DOI: https://doi.org/10.30898/1684-1719.2025.8.15
УДК: 004.934.2

ОБНАРУЖЕНИЕ РЕЗКИХ ИЗМЕНЕНИЙ ИНТЕНСИВНОСТИ РЕЧЕВОГО СИГНАЛА НА ОСНОВЕ КОНЦЕПЦИИ РЕЦЕПТИВНЫХ ПОЛЕЙ

М.М. Гуторов, В.Е. Анциперов

ИРЭ им. В.А. Котельникова РАН
125009, Москва, ул. Моховая, 11, корп.7

Статья поступила в редакцию 10 сентября 2025 г.

Аннотация. В статье рассматривается возможность автоматического выделения временных границ слов и вокализованных звуков в речевом сигнале, представленном в выборочном представлении нейроморфной модели периферического отдела биологической слуховой системы человека. Предложен методический подход, реализующий последовательную обработку речевых сигналов для детектирования временных границ речевых элементов. На первом этапе с использованием временных рецептивных полей формируются признаки резкого изменения интенсивности сигнала, позволяющие определить потенциальные границы речевых фрагментов. Далее осуществляется фильтрация событий начала и окончания речевых сегментов по интенсивности сигнала и показателю максимальной частоты спайков. Для вычисления точности определения границ слов методом рецептивных полей использовался также анализ огибающей сигнала с применением пороговых значений, обеспечивающих устойчивость и воспроизводимость результата. Объективная оценка качества расчетов временных границ звуков методом рецептивных полей проведена путём расчетов среднеквадратичной ошибки результата в сравнении с ручной разметкой.

Полученные результаты демонстрируют высокую точность детектирования, достигающую уровня десятков миллисекунд, что подтверждает практическую применимость предложенного метода. Отмечена избыточная чувствительность метода на окончаниях гласных, приводящая к ложным срабатываниям, что указывает на необходимость внедрения адаптивных контекстных правил. Представленный метод может быть использован в задачах анализа и сегментации речевых сигналов в реальном времени.

Ключевые слова: речевая сегментация; детектирование гласных; границы слов; нейроморфная обработка; модель восприятия звука; рецептивные поля; импульсная активность; детектирование событий; обработка речевого сигнала; акустический анализ.

Финансирование: Работа выполнена в рамках государственного задания (номер АААА-А19-119041590070-1) Института радиотехники и электроники им. В.А. Котельникова Российской академии наук.

Автор для переписки: Михаил Михайлович Гуторов gutorov.m.m@gmail.com

Введение

Задача автоматического анализа и сегментации речевых сигналов остается одной из ключевых в области обработки аудиоданных. Современные приложения – от голосовых помощников до автоматического индексирования и субтитрования – требуют высокой точности и скорости извлечения элементарных речевых фрагментов. Особенно важным аспектом является детектирование границ гласных и слов, поскольку эти сегменты несут основную лексическую и просодическую нагрузку. Ошибки на этом этапе могут существенно снижать эффективность последующих этапов распознавания речи. В то же время границы речевых фрагментов не всегда сопровождаются ярко выраженными акустическими переходами, что усложняет их автоматическое определение. Аналогичные сложности подробно описаны и в задачах музыкального анализа, где выявление атак звуков (on-sets) требует комплексного подхода к изменчивым частотно-временным признакам сигнала [1].

Традиционные методы детектирования границ звуков опираются на спектральные признаки и энергетические оценки сигнала, что делает их уязвимыми к вариативности произношения, шумам и изменениям речевого контекста. В связи с этим возрастает интерес к биологически-инспирированным подходам, использующим принципы, заложенные в работе слуховой системы восприятия звука человеком [2]. Аналогичные принципы устойчивого восприятия резких изменений интенсивности излучения были ранее подробно исследованы в зрительной системе человека в рамках теории Retinex, описывающей восприятие яркости независимо от меняющихся условий освещения [4]. Одним из путей исследований в этом направлении является моделирование процессов кодирования звуковых стимулов в периферии слуховой системы, где звуковой сигнал преобразуется в нейронный сигнал спайковых импульсов — дискретную последовательность отсчетов, сохраняющих критически важную временную структуру сигнала и устойчивую к нелинейным искажениям [3].

В настоящей работе предлагается метод детектирования границ гласных (вокализованных) звуков и слов (речевых пауз) на основе анализа спайкового представления речевого сигнала, полученного с использованием нейроморфной модели биологического кодирования акустического сигнала. Ключевыми элементами метода являются использование рецептивных полей, частотно-временной фильтрации и локального анализа спайков речевого сигнала. Метод включает механизмы вычисления и верификации событий, учитывая как энергетическую, так и временную структуру сигнала. Практическая реализация выполнена в виде программного прототипа и протестирована на размеченных фрагментах речи с целью оценки точности и устойчивости детектирования.

1. Цель исследования и используемые методы для ее достижения.

Целью данного исследования является выяснение способности предложенного нами метода автоматически вычислять временные границы

сигналов [5], в частности, речевых сигналов, состоящих из слов и вокализованных звуков в словах на основе применения модели рецептивных полей (Метод РП). Для достижения цели исследования нами определены:

- Входной сигнал и его характеристики;
- Требования к конструкции и классификации РП;
- Алгоритм применения классифицированных РП для определения границ слов и гласных звков;
- Метрика оценки результатов вычисления временных границ слов и гласных звуков для получения объективной оценки качества применения метода РП.

1.1. Входной сигнал.

В качестве входного сигнала был записан речевой сигнал «Дядя Саша студент» в формате *.wav, с частотой дискретизации 32000 ГЦ. Для проведения исследований использовалась программа Praat [6], с помощью которой были определены ключевые характеристики речевого сигнала:

1) Частота основного тона F0 и частоты формант F1-F5;
2) Выполнена ручная разметка временных границ каждого гласного звука в трех словах входного сигнала;
3) Определен диапазон частотных границ для каждого гласного звука.

1.2. Частота основного тона

Частота основного тона (fundamental frequency или F0) – это частота колебания голосовых связок при произнесении тоновых звуков. При произнесении не вокализованных звуков, например, шипящих или свистящих, – связки не колеблются. Деление на тоновые и не тоновые звуки не эквивалентно делению на гласные и согласные. Вариабельность частоты основного тона сильно отличается не только между людьми (для мужских голосов частота составляет 70-200 Гц, а для женских может достигать 400 Гц), но и для одного человека,

особенно в эмоциональной речи. Изменение частоты основного тона называется интонацией.

1.3. Ручная разметка временных границ каждого гласного звука в трех словах входного сигнала.

Программа Praat [6] позволяет вручную выделять временные границы желаемого фрейма звукового сигнала с последующим прослушиванием сигнала в выделенном временном интервале. На рисунках 1 и 2 показан пример выделения первого гласного звука из шести, находящиеся в трех словах входного сигнала.

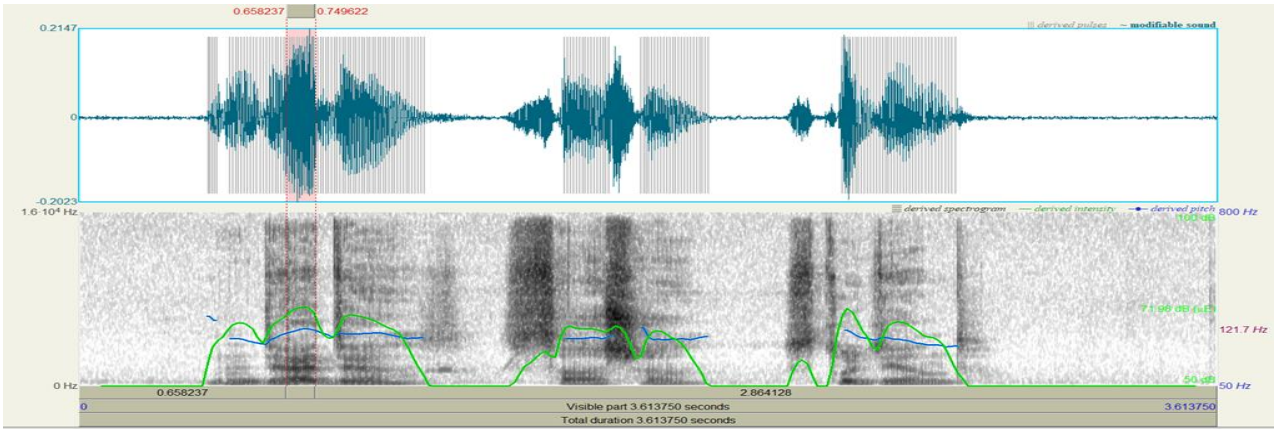

Рис. 1. Скриншот окна Praat.
В верхней половине – осциллограмма входного сигнала,
в нижней – его спектрограмма (серый цвет),
интенсивность (яркий зеленый цвет), и F0 (голубой цвет).

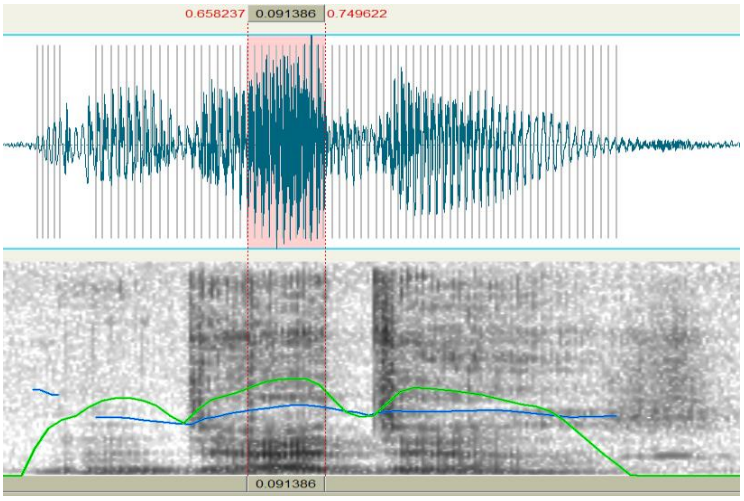

Рис. 2. Детальная визуализация выделенного вручную
первого гласного звука «Я» в слове «ДЯДЯ».

Продолжительность звукового сигнала составляет 3.6 секунды, частота дискретизации 32 кГц. На рисунке 1 вертикальными пунктирными линиями красного цвета показано выделение от руки первого гласного звука «Я» в слове «ДЯДЯ», F0 = 121.7 Гц. Серым цветом в верхнем окне обозначены те интервалы времени, которые соответствуют F0. Видно, что при произнесении шипящих звуков «С» и «Ш» во втором слове «САША» связки перестают колебаться, поэтому голубая линия имеет разрыв (звук основного тона отсутствует).

Временные границы для каждого из шести гласных звуков были определены вручную с использованием Praat, после чего записаны в конфигурационный файл. В дальнейшем они используются как эталон для вычисления метрики качества определения временных границ этих же гласных звуков методом РП. Метрикой качества вычислений выбрана RMS (квадратный корень из дисперсии случайной величины), которая в настоящем исследовании вычисляет суммарное отклонение между вычисленными временными границами гласных звуков и временными границами тех же звуков, которые записаны как эталонные в конфигурационном файле. Метрика RMS используется и для оценки качества вычисления границ слов.

1.4. Генерация спайков.

Входной звуковой сигнал является источником для генерации набора спайков. Анализ количественных и частотных характеристик спайков методом РП и методом частотного кодирования используется для достижения целей исследования.

На первом этапе входной звуковой файл был обработан в среде программной разработки Octave 9.4.0 с использованием кохлеарной модели osses2021.m [2], входящей в состав библиотеки AMToolbox v1.6.0. В результате обработки была получена амплитудная огибающая, которая сохранена в бинарный файл (mat-файл). Далее, на основе полученной огибающей с применением набора утилит, предложенных A. de Cheveigné [3], был сгенерирован набор

спайков для каждого из 31 частотного канала, определённого моделью osses2021.m. Полученные данные были сохранены во второй mat-файл.

2. Входные данные.

В качестве входных данных для предлагаемого метода, реализующего алгоритм детектирования речевых границ, использовались:

1) звуковой файл, содержащий речевую запись,

2) амплитудная огибающая, сформированная кохлеарной моделью,

3) массив спайков, сгенерированный на основе этой огибающей.

Работа алгоритма была организована в соответствии с параметрами, значения которых можно задавать в конфигурационных файлах. Для обработки использовались два различных временных масштаба, определяющих длительность элементарного окна анализа: один – для выделения границ слов, другой – для детекции гласных звуков внутри слов. Применялись также различные значения порогов, позволяющих классифицировать события рецептивных полей и выделять акустически значимые фрагменты речи.

Фильтрация шумовых фрагментов обеспечивалась с учётом контекста:

1. Ограничение минимальной длительности речевых сегментов,
2. Определение минимального временного интервала между ними,
3. Заданы начальные и конечные границы длительности гласных звуков,
4. Задан минимально допустимый интервал между гласными звуками в одном слове.

Для пространственного анализа и частотного кодирования использовались спайки из ограниченного набора частотных каналов, охватывающего область первой и второй форманты. Оценка спайковой активности производилась в скользящем временном окне с частичным перекрытием, что обеспечивало высокое временное разрешение при анализе динамики нейронной активности.

3. Конструкция метода и классификация сигналов на основе модели РП.

Во всех вариантах применения РП в настоящем исследовании используется одна и таже конструкция РП, состоящая из 4-х одинаковых интервалов времени, длительность которого задается параметром. 1-ый и 4-ый интервалы времени называются периферийными полями (ПП), а 2-ой и 3-ий – центральными полями (ЦП) этого РП [5]. Классификация РП производится по одному и тому же алгоритму во всех вариантах применения РП:

1) Для каждого РП вычисляем индивидуальный остаток, который равен сумме спайков в ЦП РП минус 0,5 умножить на сумму спайков во всем РП.
2) Для каждого РП вычисляем индивидуальный порог по формуле порог равен произведению коэффициента и корня квадратного из суммы всех спайков, которые попали в это РП.
3) Сравнение величины модуля остатка с величиной порога для каждого РП.

   – если остаток имеет положительное значение и модуль остатка больше порога, то запоминаем положительную величину остатка и присваиваем классификацию **ON** этому РП;

   – если остаток имеет отрицательное значение, но по модулю больше порога, то запоминаем отрицательную величину остатка и присваиваем классификацию **OFF** этому РП;

   – если остаток по модулю меньше порога, то это РП не классифицируем и в дальнейшем не рассматриваем.

4. Алгоритм использования классифицированных РП.

Алгоритм применения классифицированных РП для определения временных границ целевых объектов (слов, звуков) одинаковый, отличается использованием различных значений параметров относящихся, соответственно, к поиску границ слов или звуков. Интервал времени, равный продолжительности входного сигнала, разбивается на одинаковые РП, каждое длительностью 4 тика.

РП располагаются встык друг с другом. Для каждого РП определяются четыре положения на оси времени:

1) основное,
2) сдвинутое на 1 тик вправо от начального положения РП,
3) сдвинутое на 2 тика,
4) сдвинутое на 3 тика.

Для каждого положения РП определяются индивидуальные остаток и порог:

$$res = n_c - \frac{1}{2} n_{tot}, \quad (1)$$

$$T_{\text{threshold}} = d_{\text{thr}} \times \sqrt{n_{tot}}, \quad (2)$$

и проводится классификация РП для каждого положения так, как описано ранее в разделе 3. Далее из четырёх положений РП выбирается то, которое имеет максимальный остаток по модулю. Если ни одно из положений РП не было классифицировано, то такое РП далее не рассматривается.

Величина остатка *res* со знаком, который был определен при классификации РП, время середины положения РП, классификация (для положения РП, имеющего максимальный остаток по модулю) записываются в массив данных. Далее, эти данные используются в алгоритмах вычисления временных границ для слов и для звуков.

5. Результаты вычисления временных границ слов.

В работе были применены два независимых метода для определения границ слов во входном сигнале:

1) Метод огибающей Гильберта
2) Метод РП

Метод огибающей Гильберта – классический подход, основанный на энергии сигнала. Огибающая Гильберта входного сигнала сравнивается с величиной порога, величина которого подбирается вручную. Временные интервалы, где амплитуда превышает порог, считаются кандидатами

в речевые сегменты. Кандидаты проходят постобработку на предмет сравнения длительности речевых фрагментов и пауз между ними.

Данный метод обеспечивает устойчивое к шумовым и акустическим искажениям, хотя и ограниченное по точности, определение границ слов. Полученные границы далее используются в качестве эталонных значений для сопоставления с результатами, полученными методом рецептивных полей. Этим методом найдено 3 слова с границами, которые отображены на рисунке 3:

Слово 1: 0.417 – 1.081 с (длительность: 664 мс)

Слово 2: 1.382 – 1.967 с (длительность: 584 мс)

Слово 3: 2.262 – 2.805 с (длительность: 543 мс)

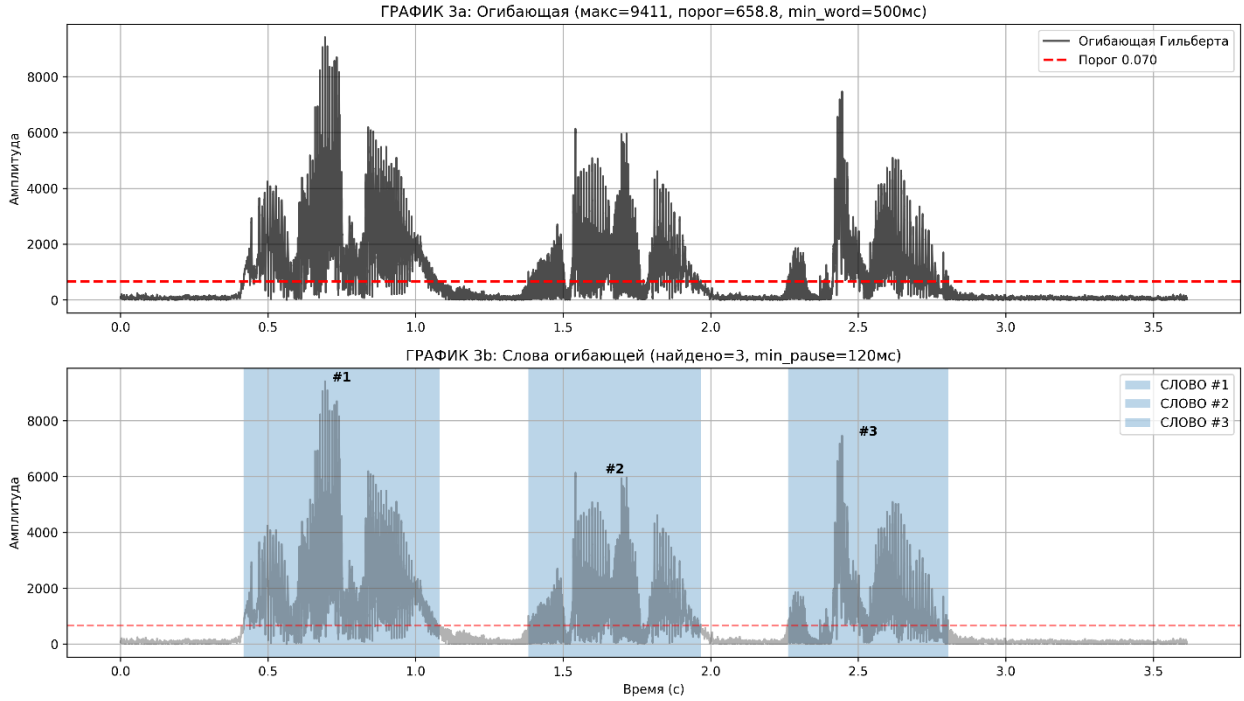

Рис. 3. Границы трех слов, найденные с использованием огибающей Гильберта и порога.

Метод РП – Метод автоматического вычисления временных границ сигналов, применительно к словам и гласным звукам в словах на основе применения модели рецептивных полей.

Определения временных границ слов во входном сигнале производится с применением РП, которые были классифицированы так, как описано в разделе 3.

Список классифицированных РП фильтруется по вычисленной плотности распределения классифицированных РП вдоль оси времени – определяются

глобальные границы речевого отрезка по плотности РП-событий, отсекая участки тишины или фонового шума. В отфильтрованной последовательности РП ищутся характерные паттерны переходов: ON-событие, за которым следует OFF-событие, которые интерпретируются как границы слов. В процессе поиска к паттернам применяются ограничения по минимальной длительности слов и паузе между ними.

При длительности РП величиной 96 мс методом РП найдено 3 слова с границами:

Слово 1: 0.444с - 1.044 с (длительность: 600 мс)

Слово 2: 1.332с - 2.100 с (длительность: 768 мс)

Слово 3: 2.244с - 2.820 с (длительность: 576 мс)

Общий **RMS = 61.9 мс**.

На рисунке 4 показаны результаты его работы.

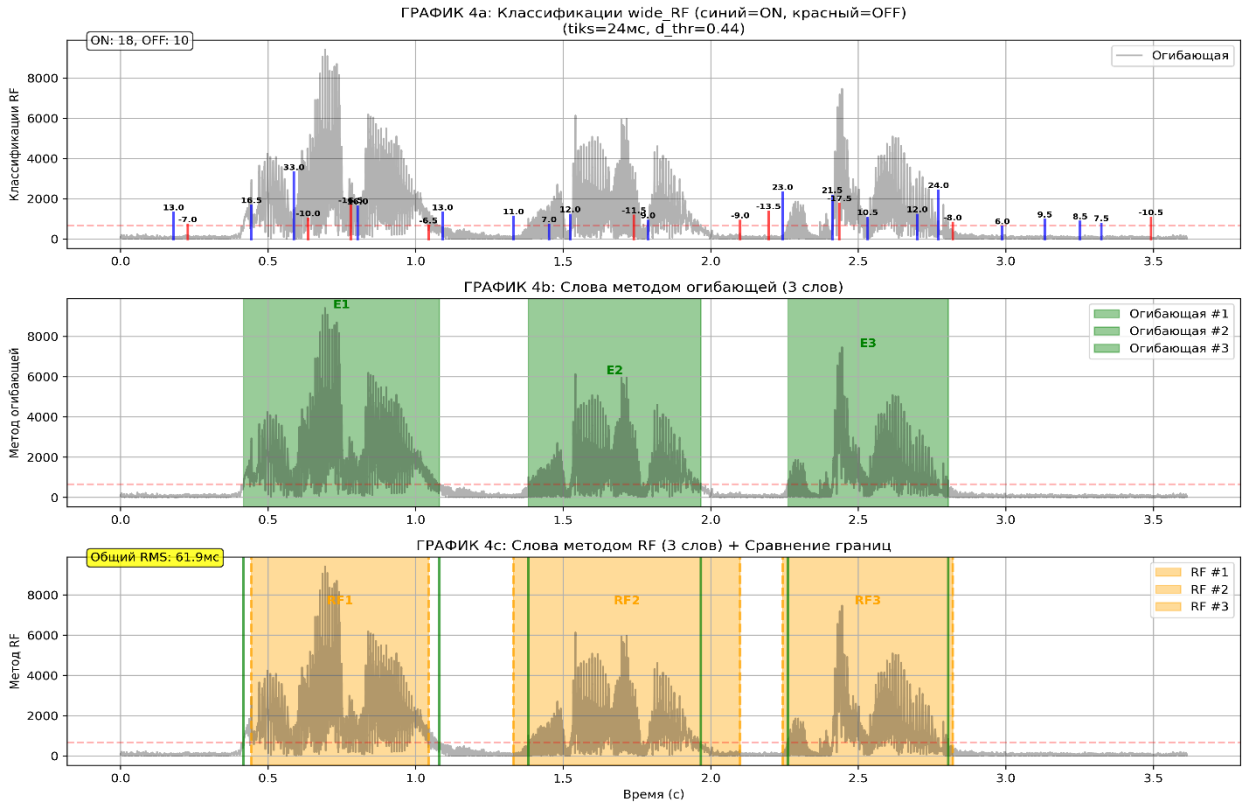

Рис. 4. Границы трех слов, найденные методом РП, при длительности РП 96 мс, общий RMS = 61.9 мс.

При длительности РП 72 мс, методом РП также найдено 3 слова с границами:

Слово 1: 0.441с - 1.125 с (длительность: 684 мс)

Слово 2: 1.342с - 2.025 с (длительность: 684 мс)

Слово 3: 2.241с - 2.745 с (длительность: 504 мс)

Общий **RMS = 44.0 мс**.

На рисунке 5 показаны результаты его работы.

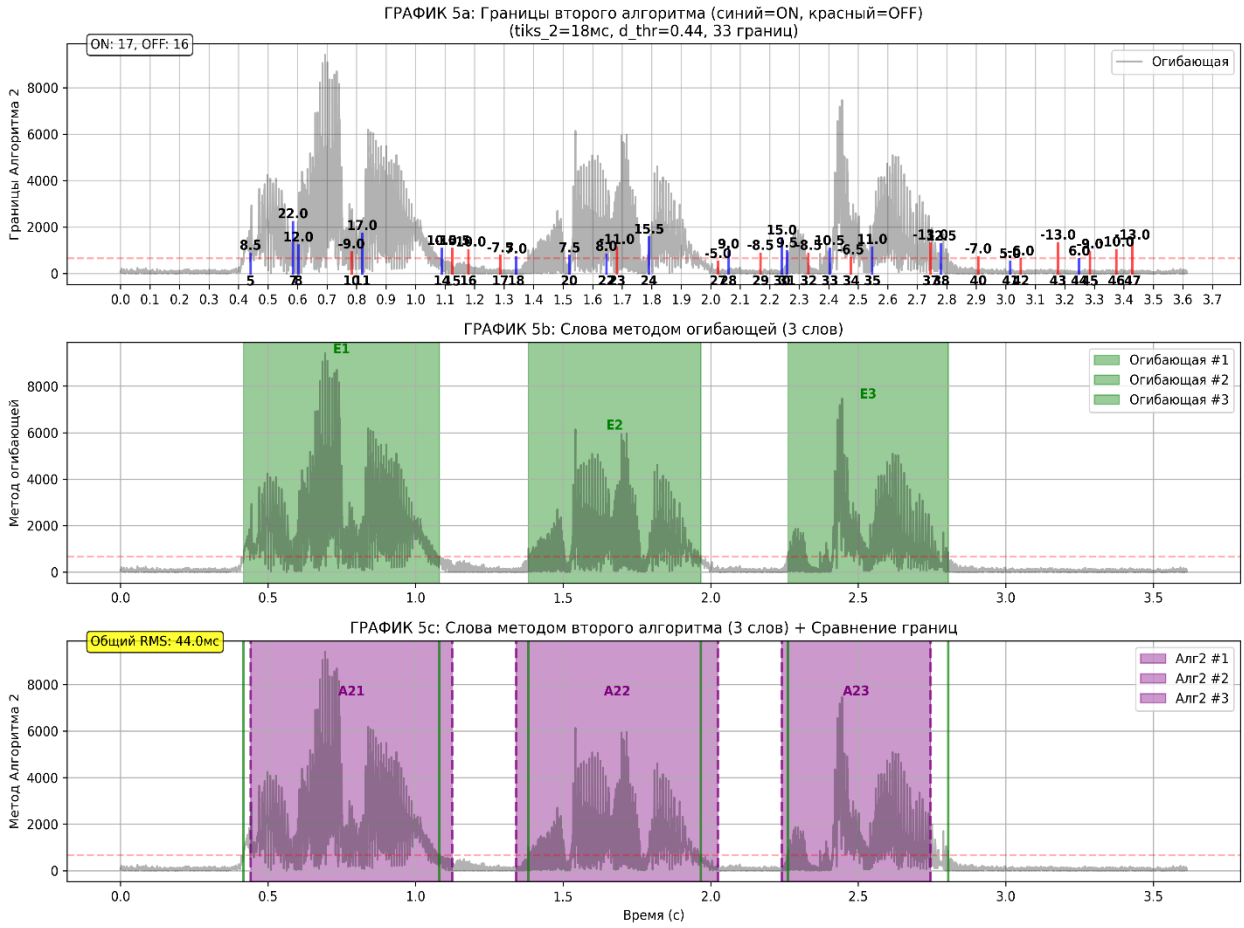

Рис. 5. Границы трех слов, найденные методом РП,
при длительности РП 72 мс, общий RMS = 44.0 мс.

Несмотря на то, что во втором варианте вычислений границ слов RMS имеет меньшее значение, для поиска гласных звуков в словах были выбраны интервалы трех слов, вычисленные по первому варианту, при значении длительности РП 96 мс - для того, чтобы использовать РП различной длительности для поиска слов и поиска звуков. Поиск границ звуков проводился при длительности РП 72 мс.

6. Результаты вычисления временных границ звуков.

6.1. Метод РП для поиска предварительных границ гласных звуков.

РП длительностью 72 мс были классифицированы и использовались для определения временных границ гласных звуков в каждом из трех слов. В последовательности РП ищутся характерные паттерны переходов: ON-событие, за которым следует OFF-событие, которые интерпретируются как границы гласных звуков. В процессе поиска к паттернам применяются ограничения по минимальной и максимальной длительности и по продолжительности паузы между соседними гласными звуками. На этом этапе формируется первичный набор кандидатов в границы гласных звуков, которые далее передаются для фильтрации с помощью частотного кодирования.

В трех словах входного сигнала были найдены методом РП 9 кандидатов во временные границы гласных звуков. Ниже приведен их перечень и графическое отображение (см. рисунок 6):

Гласный 1: 0.585с - 0.783с (198мс)
Гласный 2: 0.819с - 1.125с (306мс)
Гласный 3: 1.521с - 1.683с (162мс)
Гласный 4: 1.791с - 2.025с (234мс)
Гласный 5: 2.061с - 2.169с (108мс)
Гласный 6: 2.241с - 2.331с (90мс)
Гласный 7: 2.403с - 2.475с (72мс)
Гласный 8: 2.547с - 2.745с (198мс)
Гласный 9: 2.781с - 2.907с (126мс)
Общий **RMS = 55.6 мс**.

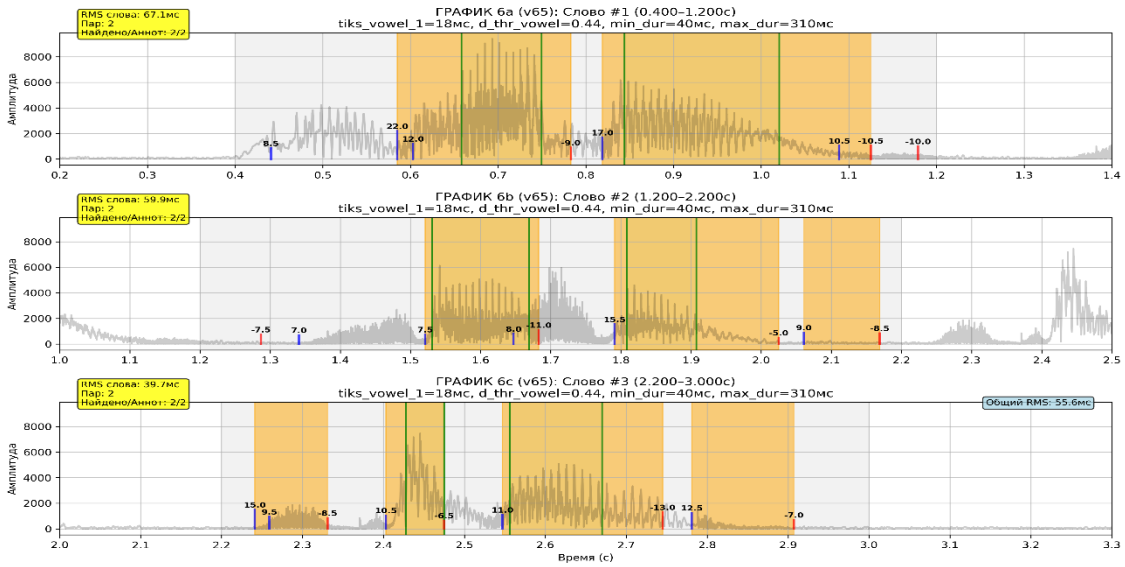

Рис. 6. Границы 9 звуков, найденные методом РП,
при длительности РП 72 мс, общий RMS = 55.6 мс.

6.2. Анализ полученных претендентов на гласный звук.

Анализ временных границ претендентов на гласный звук показан в Таблице 1.

Таблица 1. Анализ временных границ претендентов на гласный звук.

Эталон (№)	**Детекция (№)**	**Комментарий и анализ расхождений**
1	1	Частичное совпадение. Начало детектировано на 73 мс раньше, конец – на 33 мс позже. Алгоритм захватил более широкий интервал, чем при ручной разметке.
2	2	Совпадение по положению, но сильное расхождение по длительности. Детектированный интервал (306 мс) значительно длиннее эталонного (176 мс). Вероятно, произошло слияние гласного с соседним сонорным звуком.
3	3	Высокая точность. Начало смещено всего на 10 мс, конец – на 14 мс. Отличный результат, демонстрирующий потенциал метода.
4	4	Хорошее совпадение по началу. Начало смещено на 17 мс. Однако конец детектирован на 117 мс позже, что является значительной ошибкой.
-	5	Ложноположительное срабатывание (False Positive). Этот интервал (108 мс) отсутствует в ручной разметке.
-	6	Ложноположительное срабатывание (False Positive). Еще один интервал (90 мс), не соответствующий эталону.
5	7	Хорошая точность. Начало смещено на 25 мс, конец совпадает практически идеально (смещение < 1 мс).
6	8	Хорошее совпадение по началу. Начало смещено на 10 мс, но конец детектирован на 75 мс позже.
-	9	Ложноположительное срабатывание (False Positive). Третий лишний интервал (126 мс).

7. Метод частотного кодирования для фильтрации гласных звуков.

После того как с помощью метода РП были детектированы потенциальные границы гласных, для каждой из них запускается процедура верификации.

Верификация проверяет, превышает ли максимальная скорость спайков в небольшой временной окрестности (±50 мс) потенциальной границы гласной некоторый установленный порог:

- если скорость спайков в этой зоне достигает или превышает пороговое значение, граница считается валидной и сохраняется;
- в противном случае, если детектированная граница не сопровождается достаточным увеличением нейронной активности, она классифицируется как артефакт и отбрасывается.

Результат работы фильтра показан на рисунке 7.

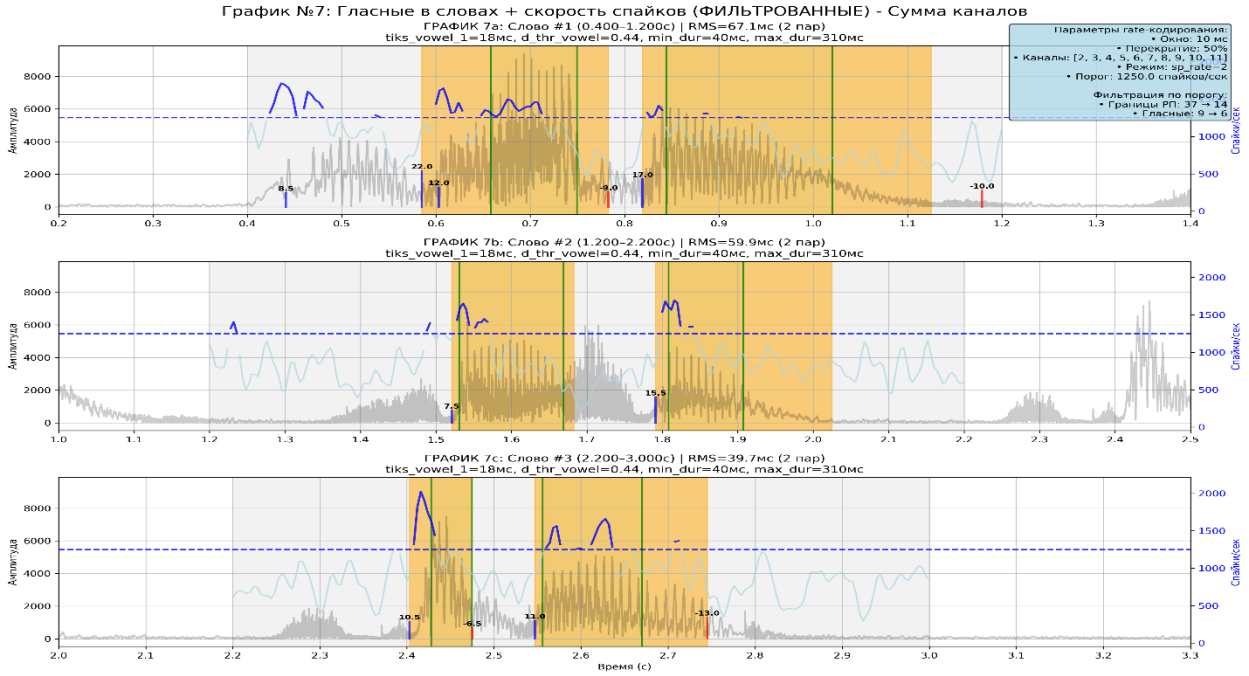

Рис. 7. Границы 6 гласных звуков, найденные методом РП, которые остались после фильтрации методом частотного кодирования. Горизонтальная пунктирная линия соответствует установленному пороговому значени 1250 спайков в секунду. Общий RMS вычислен для пары звуков своего слова и равен, соответственно перечню слов: 67.1 мс, 59.9 мс, 39.7 мс.

Основным результатом применения метода частотного кодирования является улучшение точности детекции гласных за счет значительного снижения числа ложноположительных срабатываний. Метод РП иногда может детектировать переходы, вызванные шумовыми компонентами или слабыми неречевыми звуками. Фильтрация по частотному кодированию эффективно

устраняет такие артефакты, так как они обычно не коррелируют с интенсивным увеличением скорости спайков.

В нашем случае ложноположительные детекции (№5, 6, 9) не являются случайным шумом. Они представляют собой "валидные" с точки зрения алгоритма события, так как имеют и характерную РП-структуру, и достаточно высокую скорость спайков. Это указывает не на ошибку фильтра, а на то, что сам РП-алгоритм находит паттерны, которые нейроморфная система обрабатывает как значимые, даже если они не соответствуют классическим гласным. Вероятно, это могут быть сонорные согласные или дифтонгоидные переходы, имеющие схожую нейронную репрезентацию.

Заключение

Разработанный метод РП успешно решает поставленную задачу, он способен автоматически детектировать временные границы гласных звуков на основе анализа класифицированных РП и с применением фильтрации частотного кодирования. Все 6 эталонных гласных были обнаружены, цели исследования достигнуты.

Средняя точность определения границ слов, выраженная через RMS, составила 55.6 мс при длительности РП 96 мс и 44.0 мс при длительности РП 72 мс. Этот результат является значимым и демонстрирует работоспособность метода РП.

Основная проблема – специфичность. Главной сложностью текущей версии алгоритма является его недостаточная специфичность, что проявилось в трех ложноположительных срабатываниях. Алгоритм находит больше событий, чем присутствует в ручной разметке гласных звуков.

Наибольшие отклонения наблюдаются на границах конца гласных звуков. Это может быть связано с тем, что переход от гласного к последующему согласному (затухание) имеет более сложную и вариативную нейродинамическую картину, чем начало гласного звука.

Перспективы для улучшения метода РП:

1) Необходимо усложнить правила контекста (времена длительности, паузы и т.д), чтобы отличать истинные гласные от других вокализованных звуков (например, сонорных), которые генерируют схожие РП-паттерны.

2) Для повышения специфичности можно ввести третий слой верификации, основанный на акустических признаках (например, особенности формантной структуры).

3) Использование адаптивного порога вместо фиксированного, вероятно, могло бы помочь отсеять некоторые ложные срабатывания, если их нейронный отклик слабее, чем у настоящих гласных.

Результаты исследования демонстрируют огромный потенциал метода рецептивных полей для анализа речи. Текущая версия алгоритма является мощным инструментом для детекции фонетически значимых событий, а выявленные ограничения открывают ясные пути для его дальнейшего совершенствования.

Финансирование: Работа выполнена в рамках государственного задания (номер АААА-А19-119041590070-1) Института радиотехники и электроники им. В.А. Котельникова Российской академии наук.

Литература

1. Bello J.P., Daudet L., Abdallah S., Duxbury C., Davies M., Sandler M.B. A Tutorial on Onset Detection in Music Signals // IEEE Transactions on Speech and Audio Processing. 2005. Vol. 13, No. 5. P. 1035–1047. https://doi.org/10.1109/TSA.2005.851998
2. Osses A., Varnet L., Carney L.H., Dau T., Bruce I.C., Verhulst S., Majdak P. A comparative study of eight human auditory models of monaural processing // Acta Acustica. 2022. Vol. 6. P. 17. https://doi.org/10.1051/aacus/2022008
3. de Cheveigné A. Simple and efficient auditory-nerve spike generation // bioRxiv. 2023. https://doi.org/10.1101/2023.05.02.539135

4. Land E.H., McCann J.J. Lightness and Retinex Theory // Journal of the Optical Society of America. 1971. Vol. 61, №1. P. 1–11. https://doi.org/10.1364/JOSA.61.000001
5. V.E. Antsiperov, M.M. Gutorov, Signal Intensity Change Point Detection by System of Overlapped Receptive Fields Based on Modeling Perception Mechanisms of Living Sensory Systems // Proc. 25th International Conference on Digital Signal Processing (DSP 2025), Costa Navarino, Greece. 2025. (to appear).
6. Boersma, P.; Weenink, D. Praat: doing phonetics by computer. Version 6.4.42 [Электронный ресурс]. — Amsterdam: University of Amsterdam, 1992–. — Режим доступа: http://www.fon.hum.uva.nl/praat/ (дата обращения 14.09.2025)

Для цитирования:

Гуторов М.М., Анциперов В.Е. Обнаружение резких изменений интенсивности речевого сигнала на основе концепции рецептивных полей. // Журнал радиоэлектроники. – 2025. – №. 8. https://doi.org/10.30898/1684-1719.2025.8.15