительность и людей. Измерения снимались в интервале времени, расположенном на границе двух символов OFDM и не превышающем величину задержки компоненты многолучевого распространения UE₂.

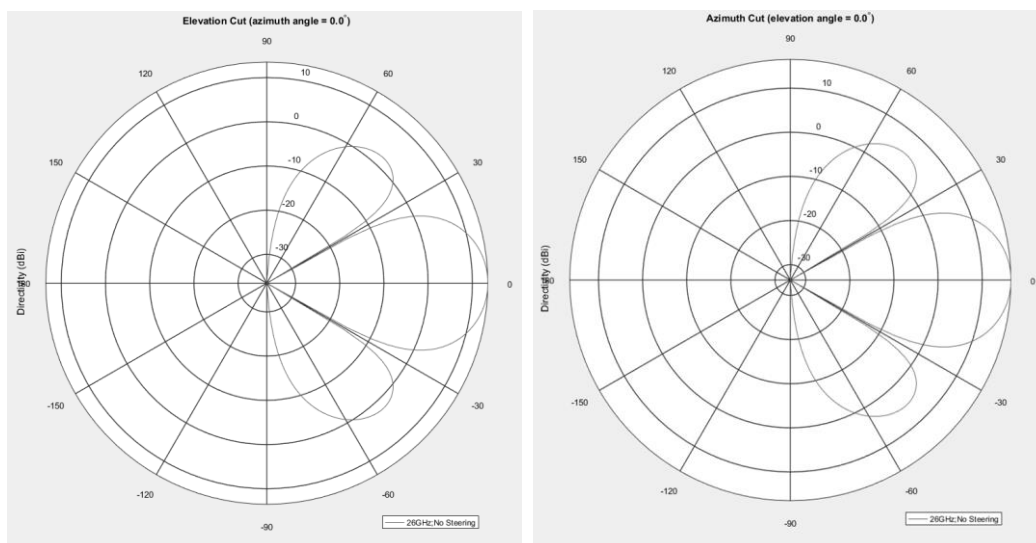
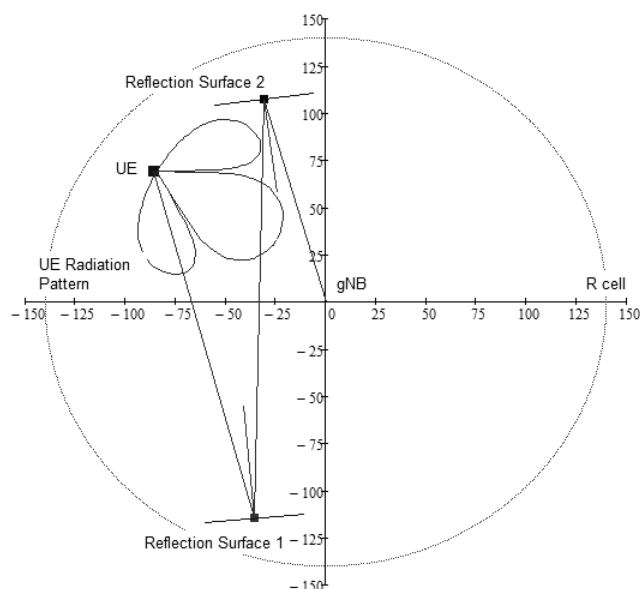


Рис. 5. Диаграмма направленности антенной решётки 4×4 (угол места и азимут)

Рис. 6. Траектория распространения компоненты отражённого луча UE_2

С целью анализа влияния метода пространственного сглаживания на декорреляцию сигналов от UE_2 и UE_6 рассматривались различные варианты формирования подрешёток, сведения о которых представлены в табл. 2.

Таблица 2. Формирование подрешёток для метода пространственного сглаживания

Параметр	Вариант разбиения AP					
	1	2	3	4	5	6
Количество элементов в подрешётке по оси Ox	16	8	16	8	8	12
Количество элементов в подрешётке по оси Oy	16	16	8	8	8	12
Перекрытие подрешёток по оси Ox Δ_x	0	0	0	0	6	8
Перекрытие подрешёток по оси Oy Δ_y	0	0	0	0	6	8
Количество подрешёток B_x	1	2	1	2	5	2
Количество подрешёток B_y	1	1	2	2	5	2
Количество подрешёток B	1	2	2	4	25	4

Влияние способа разбиения антенной решетки при выполнении процедуры пространственного сглаживания на функцию (6) можно проследить в общем виде с помощью трехмерных графиков пространственного спектра (рис. 7). Тёмным цветом на графиках выделены точки истинных значений угловых координат пользовательских устройств.

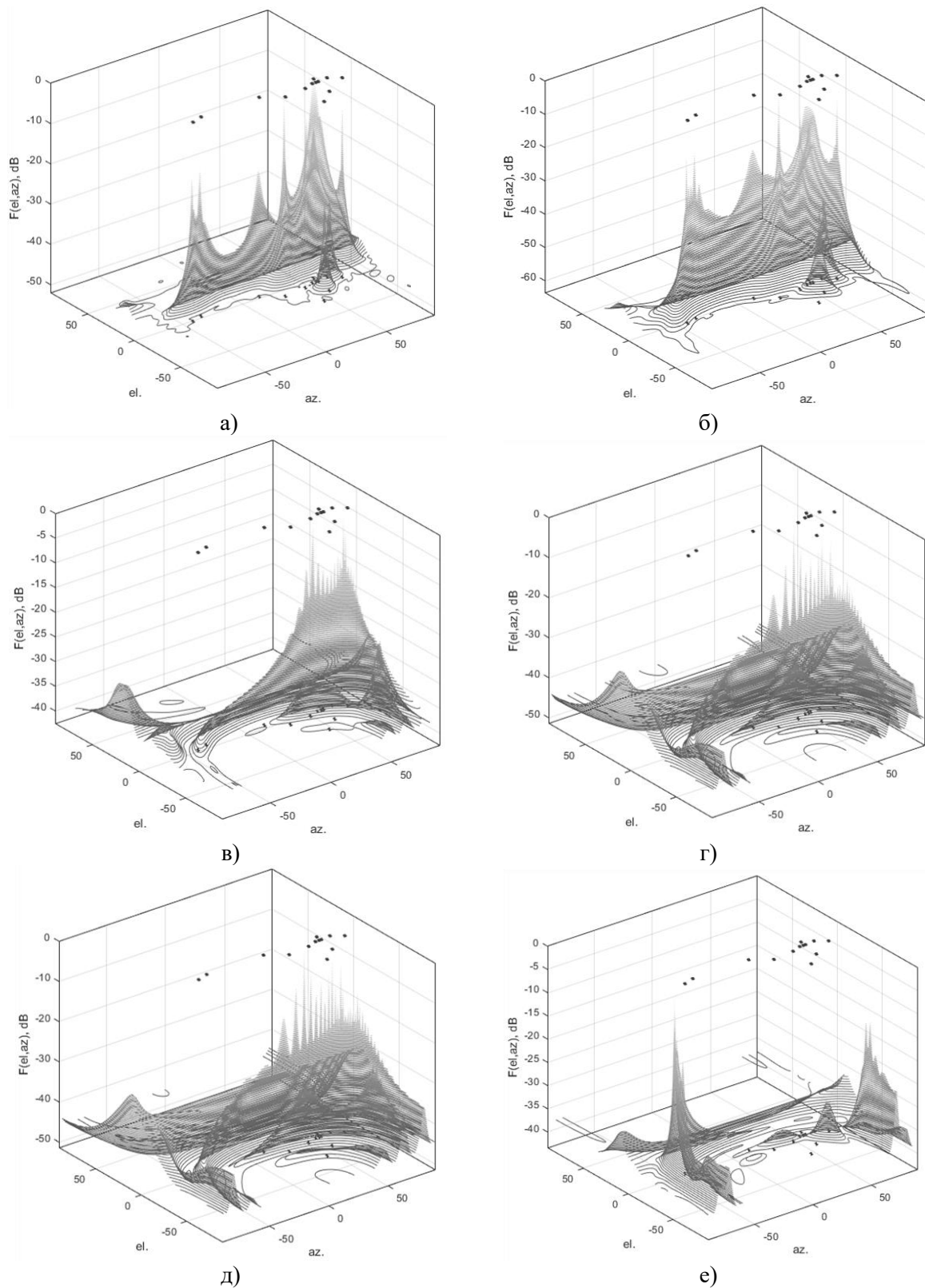


Рис. 7. Пространственный спектр сигнала при разных вариантах разбиения АР:
а) вариант 1; б) вариант 2; в) вариант 3; г) вариант 4; д) вариант 5; е) вариант 6

Результаты измерений для 6 сценариев представлены в сводной табл. 3.

Таблица 3. Оценки угловых координат пользовательских устройств

UE_k	θ°	φ°	сценарий 1			сценарий 2			сценарий 3		
			$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE,k}^o$	f , дБ	$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE,k}^o$	f , дБ	$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE}^o$	f , дБ
1	-14.7	46.1	-13.0	49.0	1.1	-13.0	47.0	-6.7	-17.0	46.0	-11.1
2	-9.7	51.1	-13.0	49.0	1.1	-10.0	51.0	-1.2	–	–	–
3	-18.1	38.3	-16.0	41.0	-7.1	-18.0	38.0	-6.8	-16.0	42.0	-0.2
4	-19.9	67.0	-18.0	69.0	-7.1	-18.0	69.0	-7.7	–	–	–
5	-10.2	6.0	-8.0	7.0	-18.3	-8.0	7.0	-1.6	–	–	–
6	-31.4	47.6	–	–	–	-24.0	45.0	-7.1	–	–	–
7	-12.8	-51.5	-11.0	-50.0	-14.3	-11.0	-50.0	-1.1	–	–	–
8	-14.6	50.6	-13.0	49.0	1.1	-13.0	47.0	-7.2	-18.0	47.0	-11.1
9	-14.4	59.3	-12.0	56.0	-9.4	-12.0	56.0	-21.4	–	–	–
10	-16.0	21.9	-14.0	24.0	0.2	-16.0	22.0	0.0	-15.0	25.0	-22.3
11	-19.6	19.9	-14.0	24.0	-15.4	-19.0	21.0	-21.1	-15.0	25.0	-22.3
12	-9.8	-42.4	-8.0	-41.0	-13.1	-8.0	-41.0	-1.1	–	–	–
13	-14.8	49.0	-13.0	49.0	1.1	-13.0	47.0	-7.2	-17.0	46.0	-11.1
14	-40.9	36.4	-39.0	38.0	-23.4	-39.0	38.0	-31.2	–	–	–
$\hat{K}$			12			14			13		
UE_k	θ°	φ°	сценарий 4			сценарий 5			сценарий 6		
			$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE,k}^o$	f , дБ	$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE,k}^o$	f , дБ	$\hat{\theta}_{UE}^o$	$\hat{\varphi}_{UE,k}^o$	f , дБ
1	-14.7	46.1	-18.0	49.0	-9.2	-18.0	49.0	-9.2	-12.0	52.0	–
2	-9.7	51.1	–	–	–	–	–	–	–	–	–
3	-18.1	38.3	-15.0	37.0	0	-15.0	37.0	0	–	–	–
4	-19.9	67.0	–	–	–	–	–	–	–	–	–
5	-10.2	6.0	-13.0	13.0	-26.3	-13.0	13.0	-26.3	–	–	–
6	-31.4	47.6	–	–	–	–	–	–	–	–	–
7	-12.8	-51.5	–	–	–	–	–	–	-12.0	-54.0	-21.2
8	-14.6	50.6	-18.0	49.0	-9.2	-18.0	49.0	-9.2	–	–	–
9	-14.4	59.3	–	–	–	–	–	–	–	–	–
10	-16.0	21.9	-13.0	21.0	-16.2	-13.0	21.0	-16.2	–	–	–
11	-19.6	19.9	-13.0	21.0	-16.2	-13.0	21.0	-16.2	–	–	–
12	-9.8	-42.4	–	–	–	–	–	–	–	–	–
13	-14.8	49.0	-18.0	49.0	-9.2	-18.0	49.0	-9.2	–	–	–
14	-40.9	36.4	–	–	–	–	–	–	–	–	–
$\hat{K}$			12			12			13		

В табл. 3 содержатся значения локальных максимумов функции пространственного спектра (7) $f(\hat{\theta}_{UE}, \hat{\varphi}_{UE})$ (в табл. 3 аргументы $\hat{\theta}_{UE}, \hat{\varphi}_{UE}$ опущены), взятые в области точки исходных координат пользовательских устройств $p_k(\theta_{UE,k}, \varphi_{UE,k})$, и значения $\hat{\theta}_{UE,k}, \hat{\varphi}_{UE,k}$, соответствующие точкам локальных максимумов. Окрестностью точки $p_k(\theta_{UE,k}, \varphi_{UE,k})$ берётся множество точек с целочисленными значениями координат $\hat{\theta}_{UE}, \hat{\varphi}_{UE}$, для которых расстояние до точки $p_k(\theta_{UE,k}, \varphi_{UE,k})$ не превышает 8° . Вычисление локальных максимумов, значения которых оказывались выше порогового уровня, определялись программой автоматически. Наличие прочерков в таблице означает, что максимумы спектра не наблюдались в области точки $p_k(\theta_{UE,k}, \varphi_{UE,k})$ или два устройства с близкими значениями угловых координат могли быть определены как одно устройство. Значения локальных максимумов пространственного спектра, соответствующие ошибочным оценкам угловых координат UE, в таблице не представлены.

С целью оценки влияния выбора подрешёток в процедуре пространственного сглаживания на разрешающие способности метода 2D-MUSIC на рис. 8а – 8е приведены графики сечений пространственного спектра плоскостями, параллельными плоскости $f(\hat{\theta}_{UE}, \hat{\phi}_{UE}) \circ \hat{\phi}_{UE}$, для значений $\hat{\theta}_{UE}$, близких значениям угла места пользовательских устройств (рис. 5а, в, д). Также на рис. (рис. 8б, г, е) приведены фрагменты сечений пространственного спектра, сделанных в области координат пары UE {1, 12}. Под разрешающей способностью метода 2D-MUSIC понимается минимальный разнос по угловым координатам, при котором между пиками пространственного спектра обеспечивается глубина провала не менее чем 5 дБ [16].

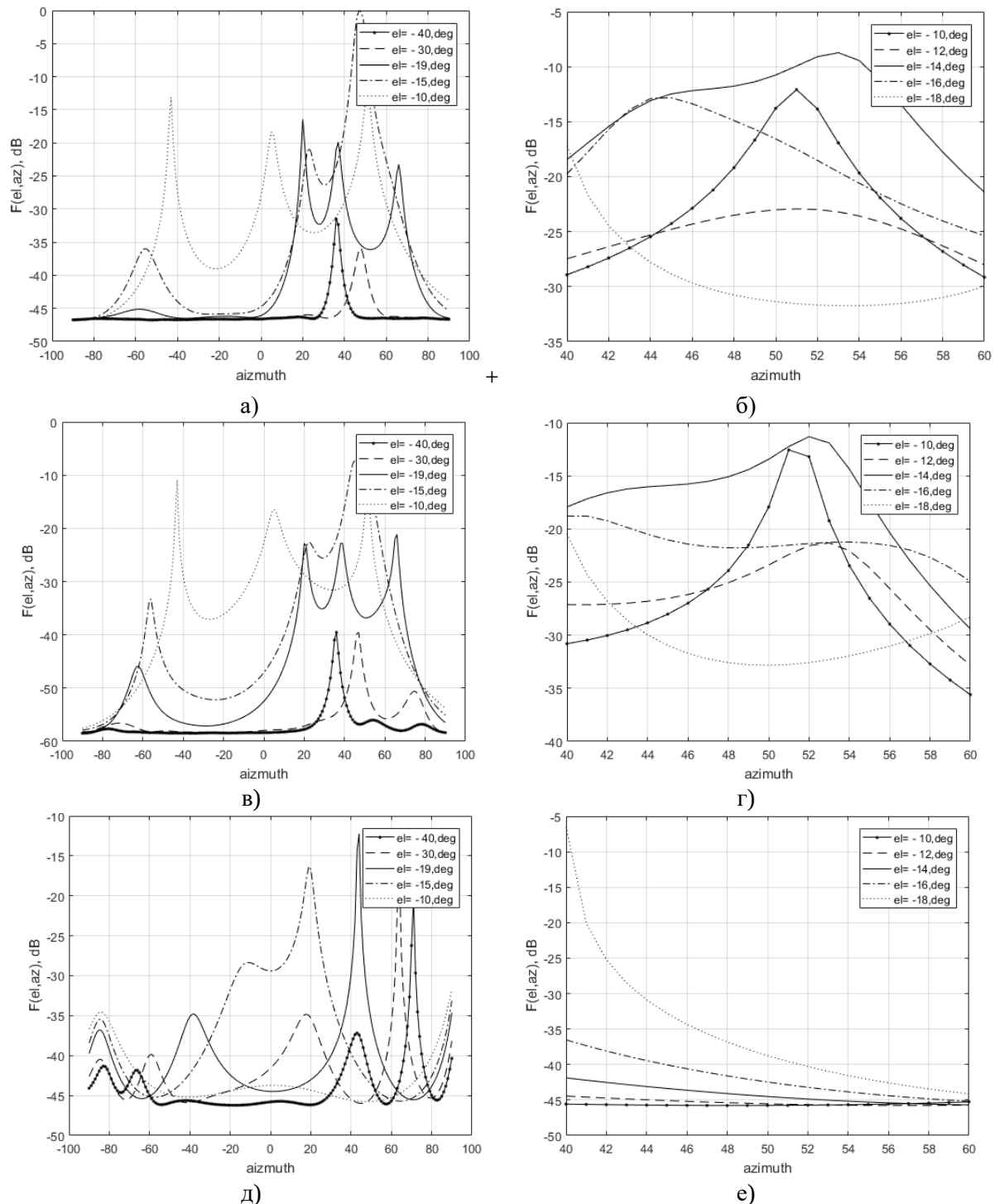


Рис. 8. Сечения пространственного спектра: Эксперимент 1. а) UE 1–14 5; б) UE 1,12; Эксперимент 2. в) UE 1–14 5; г) UE 1,12; Эксперимент 5. д) UE 1–14 5; е) UE 1,12

В табл. 4 приведены средние значения ошибок оценивания угловых координат, взятые по модулю, для пользовательских устройств, местоположение которых согласно табл. 3 можно считать обнаруженным верно.

Таблица 4. Оценки угловых координат пользовательских устройств

Параметр	номер сценария					
	1	2	3	4	5	6
$ E[\Delta\theta_{UE}] $	0.60°	0.32°	0.96°	3.35°	3.35°	4.10°
$\max \Delta\theta_{UE} $	3.70°	1.00°	3.80°	4.40°	4.40°	4.10°
$ E[\Delta\varphi_{UE}] $	0.83°	0.63°	0.87°	4.09°	3.96°	3.70°
$\max \Delta\varphi_{UE} $	3.80°	1.80°	3.80°	4.50°	4.50°	3.70°

Данные, полученные в ходе проведения экспериментов, показывают, что:

1. Рассмотренный в данной работе алгоритм пространственного сглаживания позволяет повысить точность определения размерности подпространства сигналов с помощью критерия Акаике в случае разбиения АР на подрешётки прямоугольной формы. Количество элементов в ряду большей длины должно стремиться к количеству элементов ряда АР. В случае разбиения АР на подрешётки квадратной формы применение процедуры пространственного сглаживания не привело к повышению точности определения размерности сигнального подпространства (табл. 2).

2. В случае разбиения АР на прямоугольные подрешётки согласно сценарию 2 процедура пространственного сглаживания обеспечивает декорреляцию сигналов от источников 2 и 6 и повышает разрешающую способность метода. Как видно из рис.8.г 2D-MUSIC, провал между пиками пространственного спектра, разнесенными на 3°, составляет 5 дБ. При других способах разбиения АР разрешение пользовательских устройств с номерами 1 и 2 не осуществляется. Разрешение узлов 8 и 13 с меньшим угловым разномом не было осуществлено ни в одном из проведенных экспериментов.

3. Рассмотренный в данной работе алгоритм пространственного сглаживания в случае разбиения АР на подрешётки квадратной формы приводит к искажениям тела функции пространственного спектра сигнала, схожим по эффекту воздействия с положительной дисторсией (рис. 9в), что в экспериментах 3 – 6 устраняет положительный эффект от декорреляции сигналов.

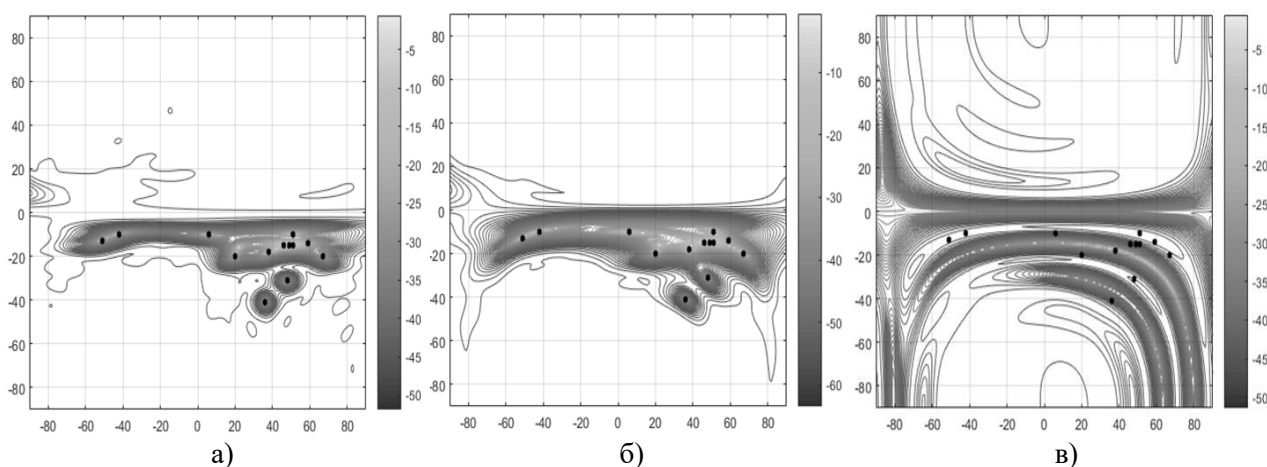


Рис. 9. Изолинии пространственного спектра сигнала для разных способов разбиения АР:
а) вариант 1 (без пространственного сглаживания); б) вариант 2; в) вариант 4

Аналогичные искажения наблюдаются при разбиении АР на подрешётки прямоугольной формы вдоль оси Oy . Данного вида искажения приводят к сдвигу пиков пространственного спектра, что значительно увеличивает погрешности оценивания угловых координат пользовательских устройств (табл. 4) по сравнению с вариантом, когда процедура пространственного сглаживания не выполняется. В случае разбиения АР на подрешётки прямоугольной формы вдоль оси Ox наблюдается снижение уровня минимальных значений пространственного спектра на 15 – 20 дБ по сравнению с методом 2D-MUSIC без применения пространственного сглаживания (рис. 8б). Искажения формы пространственного спектра наблюдаются незначительные.

На основе полученных наблюдений можно сделать вывод о том, что выбор формы подрешётки и последующий способ разбиения плоской АР базовой станции при проведении процедуры пространственного сглаживания непосредственно влияет на рабочие характеристики метода определения направления прихода сигнала 2D-MUSIC.

Алгоритм пространственного сглаживания, направленный на устранение эффекта корреляции сигналов, повышает точность определения количества источников излучения и разрешающую способность метода 2D-MUSIC при разбиении АР на подрешётки прямоугольной формы вдоль оси Ox .

5. Заключение

В настоящей работе был исследован алгоритм пространственного сглаживания, направленный на устранение корреляционных связей между сигналами, принимаемыми антенной решеткой базовой станции. С помощью вычислительных экспериментов была произведена оценка влияния процедуры пространственного сглаживания корреляционной матрицы при различных способах разбиения антенной решетки базовой станции на рабочие характеристики пространственного метода вычисления угловых координат 2D-MUSIC.

Результаты вычислительных экспериментов показали, что в случае применения процедуры пространственного сглаживания корреляционной матрицы рабочие характеристики метода 2D-MUSIC повышаются при разбиении АР вдоль оси Ox на подрешетки прямоугольной формы. Для заданного случая достигается значение величины допустимого углового разнosa между устройствами, равное 3° . Математические ожидания погрешности вычислений азимута и угла места не превысили 0.8° при максимальных значениях абсолютной величины погрешности оценивания, равных 1.8° по азимуту и 1.0° по углу места. Полученная точность определения направления прихода/ухода сигнала (DOA/DOD) устройств может считаться достаточной для управления диаграммой направленности антенной решетки базовой станции в отдельной радиолинии сверхплотной сети радиодоступа миллиметрового диапазона.

Литература

1. *Duong Trung Q. et al.* Ultra-Dense Networks for 5G and Beyond: Modelling, Analysis, and Applications. 1st ed. Wiley, 2019. ISBN: 1-119-47372-1.
2. *Gross F. B.* Smart antennas for wireless communications with MATLAB. McGraw Hills, 2005. 288 p.
3. *Григорьев В. А., Щесняк С. С., Гулюшин В. Л., Распаев Ю. А., Лагутенко О. И., Щесняк А. С.* Адаптивные антенные решетки: учебное пособие в 2-х частях. Часть 1. Санкт-Петербург: Университет ИТМО, 2016. 179 с.
4. *Григорьев В. А., Щесняк С. С., Гулюшин В. Л., Распаев Ю. А., Хворов И. А., Щесняк А. С.* Адаптивные антенные решетки: учебное пособие в 2-х частях. Часть 2. Санкт-Петербург: Университет ИТМО, 2016. 118 с.

5. *Фокин Г. А.* Моделирование сверхплотных сетей радиодоступа 5G с диаграммообразованием // Т-Сomm-Телекоммуникации и Транспорт. 2021. Т. 15, № 5. С. 4–21.
6. *Фокин Г. А.* Модели диаграммообразования в сверхплотных сетях радиодоступа 5G. Часть 1. Оценка помех // Первая миля. 2021. № 3 (95). С. 66–73.
7. *Фокин Г. А.* Модели диаграммообразования в сверхплотных сетях радиодоступа 5G. Часть 2. Оценка разнораспределения устройств // Первая миля. 2021. № 4 (96). С. 66–73.
8. *Фокин Г. А.* Процедуры выравнивания лучей устройств 5G NR // Электросвязь. 2022. № 2. С. 26–31.
9. *Фокин Г. А.* Модели управления лучом в сетях 5G NR. Часть 1. Выравнивание лучей при установлении соединения // Первая миля. 2022. № 1 (101). С. 42–49.
10. *Фокин Г. А.* Модели управления лучом в сетях 5G NR. Часть 2. Выравнивание лучей при ведении радиосвязи // Первая миля. 2022. № 3 (103). С. 62–69.
11. *Молчанов Д. А.* Разработка подходов, методов исследования и моделей обеспечения показателей качества обслуживания в беспроводных сетях пятого поколения: специальность 05.12.13 «Системы, сети и устройства телекоммуникаций»: диссертация на соискание ученой степени доктора технических наук / Молчанов Дмитрий Александрович; Российский университет дружбы народов. Санкт-Петербург, 2020. 315 с.
12. Многолучевое распространение и параметризация его характеристик: рекомендация МСЭ-R P.1407-6: https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1407-8-202109-I!!PDF-R.pdf (дата обращения: 12.07.2023).
13. *Van Trees H. L.* Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. Wiley 2002. ISBN 978-0-47109390-9.
14. *Chen Zh., Gokeda G., Yu Y.* Introduction to Direction-of-Arrival Estimation. Artech, 2010. ISBN: 9781596930896.
15. 3GPP TR 38.901 V16.1.0 Study on channel model for frequencies from 0.5 to 100 GHz. Release 16. Dec. 2019.
16. *Ратынский М. В.* Адаптация и сверхразрешение в антенных решётках. М.: Радио и связь, 2003. 200 с.

Фокин Григорий Алексеевич

д.т.н., доцент, профессор кафедры радиосвязи и вещания, Санкт-Петербургский университет телекоммуникаций им. проф. М. А. Бонч-Бруевича (СПбГУТ, 193232, Санкт-Петербург, пр. Большевиков д. 22, к.1), тел. +7 812 305-12-68, e-mail: fokin.ga@sut.ru, ORCID ID: 0000-0002-5358-1895.

Гришин Илья Владимирович

к.т.н., доцент кафедры сетей связи и передачи данных, Санкт-Петербургский университет телекоммуникаций им. проф. М. А. Бонч-Бруевича (СПбГУТ, 193232, Санкт-Петербург, пр. Большевиков д. 22, к.1), тел. +7 812 305-12-74, e-mail: grishin.iv@sut.ru, ORCID ID: 0000-0003-2161-731X.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

Evaluation of the Influence of the Spatial Smoothing Procedure of the Correlation Matrix on the Efficiency of the 2D-MUSIC Method

Grigoriy A. Fokin, Ilya V. Grishin

The Bonch-Bruевич Saint Petersburg State University of Telecommunications (SPbSUT)

Abstract: The occurrence of correlations between radio signals arriving at the antenna array of the base station reduces the effectiveness of subspace methods for positioning user devices in wireless communication networks. For the purpose of radiation sources decorrelation, a spatial smoothing algorithm can be applied dividing the base station antenna array into sub-arrays and then averaging the signal correlation matrices obtained for these sub-arrays. In this paper, we investigate the impact of various ways of dividing a flat equidistant antenna array into subarrays on the performance of the 2D-MUSIC multiple signal classification method in the scenario of ultra-dense networks operating in the FR2 band.

Keywords: algorithm for multiple signal classification, azimuth, elevation angle, spatial signal spectrum, spatial smoothing.

For citation: Fokin G. A., Grishin I. V. Evaluation of the influence of the spatial smoothing procedure of the correlation matrix on the efficiency of the 2D-MUSIC method (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 91-108. <https://doi.org/10.55648/1998-6920-2024-18-1-91-108>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Fokin G. A., Grishin I. V., 2024

The article was submitted: 26.07.2023;
revised version: 01.10.2023;
accepted for publication 01.10.2023.

References

1. Duong Trung Q. et al. *Ultra-Dense Networks for 5G and Beyond: Modelling, Analysis, and Applications*. 1st ed., Wiley, 2019.
2. Gross F.B. *Smart antennas for wireless communications with MATLAB*. McGraw Hills, 2005. 288 p.
3. Grigor'ev V. A., Shchesnyak S. S., Gulyushin V. L., Raspaev Yu. A., Lagutenko O. I., Shchesnyak A. S. *Adaptivnye antennye reshetki. Chast' 1* [Adaptive antenna arrays. Part 1]. Saint Petersburg, ITMO University, 2016. 179 p.
4. Grigor'ev V. A., Shchesnyak S. S., Gulyushin V. L., Raspaev Yu. A., Khvorov I. A., Shchesnyak A. S. *Adaptivnye antennye reshetki. Chast' 1* [Adaptive antenna arrays. Part 1]. Saint Petersburg, ITMO University, 2016. 118 p.
5. Fokin G. A. Modelirovanie sverkhplotnykh setei radiodostupa 5G s diagrammoobrazovaniem [Modeling of ultra-dense 5G radio access networks with beamforming]. *T-Comm-Telekommunikatsii i Transport*, 2021, vol. 15, no. 5, pp. 4-21.
6. Fokin G. A. Modeli diagrammoobrazovaniya v sverhplotnykh setjah radiodostupa 5G. Chast' 1. Ocenka pomeh [Diagram formation models in ultra-dense 5G radio access networks. Part 1. Evaluation of interference]. *Pervaja milja*, 2021, no. 3 (95), pp. 66-73.
7. Fokin G. A. Modeli diagrammoobrazovaniya v sverhplotnykh setjah radiodostupa 5G. Chast' 2. Ocenka raznosa ustroystv [Diagram formation models in ultra-dense 5G radio access networks. Part 2. Evaluation of device spacing]. *Pervaja milja*, 2021, no. 4 (96), pp. 66-73.
8. Fokin G. A. Procedury vyравnivaniya luchej ustroystv 5G NR [Beam alignment procedures for 5G NR devices]. *Jelektrosvjaz'*, 2022, no. 2, pp. 26-31.
9. Fokin G. A. Modeli upravleniya luchom v setjah 5G NR. Chast' 1. Vyravnivanie luchej pri ustanovlenii soedineniya [Beam Steering Models in 5G NR Networks. Part 1. Alignment of beams when establishing a connection]. *Pervaja milja*, 2022, no. 1 (101), pp. 42-49.

10. Fokin G. Modeli upravljenja luchosm v setjah 5G NR. Chast' 2. Vyravnavanie luchej pri vede-nii radi-osvjazi [Beam control models in 5G NR networks. Part 2. Alignment of beams in radio communications]. *Pervaja milja*, 2022, no. 3 (103), pp. 62-69.
11. Molchanov D. A. *Razrabotka podhodov, metodov issledovanija i modelej obespechenija pokazatelej kachestva obsluzhivaniya v besprovodnyh setjah pjatogo pokolenija* [Development of approaches, research methods and models for ensuring quality of service indicators in wireless networks of the fifth generation]. Doctor's degree dissertation, Saint Petersburg, 2020, 315 p.
12. *Mnogoluchevoe rasprostranenie i parametrizacija ego harakteristik: rekomendacija MSJe-R P.1407-6* [Recommendation ITU-R P.1407-8: Multipath propagation and parameterization of its characteristics], available at: https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1407-8-202109-I!!PDF-R.pdf (accessed 12.07.2023).
13. Van Trees H. L. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley, 2002.
14. Chen Zh., Gokeda G., Yu Y. *Introduction to Direction-of-Arrival Estimation*. Artech, 2010.
15. *3GPP TR 38.901 V16.1.0 Study on channel model for frequencies from 0.5 to 100 GHz*. Release 16. Dec, 2019.
16. Ratynskij M. V. *Adaptacija i sverhrazreshenie v antennyh reshjotkah* [Adaptation and Super Resolution in Antenna Arrays]. Moscow, Radio i svjaz', 2003. 200 p.

Grigoriy A. Fokin

Dr. of Sci. (Engineering), Professor of the Department of Radioreception and Broadcasting, The Bonch-Bruевич Saint Petersburg State University of Telecommunications (SPbSUT, Russia, 193232, Saint-Petersburg, Bolshhevikov Ave. 22), phone: +7 812 305-12-68, e-mail: fokin.ga@sut.ru, ORCID ID: 0000-0002-5358-1895.

Ilya V. Grishin

Cand. of Sci. (Engineering), Associate Professor of the Department of Communication Networks and Data Transmission, The Bonch-Bruевич Saint Petersburg State University of Telecommunications (SPbSUT, Russia, 193232, Saint-Petersburg, Bolshhevikov Ave. 22), phone: +7 812 305-12-74, e-mail: grishin.iv@sut.ru, ORCID ID: 0000-0003-2161-731X.

Анализ характеристик алгоритмов прекодирования сигналов в системе MU-MIMO с оцениванием канала

А. А. Калачиков

Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: В статье представлены результаты имитационного моделирования алгоритма прекодирования ZF с использованием группирования абонентов по алгоритмам ортогонального выбора и максимизации взаимной информации в нисходящей системе MU-MIMO. При количестве пользователей больше, чем количество антенн на базовой станции, возникает взаимная корреляция между каналами пользователей, что снижает суммарную спектральную эффективность системы MU-MIMO. Для снижения эффекта взаимной корреляции применяется подбор пользователей на основе максимальной ортогональности между ними. Суммарная спектральная эффективность в системе MU-MIMO зависит от условий реального распространения сигналов, и для каналов с пространственной корреляцией необходимо использовать выбор подмножества абонентов с низкой корреляцией между их векторами каналов. В данной статье для исследования эффективности выбора подмножества абонентов используется модель канала с открытым исходным кодом, позволяющая получать реалистичные реализации канала.

Ключевые слова: 5G new radio, QuaDRiGa, модель канала 3GPP, прекодирование ZF, massive MIMO, группирование абонентов.

Для цитирования: Калачиков А. А. Анализ характеристик алгоритмов прекодирования сигналов в системе MU-MIMO с оцениванием канала // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 109–120. <https://doi.org/10.55648/1998-6920-2024-18-1-109-120>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Калачиков А. А., 2024

Статья поступила в редакцию 26.02.2023;
переработанный вариант – 21.07.2023;
принята к публикации 17.12.2023.

1. Введение

В многопользовательской системе связи с применением многоантенных технологий (MU-MIMO – multiuser multiple input multiple output) выигрыш от пространственного мультиплексирования может быть получен при помощи управления направлением излучения передаваемого сигнала для пользователя, повышая мощность принимаемого пользователем сигнала. Выигрыш от пространственного мультиплексирования – увеличение количества независимых пространственных потоков данных при одновременном использовании частотно-временных ресурсов системы для передачи данных нескольких пользователей. Одновременно необходимо частично или полностью подавлять воздействие сигналов интерференции от других пользователей. Реализуется пространственное мультиплексирование при прекодировании сигналов. Прекодирование состоит в обработке независимых сигналов пользователей при формировании передаваемого сигнала многоантенной системой. При обработке происходит умножение передаваемых символов на комплексные весовые векторы, что определяет пространственные свойства передаваемых сигналов.

В общем виде сценарии исследования MU-MIMO-систем можно разделить на сценарии с одной базовой станцией (БС) и с несколькими БС. Рассматриваемые алгоритмы прекодирования и группирования абонентов (распределения частотно-временных ресурсов) зависят от числа БС, числа антенн на БС, количества абонентов, расположения абонентов относительно БС (степени удаленности и уровня отношения сигнал/шум), точности представления информации о канале связи на стороне передатчика (CSIT).

Целью обработки сигналов в технологии MU-MIMO является одновременная передача символов нескольких пользователей на одном элементе частотно-временного ресурса системы (ресурсном элементе, блоке) при использовании весовых векторов прекодирования сигналов.

2. Описание системы MU-MIMO

2.1. Базовый сценарий системы MU-MIMO

Базовый сценарий системы MU-MIMO предполагает, что базовая станция оборудована антенной из M антенных элементов, БС обслуживает K абонентов, каждый из абонентов оборудован N антеннами. Матрица комплексных коэффициентов передачи канала MIMO на одной поднесущей сигнала для абонента k описывается в виде $\mathbf{H}_k \in \mathbb{C}^{M \times N}$, принятый сигнал записывается в виде $y_k = \mathbf{H}_k \mathbf{x}_k + n_k$, где $\mathbf{x}_k \in \mathbb{C}^{M \times 1}$ – вектор передаваемых символов для всех пользователей. Коэффициенты матрицы передачи (модуль) описываются релеевским распределением, пространственные статистические свойства канала описываются корреляционной матрицей канала $\mathbf{R}_k = E[\mathbf{H}_k^H \mathbf{H}_k]$ и зависят от конфигурации антенн, условий распространения сигналов и изменения положения приемника относительно передатчика.

При использовании прекодирования передаваемый сигнал запишется в виде $\mathbf{x} = \sum_{k=1}^K \mathbf{W}_k \mathbf{d}_k$,

где $\mathbf{W}_k \in \mathbb{C}^{M \times d_k}$ – матрица прекодирования для пользователя k , $\mathbf{d}_k$ – вектор передаваемых данных, d_k – количество независимых потоков передаваемых данных (пространственных слоев) пользователя k .

При этом передатчик формирует сигналы одновременно для различных пользователей в одном и том же частотно-временном ресурсе системы. Пространственное управление независимыми сигналами состоит в вычислении дополнительных значений амплитуд и фаз колебаний для их суммирования в требуемом направлении и снижения их уровня в других направлениях. При использовании информации о канале и совместном вычислении нужных амплитуд и фаз сигналов всех обслуживаемых пользователей увеличивается отношение сигнал/(шум+интерференция) на заданном приемнике при одновременном снижении уровня интерференции для других пользователей. Такое формирование сигналов в теории антенных решеток называется адаптивным формированием диаграммы направленности (beamforming). В литературе по MU-MIMO термин beamforming используется для обозначения формирования сигналов нескольких пользователей, что дает возможность для реализации множественного доступа с пространственным разделением (SDMA).

2.2. Описание канала MU-MIMO

Многоантенная технология с большим числом используемых антенн (Massive MIMO) является одной из основных системных технологий в сетях мобильной связи 5-го поколения (5G) для повышения скорости передаваемых данных и суммарной спектральной эффективности. Технология massive MIMO позволяет использовать различные схемы и конфигурации антенн на стороне базовой станции при реализации многопользовательской схемы управления

диаграммой направленности для снижения пространственной интерференции и повышения отношения сигнал/шум у каждого абонента [1].

Для технической реализации возможности управления диаграммой направленности по большому числу пользователей необходимо получить точную информацию о параметрах радиоканала между антеннами БС и антенной пользователя. При большом числе антенн БС количество и сложность сбора информации о канале (CSI – channel state information) значительно увеличивается, при этом появляется задержка получения текущей информации о канале и снижается эффективность вычисления весовых векторов в алгоритмах управления диаграммой направленности (ДН) антенны.

Основная особенность технологии massive MIMO – использование большого количества антенн (64 – 256) на стороне БС для передачи сигналов абонентам на перекрывающихся частотно-временных ресурсах. Использование большого числа антенн позволяет эффективнее снижать интерференцию между пользователями.

В данном разделе приводятся результаты численного моделирования алгоритмов прекодирования ZF (zero forcing – приведение к нулю) и метода кодовых книг DFT в системе MU-MIMO с получением информации о состоянии канала (оценки канала) методом наименьших квадратов (LS). Оценивание выполняется в частотной области, частотная характеристика канала оценивается с использованием пилотных поднесущих в составе сигнала OFDM, распределенных по частотно-временной ресурсной сетке. Оценивание канала происходит для каждого коэффициента передачи в векторе канала каждого пользователя. Для получения реализаций канала при моделировании алгоритмов оценивания и прекодирования используется модель QuaDRiGa, отражающая реальные условия распространения сигналов. Полученные реализации канала используются для оценки канала, вычисления отношения сигнал/(шум+интерференция), спектральной эффективности каждого пользователя с использованием алгоритмов прекодирования ZF и кодовых книг DFT. Прекодирование по методу кодовых книг основано на квантовании канала, и в качестве информации о состоянии канала используется вычисленное значение индекса кодового слова и норма ближайшего кодового вектора, передаваемые по каналу обратной связи на БС. Полученные характеристики прекодирования MU-MIMO ZF и кодовых книг DFT в пространственно коррелированном канале сравниваются на основе эмпирической функции распределения суммарной спектральной эффективности всех обслуживаемых пользователей. Числовые результаты показывают, что суммарная спектральная эффективность зависит от точности оценивания канала. Данные результаты могут быть использованы при моделировании алгоритмов на системном уровне.

Спектральная эффективность системы MU-MIMO в реальных условиях распространения является основным показателем в системе мобильной связи любого поколения. Система 5G NR использует технологию massive MIMO с большим количеством (до 256) антенных элементов в антенне на базовой станции, что позволяет реализовать пространственное мультиплексирование с большим выигрышем в скорости, передавая сигналы нескольких пользователей в одном частотно-временном ресурсе [2]. Технология прекодирования (beamforming) massive MIMO снижает интерференцию в системе с большим количеством пользователей и повышает спектральную эффективность сети. Количество элементов в антенной решетке БС massive MIMO намного больше числа обслуживаемых абонентов (UEs), пространственные каналы абонентов с большой вероятностью становятся ортогональными при наличии большого числа рассеивателей. Для задач прекодирования требуется точная информация о канале каждого пользователя, которая оценивается при помощи пилотных сигналов.

Первоначально системы massive MIMO были предложены с использованием временного разделения каналов (TDD). При этом для оценки канала на стороне БС используются восходящие сигналы абонентов и принцип обратимости канала. Для систем с частотным разделением канала нисходящий и восходящий каналы разнесены по частоте, нисходящий канал FDD оценивается на стороне абонента и передается на БС по каналу обратной связи.

В 5G NR оценка канала выполняется в частотной области. Для оценивания канала используется несколько видов опорных пилотных сигналов: в нисходящем направлении –

информационные опорные сигналы о состоянии канала (CSI-RS), и опорные зондирующие сигналы (SRS) в восходящем направлении. Свойство обратимости канала в системе с временным разделением позволяет использовать оценку канала в восходящем направлении для вычисления весовых коэффициентов прекодирования в нисходящем направлении. Точность вычисления коэффициентов и подавление взаимной интерференции между пользователями во многом зависит от точности оценки канала.

Теоретические исследования алгоритмов прекодирования преимущественно основаны на предположении о наличии точной информации о канале на стороне БС. Для определения практических характеристик прекодирования в MU-MIMO с использованием алгоритмов оценивания канала используются модели каналов с параметризацией по результатам измерений реальных радиоканалов в различных сценариях развертывания сети [3].

Основная задача данной статьи состоит в рассмотрении числовых характеристик методов прекодирования в типовом сценарии, используя реалистичную модель канала. Реализация модели канала на основе открытого кода QuaDRiGa позволяет получить реалистичные реализации канала в различных сценариях развертывания. Представлено численное моделирование двух типов прекодирования MU-MIMO, основанные на получении информации о канале путем оценивания канала при помощи алгоритма LS в типовых условиях городской застройки Urban micro.

На стороне БС системы связи используется антенна из N_T элементов, обслуживающая K пользователей, каждый из которых оборудован одной антенной. Коэффициенты передачи канала описываются вектором коэффициентов $\mathbf{h}_k \in C^{N_T \times 1}$. Вектор передаваемых символов

$\mathbf{x}_k \in C^{N_T \times 1}$ составляется в виде $\mathbf{x}_k = \sum_{k=1}^K \mathbf{w}_k s_k$, где s_k – передаваемый символ данных пользо-

вателя k , $\mathbf{w}_k \in C^{N_T \times 1}$ – вектор прекодирования пользователя k . Принятый вектор сигналов пользователя k на поднесущей с номером s и номером символа n запишется в виде

$$y_{k,n,s} = \mathbf{h}_{k,n,s}^T \mathbf{x}_{n,s} + n_{k,n,s} \text{ для } k = 1, \dots, K. \quad (2.1)$$

где $n_{k,n,s}$ – комплексный гауссовский шум с нулевым математическим ожиданием и дисперсией σ_k^2 .

В матричной форме матрица канала MU-MIMO составляется из векторов каналов пользователей $\mathbf{H}_{n,s} = [\mathbf{h}_{1,n,s} \dots \mathbf{h}_{K,n,s}]^T$, и вектор принятых сигналов пользователей определяется как

$$\mathbf{y}_{n,s} = \mathbf{H}_{n,s}^T \mathbf{x}_{n,s} + n_{n,s}. \quad (2.2)$$

В системе 5G NR оценивание канала выполняется при использовании опорных сигналов на пилотных поднесущих в составе OFDM-сигнала. Различные конфигурации опорных сигналов SRS в восходящем направлении приведены в [4]. Сигналы SRS передаются в составе OFDM-символов с определенным расположением по поднесущим и номерам слотов. Параметры SRS включают конфигурации SRS с различной периодичностью SRS и расположением по полосе частот сигнала. Для каждой последовательности SRS настраиваются: количество антенных портов SRS, символы в слоте, соответствующие каждой последовательности SRS sequence, слоты в периоде передачи SRS transmission, плотность поднесущих SRS в полосе частот сигнала.

Последовательность SRS генерируется путем циклического сдвига базисной последовательности, которая формируется из последовательности Задова–Чу. Базовая последовательность определяется как $r_{u,v}^{SRS} = r_{u,v}^\alpha(n)$, $n = 0, \dots, M_{SC}^{RS}$, M_{SC}^{RS} – длина последовательности опорного сигнала, $u = 0, \dots, 29$ – номер группы базисной последовательности, $v = 0, 1$ – номер последовательности в группе [5]. Базовая последовательность циклически сдвигается для увеличения общего числа доступных последовательностей. Сигналы SRS от различных пользователей

могут быть мультиплексированы по частоте в пределах полосы частот сигнала, используя различные комбинированные шаблоны в соответствии с частотным смещением.

При использовании прекодирования принятый сигнал для пользователя k определяется в виде

$$y_k = \mathbf{h}_k^T \mathbf{w}_k s_k + \sum_{j \neq k} \mathbf{h}_k^T \mathbf{w}_j s_j + \mathbf{n}_k, \text{ for } k = 1, \dots, K, \quad (2.3)$$

где сумма соответствует сигналам интерференции со стороны других пользователей [6].

Суммарная спектральная эффективность вычисляется как сумма спектральных эффективностей по всем пользователям по всем поднесущим и зависит от отношения сигнал/(шум + интерференция, SINR) каждого пользователя. Величина SINR пользователя k на одной поднесущей определяется как

$$\text{SINR}_k = \frac{|\mathbf{h}_k^T \mathbf{w}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^T \mathbf{w}_j|^2 + K\sigma^2 / P}. \quad (2.4)$$

Достижимая суммарная спектральная эффективность по всем пользователям вычисляется в виде $R_{BF} = \sum_{k=1}^K (\log_2(1 + \text{SINR}_k))$, бит/с/Гц.

Данная величина используется как метрика, показатель качества при моделировании алгоритмов прекодирования.

2.3. Оценивание канала по методу LS

Оценивание канала выполняется на поднесущих пилот сигналов в частотной области с использованием сформированных последовательностей SRS. Последовательности равномерно распределяются по полосе сигнала для всех пользователей. Оценка канала на каждой пилотной поднесущей вычисляется в виде $\hat{H}_{LS}[m] = \frac{Y[m]}{X[m]}$, $m = 0, \dots, M - 1$.

Среднеквадратическая ошибка (MSE) оценки канала LS вычисляется в виде $\text{MSE}_{LS} = \frac{\sigma_n^2}{\sigma_x^2}$

, она обратно пропорциональна отношению сигнал/шум [7].

2.4. Прекодирование по алгоритму ZF

Алгоритм обнуления сигналов интерференции (ZF) вычисляет весовые векторы прекодирования для снижения взаимной интерференции между всеми пользователями данного частотно-временного ресурса. Алгоритм использует псевдообращение матрицы канала и на ее основе вычисляет векторы прекодирования. Вектор прекодирования $\text{ZF } \mathbf{w}_k$ пользователя k ортогонален комплексно сопряженным векторам каналов всех остальных пользователей $\mathbf{h}_k^H \mathbf{w}_j = 0$ для $j \neq k$. Матрица прекодирования $\mathbf{W}_{ZF}$ составляется из векторов прекодирования всех пользователей и вычисляется на основе матрицы канала $\mathbf{H}_{n,s}$ как ее псевдообращение $\mathbf{W}_{ZF} = \mathbf{H}(\mathbf{H}^H \mathbf{H})^{-1}$. Матрица канала $\mathbf{H}_{n,s}$ оценивается на пилотных поднесущих.

В реальных условиях распространения радиоканал является пространственно коррелированным, при этом элементы матрицы канала $\mathbf{H}_{n,s}$ взаимно коррелированы, и это свойство влияет на характеристики системы MU-MIMO. Вычисление обратной матрицы коррелированного канала бывает затруднительно, и матрица канала $\mathbf{H}_{n,s}$ становится плохо обусловленной, что приводит к увеличению уровня интерференции в пространственных каналах с большим значением собственных чисел [8].

2.5. Прекодирование по методу кодовых книг DFT

При использовании прекодирования по методу кодовых книг вектор прекодирования выбирается из набора предварительно вычисленных кодовых слов, образующих кодовую книгу. На стороне абонента вычисляются индекс и норма вектора как информация по качеству канала (CQI) и передаются на базовую станцию для прекодирования и распределения ресурсов абонентов.

Кодовая книга ДПФ (DFT) может быть рассчитана для любого количества антенн и кодовых слов [8]. Кодовая книга DFT CB_{DFT} для N_T антенн на БС и N_{CB} кодовых слов описывается в виде $CB_{(DFT)} = [w_1 \dots w_{N_{cb}}]$, где кодовое слово w_i вычисляется как

$$w_i = \frac{1}{\sqrt{N_T}} [1 e^{j \frac{2\pi}{N_{cb}} i} \dots e^{j \frac{2\pi}{N_{cb}} (N_T - 1) i}]^T. \quad (2.5)$$

На приемной стороне пользователь с номером k выбирает кодовое слово w_k из кодовой книги C_{CB} предварительно рассчитанных весовых векторов $w_i \in C_{cb}$, которое максимизирует норму $w_k = \operatorname{argmax} |h_k^H w_i|^2$. Индекс выбранного кодового слова и соответствующая норма отправляются на базовую станцию как показатели качества канала и используются для прекодирования и планирования нагрузки абонентов.

Оценка канала LS всегда имеет ошибку оценивания, что приводит к неточному значению CSI, передаваемому по каналу обратной связи. Это нарушает ортогональность между каналами пользователей и вычисленными весовыми векторами прекодирования, что снижает суммарную спектральную эффективность.

2.6. Модель канала QuaDRiGa

Для решения задач моделирования и определения характеристик прекодирования MU-MIMO в практических сценариях развертывания используется модель канала 5G NR. Модель канала с открытым исходным кодом QuaDRiGa является 3GPP-3D геометрической вероятностной моделью, реализованной в соответствии с требованиями стандарта [3] с параметризацией по результатам многочисленных измерений радиоканалов.

Модель канала QuaDRiGa параметризуется в соответствии с необходимым сценарием развертывания сети. Параметры крупномасштабных и мелкомасштабных замираний генерируются вероятностным методом для каждой реализации канала.

Основные параметры системы моделирования определяются перед генерированием реализаций канала. Положение БС, конфигурация антенн, траектория перемещения абонентов и соответствующие сценарии распространения являются исходными параметрами. Генерация реализаций канала состоит из генерации параметров крупномасштабных замираний (рассеяние задержек, угловое рассеяние) и положения рассеивающих кластеров объектов [12].

Модель многолучевого распространения вычисляется для каждого пользователя в соответствии с распределением рассеивателей в пространстве. Вероятностно сгенерированные значения рассеяния задержек и углового рассеяния вычисляются для многолучевых компонент, каждая с соответствующим уровнем. Полученные коэффициенты канала имеют единые пространственные и временные характеристики с данными измерений, использованными в качестве параметров модели.

3. Результаты моделирования

В данном разделе представлены результаты моделирования двух методов прекодирования MU-MIMO с использованием оценки канала LS. Эффективность прекодирования сравнивалась по величине средней спектральной эффективности. Для имитационного моделирования использовались реализации канала, полученные в модели QuaDRiGa. Параметры моделирования приведены в табл. 1.

Таблица 1. Параметры модели

Параметр модели	Значение
Тип модели	QuaDRiGa v.2.2
Сценарий	3GPP 38.901 UMi NLoS
Количество многолучевых компонент	47
Центральная частота	3.6 ГГц
Полоса	20 МГц
Количество антенн БС	16
Количество абонентов	4

В соответствии со сценарием городской застройки и малой размерности зоны покрытия БС (UMi) пользователи распределены равномерно вокруг БС на удалении до 500 м. Реализация канала для данного пользователя и элемента антенны на БС состоит из многолучевых компонент. Частотная характеристика каждого канала в полосе 10 МГц делится на 400 поднесущих передаваемого сигнала. Для каждой поднесущей вычисляется вектор прекодирования по алгоритмам ZF и DFT.

Канал каждого пользователя и весовые векторы прекодеров нормируются к единичной мощности. Использование нормирования матриц канала MU-MIMO и весовых векторов позволяет исключить из рассмотрения потери на распространение сигнала, характеристики усиления антенн, мощность передатчика и учитывать только свойства пространственной корреляции канала. При моделировании мощность передатчика $P = 1$, мощность шума σ_n^2 , отношение сигнал/шум $SNR = 1/\sigma_n^2 = 12$ дБ. Для сравнения эффективности прекодирования при различных условиях распространения были выбраны две величины числа обусловленности, соответствующие различной степени пространственной корреляции. Число обусловленности $CN = 50$ соответствует каналу с меньшей пространственной корреляцией, число $CN = 500$ соответствует сильно коррелированному каналу.

Результаты моделирования показаны на рис. 1 в виде функции распределения суммарной скорости при прекодировании методом ZF с использованием непосредственных реализаций канала и с использованием оценки канала для вычисления весовых коэффициентов. Максимальное значение числа обусловленности в данном случае $CN = 50$. Для данных условий распространения прекодер ZF с точным знанием канала обеспечивает среднее значение спектральной эффективности 9.2 бит/сек/Гц. Прекодер ZF с оценкой канала LS в тех же условиях обеспечивает спектральную эффективность 8 бит/сек/Гц для 16 антенн на стороне БС.

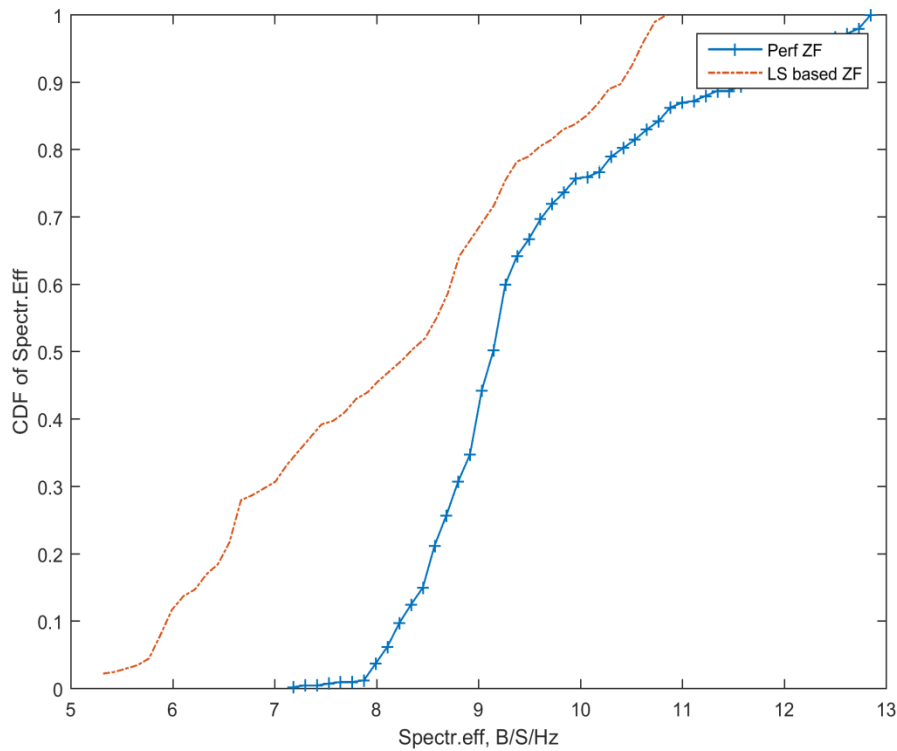


Рис. 1. Распределение спектральной эффективности при прекодировании ZF и $CN = 50$

На рис. 2 показано распределение суммарной скорости при прекодировании методом кодовых книг DFT с использованием непосредственных реализаций канала и с использованием оценки канала для поиска кодового слова. Максимальное значение числа обусловленности $CN = 50$. Для данных условий распространения прекодер DFT с точным знанием канала обеспечивает среднее значение спектральной эффективности 7.2 бит/сек/Гц. Прекодер DFT с оценкой канала LS в тех же условиях обеспечивает спектральную эффективность, равную 6.8 бит/сек/Гц, для 16 антенн на стороне БС.

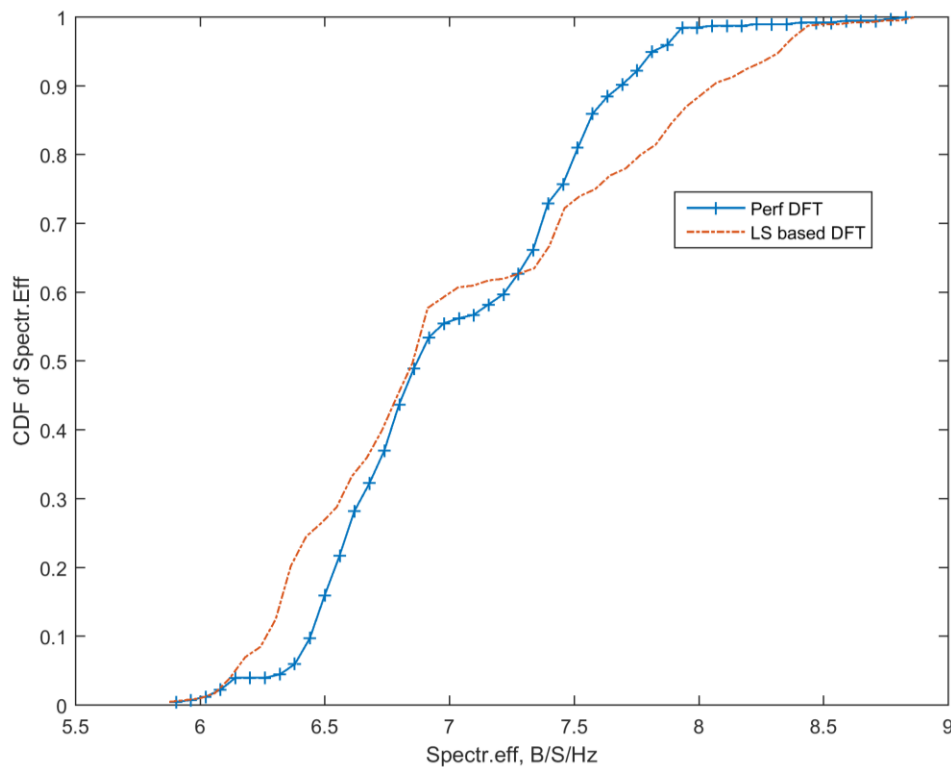


Рис. 2. Распределение спектральной эффективности при прекодировании DFT и $CN = 50$

На рис. 3 показано распределение спектральной эффективности прекодирования ZF при значении числа обусловленности $CN = 500$. Для данных условий распространения прекодер ZF с точным знанием канала обеспечивает среднее значение спектральной эффективности, равное 6 бит/сек/Гц. Прекодер ZF с оценкой канала LS в тех же условиях обеспечивает спектральную эффективность, равную 5.4 бит/сек/Гц, для 16 антенн на стороне БС.

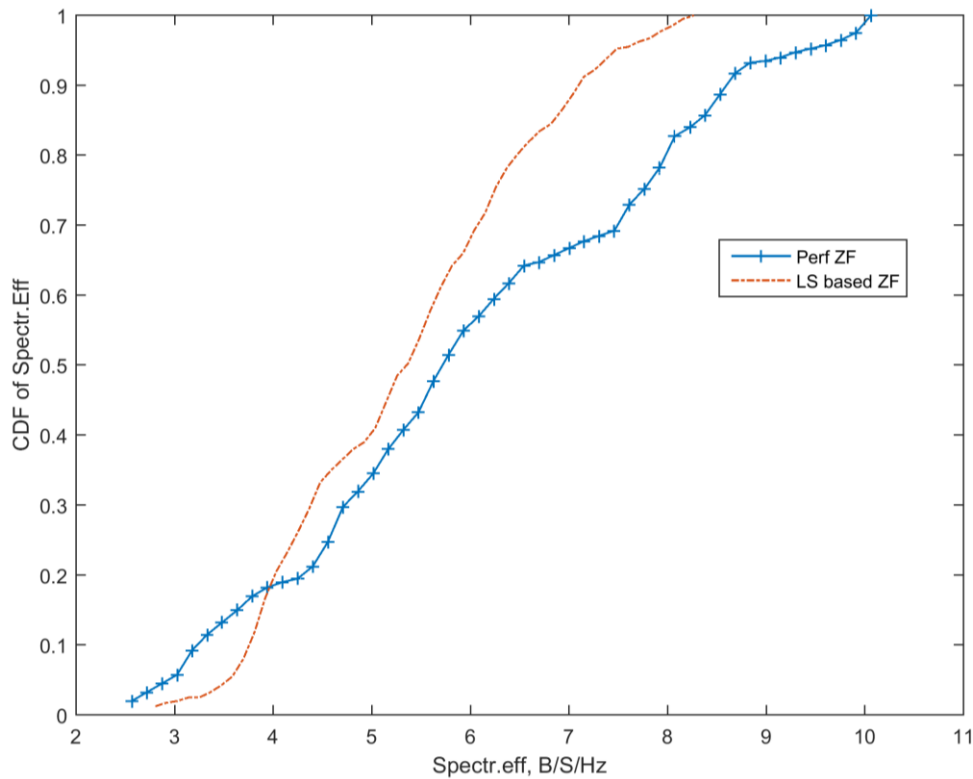


Рис. 3. Распределение спектральной эффективности при прекодировании ZF и $CN = 500$

На рис. 4 показано распределение суммарной скорости при прекодировании методом кодовых книг DFT с использованием непосредственных реализаций канала и с использованием оценки канала для поиска кодового слова. Максимальное значение числа обусловленности $CN = 500$. Для данных условий распространения прекодер DFT с точным знанием канала обеспечивает среднее значение спектральной эффективности, равное 6.8 бит/сек/Гц. Прекодер DFT с оценкой канала LS в тех же условиях обеспечивает спектральную эффективность, равную 6.5 бит/сек/Гц, для 16 антенн на стороне БС.

Можно заметить, что в этих условиях распространения при плохой обусловленности канала связи прекодер ZF показывает ухудшение характеристик спектральной эффективности по сравнению с прекодером DFT. Прекодер DFT более устойчив к ухудшению условий распространения и наличию ошибок оценки канала, что соответствует результатам, опубликованным в [9, 10].

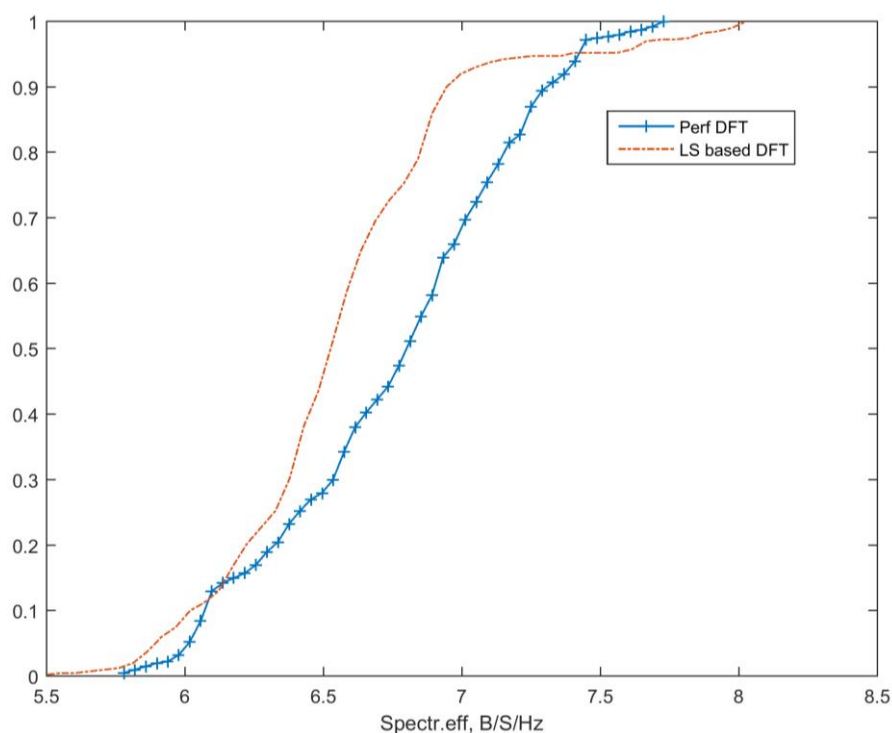


Рис. 4. Распределение спектральной эффективности при прекодировании DFT и $CN = 500$

4. Заключение

В работе представлены результаты численного моделирования двух видов прекодирования сигналов MU-MIMO в типовом сценарии Urban micro на основе модели канала QuaDRiGa. Оценивание канала на стороне абонентов выполнялось по методу LS, и результаты оценки канала используются для вычисления весовых векторов прекодирования. Наличие ошибок оценивания канала и пространственной корреляции канала снижает спектральную эффективность системы связи. При данных условиях распространения прекодер на основе кодовых книг DFT более устойчив к наличию ошибок оценивания и пространственной корреляции по сравнению с прекодером ZF по величине спектральной эффективности.

Литература

1. Bjornson E., Hoydis J., Sanguinetti L. Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency // *Foundations and Trends in Signal Processing*. 2017. V. 11, № 3–4. P. 154–655.
2. Castaneda E., Silva A., Gameiro A., and Kountouris M. An overview on resource allocation techniques for multi-user MIMO systems // *IEEE Communications Surveys and Tutorials*. 2017. V. 19, № 1. P. 239–284.
3. 3GPP TS 38.211. NR Physical channels and modulation – Release 15 Technical Report 38.211 V15.8.0, 2020.
4. Bayesteh A. and Khandani A. K. On the user selection for MIMO broadcast channels // *IEEE Trans. Inform. Theory*. 2008. V. 54, № 3. P. 1086–1107.
5. Yoo T. and Goldsmith A. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming // *IEEE J. Select. Areas Commun.* 2006. V. 24, № 3. P. 528–541.
6. Dimic G. and Sidiropoulos N. D. On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm // *IEEE Trans. Signal Processing*. 2005. V. 53, № 10. P. 3857–3868.

7. Kaltenberger F., Gaspert D. Performance of Multi-User MIMO Precoding with Limited Feedback over Measured Channels // Proc. IEEE GLOBECOM 2008, New Orleans, LA, USA.
8. Yong Soo Cho, Jaekwon Kim, Won Y. Yang, Chung G. Kang MIMO-OFDM Wireless Communications with MATLAB. John Wiley and Sons, 2010.
9. 3GPP. Study on Channel Model for Frequencies from 0.5 to 100 GHz, V16.1.0. Tech. Rep. 38.901, 2020.
10. Clerckx B., Kim G., and Sung J. Correlated fading in broadcast MIMO channels: curse or blessing? // Proc. IEEE IEEE GLOBECOM 2008, New Orleans, LA, USA.
11. Jaeckel S., Raschkowski L., Boerner K. and Thiele L., Burkhardt F., Eberlein E. QuaDRiGa – Quasi Deterministic Radio Channel Generator. User Manual and Documentation. Tech. Rep. v2.2.0, Fraunhofer Heinrich Hertz Institute, 2019.
12. Jaeckel S., Raschkowski L., Boerner K., and Thiele L. QuaDRiGa: A 3-D Multicell Channel Model with Time Evolution for Enabling Virtual Field Trials // IEEE Transactions on Antennas Propagation, 2014.

Калачиков Александр Александрович

к.т.н., доцент кафедры радиотехнических систем, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 267, e-mail: kalatch@sibguti.ru, ORCID ID: 0000-0003-1235-6314.

Автор прочитал и одобрил окончательный вариант рукописи.

Автор заявляет об отсутствии конфликта интересов.

Numerical Evaluation of the MU-MIMO Beamforming Performance with Channel Estimation

Aleksander A. Kalachikov

Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: This paper presents the numerical evaluation of the ZF beamforming algorithm and DFT precoding using the LS channel estimation in the multiuser multiantenna (MU-MIMO) downlink system. The sum rate performance depending on a number of users are presented. The arising inter user correlation degrades the sum rate (spectral efficiency) performance of multiuser MIMO system especially in scenarios where the number of users is larger than the number of antennas at the BS. For MIMO channel simulation the QuaDRiGa channel model reflecting the real propagation conditions is used. The obtained performance of MU-MIMO ZF precoding in spatially correlated channel are compared with DFT precoding based on the empirical cumulative density function of the sum rate of multiple users. Numerical results show that the ZF precoder outperforms the DFT precoder in channel with less spatial correlation. The DFT precoder in spatially more correlated channel has advantage over ZF precoder.

Keywords: 5G new radio, QuaDRiGa, 3GPP channel model, ZF precoding, DFT precoder, LS channel estimation.

For citation: Kalachikov A. A. Numerical evaluation of the MU-MIMO beamforming performance with channel estimation (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 109-120. <https://doi.org/10.55648/1998-6920-2024-18-1-109-120>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Kalachikov A. A., 2024

The article was submitted: 26.02.2023;
revised version: 21.07.2023;
accepted for publication 17.12.2023.

References

1. Bjornson E., Hoydis J., Sanguinetti L. Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency. *Foundations and Trends in Signal Processing*. 2017, vol. 11, no. 3-4, pp. 154-655.
2. Castaneda E., Silva A., Gameiro A., and Kountouris M. An overview on resource allocation techniques for multi-user MIMO systems. *IEEE Communications Surveys and Tutorials*, 2017, vol. 19, no. 1, pp. 239-284.
3. 3GPP TS 38.211. NR *Physical channels and modulation* - Release 15 Technical Report 38.211 V15.8.0, 3GPP, 2020.
4. Bayesteh A. and Khandani A. K. On the user selection for MIMO broadcast channels. *IEEE Trans. Inform. Theory* 54, 2008, no. 3, pp. 1086-1107.
5. Yoo T. and Goldsmith A. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming, *IEEE J. Select. Areas Commun.*, March 2006, vol. 24, no. 3, pp. 528-541.
6. Dimic G. and Sidiropoulos N. D. On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm. *IEEE Trans. Signal Processing*, October 2005, vol. 53, no. 10, pp. 3857-3868.
7. Kaltenberger F., Gespert D. Performance of Multi-User MIMO Precoding with Limited Feedback over Measured Channels. *Proc. IEEE GLOBECOM*, 2008, New Orleans, LA, USA.
8. Yong Soo Cho, Jaekwon Kim, Won Young Yang, Chung G. Kang. *MIMO-OFDM Wireless Communications with MATLAB*. John Wiley and Sons, 2010.
9. 3GPP (2020). Study on Channel Model for Frequencies from 0.5 to 100 GHz. *V16.1.0. Tech. Rep. 38.901*.
10. Clerckx B., Kim G., and Sung J. Correlated fading in broadcast MIMO channels: curse or blessing? *Proc. IEEE Global Telecommum. Conf. (GLOBECOM)*, New Orleans, November 2008.
11. Jaeckel S., Raschkowski L., Boerner K. and Thiele L., Burkhardt F., Eberlein E. QuaDRiGa - Quasi Deterministic Radio Channel Generator. User Manual and Documentation. *Tech. Rep. v2.2.0*, Fraunhofer Heinrich Hertz Institute, 2019.
12. Jaeckel S., Raschkowski L., Boerner K. and Thiele L. QuaDRiGa: A 3-D Multicell Channel Model with Time Evolution for Enabling Virtual Field Trials. *IEEE Transactions on Antennas Propagation*, 2014.

Aleksander A. Kalachikov

Cand. of Sci. (Engineering), Lecturer of the Department of Radio Systems Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), phone. +7 383 2698267, e-mail: kalatch@sibguti.ru, ORCID ID: 0000-0003-1235-6314.

Математический аппарат для оценивания некоторых показателей качества учебного вебинара посредством машинного обучения

В. В. Подколзин¹, А. Н. Полетайкин^{1,2}, М. Ю. Галкина²

¹ Кубанский государственный университет (КубГУ)

² Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: Данная статья является вторым этапом создания и исследования технологии оценивания качества контактной работы в вузе, реализуемой посредством вебинаров. Рассматривается математическое обеспечение интеллектуальной технологии анализа аудиовидеопотока записи учебного вебинара. Решается задача повышения эффективности и достоверности оценивания некоторых показателей качества контактной работы в институте заочного образования СибГУТИ, реализуемой посредством вебинаров, в ходе дистанционного обучения. Для анализа применяются методы машинного обучения и их реализация с использованием стандартных модулей и библиотек Python. Научная значимость работы состоит в разработке новых моделей и алгоритмов интеллектуального анализа аудиовидеопотока, параметризованных для обработки записи учебного мероприятия. Разработанные модели и алгоритмы позволяют повысить эффективность процесса оценивания качества контактной работы, реализуемой посредством вебинаров.

Ключевые слова: вебинар, область конференции, активность ведущего, критерии оценивания эффективности, интервалы времени, локализация текста, обработка аудиотрека, обработка изображений, машинное обучение.

Для цитирования: Подколзин В. В., Полетайкин А. Н., Галкина М. Ю. Математический аппарат для оценивания некоторых показателей качества учебного вебинара посредством машинного обучения // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 121–143. <https://doi.org/10.55648/1998-6920-2024-18-1-121-143>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Подколзин В. В., Полетайкин А. Н.,
Галкина М. Ю., 2024

Статья поступила в редакцию 15.12.2023;
принята к публикации 15.01.2024.

1. Введение

В Сибирском государственном университете телекоммуникаций и информатики (СибГУТИ) в рамках программы цифрового развития университета поставлена задача оценивания качества контактной работы, реализуемой посредством вебинаров, в ходе дистанционного обучения. Разработанная ранее математическая модель, базирующаяся на системе из 32 показателей качества вебинара и опубликованная в статье [1], работает преимущественно на методах экспертного оценивания и поэтому обладает высокой субъективностью и требует значительных затрат времени. Необходимо более простое и эффективное решение, характеризующееся приемлемой степенью объективности оценок.

Вебинар (от англ. webinar – web based seminar, онлайн-семинар, онлайн-конференция) – онлайн-трансляция с целью обучения или обеспечения корпоративной встречи [2]. В данной статье рассматриваются лишь обучающие вебинары. В вебинаре всегда участвуют две

стороны: докладчик-ведущий и слушатели-участники. Как правило, участники могут видеть ведущего, оставаясь при этом за кадром. Предполагается, что в течение вебинара все участники слушают ведущего и реагируют на его активности вовлечения слушателей в предмет обсуждения в основном в чате, реже – посредством выхода в эфир. Вебинар может проходить как в прямой трансляции, так и быть доступным в записи. К техническим возможностям вебинара можно отнести следующие:

- демонстрация презентаций;
- возможность быстро получать и сразу видеть точные результаты голосований и опросов среди участников вебинара;
- передача участникам нужных файлов и ссылок;
- обратная связь – в виде чата, куда можно написать свой вопрос или просто поделиться впечатлениями (если вебинар в прямой трансляции).

Для того чтобы оценить эффективность вебинаров, используются разнообразные критерии оценки. Чаще всего к ним относят такие общие критерии:

- количество зарегистрированных пользователей;
- распространение информации о предстоящем вебинаре;
- количество участников вебинара;
- активность участников.

Количество зарегистрированных пользователей показывает, насколько выбранная тема востребована у целевой аудитории. Распространение информации о предстоящем вебинаре, например, в социальных сетях, говорит о том, что есть пользователи, которые проявляют интерес к вебинару и готовы рекомендовать его другим людям. Количество участников на вебинаре – индикатор интереса слушателей к теме вебинара. Чем активнее участники вебинара будут взаимодействовать с ведущим, тем больше вероятность того, что они, например, оставят заявку, подпишутся на рассылку, посмотрят запись вебинара [3].

В 2018–2019 годах в СибГУТИ было проведено комплексное исследование [1], в результате которого разработана математическая модель, включающая систему из 32 показателей качества организации и проведения вебинара, позволяющих оценить с разных сторон качество вебинара с помощью 10-балльных экспертных оценок. Проблема заключается в существенных трудозатратах, которые несут эксперты при оценивании этих показателей. Кроме того, мнения экспертов субъективны. Целью дальнейших исследований в статье [1] заявлено сокращение трудоемкости работы экспертов и повышение объективности получаемых оценок качества вебинаров. Это достигается за счет сведения к минимуму влияния человеческого фактора в процедуре оценивания и применения для оценивания некоторых показателей качества вебинаров количественных математических методов, относящихся к категории Data Mining. Поэтому актуальной является задача разработки математических средств, которые позволят оценить большую часть показателей без участия человека в автоматическом режиме.

2. Постановка задачи

Предварительный анализ показателей из статьи [1], полный перечень которых представлен в приложении к данной статье, позволил выявить 17 таких (см. табл. 1), которые можно оценить без участия человека посредством компьютерного анализа аудиовидеозаписи вебинара. На рис. 1 показан стоп-кадр видеопотока аудиовидеозаписи вебинара на платформе Mirapolis (<https://www.mirapolis.ru>), где выделены характерные области, анализ которых позволяет определить величины показателей. Соответствие между показателями и областями кадра показано в табл. 1.

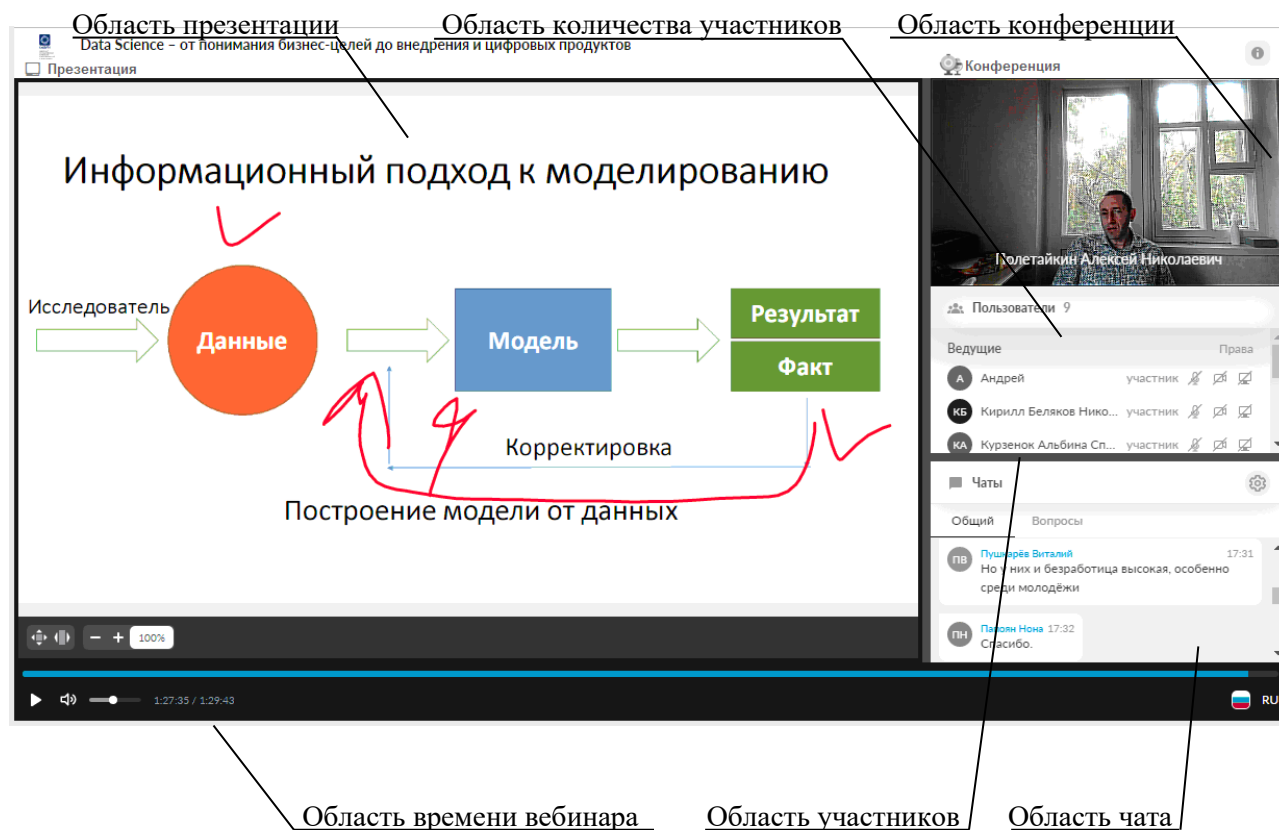


Рис. 1. Стоп-кадр аудиовидеопотока аудиовидеозаписи вебинара и его характерные области

Таблица 1. Некоторые показатели качества вебинара, которые могут быть рассчитаны на основе компьютерного анализа аудиовидеозаписи вебинара (нумерация показателей соответствует таблице в приложении, сложность оценена по шкале из табл. 2)

№	Название показателя	Сложность	Источник информации о показателе
7	Лаконичность	2	Область презентации
9	Контактные данные	1	Область презентации
10	Энергичность и эмоциональность	2	Область конференции / звуковой ряд
11	Темп изложения	2	Звуковой ряд
13	Внешний вид	1	Область конференции / звуковой ряд
14	Количество и качество информации на слайдах	1	Область презентации
15	Количество и качество иллюстраций	1	Область презентации
16	Аккуратность	1	Область презентации
17	Использование указки	3	Область презентации
18	Разнообразие презентационных инструментов	3	Область презентации / область чата
20	Проведение опросов	3	Область презентации
23	Частота взаимодействия	3	Область чата
24	Наличие вопросов слушателей	1	Область чата
25	Реактивность на вопросы	4	Звуковой ряд
26	Качество ответов на вопросы	4	Звуковой ряд
31	Точность заявленной длительности мероприятия	1	Область времени вебинара
32	Показатели посещаемости	2	Область количества участников

Таблица 2. Шкала ранжирования показателей качества вебинара по временной сложности процедуры экспертного оценивания [1]

Балл	Описание трудозатрат	Затраты времени, мин.
0	Не требует внимания экспертов	0
1	Минимальные затраты (чат, материалы, ведущий)	2–5
2	Беглый эпизодический просмотр вебинара	5–10
3	Внимательный просмотр нескольких эпизодов вебинара	10–30
4	Беглый просмотр вебинара целиком	20–40

В данной статье будут рассмотрены показатели 7, 9 – 11, 14 – 17, 31 и 32 и выработаны научно обоснованные математические решения для их эффективного оценивания. Для этого необходим в основном компьютерный анализ изображений слайдов из области презентации. Для оценивания показателей 10 и 13 необходим анализ изображений ведущего из области конференции. Оценивание показателей 31 и 32 требует распознавания чисел из областей времени вебинара и области количества участников соответственно.

Анализ области презентации сводится к анализу изображений каждого слайда презентации. Факт того, что демонстрируется именно презентация, а не видеоролик или рабочий стол ведущего, удостоверяется наличием заголовка «Презентация», а также инструментов масштабирования (— +) в нижней части области.

Показатель 7 *Лаконичность* предполагает в наиболее простом варианте идентификацию рисунков, схем, диаграмм и других изображений, а также тестовых блоков по объёму, в т.ч. в таблицах (как правило, чем меньше объём блока текста, тем больше его лаконичность). Результатом анализа должно быть число и классификация текстовых блоков по категориям лаконичности, а также общее число изображений. На основе этих данных будет рассчитано значение показателя 7, а также показатель 15 *Количество и качество иллюстраций*.

Для оценивания показателя 14 *Количество и качество информации на слайдах* необходимо для каждого слайда презентации рассчитать отношение к общей площади пространства слайда площади, занимаемой на слайде: 1) изображениями и 2) текстовыми блоками.

Показатель 16 *Аккуратность* предполагает идентификацию текстовых блоков по размеру шрифта и цвету, списков по размеру шрифта и маркерам, изображений по размеру и положению на слайде (по координатам левого верхнего угла).

Расчёт показателя 17 *Использование указки*: необходимо идентифицировать указку и сформировать массив, элементы которого соответствуют сеансам использования указки (под сеансами понимаются интервалы между паузами в использовании указки длительностью более 10 секунд) и содержат суммарную длину траектории ее перемещения (криволинейную траекторию можно рассматривать как ломаную).

Показатель 9 *Контактные данные* предполагает идентификацию на слайдах телефонных номеров и электронных адресов по их характерным признакам. Здесь, однако, существует проблема идентификации посторонних контактных данных, например, в изображениях оконного интерфейса с контактными данными. Поэтому необходимо предусмотреть этот случай и такие данные игнорировать. Результатом анализа должен быть факт наличия в презентации контактных данных ведущего.

Распознавание числовых данных требуется для анализа таких показателей, как 31 *Точность заявленной длительности мероприятия* и 32 *Показатели посещаемости*. В частности, необходимо идентифицировать в области времени вебинара момент времени:

- 1) первого включения изображения ведущего в области конференции,
- 2) последнего отключения изображения ведущего в области конференции,

а также идентифицировать в области числа участников максимальное число участников между моментами времени 1 и 2 и число участников в момент времени 2.

Показатель 10 *Энергичность и эмоциональность* требует анализа динамичности изображения ведущего в области конференции – по крайней мере, мимики лица ведущего и его

положения в области конференции.

Таким образом, для оценивания указанных показателей необходимо осуществлять распознавание изображений в видеопотоке вебинара. Для решения такого рода задач используются искусственные нейронные сети и программное обеспечение на их основе.

Показатель 11 *Темп изложения* требует анализа аудиотрека вебинара на предмет вычисления скважности его звукового ряда, для этого применяются инструменты работы с аудиоданными на основе дискретного преобразования Фурье.

Целью исследования является повышение оперативности и достоверности оценивания некоторых показателей качества вебинара. Объектом исследования является категория качества учебного вебинара, представленного аудиовидеопотоком его записи. Предмет исследования – нейросетевые математические модели и специальное программное обеспечение на их основе при решении задачи оценивания качества вебинара.

Решением задачи должна стать программная среда, разработанная на языке Python и реализующая разработанный математический аппарат в отношении заданной аудиовидеозаписи вебинара.

3. Выбор инструментария для решения задачи

Для программной реализации выбран язык программирования Python, к главным чертам которого относятся динамическая типизация, понятный синтаксис, поддержка объектно-ориентированного программирования, а также наличие обширной библиотечной базы (в том числе для задач машинного обучения) и возможность легкой установки библиотечных модулей, что значительно расширяет функционал программы и упрощает её создание. Важным отличием Python является наличие множества свободно распространяемых библиотек. В соответствии с постановкой задачи к основным выбранным модулям и библиотекам можно отнести:

- NumPy – для работы с многомерными массивами и использования некоторых функций линейной алгебры, например, для расчёта траектории движения указки или карандаша ведущего при оценивании показателя 17;
- LibROSA – для построения и анализа спектрограмм аудиопотока вебинара, например, для оценивания энергичности и эмоциональной насыщенности вебинара по показателю 10, а также темпа изложения (показатель 11);
- SciPy и Scikit-learn – для решения задач класса Data Mining, таких как кластеризация, распознавание изображений и работа с ними;
- PyQt5 – для работы с компонентами графического интерфейса пользователя, виджетами (библиотека написана на языке C++, поэтому требует установки средств разработки Visual C++);
- Cv2 – для реализации методов машинного обучения OpenCV.

4. Алгоритм работы с аудиоданными

Обработка аудиотрека вебинара используется для определения трёх видов интервалов времени:

- интервалы времени тишины;
- интервалы времени разговора ведущего;
- остальные интервалы времени.

Для извлечения аудиотрека из записи вебинара применяется модуль `moviepy.editor` и функция `VideoFileClip(filename = pathVideo + nameVideo)`, принимающая параметр `filename`, которая определяет путь к файлу аудиовидеозаписи вебинара:

```
import moviepy.editor as mp
video_clip = mp.VideoFileClip(filename = pathVideo + nameVideo)
audio_clip = video_clip.audio
```

В представленном примере с помощью библиотеки Python `moviepy.editor` сначала загружается видеотрек вебинара, затем извлекается из него аудиотрек в формате `wav`.

Первым шагом в работе с аудиопотоком является фильтрация. Для данной задачи был выбран полосовой фильтр Баттерворта [4], АЧХ которого является достаточно гладкой на частотах полосы пропускания, соответствующих спектру голоса человека. Используется фильтр с бесконечной импульсной характеристикой (БИХ-фильтр), которая вычисляется по рекурсивной формуле:

$$y(n) = \sum_{k=0}^N b_k x(n-k) - \sum_{k=1}^M a_k y(n-k),$$

где N – порядок фильтра прямой связи; M – порядок фильтра обратной связи ($M \leq N$); a_k и b_k – коэффициенты фильтра. Для вычисления коэффициентов применяется функция `butter(order, wn, btype)`, где `order` – порядок фильтра, `wn` – диапазон частот (частота среза), `btype` – тип фильтра. Для применения полученных коэффициентов используется функция `lfilter`: `filtered = lfilter(b, a, x)`, где `b`, `a` – полученные с помощью функции `butter` коэффициенты свертки; `x` – данные, которые нужно отфильтровать.

4.1. Извлечение интервалов времени тишины

К самым неинформативным частям аудиопотока относятся не только участки шума, но и участки тишины (участки, где нет никаких голосов). Речевой сигнал, содержащий одно слово, как минимум содержит паузу до и после произнесения слова. В случае с голосовым сигналом, содержащим высказывание из нескольких слов, участками такой информации выступают паузы между словами.

Среди алгоритмов выделения границ речи наибольшую популярность получил метод, основанный на энергии и количестве переходов сигнала через нулевое значение в короткие промежутки времени, и алгоритм, основанный на плотности распределения значения отсчетов паузы. Однако методы на основе кратковременной энергии речевого сигнала и числа нулей интенсивности обладают менее высоким качеством. Чаще используется метод выделения речи на основе нормального распределения.

В алгоритме на основе нормального распределения случайной величины в качестве таковой выступают значения отсчетов паузы (фоновый шум). Таким образом, некоторые свойства функций нормального распределения, которые имеются в сигнале паузы (тишины, шума), будут являться критериями и использоваться для сегментации звука на сигнал и паузу.

Для выделения интервалов времени тишины используется понятие нормального распределения и одномерного расстояния Махаланобиса. При этом алгоритм извлечения паузы состоит из следующих шагов:

- 1) вычисление статистических характеристик шума;
- 2) выполнение нормализации сигнала по амплитуде;
- 3) выполнение фрагментирования всего аудиосигнала;
- 4) вычисление расстояния Махаланобиса и выполнение проверки.

Вначале происходит вычисление статистических характеристик распределения точек c_k на множестве (мощности m) точек выделенного аудиотрека: МО $\bar{c}$ и СКО σ первых 200 мс аудиотрека:

$$\bar{c} = \frac{1}{m} \sum_{k=1}^m c_k, \quad \sigma = \sqrt{\frac{1}{m} \sum_{k=1}^m (c_k - \bar{c})^2}.$$

Этот участок аудиопотока считается паузой. После чего происходит нормализация сигнала по амплитуде. Для этого используется фильтр с конечной импульсной характеристикой (КИХ-фильтр):

$$\bar{s}(n) = s(n) - 0.95 \cdot s(n-1),$$

где $\bar{s}(n)$ – отсчет n в выходном отрезке $\bar{s}$; $s(n)$ – отсчет n во входном отрезке s . Нормализованный аудиосигнал разбивается на отрезки по 25 мс. Для каждого такого отрезка от точки x вычисляется одномерное расстояние Махаланобиса:

$$d_M(x) = \frac{|x - \bar{c}|}{\sigma}.$$

Если это расстояние больше 4, то отрезок считается сигналом (голосом), иначе – паузой.

4.2. Извлечение интервалов времени разговора ведущего

Для извлечения интервалов времени разговора ведущего из полученных интервалов времени голоса (см. подраздел 4.1) реализованы ручной и полуавтоматический способы. В ручном режиме пользователь вручную на основе уже выделенных интервалов голоса вводит интервалы времени, в которые ведущий не говорит. При работе полуавтоматического режима решается задача идентификации личности по голосу, которая основана на парном сравнении биометрических признаков (содержащихся в голосе индивидуальных значимых признаков личности), выделенных из речи в фонограмме. Структура процесса распознавания личности по голосу в данной работе изображена на рис. 2.

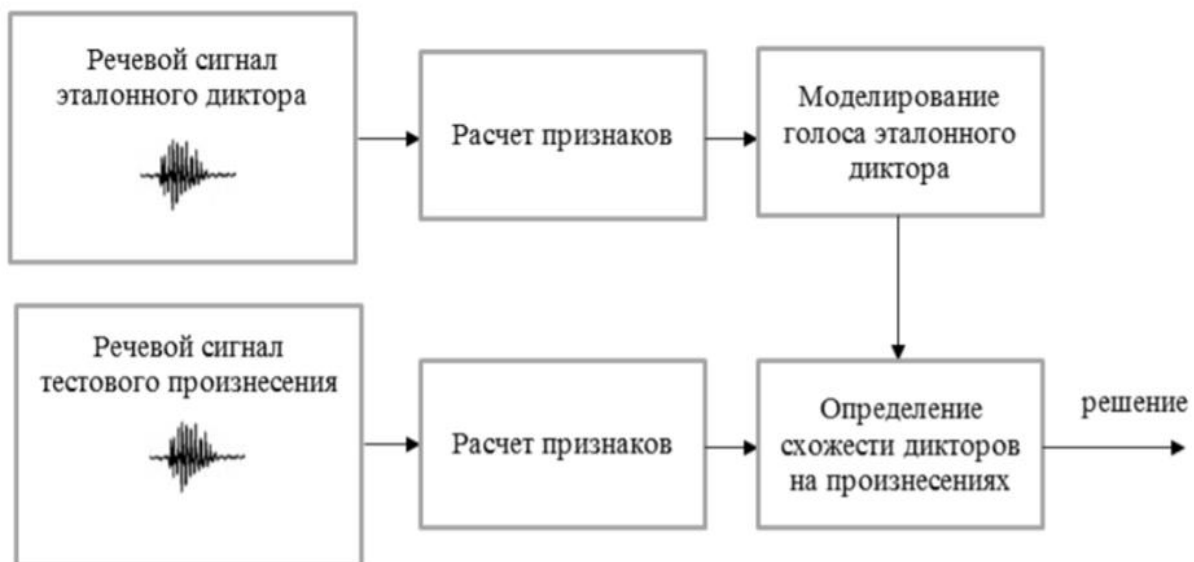


Рис. 2. Структура процесса распознавания личности по голосу

Среди методов обучения текстонезависимого (обучаемого и тестируемого на произвольных речевых данных) идентификатора наиболее используемым является метод моделей гауссовых смесей (Gaussian mixture models, GMM). Создается GMM с помощью функции `gmm = GaussianMixture(n_components, max_iter, covariance_type, n_init)`, где `gmm` – выходная GMM; `n_components` – количество компонент GMM; `max_iter` – количество итераций обучения GMM; `covariance_type` – тип ковариационной матрицы; `n_init` – количество начальных инициализаций при обучении (K-means).

Обучение GMM проводится с помощью метода `fit` построенной GMM. Её синтаксис: `gmm.fit(X)`, где `X` – обучающие данные.

В роли речевых признаков в большинстве систем идентификации личности по голосу используются мел-частотные кепстральные коэффициенты (mel-frequency cepstral coefficients, MFCC). MFCC извлекаются с помощью библиотечной функции `python_speech_features.mfcc(data, sr, winlen, winstep, numcep)`, где `data` – данные аудиопотока, из которых извлекаются коэффициенты; `sr` – частота дискретизации аудиосигнала; `winlen` – длина окна в секундах, на которые будут разбиты `data`; `winstep` – длина шага по `data` (определяет, будут ли перекрываться окна); `numcep` – количество получаемых коэффициентов на каждом окне.

Таким образом, в полуавтоматическом режиме пользователь вручную вводит один небольшой интервал времени, в котором говорит ведущий. На основе введенного интервала идет обучение GMM-модели голоса ведущего. Затем идет разбиение всего аудиопотока на отрезки по 25 мс и проверка принадлежности обученной модели каждого отрезка. После разбиения отрезков на 2 части пользователю предлагается вручную откорректировать полученные интервалы времени, в которых не говорит ведущий. Исходя из того, что в вебинаре большую часть времени говорит ведущий, корректировка таких интервалов более целесообразна.

5. Алгоритм работы с изображениями

Как отмечалось в разделе 2, большая часть показателей рассчитывается на основе анализа видеопотока вебинара. Для этого использованы средства стандартной библиотеки OpenCV, которая предоставляет множество алгоритмов для обработки изображений и распознавания текста [5].

Рассмотрим работу с областью конференции вебинара. Для начала эту область необходимо выделить из видеопотока. Для этого интервал времени между двумя последовательно взятыми кадрами вебинара выбран равным значению параметра длины окна разбиения аудиосигнала (25 мс, см. раздел 4.1). Предполагается, что все области в вебинаре имеют стандартное относительное положение. В частности, область конференции расположена в правой верхней части кадра (см. рис. 1).

5.1. Предобработка данных

Прежде чем начать работать со слайдом видеопотока вебинара, необходимо произвести его предобработку. Алгоритм предобработки слайда согласно [6]:

- слайд переводится с помощью порогового преобразования `Threshold` в бинарное изображение (`CV_THRESH_BINARY`);
- получившееся изображение переводится в цветовое пространство HSV с помощью функции `cvtColor`.

После предобработки на слайде находятся контуры. Результатом будет бинарное изображение, в котором содержатся лишь контуры. Эту операцию проводит функция `canny = cv2.Canny(image, threshold1, threshold2)`, где `image` – одноканальное изображение для обработки (в градации серого, в данном случае оно бинарное); `threshold1` – порог минимума; `threshold2` – порог максимума.

Для того чтобы оперировать контурами и хранить их, необходимо заполнить матрицу точек, которая будет хранить двумерные координаты контуров изображения. Для этого используется функция `contours, hierarchy = cv2.findContours(image, mode, method, offset)`, где `image` – исходное одноканальное изображение (ненулевые пиксели обрабатываются как 1, а нулевые как 0, для получения такого изображения из градаций серого можно, например, использовать функции `Threshold` или `Canny`); `contours` – координаты

контуров на выходе; `hierarchy` – иерархия контуров (при использовании соответствующих режимов поиска); `mode` – режим поиска; `method` – метод аппроксимации. Так как все области вебинара имеют прямоугольную форму, то в качестве режима поиска используется режим `RETR_EXTERNAL`, обеспечивающий поиск только крайних внешних контуров, а в качестве метода аппроксимации и хранения в программе используется метод `CHAIN_APPROX_SIMPLE`, так как в этом режиме они будут кодироваться лишь с помощью четырёх угловых точек. При этом анализируются не сами контуры, а минимальные прямоугольники, которые их покрывают, что исключает возможность работы с неровными фигурами контуров. Для получения минимального прямоугольника, покрывающего выделенный почти прямоугольный контур, используется стандартная функция: `minAreaRect = cv2.minAreaRect(points)`, где `points` – последовательность или массив точек (в нашем случае – контур, который необходимо обвести).

Далее необходимо произвести фильтрацию выделенных областей. Из предположения, что границы областей вебинара являются самыми большими прямоугольниками, первичная фильтрация проходит по размеру полученных прямоугольников. Вторичная фильтрация проводится по заголовкам (по надписи над полученными прямоугольниками, если она есть). Для проведения вторичной фильтрации необходимо выделить в отдельное изображение и рассмотреть небольшие прямоугольники над полученными ранее, которые должны содержать в себе название областей вебинара (на рис. 1 – презентация, конференция, пользователи, чаты). После их получения необходимо провести их предобработку, прежде чем будет произведена процедура поиска и распознавания в нём текста. Сначала необходимо увеличить размер этих изображений при помощи функции `out_im = cv2.resize(in_im, size)`, где `in_im` – входное изображение; `out_im` – выходное изображение; `size` – размер выходного изображения. Далее изображение с помощью порогового преобразования `Threshold` переводится в бинарное изображение (`CV_THRESH_BINARY`) и в цветовое пространство `Gray` с помощью функции `cvtColor`.

5.2. Обнаружение лица ведущего и дальнейшая его обработка

Для обнаружения лица ведущего была использована функция `faces = dlib.get_frontal_face_detector().detector(image)`, где `image` – входное изображение области конференции; `faces` – лист координат всех лиц на входном изображении (в нашем случае – координаты одного лица ведущего вебинара). В результате получены координаты лица ведущего в положении анфас (лицом к смотрящему) на слайде. Данная функция использует метод HOG, который эффективно идентифицирует лицо ведущего, представленное в цветовом пространстве `Gray` [7], так как данные о цвете не нужны для решаемой задачи. Если хотя бы на одном из двух анализируемых слайдов лицо ведущего не было найдено, то считается, что ведущий отвлекся и ничего не говорит. Дальнейшие действия будут происходить в прямоугольнике, в котором находится лицо ведущего вебинара.

Для анализа динамики в изображении лица ведущего используется оценка положения 68 лицевых ориентиров (ключевых точек) лица человека (см. рис. 3), предложенная Вахидом Кэземи и Джозефином Салливаном [8]. Она основана на ансамбле деревьев регрессии (деревьев решений), обучение которых осуществляется с помощью алгоритма градиентного бустинга деревьев с суммой квадратичных потерь. Для нахождения координат лицевых точек используются функции `predictor=dlib.shape_predictor(shapes_dat_file)` и `shape = predictor(image, face)`, где `shapes_dat_file` – путь к отдельному файлу формата `dat`, который содержит уже обученный на поиск лицевых точек каскад деревьев; `predictor` – загруженный детектор лицевых точек; `image` – входное изображение области конференции; `face` – координаты лица ведущего на `image`; `shape` – координаты лицевых точек на слайде.

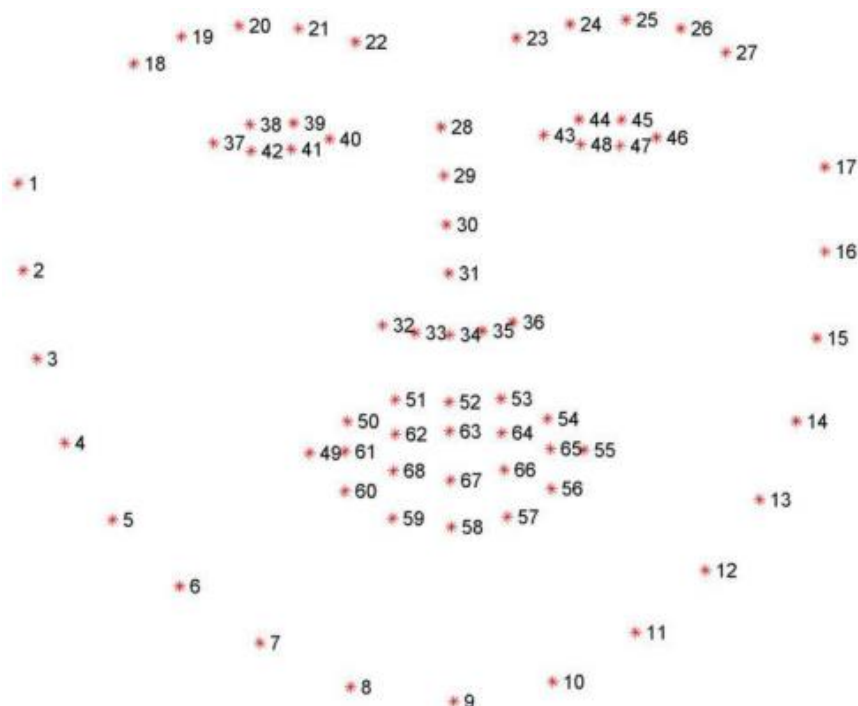


Рис. 3. Локализация антропометрических точек локализации лицевых ориентиров [8]

На рис. 3 видно, что для проведения анализа необходимо лишь 20 точек лица в области губ ведущего (с 49 по 68). Чтобы зафиксировать движение, считается расстояние Евклида между соответствующими точками в двух рассматриваемых слайдах. Если количество точек, чье расстояние попало в заданный интервал, больше половины списка анализируемых точек, то считается, что движение произошло. Интервал допустимых расстояний задается с целью фильтрации слишком больших или незримых амплитудных движений. Таким образом, были выделены все интервалы времени для решения поставленной задачи.

5.3. Анализ движения головы ведущего

Для фиксирования движения ведущего всеми остальными частями тела, кроме головы, используется оставшая часть области конференции вебинара. Чтобы понять, как изменился кадр по сравнению с предыдущим, разработана функция вычитания изображений `absdiff(im1, im2, res)`, где `im1`, `im2` – входные изображения, которые вычитаются; `res` – выходное изображение. Полученное изображение перед анализом проходит предобработку до бинарного. Для получения более точного результата и более качественного контроля выполняется морфологическая операция сужения изображения – увеличение темных областей прямоугольным ядром 5×5 . Изображение, полученное в результате, показано на рис. 4.

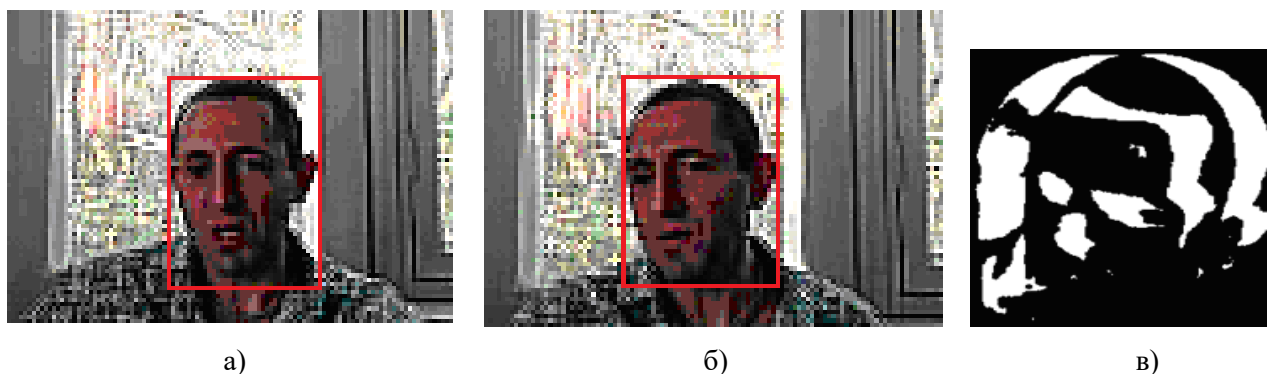


Рис. 4. Результат разности двух кадров изображения ведущего в области конференции
а) входное изображение `im1`; б) входное изображение `im2`; в) выходное изображение `res`

После завершения предобработки всё пространство, которое находится вокруг прямоугольника, описывающего голову ведущего, разделяется на 4 части: северная, южная, западная и восточная. Если хотя бы в одной из них количество ненулевых пикселей будет отличаться от нуля (это означает, что ведущий совершил движение), то счетчик движения увеличивается на 1.

Проведенный анализ области конференции видеопотока анализируемого вебинара при установленных амплитудах 10, 10, 10 по осям X, Y, Z соответственно и интервале 1000 мс между анализируемыми кадрами показал более точный результат по сравнению с интервалами времени 100 или 3000 мс. Это объясняется тем, что при установлении малого интервала времени движения проходят более плавно. Это означает, что амплитуда движения будет также малой, что приводит к потере (неучитыванию) некоторых совершённых ведущим движений. При установлении большого интервала времени также возможна потеря некоторых движений, так как они могли быть совершены в нерассматриваемых кадрах. Таким образом, наиболее подходящие интервалы времени находятся примерно на отрезке [500, 1500] мс. Так, при интервале 100 (3000) мс в течение всего вебинара ведущим было совершено 632 (1048), 360 (978), 462 (698) движений головы по осям X, Y, Z соответственно и 545 (899) движений остальных частей тела. При интервале 1000 мс в течение всего вебинара ведущим было совершено 2286, 1878, 1414 движений головы по осям X, Y, Z соответственно и 1827 движений остальных частей тела. На рис. 5 представлена зависимость точности результата от выбранного интервала времени.

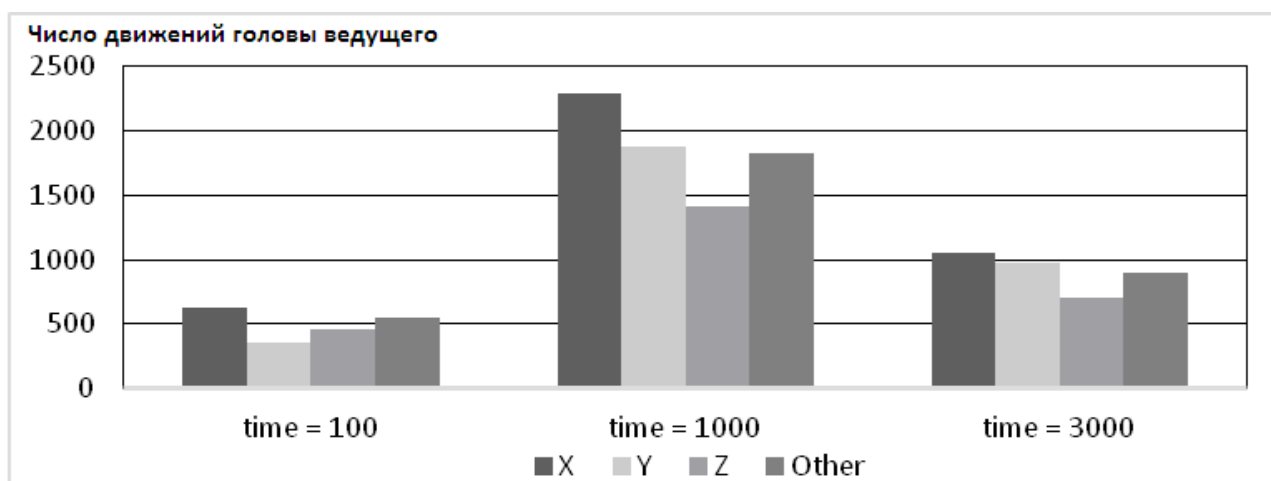


Рис. 5. Диаграмма зависимости зафиксированного числа движений головы ведущего вебинара от выбранного интервала времени в миллисекундах

5.4. Локализация текста на изображениях

На сегодняшний день алгоритмы локализации текста развиты достаточно хорошо. Тем не менее они не могут дать абсолютно точный результат, так как входные данные могут быть подвержены искажениям. Основные из них: низкое разрешение, сложный фон, неравномерное освещение, геометрические искажения, дисторсия, шумы и др. В отношении анализа слайдов презентации вебинара актуальны лишь первые два недостатка. Поэтому достоверность результатов локализации текста (а именно локализация необходима для оценивания показателей 14–16, см. табл. 1) во многом зависит от качества исполнения слайдов презентации, которые демонстрируются на кадрах видеопотока вебинара.

Первый этап большинства алгоритмов распознавания текста – его локализация. Главное отличие текста от фона и изображений – высокий контраст. Также предполагается, что текст имеет определенный цвет и яркость, что и отделяет его от фона. Кроме того, текст презентации чаще всего выполнен одинаковым шрифтом постоянного размера. Для таких задач хорошо работают градиентные методы локализации текста [9]. Идентификация блоков текста

осуществляется аналогично выделению области конференции в режиме RETR_EXTERNAL (см. подраздел 5.2). На последнем этапе алгоритм выделения текста требует выполнения процедуры фильтрации контуров, для чего используется несколько работающих последовательно фильтров. Первый из них удаляет все вложенные друг в друга контуры и те, чья площадь была меньше допустимой. Второй фильтр пытается, кроме этого, отличить текстовый блок от других изображений, основываясь на том факте, что текстовый блок можно разбить на строки, строки можно разбить на слова, а слова разбить на символы. Формы символов, в свою очередь, имеют определенные ограничения. В реализации для каждой строки вычисляется величина: *высота строки / среднее арифметическое ширины символа в строке*. Далее вычисляется медиана ряда полученных выше величин. Эта медиана m проверяется неравенством: $1 \leq m \leq 3.5$. Если неравенство выполняется, то данный блок считается текстовым, иначе отбрасывается. Для выделения строк текста для заданной области вычисляется вертикальная гистограмма как вектор, длина которого равна высоте выделенной области, а каждый элемент равен количеству небелых пикселей в соответствующей строке. Каждая непрерывная последовательность ненулевых элементов вектора определяется границей строки текста.

Пример работы алгоритма выделения текста представлен на рис. 6. С использованием полученного методом `findContours` массива точек (см. подраздел 5.1) рассчитывается площадь, занимаемая выделенными на слайде блоками. В данном примере текстовые блоки занимают 0.719 пространства слайда.

5.5. Распознавание текста на изображениях

Алгоритм распознавания текста на изображении состоит из следующих шагов:

1. Предобработка изображения посредством масштабирования, поворота и сегментации текста.
2. Применение сверточных слоёв для извлечения признаков из изображения.
3. Объединение извлеченных признаков и преобразование в единый вектор признаков.
4. Применение полносвязных слоёв для классификации полученного вектора признаков.
5. Получение результата распознавания текста.

Предобработка изображения рассмотрена в подразделе 5.1. Наряду с этим использован метод билинейной интерполяции, который позволяет уменьшить размерность изображения, сгладив пиксели, сохраняя при этом основные детали. Уменьшение размерности изображения позволяет снизить объём данных, с которыми необходимо работать, что повышает эффективность работы алгоритмов распознавания текста [10].

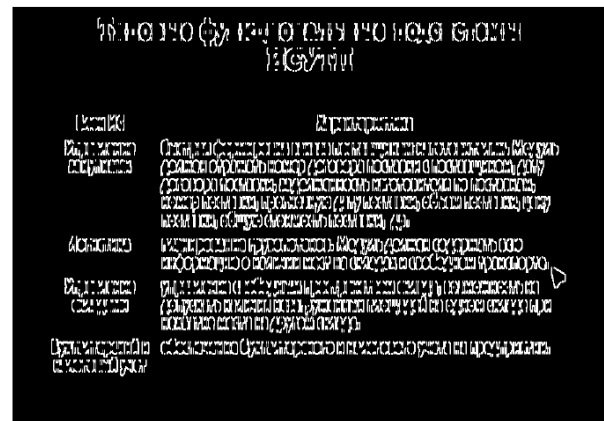
Шаги 2–4 реализуют процедуру поиска и распознавания текста на изображении с помощью библиотеки Tesseract OCR, в которой реализованы нейросетевые алгоритмы машинного обучения для распознавания текста на изображениях. Это открытое ПО, которое позволяет автоматически распознавать как единичную букву, так и сразу текст. Его существенным недостатком является плохая работа с замыленным, битым, грязным и деформированным текстом, однако авторы исходят из того, что область кадра вебинара не содержит таких повреждений текста.

Для получения текста на изображении (в нашем случае – заголовка области вебинара) используется функция `text = pytesseract.image_to_string(image, lang)`, где `image` – входное изображение; `lang` – обозначение используемого для распознавания языка в виде строки («rus», «eng» и др.); `text` – текст с входного изображения в виде строки.

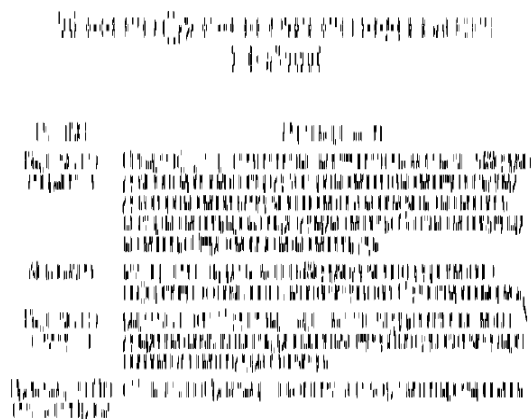
Результатом работы Tesseract являются строки, содержащие текст на входных изображениях. Прямоугольная область, в процессе обработки которой было получено, что заголовком является строка «Конференция», содержит область конференции. Информация о расположении области конференции вебинара обновляется в течение всего вебинара (установлен интервал времени обновления 5 минут). Соответственно, по строкам «Презентация», «Пользователи» и «Чаты» выполняется идентификация соответствующих областей.

Типовые функциональные подсистемы ИСУТП	
Блок ИС	Характеристика
Управление закупками	быстрая формировка заказа поставщикам-изготовителям. Модуль должен отражать номер договора поставки с поставщиком, дату договора поставки, задолженность изготовителя по поставкам, номер поставки, прогнозную дату поставки, объем поставки, цену поставки, общую стоимость поставки, др.
Логистика	планирование грузопотоков. Модуль должен содержать всю информацию о наличии мест на складах и свободном транспорте.
Управление складами	управление свободным пространством склада, возможность не допускать наличия незагруженных площадей на одном складе при нехватке места на другом складе.
Бухгалтерский и налоговый учет	обеспечение бухгалтерского и налогового учета на предприятии.

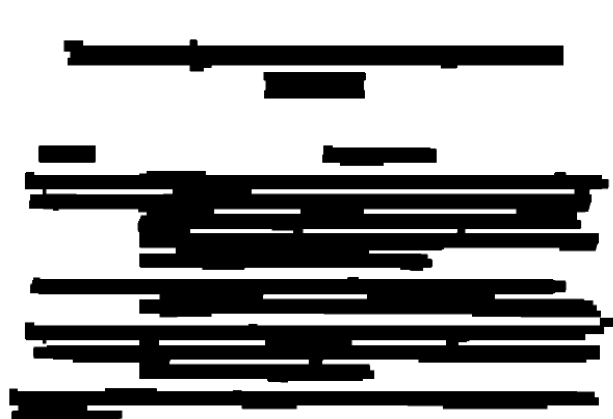
а)



б)



в)



г)

Типовые функциональные подсистемы ИСУТП	
Блок ИС	Характеристика
Управление закупками	быстрая формировка заказа поставщикам-изготовителям. Модуль должен отражать номер договора поставки с поставщиком, дату договора поставки, задолженность изготовителя по поставкам, номер поставки, прогнозную дату поставки, объем поставки, цену поставки, общую стоимость поставки, др.
Логистика	планирование грузопотоков. Модуль должен содержать всю информацию о наличии мест на складах и свободном транспорте.
Управление складами	управление свободным пространством склада, возможность не допускать наличия незагруженных площадей на одном складе при нехватке места на другом складе.
Бухгалтерский и налоговый учет	обеспечение бухгалтерского и налогового учета на предприятии.

д)

Типовые функциональные подсистемы ИСУТП	
Блок ИС	Характеристика
Управление закупками	быстрая формировка заказа поставщикам-изготовителям. Модуль должен отражать номер договора поставки с поставщиком, дату договора поставки, задолженность изготовителя по поставкам, номер поставки, прогнозную дату поставки, объем поставки, цену поставки, общую стоимость поставки, др.
Логистика	планирование грузопотоков. Модуль должен содержать всю информацию о наличии мест на складах и свободном транспорте.
Управление складами	управление свободным пространством склада, возможность не допускать наличия незагруженных площадей на одном складе при нехватке места на другом складе.
Бухгалтерский и налоговый учет	обеспечение бухгалтерского и налогового учета на предприятии.

е)

Рис. 6. Результат работы алгоритма выделения текста из изображения слайда презентации:

- а) исходное изображение, б) результат работы алгоритма поиска границ, в) результат инверсии и первого этапа морфологических операций, г) результат второго этапа морфологических операций – объединения границ символов в блоки, д) результат операции выделения контуров текстовых блоков, е) результат фильтрации контуров из предыдущего этапа

В задаче анализа качества вебинара по показателям из табл. 1 распознавание текста используется для:

- 1) точной идентификации областей в кадре видеопотока (область конференции, область презентации, область чата, область количества участников);
- 2) идентификации контактных данных ведущего и оценивания показателя 9 *Контактные данные*;

3) считывания количества участников в области количества участников для расчёта показателя *32 Показатели посещаемости*;

4) считывания текущего и общего времени вебинара из области времени вебинара для расчёта показателя *31 Точность заявленной длительности мероприятия*.

Пример работы алгоритма распознавания текста представлен на рис. 7. Функция `pytesseract.image_to_string` возвращает строку "1:27:35/1:29:43", которая при помощи функции `strptime(str, format)` библиотеки `datetime` определяет оперативное время вебинара T_o (левая часть результатной строки) и фактическую длительность вебинара T_f (правая часть результатной строки).

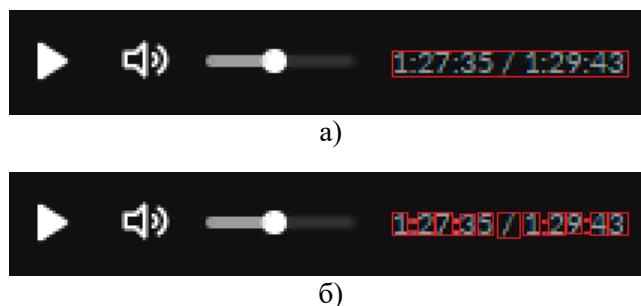


Рис. 7. Результат работы алгоритма выделения текста из изображения слайда презентации:
а) результат фильтрации контура, б) результат детализации контура для распознавания текста

Оперативное время используется для параметризации модели анализа звукового ряда (см. раздел 4), а также для расчёта некоторых показателей, отражающих динамику вебинара, например, *17 Использование указки* и *23 Частота взаимодействия*.

Фактическое время используется для расчёта точности соблюдения заявленной длительности вебинара T_b и его завершения в указанный в регламенте момент времени в соответствии с рисунком в приложении к данной статье. В рассмотренном примере при заявленном времени вебинара 1.5 часа значение показателя *31 Точность заявленной длительности мероприятия* составляет $9.96 \approx 10$ баллов.

6. Результат работы программы анализа аудиовидеопотока

Для компьютерной реализации разработанных моделей и алгоритмов написана программа на Python.

На первом этапе необходимо загрузить видеопоток вебинара. После загрузки видеопотока появится окно установки параметров (см. табл. 3). Значения этих параметров не следует устанавливать слишком малыми, так как при анализе выполняется большое число операций и расходуется много времени, а результаты излишне подробны и сложно анализируемы в дальнейшем. Также значения этих параметров не следует устанавливать слишком большими, так как возможно слияние разных типов интервалов времени, а их редактирование будет более затратно по времени. Все действия и инструкции программы на данном этапе будут написаны в текстовом окне «Статус процессов».

Таблица 3. Параметры алгоритма первичного анализа аудиовидеопотока вебинара

№	Наименование параметра	Значение по умолчанию
1	Длина одного фрейма, образованного при разбиении полученного аудиопотока, мс	100
2	Допустимая пауза между интервалами разных типов, мс	500
3	Минимальная длина фрейма после объединения соседних однотипных интервалов, мс	1000

После того как параметры будут установлены, выполняется фильтрация аудиосигнала и извлечение из него интервалов тишины. Исходя из того, что в интервалы голоса могли попасть посторонние звуки (стук, скрип, фоновый шум и др.), а в интервалы тишины могли попасть тихие голоса людей, имеется возможность редактировать образованные интервалы тишины. Из них потом будут образованы интервалы голосов. Для редактирования интервалов разработан специальный инструмент, обеспечивающий возможность изменять, удалять, склеивать имеющиеся в списке интервалы, а также добавлять новые интервалы. Все интервалы представлены в миллисекундах. Для улучшения анализа интервалов соответствующие им части аудиопотока записываются в отдельную директорию. Количество отредактированных интервалов и значения параметров зависят от оборудования записи вебинара, влияющего на качество звука и видео, а также голоса ведущего, частоты смены личности говорящего, фонового шума.

В рассматриваемом примере было выделено 226 интервалов тишины. На их основе было выделено 225 интервалов голосов. Так как в интервалы голоса могли попасть посторонние звуки (стук, скрип, фоновый шум и др.), а в интервалы тишины могли попасть тихие голоса людей, эти интервалы были отредактированы (всего на этом этапе внесено 19 изменений). После внесения изменений было получено 207 интервалов голосов. Остальные интервалы отнесены к интервалам тишины.

На втором этапе пользователь выбирает режим работы извлечения голоса ведущего из интервалов времени голосов (ручной или полуавтоматический).

При выборе ручного режима инициируется описанная выше процедура редактирования интервалов. Исходя из того, что большую часть аудиопотока занимает голос ведущего, на редактирование в качестве основного списка интервалов подаётся список интервалов не ведущего. При выборе ручного режима этот список интервалов будет пустым. Вспомогательным списком является выделенный ранее список голосов. На основе вспомогательного списка и заполняется основной список интервалов.

При выборе полуавтоматического режима работы сначала появится окно для ввода одного небольшого интервала (больше двух секунд), в котором говорит один ведущий. Этот интервал нужен для получения GMM голоса ведущего. После тренировки GMM автоматически будут выделены предположительные интервалы времени голосов не ведущего, которые затем будут доступны для редактирования.

В общем случае, если запустить программу на одном и том же видеопотоке несколько раз, то результаты получатся похожими, но разными. Связано это с тем, что среди недостатков ЕМ-алгоритма обучения GMM есть неустойчивость к выбору начальных значений параметров. От них зависит как скорость сходимости, так и результат работы, то есть алгоритм может сойтись к локальному экстремуму функции правдоподобия, который может быть существенно ниже глобального. В качестве начальных параметров обучения в программе установлены обобщенные результаты работы модели кластеризации (методом k-means) некоторого количества данных из начальной выборки. Одним из недостатков этого метода является чувствительность к выбору начальных центров кластеров. Если они выбираются случайно, то это может являться источником погрешности. Поэтому целесообразно выполнять начальную кластеризацию несколько раз для получения более точного результата. После того как интервалы голосов не ведущего были откорректированы, остальные интервалы голосов автоматически будут считаться интервалами голоса ведущего.

В рассматриваемом примере был выбран полуавтоматический режим. Исходя из введённого интервала времени голоса ведущего, на котором обучалась GMM, было выделено 433 интервала времени разговора не ведущего, что не соответствует действительности. Для уменьшения разницы между полученными и реальными интервалами происходит их корректировка (всего на этом этапе внесено 217 изменений). Также здесь происходит фильтрация интервалов по длине. Интервалы с меньшей длиной попадают в интервалы голоса ведущего, исходя из того, что ведущий в вебинаре говорит большую часть времени. После корректировки было выделено 207 интервалов голоса ведущего.

На третьем этапе выполняется анализ изображений вебинара из интервалов времени голосов не ведущего с параметрами согласно табл. 4. После установки параметров выполняется анализ слайдов вебинара. Рассчитанные интервалы доступны для редактирования пользователем.

Таблица 4. Параметры алгоритма вторичного анализа аудиовидеопотока вебинара

№	Наименование параметра	Значение по умолчанию
1	Время между анализируемыми слайдами вебинара, мс	100
2	Минимальная и максимальная амплитуды движения губ ведущего (амплитуда, находящаяся в этом интервале, свидетельствует о том, что ведущий что-то говорит)	[2, 3]
3	Требуется ли показ области вебинара во время анализа слайдов	Нет
4	Требуется ли сохранять все полученные типы интервалов в файл	Да

Значения параметров 1 и 2 не следует задавать слишком малыми и слишком большими. При слишком малых значениях разница положения ведущего в анализируемых слайдах является почти незаметной. Значит, будет пропущено много движений. При слишком больших значениях также будут пропущены некоторые движения, которые произошли между рассматриваемыми слайдами. Чтобы резкие движения головы или тела ведущего не считались движением губ, в качестве параметра 2 используется интервал.

После завершения работы программы в директории с видеофайлом вебинара появляется директория «Result» с двумя текстовыми файлами с интервалами тишины (209 интервалов) и интервалами голоса ведущего (207 интервалов).

Таким образом, на протяжении работы системы на данном примере выделено 659 интервалов. После 236 исправлений было сохранено 416 интервалов.

Следует заметить, что для увеличения достоверности полученных результатов разметки аудиовидеопотока и уменьшения затрат на корректировку в конечной реализации модели необходимо реализовать композицию анализа аудио- и видеопотоков. Однако это относится к следующему, заключительному, этапу исследований.

7. Заключение

В данной статье на основе методов Data Mining разработано и исследовано математическое обеспечение для интеллектуального анализа аудиовидеопотоков вебинаров в части идентификации и оценивания статических и динамических характеристик вебинара с целью упрощения и объективизации их оценивания. Это вторая часть исследования, имеющая целью повышение оперативности и достоверности оценивания некоторых показателей качества вебинара, допускающих интеллектуальный анализ аудиовидеопотока вебинара без активного участия эксперта и непосредственного применения экспертных методов оценивания.

На основе первой части исследования [1] была поставлена задача разработки математического аппарата, рассмотрена теоретическая часть для решения поставленной задачи, определены средства разработки для создания специального программного обеспечения. Для создания ПО задействованы функциональные возможности языка Python для машинного обучения и подключаемых к нему модулей и библиотек (OpenCV, DLib, LibROSA, PyQt5, NumPy и др.). Полученные компоненты математического обеспечения исследованы на контрольном аудиовидеопотоке вебинара и показали свою эффективность.

Третьей и завершающей частью исследования должна стать разработка формального аппарата для вычисления 10-балльных оценок показателей вебинара, а также формирования

модельно-инструментального комплекса, включающего в себя а) компоненты математической модели [1], отвечающие за формирование и отладку системы показателей качества вебинара, б) компоненты математического обеспечения на основе методов Data Mining, разработанные в данной статье, и в) формализмы для вычисления иерархической системы оценок качества вебинара. Построенный модельно-инструментальный комплекс потребует исследования адекватности и эффективности как гибридной модели в целом, так и отдельных её подмоделей, в том числе посредством сравнения результатов оценивания с полученными ранее результатами в рамках исследования [1] оценок качества вебинаров.

Приложение

В приложении содержится таблица, отражающая систему показателей качества вебинара, разработанная в ходе исследования [1]. Представлены шкала и критерии оценивания показателей, а также временная сложность процедуры их оценивания согласно табл. 2 основной части статьи. В таблице не отражены веса показателей, которые не играют существенной роли для материала статьи. В конце приложения представлен рисунок для оценивания показателя 31 *Точность заявленной длительности мероприятия*.

Таблица. Показатели качества вебинара, критерии и экспертные оценки сложности их оценивания

№	Показатель	Критерии оценивания		Оценка	Сложность
I	Сценарий: начало вебинара				
1.	Самопрезентация	Нет	Ведущий не представился	0	1
Очень краткая		1...2			
Краткая		Ведущий представился	3...5		
Достаточная	6...10				
2.	Обозначение регламента, в т.ч. правил	Нет		0	1
Краткое		3...5			
Достаточное (обозначены правила, в т.ч. дисциплина ответов на вопросы)		6...10			
3.	Обозначение цели вебинара	Нет		0	1
Краткое		3...5			
Достаточное		6...10			
4.	Обозначение плана вебинара	Нет		0	1
Краткое		3...5			
Достаточное (обозначена ценность)		6...10			
II	Сценарий: ход вебинара				
5.	Проблематика	Не обозначена		0	4
Не обозначена. Эпизодические обозначения «болевых» точек		3...5			
Обозначена в начале вебинара. Регулярные обозначения «болевых» точек		6...10			
6.	Ясность	Понятность, логичность и последовательность изложения		0...10	2
7.	Лаконичность	«Вода»		0	2
Разжиженный, преимущественно текстовый, материал		1...4			
Более-менее компактный материал с некоторым количеством «воды»		5...7			
Концентрированный материал со схемами и иллюстрациями		8...10			

№	Показатель	Критерии оценивания	Оценка	Сложность
8.	Резюмирование	В конце нет резюме В конце подводятся итоги, есть резюме Резюмирование происходило как в конце, так и несколько раз внутри вебинара	0 3...5 6...10	4
9.	Контактные данные	Не представлены Представлены	0 10	1
III Качества ведущего				
10.	Энергичность и эмоциональность ¹	Вялый, заторможенный, пессимистичный, эмоции выражаются неуместно Напряженный, агрессивный, торопливый, неуместное выражение эмоций Бодрый, раскованный, оптимистичный, эмоции выражаются уместно	0...2 3...7 8...10	2
11.	Темп изложения	Очень медленно, скучно Очень быстро Монотонно, заметные паузы Размеренно, ритмично, «как песня льется»	0...1 2...3 4...6 7...10	2
12.	Компетентность	Компетентность невысокая, опыта нет Ведущий приводит некоторые абстрактные примеры из практики Опыт ведущего налицо, приводимые примеры привязаны к реальности	0...2 3...5 6...10	3
13.	Внешний вид ¹	Опрятность, корректность визуального фона, шумы звукового фона	0...10	1
IV Качество презентационных материалов				
14.	Количество и качество информации на слайдах	Слайды сильно перегружены, на них слишком много монотонного текста Слайды слегка перегружены или недогружены, местами требуется больше текста, иллюстраций там, где есть только голос и видео Слайды использованы оптимально	0...2 3...7 8...10	1
15.	Количество и качество иллюстраций	Скучные и однообразные текстовые слайды Более-менее разнообразные слайды, мало иллюстраций, много текста Интересные и разнообразные слайды, иллюстрации преобладают над текстом	0...2 3...7 8...10	1
16.	Аккуратность	– одинаковые списки – единая цветовая гамма для всех элементов – отсутствие неаккуратностей (обрезанные края, лишние линии и т.п.) – примерно одинаковые размеры изображений и их размещение на слайдах – одинаковые размеры шрифтов (не более 2–4 размеров на всю презентацию) из диапазона 20–30 пт – одинаковые расстояния между элементами	0...10	1
17.	Использование указки	Нет Эпизодическое Регулярное и уместное Чрезмерное	0 5 10 3	3
18.	Разнообразие презентационных инструментов	Ничего Указка, рисование, доска, поднятие рук, эмоции, демонстрация рабочего стола	0 1...10	3

№	Показатель	Критерии оценивания	Оценка	Сложность
V Интерактив по инициативе ведущего				
19.	Знакомство	Нет Ведущий знакомится со слушателями, выясняет их степень знакомства с предметом изучения (обратная связь через чат или виртуальные руки, опрос)	0 3...10	2
20.	Проведение опросов	Нет 1 опрос 2 и более опросов	0 5 10	3
21.	Формат вопросов ведущего	Открытые, предполагающие длинные ответы Открытые, предполагающие короткие ответы, или закрытые, предполагающие знак в чате, виртуальную эмоцию или виртуальное поднятие руки	0...5 5...10	4
22.	Разнообразие интерактива	– вывод участников в эфир – вызов участников к доске – дискуссии – конкурсы и т.д.	0...10	3
23.	Частота взаимодействия	Нет Интервалы более 10 минут Интервалы менее 5 минут Интервалы от 5 до 10 минут	0 1...3 4...5 6...10	2
VI Интерактив по инициативе слушателей				
24.	Наличие вопросов слушателей	Нет 1–2 3–5 6 и более	0 2 5 10	1
25.	Реактивность на вопросы ²	Нет Промежуточный вариант Моментальная и в соответствии с регламентом (правилами)	0 1...9 10	4
26.	Качество ответов на вопросы ²	– озвучивание имени задающего вопрос – повторение вопроса вслух – вербальная оценка вопроса – лаконичные ответы	0...10	4
VII Оценки слушателей				
27.	Понятность ³	Оценка слушателей	0...10	
28.	Интересность ³	Оценка слушателей	0...10	
29.	Полезность ³	Оценка слушателей	0...10	
VIII Параметры результативности вебинара				
30.	Точность начала мероприятия	Пунктуальность ведущего, начало мероприятия в заявленное время T_0 с погрешностью 5 минут	0...10	1
31.	Точность заявленной длительности мероприятия	Завершение мероприятия в заявленное время (T_B – заявленная ведущим длительность вебинара ⁴)	0...10 (см. рис. ниже)	1
32.	Коэффициент результативной посещаемости вебинаров	Отношение числа слушателей, дослушавших вебинар до конца, к числу слушателей, посетивших вебинар	0...10	1
Средняя сложность экспертного оценивания:				2.87

¹ При отсутствии видеопотока число баллов по этим показателям не может быть больше 5.

² При отсутствии вопросов слушателей данный показатель принимается равным нулю.

³ При отсутствии оценок слушателей данные показатели принимаются равными нулю.

⁴ В случае, если ведущий не обозначил регламент в начале вебинара, принимается во внимание время, заявленное при планировании вебинара (в анонсе).

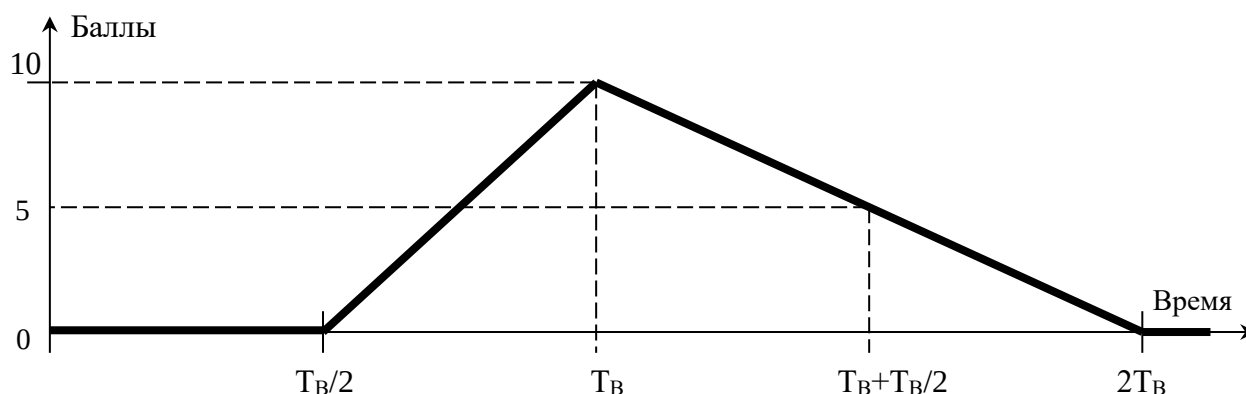


Рис. 8. График определения оценок показателя
31. Точность заявленной длительности мероприятия [1]
(T_B – заявленная длительность вебинара, мин.)

Литература

1. Полетайкин А. Н., Шевцова Ю. В., Подколзин В. В., Струкова Е. Г. Математическая модель оценивания качества контактной работы, реализуемой посредством вебинаров в ходе дистанционного обучения // Информатика и образование. 2019. № 7 (306). С. 42–53. DOI: 10.32517/0234-0453-2019-34-7-42-53.
2. Сизов Л. А. Вебинары – передовая технология непрерывного сетевого образования // Материалы XIV Международной научно-практической конференции «Проблемы экономики и информатизации образования». Тула: ТИЭИ. 2017. С. 92–95.
3. Цаллагов Н. А., Астахова Т. А. Проведение эффективных вебинаров от А до Я. ООО «Мираполис», 2015. 174 с. / [Электронный ресурс]. URL: <http://indo.kg/wp-content/uploads/2018/11/Проведение-эффективных-вебинаров-от-А-до-Я.pdf> (дата обращения: 31.12.2023).
4. Сорокин Г. А. Фильтры нижних частот // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. 2015. № 1. С. 100–107.
5. Попов А. Н., Шелупанов А. А. Машинное обучение. Эволюция методов. М.: ДМК Пресс, 2019. 304 с.
6. Bradski G. and Kaehler A. Learning OpenCV: Computer Vision with the OpenCV Library. First Edition. USA: O'Reilly Media, 2008. 543 p.
7. Андриянов Н. А., Дементьев В. Е., Ташлинский А. Г. Обнаружение объектов на изображении: от критериев Байеса и Неймана–Пирсона к детекторам на базе нейронных сетей EfficientDet // Компьютерная оптика. 2022. Т. 46, № 1. С. 139–159. DOI: 10.18287/2412-6179-CO-922.
8. Kazemi V. and Sullivan J. One millisecond face alignment with an ensemble of regression trees // Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014. P. 1867–1874. DOI: 10.1109/CVPR.2014.241.
9. Liang J., Doermann D., and Li H. Camera-based analysis of text and documents: a survey // IJDAR. 2005. V. 7. P. 84–104. DOI: 10.1007/s10032-004-0138-z.
10. Дворянкин Д. К., Кулешов Д. А., Кадурын А. В. Нейронные сети и глубокое обучение. СПб.: Питер, 2019. 352 с.

Подколзин Вадим Владиславович

к.ф.-м.н., заведующий кафедрой информационных технологий, Кубанский государственный университет (350040, Краснодар, ул. Ставропольская, 149), e-mail: vvp_kubsu@mail.ru, ORCID: 0000-0002-4491-1493.

Полетайкин Алексей Николаевич

к.т.н., доцент, доцент кафедры информационных технологий Кубанского государственного университета (350040, Краснодар, ул. Ставропольская, 149), тел. +7 861 2199-577;

доцент кафедры математического моделирования и цифрового развития бизнес-систем СибГУТИ (630102, Новосибирск, ул. Кирова, 86);

e-mail: alex.poletaykin@gmail.com, ORCID ID: 0000-0002-5128-1952.

Галкина Марина Юрьевна

доцент, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 272, e-mail: gmur7@sibguti.ru, ORCID ID: 0009-0002-4515-9923.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.

**Mathematical Apparatus for Some Quality Indicators Assessment
of a Training Webinar Using Machine Learning**

Vadim V. Podkolzin¹, Aleksey N. Poletaikin^{1,2}, Marina Yu. Galkina²

¹ Kuban State University (KubSU)

² Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: This article is the second stage of creating and researching a technology for assessing the quality of contact work at a university implemented through webinars. The mathematical support of the intelligent technology for analyzing the audio-video stream of an educational webinar recording is considered. The problem of increasing the efficiency and reliability of assessing some indicators of the quality of contact work at the Faculty of Distance Learning of SibSUTIS implemented through distance learning webinars is being solved. Machine learning methods and their implementation using standard Python modules and libraries are used for analysis. The scientific significance of the work lies in the development of new models and algorithms for intelligent analysis of audio-video streams parameterized for processing the recording of an educational event. The developed models and algorithms make it possible to increase the efficiency of the process of assessing the quality of contact work implemented through webinars.

Keywords: webinar, conference area, presenter activity, performance evaluation criteria, time intervals, text localization, audio track processing, image processing, machine learning.

For citation: Podkolzin V. V., Poletaikin A. N., Galkina M. Yu. Mathematical apparatus for some quality indicators assessment of a training webinar using machine learning (in Russian). *Vestnik SibGUTI*, 2024, vol. 18, no. 1, pp. 121-143. <https://doi.org/10.55648/1998-6920-2024-18-1-121-143>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Podkolzin V. V., Poletaikin A. N.,
Galkina M. Yu., 2024

The article was submitted: 15.12.2023;
accepted for publication 15.01.2024.

References

1. Poletaikin A. N., Shevtsova Yu. V., Podkolzin V. V., Strukova E. G. Matematicheskaya model' otsenivaniya kachestva kontaktnoi raboty, realizuemoi posredstvom vebinarov v khode distantsionnogo obucheniya [Mathematical model for assessing the quality of contact work implemented through webinars during distance learning]. *Informatika i obrazovanie*, 2019, no. 7 (306), pp. 42-53, DOI: 10.32517/0234-0453-2019-34-7-42-53.
2. Sizov L. A. Vebinary – peredovaya tekhnologiya nepreryvnogo setevogo obrazovaniya [Webinars - advanced technology for continuous online education]. *Problemy ekonomiki i informatizatsii obrazovaniya. Materialy XIV Mezhdunarodnoi nauchno-prakticheskoi konferentsii*, Tula, TIEI, 2017. pp. 92-95.
3. Tsallagov N. A., Astakhova T. A. *Provedenie effektivnykh vebinarov ot A do Ya* [Conducting effective webinars from A to Z]. ООО «Mirapolis», 2015, 174 p., available at: <http://indo.kg/wp-content/uploads/2018/11/Provedenie-effektivnykh-vebinarov-ot-A-do-Ya.pdf> (accessed: 31.12.2023).
4. Sorokin G.A. Fil'try nizhnikh chastot [Low pass filters]. *Vestnik YuUrGU, Seriya: Komp'yuternye tekhnologii, upravlenie, radioelektronika*, 2015, no. 1, pp. 100-107.
5. Popov A. N., Shelupanov A. A. *Mashinnoe obuchenie. Evolyutsiya metodov*. [Machine learning. Evolution of methods]. Moscow, DMK Press, 2019. 304 p.
6. Gary Bradski, Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*, First Edition, USA, O'Reilly Media, 2008. 543 p.
7. Andriyanov N. A., Dement'ev V. E., Tashlinskii A. G. Obnaruzhenie ob"ektov na izobrazhenii: ot kriteriev Baiesa i Neimana–Pirsona k detektoram na baze neironnykh setei EfficientDet [Detection of objects in an image: from Bayes and Neyman-Pearson criteria to detectors based on EfficientDet neural networks]. *Komp'yuternaya optika*, 2022, vol. 46, no. 1, pp. 139-159. DOI: 10.18287/2412-6179-CO-922.
8. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1867-1874, DOI: 10.1109/CVPR.2014.241.
9. Liang J., Doermann D., Li H. Camera-based analysis of text and documents: a survey. *IJDAR*, 2005, vol. 7, pp. 84-104, DOI: 10.1007/s10032-004-0138-z.
10. Dvoryankin D. K., Kuleshov D. A., Kadurin A. V. Neironnye seti i glubokoe obuchenie [Neural networks and deep learning]. Saint Petersburg, Piter, 2019. 352 p.

Vadim V. Podkolzin

Cand. of Sci. (Physical and Mathematical), Head at the Department of Information Technologies, (KubSU, Krasnodar, Russia). e-mail: vvp_kubsu@mail.ru, ORCID: 0000-0002-4491-1493, Scopus AuthorID: 57215326416, Researcher ID: AAU-8042-2021.

Aleksey N. Poletaikin

Cand. of Sci. (Engineering), Assistant Professor at the Information Technologies Department, Kuban State University (KubSU, Krasnodar, Russia).

Assistant professor at the Mathematical Modeling and Digital Development of Business System Department, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86). e-mail: alex.poletaykin@gmail.com, ORCID ID: 0000-0002-5128-1952, Scopus AuthorID: 57213829361, ResearcherID: ABF-6799-2020.

Marina Yu. Galkina

Assistant Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86). e-mail: gmur7@sibgut.ru, ORCID ID: 0009-0002-4515-9923.